UNIVERSITY OF CAPE COAST

DIFFERENTIAL ITEM FUNCTIONING OF WEST AFRICAN SENIOR SCHOOL CERTIFICATE EXAMINATION IN CORE SUBJECTS IN SOUTHERN GHANA

RUTH ANNAN-BREW

2020

UNIVERSITY OF CAPE COAST

DIFFERENTIAL ITEM FUNCTIONING OF WEST AFRICAN SENIOR SCHOOL CERTIFICATE EXAMINATION IN CORE SUBJECTS IN SOUTHERN GHANA

BY

RUTH ANNAN-BREW

Thesis submitted to the Department of Education and Psychology of the Faculty of Educational Foundations, College of Education Studies, University of Cape Coast, in partial fulfilment of the requirements for the award of Doctor of Philosophy Degree in Educational Measurement and Evaluation.

MAY 2020

DECLARATION

Candidate’s Declaration

I hereby declare that this thesis is the result of my own original research and that no part of it has been presented for another degree in this university or elsewhere.

Candidate’s Signature …………………………….. Date ……………………..

Name: ………………………………………………………………………

Supervisors’ Declaration

We hereby declare that the preparation and presentation of thesis were supervised in accordance with the guideline on supervision of thesis laid down by the University of Cape Coast.

Principal Supervisor’s Signature …………………….. Date ………………

Name: ……………………………………………………………

Co- Supervisor’s Signature …………………………… Date …………………

Name: …………………………………………………………….

ABSTRACT

The research aimed at examining whether the 2012-2016 May/June West African Senior Secondary Certificate Examination (WASSCE) in core subjects exhibited gender and location differential item functioning (DIF) in Ghana using the cross-sectional design. Six research hypotheses and one research question were formulated for the study. A sample of 36,035 candidates consisting of 8,994(English Language), 8,935(Mathematics), 9,089(Integrated Science) and 9,017 (Social Studies) candidates was selected from a population of 273,289 candidates each who sat for the examination from 2012-2016. The instrument for the study was the 50 multiple- choice test items each for Science, Mathematics and Social Studies and 78 (2015), 80 (2016) and 100 (2012, 2013 and 2014) English Language. MH, LR and IRT DIF detection methods were used to identify items that exhibited DIF. The findings showed that there was a significant gender differential item functioning. There was also a significant location differential item functioning as all three methods detected items that function differentially among the five regions under study. There was a high degree of agreement between the Logistic regression, Mantel Haenszel and 3PL Item Response Theory in identifying items with DIF. It was concluded that some items in exams used by WAEC exhibited significant DIF and it was recommended that DIF studies should be conducted by test developers on their test so that the items exhibiting Differential Item Functioning (DIF) could be revised or eliminated to enhance fairness.

KEYWORDS

Differential Item Functioning (DIF)

Core subjects

Mantel Haenszel (MH)

Logistic Regression (LR)

Item Response Theory (IRT)

Unidimensional

Local independence

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisors, Prof. Y. K. A. Etsey and Dr K. Asamoah-Gyimah both in the Department of Education and Psychology, for their professional guidance, advice, encouragement and the goodwill with which they guided this work. I am grateful.

I am also grateful to Dr Eric Anane and all lecturers of Education and Psychology Department for their generous contribution to make this work better. I am again grateful to Mr Jeffrey K.O. Akurang for his unflinching support throughout my life. Finally, I wish to thank my family and friends for their support, especially, my mother Janet Emily Abaidoo, my husband Thompson, Madam Riggie, Theo, Nana Yaw Nsiah, Sandra Sikanku, Justice, Ruth and Benedicta.

DEDICATION

To my family

TABLE OF CONTENTS

Page

DECLARATION ii

ABSTRACT iii

KEYWORDS iv

ACKNOWLEDGEMENTS v

DEDICATION vi

LIST OF TABLES xii

LIST OF FIGURES xx

CHAPTER ONE: INTRODUCTION

Background to the Study 2

DIF across Gender 10

DIF across Nations 13

Statement of the Problem 14

Purpose of the Study 19

Research Hypotheses 20

Research Question 22

Significance of the Study 22

Delimitations 23

Limitations 23

Definition of Terms 23

The organisation of the Study 25

CHAPTER TWO: LITERATURE REVIEW

Introduction 26

Conceptual Framework 26

Test and Measurement 28

Purpose of Assessment 31

Theoretical Review 32

Classical Test Theory 34

Assumptions of Classical True Score Theory 35

Item Properties Under Classical Test Theory 39

Difficulty level 41

Discrimination index 42

Differential item weighting 44

Implications and Limitations of Classical Test Theory Assumptions 46

Item Analysis within Classical Test Theory 47

Strengths of Classical Test Theory 49

Limitations of Classical Test Theory 50

Item Response Theory (IRT) 51

Assumptions of Item Response Theory 52

IRT Models 57

Item Characteristic Curve Models 59

The Rasch Model 59

One-parameter logistic model (1 PL) 61

The item information function of the 1PL model 63

The three-parameter logistic (3PL) model 66

The four-parameter logistic (4PL) model 67

Uses of IRT 67

Limitations 71

Differential Item Functioning (DIF) 71

Test bias, impact, DIF, and item bias 73

Mantel-Haenszel 77

Logistic Regression 84

Measurement Issues 86

Reliability 87

Validity 88

Evidence-based on content 89

Evidence-based on the internal structure 89

Evidence-based on criterion 90

Test Validation and Testing Fairness 91

Measurement Invariance 91

The Importance of Detecting DIF 93

Empirical Review of Related Studies 94

DIF detection methods 94

DIF studies based on observed scores 95

DIF across Gender 103

DIF and Region/Nation/Location 115

Summary of Literature review 119

CHAPTER THREE: RESEARCH METHODS

Introduction 122

Research Paradigm 122

Research Design 126

Population 128

Sampling Procedure 128

Data Collection Instrument 131

Data Processing and Analysis 138

Detecting DIF using DIFAS Mantel-Haenszel method (M-H) 140

Analysis Using Logistic Regression (LR) 143

CHAPTER FOUR: RESULTS AND DISCUSSION

Introduction 146

Figure 15: Plot of English Language scores for 2012 147

Checking unidimensionality assumption 159

Assessing unidimensionality using principal component 3PL analysis 159

Checking of Local Independence Assumption of the 3PL model 181

Checking measurement invariance assumption 193

Tests for invariance 194

IRT Model-Data Fit 201

Item fit 202

Results 206

Effect Size 207

Results of Research hypotheses one and two 207

Summary 219

Research hypotheses three and four 220

Summary 232

Research hypotheses five and six 233

Summary 248

Research question 249

Discussion 262

Differential Item Functioning Across Gender 264

Differential Item Functioning Across Regions 266

Comparison of MH, LR and 3PL IRT in detecting DIF 268

CHAPTER FIVE: SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

Summary 269

Conclusions 270

Recommendations 271

Contribution to knowledge 272

Methodology 272

Implications 273

Suggestions for Further Research 274

REFERENCES 275

APPENDICES

A PAPER 1 OF WASSCE 2012-2016 ENGLISH LANGUAGE 316

B PAPER 1 OF WASSCE 2012-2016 CORE MATHEMATICS 410

C PAPER 1 OF WASSCE 2012-2016 INTEGRATED SCIENCE 464

D PAPER 1 OF WASSCE 2012-2016 SOCIAL STUDIES 522

E ETHICAL CLEARANCE 562

F INTRODUCTORY LETTER 563

LIST OF TABLES

Table Page

Table 1: WASSCE Pass Rate at Credit in Core Subjects by Gender, 2012 3

Table 2: WASSCE Pass Rates at Credit in Core Subjects by Gender, 2014 4

Table 3: WASSCE Pass Rates at Credit level in Core Subjects 5

by Location 2012 5

Table 4: WASSCE Pass Rates at Credit level in Core Subjects by Location 6

Table 5: Example of Item Discrimination Indices 43

Table 6: Contingency Table for a Dichotomous Item with Total Score k 78

Table 7: Sample Distribution of Core Subjects by Gender (2012-2016) 129

Table 8: Sample Distribution of Core Subjects by Location (2012-2016) 130

Table 9: Sample Size Distribution by Subject and Year 158

Table 10: English Language 2012 Principal Components 3PL results 161

Table 11: Mathematics 2012 Principal Components 3PL results 162

Table 12: Integrated Science 2012 Principal Components 3PL results 163

Table 13: Social Studies 2012 Principal Components 3PL results 164

Table 14: English Language 2013 Pricipal Componets 3PL results 165

Table 15: Mathematics 2013 Principal Components 3PL results 166

Table 16: Integrated Science 2013 Principal Components 3PL results 167

Table 17: Social Studies 2013 Principal Components 3PL results 168

Table 18: English Language, 2014 Principal Components 3PL results 169

Table 19: Mathematics, 2014 Principla Components 3PL results 170

Table 20: Integrated Science 2014 Principla Components 3PL results 171

Table 21: Social Studies 2014 Principla Components 3PL results 172

Table 22: English Language 2015 Principla Components 3PL results 173

Table 23: Mathematics 2015 Principla Components 3PL results 174

Table 24: Integrated Science 2015 Principla Components 3PL results 175

Table 25: Social Studies 2015 Principal Components 3PL results 176

Table 26: English Language 2016 Principal Components 3PL results 177

Table 27: Mathematics 2016 Principal Components 3PL results 178

Table 28: Integrated Science 2016 Principal Components 3PL results 179

Table 29: Social Studies 2016 Pricipal Components 3PL results 180

Table 30: Standardized Residual Correlations for English Language 2012 184

Table 31: Standardized Residual Correlations for Mathematics 2012 184

Table 32: Standardized Residual Correlations for Integrated Science 2012 185

Table 33: Standardized Residual Correlations for Social Studies, 2012 185

Table 34: Standardized Residual Correlations for English Language 2013 186

Table 35: Standardized Residual correlations for Mathematics, 2013 186

Table 36: Standardized Residual Correlations for Integrated Science, 2013 187

Table 37: Standardized Residual Correlations for Social Studies, 2013 187

Table 38: Standardized Residual Correlations for English Language 2014 188

Table 39: Standardized Residual Correlations for Mathematics 2014 188

Table 40: Standardized Residual Correlations for Integrated Science 2014 189

Table 41: Standardized Residual Correlations for Social Studies, 2014 189

Table 42: Standardized Residual Correlations for English Language 2015 190

Table 43: Standardized Residual Correlations for Mathematics 2015 190

Table 44: Standardized Residual for Integrated Science 2015 191

Table 45: Standardized Residual Correlations for Social Studies 2015 191

Table 46: Standardized Residual Correlations for English Language 2016 191

Table 47: Standardized Residual Correlations for Mathematics 2016 192

Table 48: Standardized Residual Correlations for Integrated Science 2016 192

Table 49: Standardized Residual Correlations for Social Studies, 2016 193

Table 50: Test of Measurement Invariance for 2012-2016
English_Language Dataset 195

Table 51: Test of Measurement Invariance for 2012-2016
Mathematics Dataset 197

Table 52: Test of Measurement Invariance for 2012-2016
Integrated_Science Dataset 198

Table 53: Test of Measurement Invariance for 2012-2016
Social Studies Dataset 199

Table 54: Summary of Data Fit Statistics of 2012 Items 203

Table 55: Summary of Data Fit Statistics of 2013 Items 203

Table 56: Summary of Data Fit Statistics of 2014 Items 204

Table 57: Summary of Data Fit Statistics of 2015 Items 205

Table 58: Summary of Data Fit Statistics of 2016 Items 205

Table 59: Distribution of DIF Items in 2012-2016 English Language
Exams Using Gender 208

Table 60: Distribution of DIF Items in 2012-2016 English Language
ExamsUsing Location with CR as Reference Group (RF) 209

Table 61: Distribution of DIF Items in 2012-2016 English Language
Exams Using Location with GAR as RG 210

Table 62: Distribution of DIF Items in 2012-2016 Mathematics
Exams Using Gender 211

Table 63: Distribution of DIF Items in 2012-2016 English Language
ExamsUsing Location with CR as the RG 212

Table 64: Distribution of DIF Items in 2012-2016 Mathematics

Exams Using Location with GAR as the RG 213

Table 65: Distribution of DIF Items in 2012-2016 Integrated Science 214

Exams Using Gender 214

Table 66: Distribution of DIF Items in 2012-2016 Integrated Science 215

Exams Using Location with CR as the RG 215

Table 67: Distribution of DIF Items in 2012-2016 Integrated Science 216

Exams Using Location with GAR as the RG 216

Table 68: Distribution of DIF Items in 2012-2016 Social Studies

Exams Using Gender 217

Table 69: Distribution of DIF Items in 2012-2016 Social Studies

Exams Using Location with CR as the RG 218

Table 70: Distribution of DIF Items in 2012-2016 Social Studies

Exams Using Location with GAR as the RG 219

Table 71: Distribution of DIF Items in 2012-2016 English Language

Exams Using Gender 221

Table 72: Distribution of DIF Items in 2012-2016 English Language 222

Exams Using Location with CR as the RG 222

Table 73: Distribution of DIF Items in 2012-2016 English Language

Exams Using Location with GAR as the RG 223

Table 74: Distribution of DIF Items in 2012-2016 Mathematics

Exams Using Gender 224

Table 75: Distribution of DIF Items in 2012-2016 Mathematics

Exams Using Location with CR as the RG 225

Table 76: Distribution of DIF Items in 2012-2016 Mathematics

Exams Using Location with GAR as the RG 226

Table 77: Distribution of DIF Items in 2012-2016 Integrated Science 227

Exams Using Gender 227

Table 78: Distribution of DIF Items in 2012-2016 Integrated Science 228

Exams Using Location with CR as the RG 228

Table 79: Distribution of DIF Items in 2012-2016 Integrated Science 229

Exams Using Location with GAR as the RG 229

Table 80: Distribution of DIF Items in 2012-2016 Social Studies

ExamsUsing Gender 230

Table 81: Distribution of DIF Items in 2012-2016 Social Studies

Exams Using Location with CR as the RG 231

Table 82: Distribution of DIF Items in 2012-2016 Social Studies

Exams Using Location with GAR as the RG 232

Table 83: Distribution of DIF Items in 2012-2016 English Language

Exams Using Gender 235

Table 84: Distribution of DIF Items in 2012-2016 Mathematics

Exams Using Gender 237

Table 85: Distribution of DIF Items in 2012-2016 Integrated Science 238

Exams Using Gender 238

Table 86: Distribution of DIF Items in 2012-2016 Social Studies

Exams Using Gender 239

Table 87: Distribution of DIF Items in 2012-2016 English Language

Exams Using Location with CR as the RG 240

Table 88: Distribution of DIF Items in 2012-2016 English Language
Exams Using Location with GAR as the RG 241

Table 89: Distribution of DIF Items in 2012-2016 Mathematics 242
Exams Using Location with CR as the RG 242

Table 90: Distribution of DIF Items in 2012-2016 Mathematics
Exams Using Location with GAR as the RG 244

Table 91: Distribution of DIF Items in 2012-2016 Integrated Science 245
Exams Using Location with CR as the RG 245

Table 92: Distribution of DIF Items in 2012-2016 Integrated Science 246
Exams Using Location with GAR as the RG 246

Table 93: Distribution of DIF Items in 2012-2016 Social Studies
Exams Using Location with CR as the RG 247

Table 94: Distribution of DIF Items in 2012-2016 Social Studies
Exams Using Location with CR as the RG 248

Table 95: Distribution of Number of Gender DIF Items Using MH, LR
and 3PL IRT Detecting Methods for 2012-2016
English Language with Females as Reference Group 250

Table 96: Distribution of Number of Gender DIF Items Using MH, LR
and 3PL IRT Detecting Methods for 2012-2016
Mathematics with Females as Reference Group 251

Table 97: Distribution of Number of Gender DIF Items Using MH, LR
and 3PL IRT Detecting Methods for 2012-2016
Integrated Science with Females as Reference Group 252

Table 98: Distribution of Number of Gender DIF Items Using MH, LR and 3PL IRT Detecting Methods for 2012-2016 Social Studies with Females as Reference Group. 253

Table 99: Distribution of Number of Location DIF Items Using MH, LR and 3PL IRT Detecting Methods for 2012-2016 English Language with CR as Reference Group 254

Table 100: Distribution of Number of Location DIF Items Using MH, LR and 3PL IRT Detecting Methods for 2012-2016 English Language with GAR as Reference Group 255

Table 101: Distribution of Number of Location DIF Items Using MH, LR and 3PL IRT Detecting Methods for 2012-2016 Mathematics with CR as Reference Group 256

Table 102: Distribution of Number of DIF Items Using MH, LR and 3PL IRT Detecting Methods for 2012-2016 Mathematics with GAR as Reference Group 257

Table 103: Distribution of Number of Location DIF Items Using MH, LR and 3PL IRT Detecting Methods for 2012-2016 Integrated Science with CR as Reference Group 258

Table 104: Distribution of Number of Location DIF Items Using MH, LR and 3PL IRT Detecting Methods for 2012-2016 Integrated Science with GAR as Reference Group 259

Table 105: Distribution of Number of Location DIF Items Using MH, LR and 3PL IRT Detecting Methods for 2012-2016 Social Studies with CR as Reference Group 260

Table 106: Distribution of Number of Location DIF Items Using MH, LR and 3PL IRT Detecting Methods for 2012-2016 Social Studies with GAR as Reference Group 261

LIST OF FIGURES

Figure Page

Figure 1: ICC for Uniform DIF 8

Figure 2: ICC for Nonuniform DIF 9

Figure 3: WASSCE pass rates by gender, 2006. 16

Figure 4: WASSCE pass rates by gender, 2009. 17

Figure 5: Measurement theory of invariance model 27

Figure 6: The item response function of the one-parameter logistic (1PL) model 62

Figure 7: Locating the difficulty of an item on the ability or difficulty axis 63

Figure 8: Item response function and item information function of the 1PL model 64

Figure 9: The item response functions of two 2PL items 65

Figure 10: Item response function of a 3PL item 66

Figure 11: ICC showing no DIF item 83

Figure 12: An example of an item that does not display DIF. 85

Figure 13: An example of an item that displays substantial uniform DIF 86

Figure 14: Interconnections of four Basic Elements in Research Design 127

Figure 15: Plot of English Language scores for 2012. 147

Figure 16: Plot for English Language scores for 2013. 147

Figure 17: Plot of English Language scores for 2014 148

Figure 18: Plot of English Language scores for 2015 148

Figure 19: Plot for English Language scores for 2016 149

Figure 20: Plot of Mathematics scores for 2012 149

Figure 21: Plot for Mathematics scores for 2013 150

Figure 22: Plot of Mathematics for 2014. 150

Figure 23: Plot of Mathematics for 2015 151

Figure 24: Plot of Mathematics scores for 2016 151

Figure 25: Plot for Integrated Science scores for 2012 152

Figure 26: Plot of Integrated Science scores for 2013 152

Figure 27: Plot of Integrated Science scores for 2014 153

Figure 28: Plot of Integrated Science scores for 2015 154

Figure 29: Plot of Integrated Science scores for 2016. 154

Figure 30: Plot of Social Studies scores for 2012 155

Figure 31: Plot of Social Studies scores for 2013 155

Figure 32: Plot of Social Studies scores for 2014 156

Figure 33: Plot of Social Studies scores for 2015 156

Figure 34: Plot of Social Studies for 2016 157

Figure 35: Level of agreement among MH, LR and 3PL IRT DIF detecting methods for 2012-2016 English Language. 250

Figure 36: Level of agreement among MH, LR and 3PL DIF detecting methods for 2012-2016 Mathematics. 251

Figure 37: Level of agreement among MH, LR and 3PL IRT DIF detecting methods for Integrated Science. 252

Figure 38: Level of agreement among MH, LR and 3PL IRT DIF detecting methods for Social Studies. 253

Figure 39: Level of agreement among MH, LR and 3PL IRT DIF detecting methods for 2012-2016 English Language with CR as RG. 254

Figure 40: Level of agreement among MH, LR and 3PL IRT DIF detecting methods for the English Language with GAR as RG. 255

Figure 41: Level of agreement among MH, LR and 3PL IRT DIF detecting methods for 2012-2016 Mathematics with CR as RG. 256

Figure 42: Level of agreement among MH, LR and 3PL IRT detecting methods for 2012-2016 Mathematics using GAR as RG. 257

Figure 43: Level of agreement among MH, LR and 3PL IRT DIF detecting methods for 2012-2016 Integrated Science using CR as RG. 258

Figure 44: Level of agreement among MH, LR and 3PL IRT DIF detecting methods for 2012-2016 Integrated Science using GAR as RG. 259

Figure 45: Level of agreement among MH, LR and 3PL IRT DIF detecting methods for 2012-2016 Social Studies using CR as RG. 260

Figure 46: Level of agreement among MH, LR and 3PL IRT DIF detecting methods for 2012-2016 Social Studies using GAR as RG. 261

CHAPTER ONE

INTRODUCTION

Results from West African Senior School Certificate Examinations (WASSCE) from 2012 to 2016 in Ghana have not been encouraging, especially in the four core subjects namely, Mathematics, English Language, Integrated Science and Social Studies. It is believed that the items used to examine these students are not free of bias and might be causing misleading decisions at the policy level as policymakers review educational policies, change their curriculum, teaching methods and assessments based on the results of the WASSCE. Such uses or decisions are valid only to the extent that the test items in the test are bias-free. There is a wide range of literature on performance differences in Mathematics and Integrated Science by gender reported in several meta-analytic studies (Friedman, 1989; Hyde, Fennema, & Lamon, 1990) and reviews (Fennema & Leder, 1990; Leder, 1992) that males and females are endowed with about an equal percentage of the human potentials in the world (Siamisang & Nenty, 2012). Research, however, generally shows that males perform better than females in Mathematics, particularly at the upper end of secondary schools and university (Hyde, Fennema, & Lamon, 1990).

Given the gender researches done so far, the purpose of this study is to investigate differential item functioning (DIF) by gender and location(region) of WASSCE (Core Subjects) using Mantel Haenszel (MH), Item Response Theory Model (IRT) and Logistic Regression (LR) techniques.

Background to the Study

Hardcastle (as cited in Victorino, 2011) sees education as the principal vehicle by which economically and socially marginalized adults and children can lift themselves out of poverty and obtain the means to participate fully in their communities.

Ghana as a middle-level income country requires human resources with capabilities in abstract and problem-solving skills to tackle the increasingly technological environment of production and trade. Essentially, these capabilities mentioned above begin their development mainly from the secondary education level where it is believed, the returns to the individual and society are much higher (Lewin & Sayed, 2005; World Bank, 2007). Analysis of the rate of return by the level of education in Ghana has indicated that senior high school produces a higher rate of private and social return than the junior high school level (Canagarajah & Coulombe, 1997). The relatively low rates of return to Junior High School (JHS) are also an indication that overall, Junior High School has not been very efficient in preparing many students who complete, to qualify for Senior High School or actively participate in the labour market. In contrast, the high rates of return to senior high indicate that it functions better as terminal education for entry into the labour market.

According to Institute of Statistical, Social and Economic Research ISSER (2008), the performance of many children in Ghana is failing to meet the minimum learning requirements and to acquire basic skills and competencies. In 2006, the Basic Education Certificate Examination (BECE) results released by the West African Examination Council (WAEC) showed that out of the 308,379 candidates who sat for the examination, only 190,921 candidates were able to obtain aggregates between 06 and 30 (the required national pass aggregates), which represented 62 per cent (WAEC, 2006).

Notwithstanding, the results of WASSCE 2014 and 2015 results released points to the fact that students performed poorly especially in the three core subject areas of English Language, Mathematics and Integrated Science. These subjects serve as the basis for admission into many tertiary institutions.

In Ghana, WASSCE serves as the link for senior high school graduates to enter or get admission into tertiary institutions. Performance of students in the core subjects (WASSCE) by gender has also not been too good as shown in Tables 1 and 2.

Table 1: WASSCE Pass Rate at Credit in Core Subjects by Gender, 2012

Subjects	Gender		
	Male	Female	Total
English Language	69%	67%	68%
Mathematics	55%	44%	48%
Integrated Science	61%	52%	57%
Social Studies	88%	86%	87%

From Table 1, out of the four core subjects, Social Studies had the highest pass rate, with 88% for male candidates and 86% for female candidates. This was followed by English Language where 69% of candidates passed, then by Mathematics where 55% of students passed, and finally by Integrated Science where 61% of male students passed. Male students performed better than females in all subjects with the largest disparity in Mathematics where the average pass rate for females was 44% compared to 55% for males.

Table 2 presents information on WASSCE pass rates at credit in core subjects by gender for 2014.

Table 2: WASSCE Pass Rates at Credit in Core Subjects by Gender, 2014

Subjects	Gender		
	Male	Female	Total
English Language	51%	44%	48%
Mathematics	63%	65%	64%
Integrated Science	48%	43%	46%
Social Studies	73%	69%	71%

From Table 2, out of the four core subjects, Social Studies had the highest pass rate, with 73% of candidates. This was followed by English Language where 63% of candidates passed, then by Mathematics where 51% of students passed, and finally by Integrated Science where 48% of students passed. Male students performed better than females in all subjects except English Language, with the largest disparity in Mathematics where the average pass rate for females was 44% compared to 51% for males.

Table 3 presents information on pass rates at a credit (A1-C6) in core subjects by location in 2012.

Table 3: WASSCE Pass Rates at Credit level in Core Subjects by Location 2012

Region	English Language	Mathematics	Int. Science	Social Studies	Total
Ashanti	68%	53%	57%	90%	67%
Brong Ahafo	70%	56%	58%	90%	69%
Central	68%	48%	54%	88%	65%
Eastern	76%	57%	64%	91%	72%
Greater Accra	73%	43%	57%	83%	64%
Northern	48%	34%	41%	72%	49%
Upper East	62%	54%	70%	89%	69%
Upper West	65%	54%	70%	89%	70%
Volta	68%	38%	50%	79%	59%
Western	66%	55%	56%	91%	67%
Total	64%	49%	58%	86%	

The results varied by administrative regions (Table 3). In all four core subjects, the Northern region had the lowest pass rate. In English Language, the pass rates ranged from 76% in the Eastern region to 48% in the Northern region. Notably, Upper West and Upper East performed far above average in Integrated Science, with a 70% pass rate compared with 57% nationally. In Mathematics, all the pass rates were lower than for other subjects, ranging from 57% (Eastern) to 34% (Northern).

Table 4 gives a breakdown of WASSCE pass rates in core subjects by region in 2014, which provides some telling insights.

Table 4: WASSCE Pass Rates at Credit level in Core Subjects by Location 2014

Region	English Language	Mathematics	Integrated Science	Social Studies	Total
Ashanti	69%	59%	55%	72%	64%
Brong Ahafo	70%	76%	65%	86%	74%
Central	59%	36%	36%	68%	50%
Eastern	70%	47%	45%	73%	59%
G. Accra	75%	45%	46%	74%	60%
Northern	33%	23%	21%	52%	32%
Upper East	47%	28%	34%	66%	44%
Upper West	59%	28%	42%	80%	52%
Volta	61%	33%	36%	65%	49%
Western	63%	54%	47%	71%	59%
Total	64%	48%	46%	71%	

As Table 4 shows, in Mathematics, Integrated Science and Social Studies, Brong Ahafo have outperformed all other regions, scoring 76%, 65%, and 86% respectively; all far surpassing the national averages. In English Language, Brong Ahafo performed second best after Greater Accra with a score of 70%. The northern region, on the other hand, had the lowest performance across all four core subjects with 33% for English Language, 23% for Mathematics, 21% for Integrated Science, and 52% for Social Studies. It is also noted that the greatest disparity in pass rates was for Mathematics with a 53%-point difference between the Northern region and Brong Ahafo.

As a part of the determination of validity for these tests, differential item analysis should be employed to evaluate the degree to which measurements distinguish true abilities among examinees in an unbiased manner. Psychometricians and test developers use differential item functioning (DIF) analysis to determine if there is a possible bias in a given test item (Zumbo, 2003).

DIF is said to be present when examinees from different groups have different probabilities of success on an item after controlling for overall ability (Clauser & Mazor, 1998). If an item is free of bias, responses to that item are related only to the level of the underlying trait that the item is trying to measure. If item bias is present, responses to the item are related to some other factors as well as the level of the underlying trait (Camilli & Shepard, 1994). The tight relationship between the probability of correct responses and ability trait levels is an explicit assumption of IRT (Edelen & Reeve, 2007) and an implicit assumption of classical test theory by McDonald's (Erguven & Erguven, 2014). The presence of large numbers of items with DIF is a severe threat to the construct validity of tests and the conclusions based on test scores derived from items with DIF.

According to Walker (2011), two types of DIF can occur in items, namely uniform and non-uniform DIF as shown in Figures 1 and 2. Uniform DIF is the simplest type of DIF where the magnitude of conditional dependency is relatively invariant across the latent trait continuum (θ). The item of interest consistently gives one group an advantage across all levels of ability (θ). For non-uniform DIF, the difference in the probabilities of answering an item correctly can vary in different directions for different ability levels. In this case, there is an interaction between ability levels and group membership. Within an item response theory (IRT) framework this would be evidenced when both item characteristic curves (ICC) are equally discriminating yet exhibit differences in the difficulty parameters (i.e., $a_r = a_f$ and $b_r < b_f$) as depicted in Figure 1. (Mellenbergh, 1982).

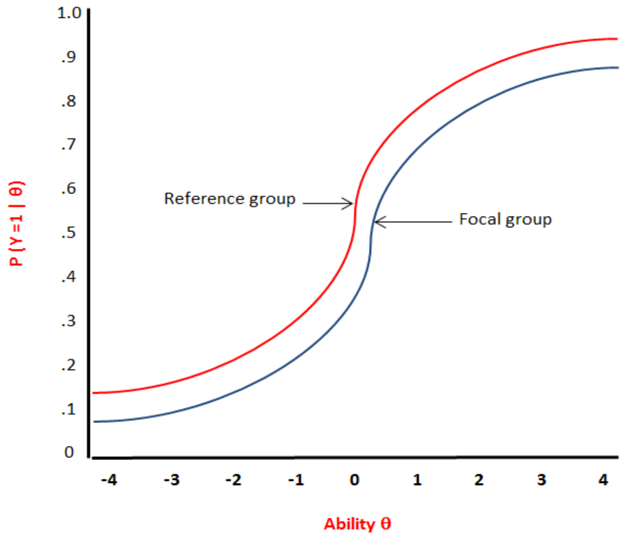

Figure 1: ICC for Uniform DIF

Figure 1 is an example of an item that displays substantial DIF with a small area between the two ICCs, one for the focal group and the other for the reference group. This type of DIF is known as uniform DIF because the ICCs do not cross. An item such as the one shown in Figure 1 may not be an equivalent measure of the same latent variable for both groups.

However, nonuniform DIF presents an interesting case. Rather than a consistent advantage being given to the reference group across the ability continuum, the conditional dependency moves and changes direction at different locations on the θ continuum (Walker & Beretvas, 2003). For instance, an item may give the reference group a minor advantage at the lower end of the continuum while a major advantage at the higher end. Also, unlike uniform

DIF, an item can simultaneously vary in discrimination for the two groups while also varying in difficulty (i.e., $a_r \neq a_f$ and $b_r < b_f$). Even more complex is "crossing" nonuniform DIF. As demonstrated in Figure 2, this occurs when an item gives an advantage to a reference group at one end of the θ continuum while it favours the focal group at the other end. Differences in ICCs indicate that examinees from the two groups with identical ability levels have unequal probabilities of correctly responding to an item. When the curves are different but do not intersect, this is evidence of uniform DIF. However, if the ICC cross at any point along the θ scale, there is evidence of non-uniform DIF.

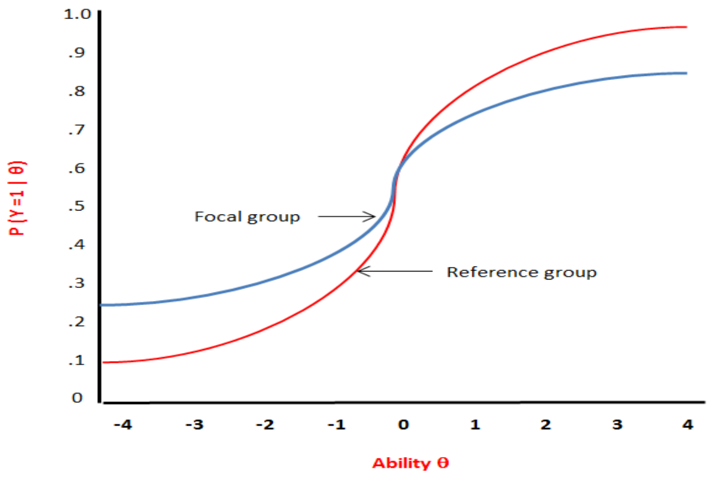

Figure 2: ICC for Nonuniform DIF

Figure 2 is an example of an item that displays substantial non-uniform DIF (i.e., the ICC cross over one another). It depicts non-uniform DIF because for those individuals who score at or below the mean (i.e., $z \leq 0$), the focal group is favoured whereas for those scoring above the mean (i.e., $z>0$) reference group is favoured.

DIF across Gender

There is a wide range of literature of performance differences in Mathematics by gender reported in several meta-analysis studies (Ross, Xu & Ford, 2008; Hyde, Fennema & Lamon, 1990) and reviews (Fennema & Leder, 1990; Leder, 1992; Young & Fisler, 2000). The consensus in the existing literature is that gender differences in Mathematics achievement are thought to emerge at age 14. Hyde et al. (1990), for example, found in a meta-analysis that differences in mathematical problem solving do not exist below age 14, but do exist beyond age 14. However, a longitudinal study of kindergarten children showed that gender differences seemed to emerge when boys and girls enter kindergartens. Initially, boys performed better at the top of the distribution and worse at the bottom, but by third-grade boys performed as well or better than girls throughout the distribution (Penner & Paret, 2008). In the United Kingdom, the 1999 Trends in International Mathematics and Integrated Science Study (TIMSS) results showed that the boys scored higher than girls and this difference was significant for fourth and eighth-grade pupils (Mullis et al, 2000). In TIMSS 2003, fourth-grade boys performed better than girls but not statistically significant (Mullis, Martin, Gonzales, & Chrostowski, 2004). In TIMSS 2007 (Sturman, et al., 2008) no gender differences were found in either grade. The results of the Programme for International Student Assessment (PISA) 2006, Organisation for Economic Co-operation and Development (OECD, 2007), indicated that boys outperformed girls in Mathematics for pupils from the UK.

PISA 2006 results reported that boys outperformed girls by 17 points and the UK is ranked as one of the four countries (after Austria, Japan and

Germany) that showed a big gender gap in Mathematics performance. The results of the PISA studies consistently showed that boys performed better than girls compared to the TIMSS studies. It can be argued that the TIMSS tests are focused on basic content and cognitive mathematical knowledge, whereas the PISA tests assessed students' ability to apply their mathematical skills in solving problems in some real-world contexts. Notwithstanding, in a study of SAT Mathematics test, Harris and Carlton (1993) found that abstract algebra items and items requiring low cognitive processing favoured females whereas, on geometry, measurement, number, computation, data analysis, and proportional reasoning items DIF favoured males. Later, however, Mendes-Barnett and Ercikan (2006) concluded that boys performed better on items requiring problem-solving, high cognitive complexity, visual reasoning, and application of Mathematics principles to word problems. Other researchers have identified no systematic gender DIF for Mathematics items across different testing application contexts such as California Achievement Tests (Ahmadi & Bazvand, 2016), and Iowa Tests of Basic Skills Mathematics problem solving and Mathematics concepts items (Plake, 1980). In this context, it seemed that boys were generally better in applied problem-solving in Mathematics compared to girls.

As regards the second stream of gender research which is pertinent to subject matter studies, some investigations have been made. A lot of research seems to have been carried out on gender in terms of language studies. Specifically, some gender DIF studies have been carried out in testing literature. For instance, Ryan and Bachman (1992) found gender differences across Test of English Language as a Foreign Language (TOEFL) and First Certificate of

English Language (FCE) using Mantel-Haenszel (MH) procedure. Regarding TOEFL, four of the items favoured males and two items were biased toward females. As regards the FCE, one item favoured males and the other one in favour of females. In the same line, Amirian, Alavi and Fidalgo (2014) detected gender DIF in a language proficiency test in Iran known as University of Tehran English Language Proficiency Test (UTEPT) using Mantel-Haenszel and Logistic Regression (LR) methods. Results indicated that 28% of the items displayed DIF, suggesting that humanities-related topics were more in favour of females, while Integrated Science oriented texts were biased for males.

Gender DIF studies across reading comprehension tests have also attracted the attention of researchers. Using MH procedures, Pae's (2004b) detected gender DIF across the English Language subtest of Korean College Scholastic Ability Test (KCSAT) and found that logical inference items were more likely to favour males, while items dealing with impressions, mood and tone of a given passage tended to favour females. Similarly, Pae (2012) systematically examined the same sub-test but on a long-term basis and across three regular forms (1999, 2003, 2007), applying MH procedures and IRT-LR methods. It was reported that item type is a more reliable predictor of gender DIF than item content, thus being consistent with his previous (2004b) study. Ahmadi and Jalili (2014) also applied two DIF detection methods of LR and IRT across an Iranian reading comprehension test. Consistent with Pae (2004b, 2012) findings, this study revealed that 17% of the items displayed DIF, suggesting that item types such as reference and vocabulary were better predictors of gender DIF (mostly favouring females) than test content.

As regards gender DIF in Integrated Science, many studies have been carried out in this field as well. For example, some have examined item format effect (Bolger & Kellaghan, 1990; Hamilton, 1999; Cole,1997; Zenisky, Hambleton, & Robin, 2003), suggesting that multiple-choice items seem to favour males, while open-ended items are more biased for females. Others have studied the effect of item contents (Becker, 1989; Burkam, Lee & Smerdon, 1997; Jovanovic, Solano-Flores, & Shavelson, 1994; Young & Fraser, 1994), concluding that males seem to outperform females on physical, earth, and space Integrated Science items.

On the effect of cognitive domain items, some evidence was found that male examinees performed differentially better than female examinees (when matched on total test score) on items requiring spatial reasoning or visual content (Linn & Hyde, 1989; Halpern, 1992). Consistently, items requiring spatial reasoning or visual content favoured males (Halpern, 1992).

Nevertheless, one study has been identified that examined gender DIF in Social Studies. This study was done by Osadebe and Agbure (2018) to examine differential item functioning in Social Studies multiple choice questions in the Basic Education Certificate Examination. The study used all Junior Secondary class three students in Delta Central Senatorial District in Nigeria. The finding revealed that there is the incidence of gender, location, socio-economic, school type and school ownership differential functioning in 2014 BECE Social Studies multiple-choice test.

DIF across Nations

Zumbo (2003) conducted a study to investigate the extent to which English language items function differently among students from America, Canada and New Zealand. Results showed a significant country and test language effect on nation related DIF in the international data. Nenty (2010) conducted an item bias study comparing the DIF across three mutually remoted and culturally disparate groups (600 Americans, 231 Indians and 800 Nigerians). This study was conducted using four techniques: SSX^2, Item Characteristics Curve (ICC), The Cochran's Test Method (CTX^2) and Transformation Item Difficulty (TID-45^0). A summary of the results showed that 27 out of 46 items of the Cattell Culture Fair Intelligence Test tended to be relatively biased which might be due to DIF.

Statement of the Problem

As a result of this, there are some instances where an item in these examinations could be more difficult for a particular group of examinees who are of the same ability level but from different subgroups to perform differently, such an item can be said to be showing differential item functioning (DIF). Ability is the quality of being able to do something. Hence, the recorded level of accomplishment an individual reaches is referred to as ability level. There is perhaps, no issue more visible among national examinations conducted in a heterogeneous country like Ghana than differential item functioning (DIF). The problem that necessitated this study centres around the effect of a test item differentially functioning.

Differential item functioning can simply be said to occur when test takers from different groups that have been matched on similar ability levels are performing differently on test items. The effect is that some examinees will be doing well while some will not be doing well. This has created the problem of unequal opportunity among the examinees. In Ghana, the inclusive education policy under the Universal Design for Learning (UDL) and the Child-Friendly Schools (CFS) model, emphasize that education services are to respond to the diverse needs of all pupils/students within the framework of universal design for learning. This has led to improving equitable access to quality education for all children of diverse educational needs, provision of requisite teaching and learning materials, capacity development for professional and specialised teachers and managers as well as improvements in education service delivery.

Despite the strategic importance of examination or test-taking for diagnostic, placement, classification and quality control in Ghanaian institutions, the performance of Senior High School (SHS) students in the English language, Mathematics, Integrated Science and Social Studies seems to differ in terms of gender and school location. For example, several research studies have shown that gender differences in Mathematics learning are not clear during the elementary school years (Hyde & Geiringer, 1975; Mann, Sasanuma, Sakuma, & Masaki, 1990), but girls begin to fall behind boys during the intermediate school years, and they fall further behind during the high school years (Fennema, 1974, 1980; Leder, 1992).

Kimball (1989) cited many studies showing that boys in high school, generally, achieved higher scores than girls on standardized tests. Studies of gender differences in Mathematics achievement (Hedges & Nowell, 1995; Peterson & Fennema, 1985; Randhawa, 1994) found that, in general, males outperformed females in Mathematics during the high school years. Other studies (Fox, Brody, & Tobin, 1985) emphasized high Mathematics achievement being dominated by males. Leder (1992) has also reported the existence of gender differences in Integrated Science subjects, in general, as well as in Mathematics. Gessell (2004) asserted that girls under the age of fourteen years usually perform better in the English Language than boys of the same age. Also, after that age, the boys usually overtake the girls whereas Denga (1998) posited that no evidence is clear as to whether differences exist between males and females in academic achievement. He, however, stated that girls tend to do better than boys in language Arts like English Language and Music while the boys tend to outperform the girls in Mathematics and Integrated Sciences. Many studies indicate that women are better than men in verbal skills whereas men tend to outperform women in geometry and arithmetic and algebra reasoning questions (Geary,1996).

The literature reviewed so far is not quite different from the WASSCE results for the core subjects in terms of gender in Ghana. There are gender disparities in the performance of students who sit for the WASSCE as depicted by Figures 3 and 4.

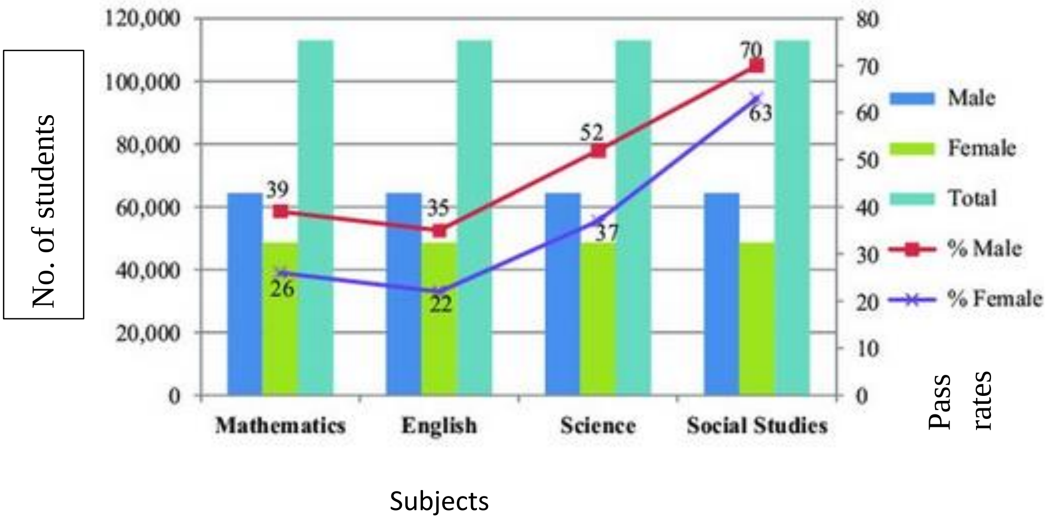

Figure 3: WASSCE pass rates by gender, 2006.

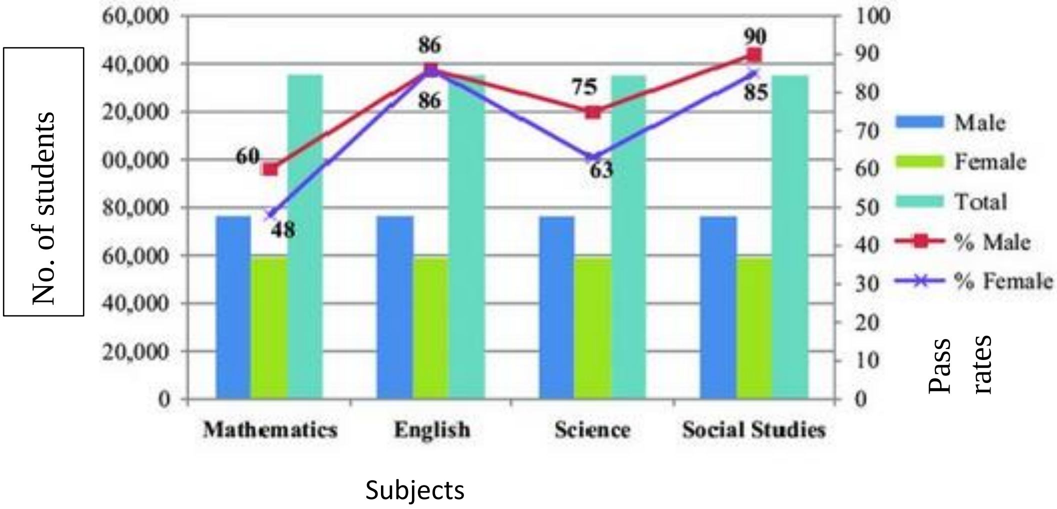

Figure 4: WASSCE pass rates by gender, 2009.

From Figures 3 and 4, it can be concluded that males perform little better in Mathematics and Integrated Science than females whereas there seems to be a virtually equal performance by male and female in English Language and Social Studies.

The problem of testing not providing equal opportunity for examinees has been created as a result of the test items functioning differentially. There is no question that education is a key element in improving the lives of children, families, communities and nations. When some examinees are failing while some are passing as a result of the difficulty posed by the test items, it has distorted the chance of those who failed to be promoted. This has proved the notion of how DIF could be harmful and threatening.

There is also the problem of locational differentiation. Location differentiation in this context deals with dividing students in the schools into groups based on the school location such that they have certain economic or/and social characteristics in common. The focus of education is to bridge the gap between groups. This could lead to the purpose of education being subjugated as a result of the threat being posed by the effect of DIF. The presence of DIF coincides with differential drop out in schools since the test items are proving difficult to the examinees. Examinees' failure may not be because of their inability to answer the items correctly but because of the unfairness of the test; it can result in many examinees withdrawing out of school. According to Odili (2010), the interest in the analysis of differential item functioning in test derives from the consideration that education is perceived as an instrument for achieving equity among persons. In achieving this requires test items to measure traits which are taught in schools and not those that are foreign to it.

Results of WASSCE may have test items that are not free from item bias which might be causing misleading decisions at the policy level as policymakers review educational policies, change their curriculum, teaching methods and assessment methods based on the results of these external examinations. Such uses or decisions are valid only to the extent that the items in the test are bias-free. This study, therefore, seeks to investigate gender and location DIF in multiple-choice test items of English Language, Core Mathematics, Integrated Science and Social Studies for the 2012-2016 WASSCE.

The basis for this study was the literature claim of the disadvantage faced by females in Mathematics, Integrated Science, Social Studies and English Language learning and assessments, which may hinder their choice of a programme at higher levels or career selection with higher demand for these core subjects. The contrasting results of international comparative studies and poor national examination results for the core subjects in Ghana may suggest the invalidity of test with regards to curriculum development, test construction, administration, scoring and analysis. Hence the test scores of these examinees may not reflect student' true performance.

The literature on gender and nation DIF mostly concerns Mathematics and English Language leaving Integrated Science and Social Studies. Also, previous studies seem to concentrate more on one-year group test items for DIF analysis. This study aimed at filling these gaps in examining gender and location DIF in all core subjects which includes Mathematics, Integrated Science, English Language and Social Studies across five years items results using a three-step procedure involving the LR, the MH procedure and the 3PL IRT model analysis.

Purpose of the Study

The purpose of this research is to determine whether Core Mathematics, Integrated Science, English Language and Social Studies test items in the 2012-2016 WASSCE items exhibited any significant differential item functioning based on gender and location. Specifically, this study is to find out whether the 2012-2016 WASSCE (Core subjects) examination items exhibited gender and location (Eastern, Volta, Western, Central and Greater Accra) differential item functioning.

Assumptions

It is assumed that:

1. the raw score of an individual is made up of a true score and a random error.
2. random errors around a true score are normally distributed.
3. random errors are uncorrelated with each other.
4. every test measure only one construct (unidimensional)

5. responses are independent given a subject's latent trait value in terms of conduct and administration of examinations. (local independence)

Research Objectives

In line with the study's main purpose, the following objectives guided the research:

1. To examine the differential item functioning in the core subjects based on gender.
2. To examine the differential item functioning in the core subjects based on location.
3. To determine which of the DIF detecting methods is most effective against identifying DIF.

Research Hypotheses

The study sought to answer the following research hypotheses and question:

1. H_0: There is no statistically significant gender differential item functioning in 2012-2016 WASSCE core subjects' examinations using MH DIF detecting procedure.

   H_1: There is a statistically significant gender differential item functioning in 2012-2016 WASSCE core subjects' examinations using MH DIF detecting procedure.

2. H_0: There is no statistically significant location differential item functioning in 2012-2016 WASSCE core subjects' examinations using MH DIF detecting procedure.

   H_1: There is a statistically significant location differential item functioning in 2012-2016 WASSCE core subjects' examinations using MH DIF detecting procedure.

3. H_0: There is no statistically significant gender differential item functioning in 2012-2016 WASSCE core subjects examinations using LR DIF detecting procedure.

   H_1: There is a statistically significant gender differential item functioning in 2012-2016 WASSCE core subjects examinations using LR DIF detecting procedure.

4. H_0: There is no statistically significant location differential item functioning in 2012-2016 WASSCE core subjects examinations using LR DIF detecting procedure.

   H_1: There is a statistically significant location differential item functioning in 2012-2016 WASSCE core subjects examinations using LR DIF detecting procedure.

5. H_0: The 2012-2016 WASSCE core subjects examinations do not statistically significantly exhibit gender and location differential item functioning using 3PL IRT model.

   H_1: The 2012-2016 WASSCE core subjects examinations statistically significantly exhibit gender and location differential item functioning using 3PL IRT model.

6. H_0: The 2012-2016 WASSCE core subjects examinations do statistically significantly exhibit location differential item functioning using 3PL IRT model.

   H_1: The 2012-2016 WASSCE core subjects examinations do not statistically significantly exhibit location differential item functioning using 3PL IRT model.

Research Question

What is the level of agreement among the MH, LR and 3PL IRT DIF detection methods?

Significance of the Study

It was envisaged that the findings of the study would contribute to the theories of differential item functioning as well as literature on DIF in Ghana and the rest of the world. The findings would theoretically attempt to fill the existing gap in research in the field of differential item functioning.

This study would give feedback to test developers both internally and externally, especially test developers and administrators of West Africa Examination Council on the need to consider ways of reducing or eliminating possible DIF items during test construction.

It is anticipated that the results of this study should be of interest to test developers, practitioners, policymakers, stakeholders and those who use tests to inform and make various decisions in Ghana. Test developers and test users must ensure the validity of the test and should justify the interpretation of the proposed use of the test. DIF analyses as a utility of construct validation may provide evidence to support the interpretation of the test scores. Besides, DIF data may be employed in the test development process.

With a good understanding of gender DIF in the core subjects testing, teachers as part of stakeholders may be better equipped to write test items, and further provide instruction that targets the weaknesses of both males and females.

The findings would help researchers and practitioners to understand the issue about gender and location differences in Mathematics, English Language, Social Studies and Integrated Science education and to encourage more research to be conducted, in particular into the instructional strategies in the teaching and learning of these subjects to cater for the needs of every student.

Delimitations

The scope of the study was limited in the following ways. Firstly, this study did not attempt to study the school effects and other characteristics of students such as ethnicity and language proficiency. It was also limited to Integrated Science, Social Studies, English Language and Mathematics subjects only. This study examined only DIF based on gender and location of the schools in regions within Ghana and not the type of ethnic groups.

Limitations

The Mantel-Haenszel (M-H) technique is ideal but lacks the power to detect DIF that is not uniform across the range of θ scores and this is somewhat arbitrary and may affect the statistical decision regarding DIF. (Hambleton & Rogers, 1989; Swaminathan & Rogers, 1990; Uttaro & Millsap, 1994). Logistic Regression method also has two major weaknesses: 1) The Type I error or false positive rate is higher than expected, and 2) the lack of an effect size measure which might also affect the meaning of results of this study.

Definition of Terms

The following are definitions of key terms used in this study.

Differential functioning (DF)

Differential functioning here refers to an item, a bundle of items or a test that functions differentially for different groups with the same (or comparable) ‘ability’.

Differential item functioning (DIF)

DIF occurs when examinees from different subgroups of the same population have differing probabilities of responding correctly to (or endorsing) an item because there are true differences between the groups in the underlying ability being measured.

Uniform differential item functioning (Uniform DIF)

Uniform DIF occurs when the probability of answering an item correctly is consistently higher for one group than the other overall ability levels. It is characterised by two parallel ICCs. In this sense, there is no interaction between ability levels and group membership.

Nonuniform differential item functioning (Nonuniform DIF)

Nonuniform DIF occurs when the differences in the probabilities of answering an item correctly vary in different directions for different ability levels for different groups. It is characterized by two intersecting ICCs. In this case, there is an interaction between ability levels and group membership.

Crossing DIF

For crossing DIF, the two ICCs are more likely to cross at average ability, where the probability of correct response is likely to be equal to 0.5 and the magnitude of the area under the curves of two groups, in this case, maybe cancelled out.

Non- crossing DIF

For this type of DIF, the two ICCs are more likely to cross at the lower or higher ability values and the area under the curves for two groups may not cancel out and so may show a uniform DIF effect.

Fairness

Fairness refers to the absence of DIF and equitable treatment of all examinees.

Test bias

Test bias occurs when examinees of one group are less likely to answer an item or items on a test correctly than examinees of another group because of some characteristic of the test item or testing situation that is not relevant to the test purpose. DIF is required but not sufficient, for item bias.

The organisation of the Study

The study comprises five chapters. Chapter one is the general introduction which throws light on what to expect in the study. It considers the Background to the Study, Statement of the Problem, Purpose of the study, Objectives of the Study, Significance of the Study, Research Questions, Delimitation, Limitations and Organization of the Study. The second chapter deals with the empirical, theoretical and conceptual review of relevant literature on the study. Related literature from books, the internet, journals, articles and periodicals are reviewed. Chapter three focuses on the methodology. It includes the study design, study population, research instruments, methods of data collection and analysis. The fourth chapter presents the analysis of the data obtained from the field and discusses the result of the study. The last chapter contains a summary of the findings, conclusion, recommendations and suggestions for further research.

CHAPTER TWO

LITERATURE REVIEW

Introduction

The concern of this chapter is to review literature related to the study. It considers issues relating to measurement and differential item functioning (DIF), and to identify a gap in the literature that this research aims to fill. This review is structured as follows:

1. conceptual framework and reviews, issues relating to testing and measurement, performance differences, DIF by gender and location, WASSCE performance of students in Mathematics, English Language, Social Studies and Integrated Science.
2. the theoretical review of the two test theories namely, the classical test theory and the item response theory about the DIF.
3. DIF detection methods.
4. Summary.

Conceptual Framework

A conceptual framework contributes to the identification and the classification of the relationship between the research variables. (Lewis, 2001). Miles and Huberman (1994) defined a conceptual framework as a visual or written product, one that "explains, either graphically or in narrative form, the main things to be studied (i.e., the key factors, concepts, or variables) and the presumed relationships among them" (p. 18). Haralambos and Holbom (2008) also asserted that a conceptual framework enables the researcher to establish the relationship between the existing literature and research goals.

This study was to determine whether the 2012-2016 WASSCE core subjects items exhibit gender/location DIF as shown in Figure 5. It demonstrates the conceptual framework of DIF based on the measurement theory of invariance.

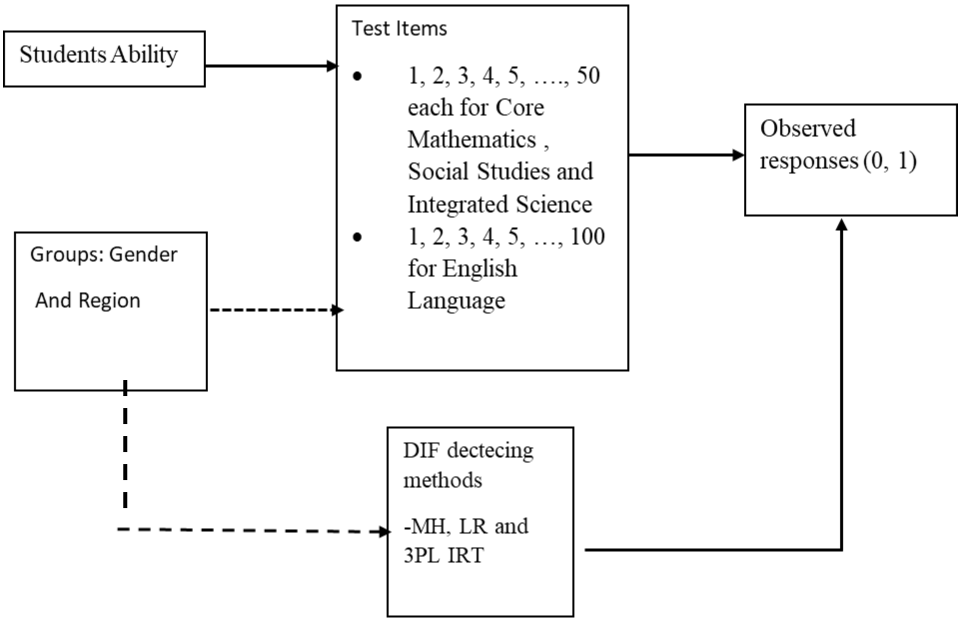
Students Ability
Test Items
1, 2, 3, 4, 5, …., 50 each for Core Mathematics , Social Studies and Integrated Science
1, 2, 3, 4, 5, …, 100 for English Language
Observed responses (0, 1)
Groups: Gender And Region
DIF dectecing methods
-MH, LR and 3PL IRT

Figure 5: Measurement theory of invariance model

Adapted from Engelhard (2009, p. 591)

Figure 5 depicts a student' achievement on a Mathematics, English Language language, Integrated Science and Social Studies test is measuring the latent trait (i.e., mathematical ability or Integrated Science ability) that is observed through their responses to a set of 50 items each for Integrated Science, Social Studies and Mathematics and 78-100 items for English Language that these items vary in their difficulty.

However, the presence of DIF seems to suggest that the group membership (i.e., gender/location) may influence the calibration of items and consequently may result in the mismeasurement of a persons' ability.

The dashed arrows from the group shown are modelled as interaction effects that may change the item difficulty of test items for different groups. The interaction effects between groups and items may be signalling DIF. In other words, the item difficulties are non-invariant over groups. The term non-invariant over groups means that the item difficulties vary when the test is administered to different groups with a similar latent ability (or matching variable).

Analogously, the 'non-invariance' or 'lack of invariance' term over groups indicates DIF. To detect the interaction effects between items and groups at item-level, this study employed three DIF detection methods, namely the MH, LR and the 3PL IRT procedures. At the item-level, the observed responses are usually (or can be made into) dichotomies and these responses are a function of both person ability and item difficulty.

Test and Measurement

A test is defined in this study as a measurement instrument, which measures an attribute that is not clearly observable. For example, the mathematical ability of students cannot be observed clearly. In other words, a test is a collection of items, which appear to be a representative of a construct that is being measured. A test may be administered verbally, on paper, on a computer, or in a predetermined area that requires a test taker to demonstrate or perform a set of skills.

Tests vary in style, rigour and requirements. For example, in a closed book test, a test taker is usually required to rely upon memory to respond to specific items whereas, in an open book test, a test taker may use one or more supplementary tools such as a reference book or calculator when responding. A test may be administered formally or informally. An example of an informal test would be a reading test administered by a parent to a child. A formal test might be a final examination administered by a teacher in a classroom or an intelligent quotient test administered by a psychologist in a clinic. Formal testing often results in a grade or a test score (Thissen & Wainer, 2001).

A test score may be interpreted with regard to a norm and criterion, or occasionally both. The criterion (interpret a student’s performance against a goal, specific objective, or standard) may be established independently, or by statistical analysis of a large number of participants whereas norm compares a student’s performance against other students (a national group or other “norm”).

A standardized test is any test that is administered and scored consistently to ensure legal defensibility (Kaplan & Saccuzzo, 2009). Standardized tests are often used in education, professional certification, psychology (e.g., SAT, GRE), the military, and many other fields.

A non-standardized test is usually flexible in scope and format, variable in difficulty and significance. Since these tests are usually developed by individual instructors, the format and difficulty of these tests may not be widely adopted or used by other instructors or institutions. A non-standardized test may be used to determine the proficiency level of students, motivate students to study, and to provide feedback to students. In some instances, a teacher may develop non-standardized tests that resemble standardized tests in scope, format, and difficulty to prepare their students for an upcoming standardized test (Goswami,1991). Finally, the frequency and setting by which non-standardized tests are administered are highly variable and are usually constrained by the duration of the class period. A class instructor may, for example, administer a test weekly or just twice a semester. Depending on the policy of the instructor or institution, the duration of each test itself may last for only five minutes to an entire class period.

In contrasts to non-standardized tests, standardized tests are widely used, fixed in terms of scope, difficulty and format, and are usually significant in consequences. Standardized tests are usually held on fixed dates as determined by the test developer, educational institution, or governing body, which may or may not be administered by the instructor, held within the classroom or constrained by the classroom period. Although there is little variability between different copies of the same type of standardized test (e.g., SAT or GRE), there is variability between different types of standardized tests.

Any test with important consequences for the individual test taker is referred to as a high-stakes test. A test may be developed and administered by an instructor, a clinician, a governing body, or a test provider. In some instances, the developer of the test may not be directly responsible for its administration. For example, Educational Testing Service (ETS), a nonprofit educational testing and assessment organization, develops standardized tests such as the SAT but may not directly be involved in the administration or proctoring of these tests. As with the development and administration of educational tests, the format and level of difficulty of the tests themselves are highly variable and there is no consensus or invariable standard for test formats and difficulty. Often, the format and difficulty of the test are dependent upon the educational philosophy of the instructor, subject matter, class size, policy of the educational institution, and requirements of accreditation or governing bodies. In general, tests developed and administered by individual instructors are non-standardized whereas tests developed by testing organizations are standardized.

Purpose of Assessment

Educational assessment is conducted for a variety of reasons and the nature of assessment often reflects the purpose for which it is being carried out. The assessment provides information for decisions about students, curriculum and programmes, and educational policy. The decisions are:

1. Instructional management decisions: These decisions include providing knowledge about the readiness of individuals to learn a new set of curricula content, equips teachers in setting realistic instructional goals and objects for the class as well as individual students, in discovering learning difficulties of students and provides remedial action.
2. Selection decision: Assessment provides information to select the right calibre of students for admission, promotion and awards of prizes. Those not acceptable during the selection processes are rejected.
3. Placement decision: Assessment provides information to place students in courses and classes where they are likely to succeed in the future. It also provides the basis for grouping individuals for instruction given individual differences.
4. Guidance and counselling decisions: Assessment aids in providing guidance and counselling in social and psychological adjustment problems that affect the student's performances in the class, assisting students to explore and choose careers and in directing them to prepare for the careers they select.

5. Credentialing and certification decisions: Assessment enables the students to acquire certificates that are needed for employment in the world of work. (McMillan, 2003).

Theoretical Review

The theoretical models for the study include measurement theories of Classical Test Theory (CTT) and Item Response Theory (IRT).

Measurement Theories

Measurement theory is a branch of applied mathematics that is useful in measurement and data analysis (Hand, 1996). The fundamental idea of measurement theory is that measurements are not the same as the attribute being measured. Hence, if any conclusions about the attribute will be drawn, it must be based on the nature of the correspondence between the attribute and the measurements (Hand, 1996).

Michell (2014) states that measurement involves the assignment of numerals to objects or events to represent certain attribute or properties of those objects and events. Any general measurement must come to grips with three basic problems namely error, representation and uniqueness. Various systems of axioms, or basic rules and assumptions, have been formulated as a basis for measurement theory.

Michell (2014), once again indicates that some of the most important types of axioms include (1) axioms of order (2) axioms of extension (3) axioms of difference (4) axioms of conjointness and (5) axioms of geometry. Axioms of order ensure that the order imposed on objects by the assignment of numbers is the same order attained in actual observation or measurement. Axioms of extension deal with the representation of such attributes as time duration, length, and mass, which can be combined, or concatenated, for multiple objects exhibiting the attribute in question.

Axioms of difference govern the measuring of intervals. Axioms of conjointness postulate that attributes that cannot be measured empirically (for example, loudness, intelligence, or hunger) can be measured by observing the way their component dimensions change about each other. Axioms of geometry govern the representation of dimensionally complex attributes by pairs of numbers, triples of numbers, or even *n*-tuples of numbers.

According to Streiner, Norman and Cairney (2015), the problem of error is one of the central concerns of measurement theory. At one time it was believed that errors of measurement could eventually be eliminated through the refinement of scientific principles and equipment. This belief is no longer held by most scientists, and almost all physical measurements reported today are accompanied by some indication of the limitation of accuracy or the probable degree of error. Among the various types of error that must be taken into account are errors of observation (which include instrumental errors, personal errors, systematic errors, and random errors), errors of sampling, and direct and indirect errors (in which one erroneous measurement is used in computing other measurements). For example, the data obtained from the WAEC on the 2012-2016 WASSCE results are measuring the mathematical ability, English Language language, Integrated Science ability and Social Studies skills of candidates that participated in the examination. These abilities are attributes which can only be ascertained through measurements which are without error which is one of the central issues in educational measurement (Streiner, Norman & Cairney, 2015)

Classical Test Theory

According to Eleje, Onah and Abanobi (2018), Classical Test Theory (CTT) has been the foundation for measurement theory for decades. The conceptual foundations, assumptions and extensions of the basic premises of CTT have allowed for the development of psychometrically sound scales in the assessment practices of educational bodies. This is due to the simplicity of interpretation which can usefully be applied to examinees achievement and aptitude test performance

Classical test theory was born only after the following three achievements or ideas were conceptualized: one, a recognition of the presence of errors in measurements, two, a conception of that error as a random variable, and third, a conception of correlation and how to index it. In 1904, Charles Spearman was responsible for figuring out how to correct a correlation coefficient for attenuation due to measurement error and how to obtain the index of reliability needed in correcting (Traub, 1997). Spearman's finding is thought to be the beginning of Classical Test Theory (Traub, 1997).

Classical test theory assumes that each person has a *true score*, *T*, that would be obtained if there were no errors in measurement. A person's true score is defined as the expected number-correct score over an infinite number of independent administrations of the test. Unfortunately, test users never observe a person's true score, but only an *observed score*, *X*. It is assumed that *observed score* = *true score* plus some *error*:

$$\underbrace{X}_{\text{observed score}} = \underbrace{T}_{\text{true score}} + \underbrace{E}_{\text{error score}}$$

Classical test theory is concerned with the relations between the three variables X, T and E in the population. These relations are used to describe the quality of test scores. In this regard, the most important concept is that of *reliability*. The reliability of the observed test scores X, which is denoted as ρ_{XY}^2 is defined as the ratio of true score variance $\sigma_T^2$to the observed score variance σ_X^2. That is, $\rho_{XY}^2 = \frac{\sigma_T^2}{\sigma_X^2}$. The variance of the observed scores can be shown to equal the sum of the variance of true scores and the variance of error scores, this is equivalent to $\rho_{XY}^2 = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2+\sigma_E^2}$

This equation, which formulates a signal-to-noise ratio, has an intuitive appeal. The reliability of test scores becomes higher as the proportion of error variance in the test scores becomes lower and vice versa. The reliability is equal to the proportion of the variance in the test scores that could be explained if the true scores are known. The square root of the reliability is the correlation between true and observed scores (Traub,1997).

Assumptions of Classical True Score Theory

According to Allen and Yen (1979, 2001) and Crocker and Algina (1986), there are seven (7) assumptions under Classical True Score Theory. These are as follows;

1. X =T + E.

Assumption 1: X=T + E, states that this observed score is the sum of two parts: T, the true score, and E, the error score, or an error of measurement.

For example, if in the WASSCE examination, an examinee named Kate’s true score is 69 but her observed score is 75 in English Language, then X is 75, T is 69 and E is +6. Thus, for any given examinee and test, T is assumed to be a fixed value, although E and X vary for the examinee on different testing occasions. In classical true-score theory, the true scores and the error scores are assumed to add (rather than to have some other relationship, such as a multiplicative one)

2. Ɛ(X)=T

Assumption 2: Ɛ(X)=T, states that the expected value (population mean) of X is T. This assumption is the definition of T. T is the mean of the theoretical distribution of X scores that would be found in repeated independent testing of the same person with the same test. For example, if Kate had had the examination an infinite number of times, then the mean of her observed scores would be 69. For this definition of T, it is assumed that the test results are independent, that is, that each testing does not influence any subsequent testing. Because this lack of contamination among test results is impossible in practice and an infinite number of tests are not available, T must remain a theoretical construct.

In the classical model, the true score is the theoretical mean of the results of repeated independent testing. Whether this true score accurately reflects some theoretical ability or characteristic is a question of test validity. That is, it deals with a theoretical distribution of observed scores over different testing occasions for one examinee on one test.

3. $\rho ET=0$

Assumption 3: $\rho ET = 0$, is extremely important for further derivations. It states that the error scores and the true scores obtained by a population of examinees on one test are uncorrelated. This assumption implies that examinees with high true scores do not have systematically more positive or negative errors of measurement than examinees with low true scores. This assumption would be violated if, for example, on one administration of a WASSCE exam, a student with low true score copied answers from those students with high true scores. This situation would create a negative correlation between true scores and error scores. If the situation were reversed a positive correlation between true scores and error scores would be produced.

4. $\rho E_1E_2 = 0$

In Assumption 4: $\rho E_1E_2 = 0$**,** E_1 is the error score for Test 1 and E_2 is the error score for Test 2. This assumption states that the error scores on two different tests are uncorrelated. That is, if a student has a positive error score on Test 1, he or she is *not* more likely to have a positive or a negative error score on Test 2. This assumption is not reasonable if the test scores are greatly affected by factors such as fatigue, practice effect, the examinee's mood, or effects of the environment.

For example, if two tests are taken in a room with many interruptions or distractions, some examinees will tend to have negative errors of measurement on both tests. Nevertheless, for an assessor to apply classical true-score theory to tests that are greatly influenced by practice effects, fatigue, or environmental conditions, it should be noted that the examiner should attempt to ensure that the testing conditions are as homogeneous as possible for all examinees on all tests over all testing occasions. This control will reduce the sizes of the errors of measurement on each test as well as the correlations of errors of measurement between tests.

5. $\rho E_1T_2 = 0$

Assumption 5: $\rho E_1T_2 = 0$, states that the error scores on one test (E_1) are uncorrelated with the true scores on another test (T_2). This assumption would be violated if Test 2 measures a personality trait or ability dimension that influences errors in Test 1. It would also be violated under the same conditions that lead to violation of Assumption 3.

6. If two tests have observed scores X_1 and X_2 that satisfy Assumptions 1 through 5, and if, for every population of examinees, $T_1 = T_2$, and $\sigma^2E_1 = \sigma^2E_2$, then the tests are called *parallel tests.*

Assumption 6 presents the definition of parallel tests. X is an observed score for one test, T is the true score, and σ^2E is the error variance. The error variance is the variance of error scores for the test among the examinees in a population. X_1, T_2, and σ^2E_2 are the observed score, the true score, and the error variance, respectively, for a second test. Assumption 6 states that the tests are parallel if $T_1 = T_2$ and $\sigma^2E_1 = \sigma^2E_{2'}$ for every population of examinees taking both tests.

Parallel tests are sometimes called *parallel test forms or parallel forms*. For σ^2E_1 to be equal to σ^2E_2, the conditions leading to errors of measurement, such as mood and environmental effects, must vary in the same way for the two tests. The definition of parallel tests also implies that parallel tests will have equal observed score means, variances, and correlations with other observed test scores.

It must be noted, however, that scores on two parallel tests are not necessarily perfectly correlated with each other. For example, a parallel test of mathematical ability will yield the same true scores, error variances, observed-score variances, and relationships with other scores, but the observed mathematical ability-test scores will not be perfectly correlated with each other unless there is no error variance. Error variance is not predictable, so if the error variance is not 0, the parallel aggression-test scores cannot be perfectly correlated.

7. If two tests have observed scores X_1 and X_2 that satisfy Assumptions 1 through 5, and if, for every population of examinees, $T_1 = T_2 = C_{12}$, where c_{12} is a constant, then the tests are called *essentially τ–equivalent tests.*

Assumption 7 states the definition of *essentially τ–equivalent tests.* The Greek letter *τ (tau)* represent the true score, T. Tests that are *essentially τ-equivalent* have true scores that are the same except for an additive constant, c_{12}. For example, on one test four examinees have a true score of 10, 11, 13, and 18. if this test and a second test are *essentially τ-equivalent* with $c_{12} = 3$, the examinees have true scores of 13, 14,16, and 21 on the second test. Unlike parallel tests, *essentially τ-equivalent tests* unequal error variances; true scores may be measured more accurately by one of the *τ-equivalent* tests than by the other.

Item Properties Under Classical Test Theory

Item Reliability

According to Davidshofer, Murphy and Charles (as cited in Onyeneke, Olorunju, Eta & Nwaonu, 2018), the goal of reliability theory is to estimate errors in measurement and to suggest ways of improving tests so that errors are minimized. The central assumption of reliability theory is that measurement errors are essentially random. This does not mean that errors arise from random processes. For any individual, an error in measurement is not a completely random event. However, across many individuals, the causes of measurement error are assumed to be so varied that measured errors act as random variables

If errors have the essential characteristics of random variables, then it is reasonable to assume that errors are equally likely to be positive or negative and that they are not correlated with true scores or with errors on other tests. According to Gulliksen (2013), it is assumed that:

1. Mean error of measurement = 0
2. True scores and errors are uncorrelated
3. Errors on different measures are uncorrelated

Reliability theory shows that the variance of obtained scores is simply the sum of the variance of true scores plus the variance of errors of measurement, that is,

$\sigma_X^2 = \sigma_T^2 + \sigma_E^2$ (Davidshofer, Murphy & Charles, 2005).

This equation suggests that test scores vary as the result of two factors:

1. Variability in true scores
2. Variability due to errors of measurement.

The reliability coefficient ρxx^1 provides an index of the relative influence of true and error scores on attained test scores. In its general form, the reliability coefficient is defined as the ratio of true score variance to the total variance of test scores. Or, equivalently, one minus the ratio of the variation of the error score and the variation of the observed score:

$$\rho xx^1 = \frac{\sigma_T^2}{\sigma_X^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2}$$

Unfortunately, there is no way to directly observe or calculate the true score, so a variety of methods are used to estimate the reliability of a test. Examples of the methods to estimate reliability include test-retest reliability, internal consistency reliability, and parallel-test reliability. Each method comes at the problem of figuring out the source of error in the test somewhat differently. The true score is assumed to be equal to the obtained score collected over two or more occasions or under two or more conditions.

Difficulty level

The difficulty level of a test is defined as the proportion of examinees who endorse or pass a dichotomous item and is termed its *p*-value. While *p* is useful as a descriptive statistic, it is also called the item's difficulty level in CTT (Lord & Novick, 2008). Items with high *p* values are easy items and those with low *p* values are difficult items. This carries very useful information for designing tests of ability or achievement. When items of varying *p* values are added up across all items, the total (also called composite) score for any individual will be based on how many items she or he endorsed or passed. Items that have *p* levels of 1.00 or 0.00 are not useful because they do not differentiate between individuals. That is, if everyone passes an item, it acts the same as does adding a constant of 1 to everyone's total score. If everyone fails an item, then a constant of 0 is added to everyone's score. The time taken to write the item, produce it, respond to it, and score it, is wasted.

Items with *p* values of 0.50, that is, 50% of the group passes the items provide the highest levels of differentiation between individuals in a group. For example, if 100 individuals are taking a test and an item has a *p*-value of 0.50, then there will be 50 × 50 (2,500) differentiations made by that item, as each person who passed is differentiated from each person who failed the item. An item with a *p*-value of 0.20 will make 20 × 80 (1,600) differentiation among the 100 test-takers. Thus, the closer the *p*-value is to 0.50, the more useful the item is at differentiating among test takers. The one caveat about the *p*-value of 0.50 being the best occurs when items are highly intercorrelated. In this, the same 50% of respondents will pass all of the items and one item, rather than the entire test, would have sufficed to differentiate the test-takers into two groups. For example, assume 200 examinees taking a 50-item test comprising of very homogeneous items. Further, assume that the *p*-value for 10 items is 0.50. The same 50% of the 200 students would pass all the items as would pass only one item.

Therefore, this test of 50 items is not any better than a test of one item at differentiating the top and bottom 50% of the class. It is because of this characteristic that test developers usually attempt to create items of varying difficulty with an average *p*-value across the items of 0.50 (Ghiselli, Campbell, & Zedek, 1981). Some tests are developed deliberately to get progressively more difficult. That is, easy questions are placed at the beginning of a test and the items become more and more difficult. The individual taking the test completes as many items as possible.

Sometimes examiners deliberately put a few easy items at the beginning of a test to get students relaxed and confident so that they continue and do as well as possible. A lot of examinees have had the negative experience of being daunted by the first question on a test with the effects of lowered motivation and heightened anxiety this can bring. Thus, examiners should be quite conscious of the difficulty level of items presented early in a testing situation.

Discrimination index

Using the *p* values (difficulty indices), discrimination indices (*D*) can be calculated for each dichotomous item. The higher the *D*, the more the item discriminates. Items with *p* levels in the midrange usually have the best *D* values and, the opportunity for *D* to be highest occurs when the *p* level for the item is at 0.50. The extreme group method is used to calculate *D*. There are three simple steps to calculating *D*.

According to Cureton (as cited in Allen, & Yen, 2001), first, those who have the highest and lowest overall test scores are grouped into upper and lower groups. The upper group is made up of the 25%–33% who are the best performers (have the highest overall test scores), and the lower group is made up of the bottom 25%–33% who are the poorest performers (have the lowest overall test scores). The most appropriate percentage to use in creating these extreme groups is to use the top and bottom 27% of the distribution, as this is the critical ratio that separates the tail from the mean of the standard normal distribution of response error.

Step two is to examine each item and determine the *p* levels for the upper and lower groups, respectively.

Step three is to subtract the *p* levels of the two groups; this provides *D*. Table 5. shows an example for a set of four items. Assume that these data are based on 500 individuals taking a test that is 50 items in length. The highest scoring 135 individuals (500 ×0.27) for the entire test and lowest scoring 135 individuals for the entire test now make up our upper and lower extreme groups.

Table 5: Example of Item Discrimination Indices

Item	p-level for the upper group	p-level for the lower group	D
1	.80	.20	.60
2	.90	.10	.80

3	.60	.55	.05
4	.10	.70	-.60

For Item 1, the upper group has a *p* level of 0.80 and the lower group has a *p* level of 0.30. The *D*, then, is 0.80 – 0.20 = 0.60. For Item 2, the *D* is 0.80; for Item 3, it is 0.05; and for Item 4, it is −0.60. Items 1 and 2 have reasonable discrimination indices. The values indicate that those who had the highest test scores were more likely to pass the items that individuals with low overall scores. Item 3 is very poor at discriminating. Although 60% of those in the upper group passed the item, almost as many (55%) in the lower group passed the item. Item 4 is interesting because it has a negative *D* value. In tests of achievement or ability, this would indicate a poor item in that those who scored most highly on the test overall were not likely to pass the item, whereas those with low overall scores were likely to pass the item. However, in assessment tools of personality, interests, or attitudes, this negative *D* is not problematic. In these types of tests, it is often of interest to differentiate between types or groups, and items with high *D* values (positive or negative) will help in differentiating those groups (Kline, 2014).

Differential item weighting

Differential item weighting occurs when items are given weight when being combined into a total score. This contrasts with unit-weighting items, where each item has a weight of 1.0 (i.e., effectively contributing equally to the total score). There are several different options for assigning weights to items (e.g., Ghiselli, Campbell & Zedek, 1981).

The first group of techniques is based on statistical grounds. For example, the reliability of items can be calculated and then the reliabilities can be used to assign different weights to the items. Items with higher reliabilities carry more weight in the total score. Another option would be to use a criterion measure and regress the criterion on the items and use the resulting regression weights as the basis for item weighting. Those with higher weights are, in turn, weighted more heavily in generating the total score. Another way to decide on weights is to run a factor analysis and use the factor loadings to assign weights to the items. Finally, item-to-total correlation coefficients can be used to weight the items (Kline,2014).

Alternatively, theory or application may drive the decision making, and items that are deemed by some decision rule (e.g., majority or consensus) to be more important or meaningful are given more weight. For example, if there is a 10-item assessment of instruction for a course, and ‘organization’ and ‘fairness’ are perceived by stakeholders to be more important than ‘punctuality’ or ‘oral skills’ then the items can be weighted accordingly when obtaining a total score on teaching effectiveness.

While much effort goes into discussing and determining differential item weights, Ghiselli, Campbell and Zedek (1981) are persuasive in arguing that differential item weighting has virtually no effect on the reliability and validity of the overall total scores. Specifically, they say that “empirical evidence indicates that reliability and validity are usually not increased when nominal differential weights are used” (p. 438). The reason for this is that differential weighting has its greatest impact when there (a) is a wide variation in the weighting values, (b) is little inter-correlation between the items, and (c) are only a few items. All three are usually the opposite of what is likely to occur in test development. That is if the test is developed to assess a single construct, then if the developer has done the job properly, items will be intercorrelated.

As a result, the weights assigned to one item over another are likely to be relatively small. Besides, tests are often 15 or more items in length, thus rendering the effects of differential weighting to be minimized. Finally, the correlation between weighted and unit-weighted test scores is almost 1.0. Thus, the take-home message is simple—don't bother to differentially weight items. It is not worth the effort.

Implications and Limitations of Classical Test Theory Assumptions

Embretson and Reise (2000) review the implications (ramifications) of CTTs. These are as follows:

1. The standard error of measurement of a test is consistent across an entire population. That is, the standard error does not differ from person to person but is instead generated by large numbers of individuals taking the test, and it is subsequently generalized to the population of potential test-takers. Besides, regardless of the raw test score (high, medium, or low), the standard error for each score is assumed to be the same.
2. As tests become longer, they become increasingly reliable. This happens because, in domain sampling, the sample of test items that makes up a single test comes from an infinite population of items. Larger numbers of items better sample the universe of items and statistics generated by them (such as mean test scores) are more. Multiple forms of a test (e.g., Form A and Form B) are considered to be

parallel only after much effort has been expended to demonstrate their equality Gulliksen (as cited in Guion, 2011).

3. The important statistics about test items (e.g., their difficulty) depend on the sample of respondents being representative of the population. Statistics generated from the sample can only be confidently generalized to the population from which the sample was drawn.
4. True scores in the population are assumed to be (a) measured at the interval level and (b) normally distributed. When these assumptions are not met, test developers convert scores, combine scales, and do a variety of other things to the data to ensure that this assumption is met.
5. In CTT, if item responses are changed (e.g., a test that had a 4-point Likert-type rating scale for responses now uses a 10-point Likert-type rating scale for responses), then the properties of the test also change.
6. If item responses are dichotomous, CTT suggests that they should not be subjected to factor analysis. This poses problems in establishing the validity for many tests of cognitive ability, where answers are coded as correct or incorrect.
7. Once the item stems are created and subjected to content analysis by the experts, they often disappear from the analytical process. Individuals may claim that a particular item stem is biased or unclear, but no statistical procedures allow for comparisons of the item content, or stimulus, in CTT.

Item Analysis within Classical Test Theory

DeVellis (2016) notes that assessment of test items under CTT uses approaches that have been developed within the theoretical framework of CTT.

At the outset, it has been assumed that a test is composed of several items and has been administered to a sample of examinees. Once the respondents have completed the test, the analyses can begin. There are several pieces of information that can be used to determine if an item is useful and/or how it performs concerning the other items on the test.

Whenever a dataset is examined, descriptive statistics come first, and the most common of these are the mean and variance. The same is true for test items. The means and standard deviations of items can provide clues about which items will be useful and which ones will not. For example, if the variance of an item is low, this means that there is little variability on the item, and it may not be useful. It is not common to examine item-level descriptive statistics in most research applications but in creating and validating tests it is a crucial first step.

Generally, the higher the variability of the item and the more the mean of the item is at the centre point of the distribution, the better the item will perform. For dichotomous items, the mean is equal to the proportion of individuals who endorsed/passed the item (denoted p). The variance of a dichotomous item is calculated by multiplying $p \times q$ (where q is the proportion of individuals who failed, or did not endorse, the item). The standard deviation, then, of dichotomous items is simply the square root of $p \times q$.

For example, if 500 individuals respond to a yes/no item and 200 respond ‘yes’ then the p-value for that item is 200/500, or 0.40. The q is 0.60 (1.0 – 0.40 = 0.60). The variance of the item is 0.24 (0.40 × 0.60 = 0.24) and the standard deviation is the square root of 0.24, or 0.49.

Strengths of Classical Test Theory

CTT has remained popular despite the emergence of newer measurement approaches, because of several advantages. One is familiarity with its basic concepts. That is, researchers who have had any exposure to measurement theory are likely to have encountered CTT.

Also, most of the scales that are available and most of the descriptions of those scales are based on principles of CTT. The nearly ubiquitous use of coefficient alpha as an indicator of reliability illustrates this point.

Another advantage is that the methods are reasonably tractable. For example, programs for performing factor analyses and for computing coefficient alpha are widely available and relatively easy to use. Major statistical packages routinely include components for performing those analyses.

A third advantage is that the underlying model fits certain types of instruments well, for example, a scale that adds together the scores from items designed as roughly equivalent indicators of a common underlying variable. A scale that has been developed under CTT should consist of items that do an equally good job of detecting the true score of the variable of interest. That is, a score value on one item should mean the same thing as the same score value on another item of the same scale. By adding the multiple items together, the effect of errors associated with each item is attenuated. This set of scale characteristics is common. CTT has been widely used in the social sciences because the data of interest often fit this pattern. Using CTT-based measures of this sort will often yield satisfactory results for many types of research investigations.

Another important advantage of the CTT approach is that individual items need not be optimal. Items that relate only modestly to the underlying variable can be used successfully by having many of them. Sometimes it can be difficult to create items that, individually, optimally capture the underlying variable. If the correlations among items are weak (like they will be when the individual items are related only modestly to the underlying variable), adding items can offset this problem and theoretically, just about any desired level of reliability can be achieved. Finally, some limitations of CTT are better documented in theory than in fact.

Limitations of Classical Test Theory

The classical approach also has some notable disadvantages. One is that because redundancy is the root of precision under this model, scales are typically long and items often seem quite similar. In some cases, the effort to develop items that correlate strongly with each other can result in superficial similarities. When this occurs, not only the variable of interest but irrelevant item characteristics such as grammatical structure may be common across items.

The "true score" becomes an undifferentiated mixture of all the characteristics the items have in common, including both the substantive variable of interest and superficial features that are not of interest and were not the intended target of the measure. CTT methods have difficulty differentiating between common themes across items that are important to the variable of interest and common themes of this more superficial type.

CTT-based methods do not involve the rigorous scrutiny of item characteristics that certain other methods involve. This can be a shortcoming.

For example, CTT-based scales may be prone to differential sensitivity at the centre, relative to the extremes, of the score range. Thus, a 2-point difference at the centre of the range of scores may represent a smaller true score difference than a 2-point spread at one of the extremes.

Another disadvantage is that parameter estimates under CTT depend on the sample of individuals studied. Properties of items (eg., difficulty and discrimination) and scales (eg., coefficient alpha) developed under CTT are based on correlations computed on the sample.

Different samples with different variances will not yield equivalent data or data that can easily be compared across samples. The disadvantages of CTT may pose a greater problem in some contexts than in others. One domain that merits special consideration is research involving comparisons across populations.

Item Response Theory (IRT)

IRT is also sometimes called latent trait theory. This is a modern test theory (as opposed to classical test theory). It is not the only modern test theory, but it is the most popular one and is currently an area of active research. IRT requires stronger assumptions than classical test theory.

Nevertheless, IRT was originally developed to overcome the limitation of Classical Test Theory. While the concept of the item response function has been evolving since 1950, the pioneering work on IRT as a theory began after the 1960s. Two of the pioneers were Frederic M. Lord, from the Educational Testing Service (ETS) Princeton, NJ, US, and a Danish Mathematician George Rasch.

In IRT, the true score is defined on the latent trait of interest rather than on the test, as is the case in classical test theory. IRT is popular because it provides a *theoretical justification* for doing lots of things that classical test theory does not do, according to Lord (1980) and Hambleton and Swaminathan, (as cited in Retnawati, 2008)). Any theory of item responses supposes that, in testing situations, examinee performance on a test can be predicted (or explained) by defining examinee characteristics, referred to as *traits,* or *abilities;* estimating scores for examinees on these traits (called ability scores) and using the scores to predict or explain item and test performance according to Lord and Hickem Novick, (as cited in DeVellis, 2016).

Since traits are not directly measurable, they are referred to as *latent traits* or *abilities.* An item response model specifies a relationship between the *observable* examinee test performance and the *unobservable* traits or abilities assumed to underlie performance on the test. Within the broad framework of item response theory, many models can be operationalized because of a large number of choices available for the mathematical form of the item characteristic curves. But whereas item response theory cannot be shown to be correct or incorrect, the appropriateness of particular models with any set of test data can be established by conducting suitable goodness of fit investigation (Hambleton & Swaminathan,1985).

Assumptions of Item Response Theory

Unidimensionality

It is commonly assumed that only one ability or trait is necessary to "explain," or "account" for examinee test performance. Item response models that assume a single latent ability are referred to as *unidimensional.*

Unidimensionality in item response models (Wright, 1968; Bock & Wood as cited in Rowntree, 2015) implies that in any given set of test items that have been fitted to an item response model (that is, items designed to measure a known trait or characteristic), it should be possible to estimate an examinee's ability on the same ability scale from *any* subset of items in the domain of items that have been fitted to the model. The domain of items needs to be homogeneous in the sense of measuring a single ability: If the domain of items is too heterogeneous, the ability estimates will have little meaning. Regardless of the number of items administered (if the number is not too small) or the statistical characteristics of the items, the ability estimate for each examinee will be an asymptotically unbiased estimate of true ability, provided the item response model fits the dataset. Any variation in ability estimates obtained from different sets of test items is due to measurement errors only.

Ability estimation independent of the choice (and number) of items represents one of the major advantages of item response models. Hence, item response models provide a way of comparing examinees even though they may have taken quite different subsets of test items. Once the assumptions of the model are satisfied, the advantages associated with the model can be gained (Hambleton & Swaminathan,1985).

However, this assumption cannot be strictly met because there are always other cognitive, personality, and test-taking factors that impact on test performance, at least to some extent. These factors might include the level of motivation, test anxiety, ability to work quickly, knowledge of the correct use of answer sheets, and other cognitive skills in addition to the dominant one measured by the set of test items. What is required for this assumption to be met adequately by a set of test data is a "dominant" component or factor that influences test performance. This dominant component or factor is referred to as the ability measured by the test (Hambleton & Swaminathan,1985).

Hambleton and Swaminathan (1985) add that, at times, researchers are interested in monitoring the performance of individuals or groups on a trait over some time. For example, at the individual (or group) level, interest may be centred on the amount of individual (group) change in English Language or mathematical ability over several years. Traub (1983) has indicated that the nature of training and education can influence the dimensionality of a set of test items. For example, concerning education,

Traub (1983) has noted:

> The curriculum and the method by which it is taught vary from student to student, even within the same class. Out-of-school learning experiences that are relevant to in-school learning vary widely over students. Individual differences in previous learning, quality of sensory organs, and presumably also the quality of neural systems contribute if they do not define, individual differences in aptitude and intelligence (p.70).

It seems reasonable then to expect differences of many kinds, some obvious, some subtle, in what different students learn, both in school and outside. How these differences are translated into variation in the performance of test items that themselves relate imperfectly to what has been taught and learned and thus into the dimensionality of inferred latent space. Hambleton and Swaminathan (1985) continue to say that, the assumption of a unidimensional latent space is a common one for test developers since they usually desire to construct unidimensional tests to enhance the interpretability of a set of test scores. Factor analysis can be used to check the reasonableness of the assumption of unidimensionality with a set of test items (Hambleton & Traub, 1973).

Local Independence

This assumption states that an examinee's responses to different items in a test are statistically independent. For this assumption to be true, an examinee's performance on one item must not affect, either for better or for worse, his or her responses to any other items in the test. For example, the content of an item must not provide clues to the answers of other test items. When local independence exists, the probability of any pattern of item scores occurring for an examinee is simply the product of the probability of occurrence of the scores on *each* test item. For example, the probability of the occurrence of the five-item response pattern U = (1 0 1 1 0), where 1 denotes a correct response and 0 an incorrect response, is equal to P_i (1 - $P2)'$ $P3$ • $P4$.(1 - $P5)$, where P_i is the probability that the examinee will respond correctly to item i and 1 – P_i is the probability that the examinee will respond incorrectly. But test data must satisfy other properties if they are to be consistent with the assumption of local independence.

Also, the test data must be unidimensional. Performance across test items at a fixed ability level will be correlated when a second ability or more than two abilities are being measured by the test items. For examinees located at an ability level, examinees with higher scores on a second ability measured by a set of test items are more apt to answer items correctly than examinees with lower scores on the second ability.

If U*i* = 1, 2, ..., *n*, represent the binary responses (1, if correct; 0 if incorrect) of an examinee to a set of *n* test items, *Pi* = the probability of a correct answer by an examinee to item *i*, and *Qi* = 1 - *Pi'* then the assumption of local independence leads to the following statement:

Prob [U_1=u_1, U_2 =u_2,Un=$u_n|\theta$] =*Prob* [U_1 = $u|\theta$]

Prob [U_2 = $u_2|\theta$] ... *Prob* [Un = u_n $|\theta$].

If we set Pi (θ) = *Prob* [U_i. = $1|\theta$] *and Qi*(θ) = *Prob* [U_i. =$0|\theta$],

then Prob [U_1=u_1, U_2 =u_2, ...,Un=$u_n|\theta$]

$$=P_I(\theta)^{u_i})Q_1(\theta)^{1-u_i})P_2(\theta)^{u_2})Q_2(\theta)^{1-u_2})\ldots P_n(\theta)^{u_n})Q_n(\theta)^{1-u_n})$$

$$= \prod_{i=1}^{n} P_i(\theta)^{u_i}Q_i(\theta)^{1-u_i})$$

In other words, the assumption of local independence applies when the probability of the response pattern for each examinee is equal to the product of the probabilities associated with the examinee response to each item.

One should note that the assumption of local independence for the case when θ is unidimensional and the assumption of a unidimensional latent space are equivalent.

Based on the unidimensionality assumption, a set of test items should measure a common ability. Then, for examinees at a fixed ability level θ, item responses are statistically independent. For fixed ability level *θ,* if items were not statistically independent, it would imply that some examinees have higher expected test scores than other examinees of the same ability level. Consequently, more than one ability would be necessary to account for examinee test performance. This is a clear violation of the original assumption that the items were unidimensional.

For the assumption of local independence, item responses are statistically independent for examinees at a fixed ability level. Therefore, only one ability is necessary to account for the relationship among a set of test items. It is important to note that the assumption of local independence does *not* imply that test items are uncorrelated over the total group of examinees (Lord & Novick, 1968). Positive correlations between pairs of items will result whenever there is variation among the examinees on the ability measured by the test items. But item scores are uncorrelated at a fixed ability level. Because of the equivalence between the assumptions of local independence and of the unidimensionality of the latent space, the extent to which a set of test items satisfies the assumption of local independence can also be studied using factor analytic techniques.

Three Basic Components of IRT

According to Embretson and Reise (2013), IRT is, generally, made up of three basic components.

1. Item Response Function (IRF): Mathematical function that relates the latent trait to the probability of endorsing an item.
2. Item Information Function: An indication of item quality. That is, the item's ability to differentiate among respondents.
3. Invariance: Position on the latent trait that can be estimated by any items with known IRFs.

IRT Models

There is a wide array of mathematical models that have been used in the analysis of educational and psychological test data. Each model consists of (1) an equation linking (observable) examinee item performance and a latent (unobservable) ability and (2) several of the assumptions. To date, most of the IRT models have been developed for use with binary-scored aptitude and achievement test data. One of how IRT models can be classified is based on the examinee responses to which they can be applied. Three response levels are common: dichotomous, polytomous and continuous.

Over the years multiple-choice test items with dichotomous scoring have become the main mode through which educational assessments have been made (Hambleton & Swaminathan,1985). However, there are other types of items for which dichotomous scoring systems are used. These are true-false, short answer, sentence completion, and matching items. With psychological assessments, dichotomous data are often obtained from "true-false," "forced-choice" or "agree-disagree" rating scales. Even free-response data can be subjected to a dichotomous scoring system. The majority of the presently available item response models handle binary-scored data. To use these models, it is sometimes a force to use a binary scoring system on polytomous response data. This may be done by combining the available scoring categories so that only two are used.

Somewhat less common in present measurement practices are polytomous or polychotomous scoring systems. These systems arise, for example, when scoring weights are attached to the possible responses to multiple-choice test items. The scoring system for essay questions is usually polychotomous as is the scoring system for Likert scales. With essay questions, points are assigned either to reflect the overall quality of an essay or to reflect the presence of desirable characteristics such as correct spelling, grammatical structure, originality, and so on. The nominal response and graded response models are available to handle polychotomous response data (Hambleton & Swaminathan,1985).

Finally, continuous scoring systems occasionally arise in practice. Here, an examinee or rater places a mark (V) at a point on some continuous rating scale according to Samejima (as cited in Hambleton & Swaminathan, 2013). Even though the responses from this type of rating scale can easily be categorized and fit a polytomous response model, some information is lost in the process if item response theory relates characteristics of items (item parameters e.g., difficulty level and discrimination) and characteristics of individuals (individual parameters e.g. ability) to the probability of a student giving a correct response to an item. IRT is a model based in which probability of answering an item correctly is related to the cognitive ability of a student through Item Characteristic Curve (ICC).

Item Characteristic Curve Models

These include the one-, two-, three- and four-parameter logistic (PL) models. All these models assume a single underlying latent trait or ability. All four of these models have an item difficulty parameter denoted *b*

The Rasch Model

The Rasch model is often considered to be the 1PL IRT model. However, proponents of Rasch modelling prefer to view it as a completely different approach to conceptualizing the relationship between data and theory (Andrich, 1989). Like other statistical modelling approaches, IRT emphasizes the primacy of the fit of a model to observed data, while the Rasch model emphasizes the primacy of the requirements for fundamental measurement, with adequate data-model fit being an important but secondary requirement to be met before a test or research instrument can be claimed to measure a trait. Operationally, this means that the IRT approaches include additional model parameters to reflect the patterns observed in the data (e.g., allowing items to vary in their correlation with the latent trait), whereas in the Rasch approach, claims regarding the presence of a latent trait can only be considered valid when both (a) the data fit the Rasch model, and (b) test items and examinees conform to the model. Therefore, under Rasch model, misfitting responses require diagnosis of the reason for the misfit and may be excluded from the dataset if one can explain substantively why they do not address the latent trait (Smith, 1990) Thus, the Rasch approach can be seen to be a confirmatory approach, as opposed to exploratory approaches that attempt to model the observed data. As in any confirmatory analysis, care must be taken to avoid confirmation bias. The presence or absence of a guessing or pseudo-chance parameter is a major and sometimes controversial distinction. The IRT approach includes a left asymptote parameter to account for guessing in multiple-choice examinations, while the Rasch model does not because it is assumed that guessing adds randomly distributed noise to the data. As the noise is randomly distributed, it is assumed that, provided enough items are tested, the rank-ordering of persons along with the latent trait by raw score will not change, but will simply undergo a linear rescaling. By contrast, three-parameter IRT achieves data-model fit by selecting a model that fits the data, (Zwick, Thayer, & Wingersky, 1995), at the expense of sacrificing specific objectivity.

In practice, the Rasch model has at least two principal advantages in comparison to the IRT approach. The first advantage is the primacy of Rasch's specific requirements, which (when met) provides *fundamental* person-free measurement (where persons and items can be mapped onto the same invariant scale). Another advantage of the Rasch approach is that estimation of parameters is more straightforward in Rasch models due to the presence of enough statistics (Rasch, 1980).

One-parameter logistic model (1 PL)

The one-parameter logistic model which is also known as Rasch model is the simplest and is one of the most widely used IRT models. Item characteristics curves for the one-parameter logistic model are given by the equation: $P_i(\theta) = \frac{e^{\theta - b_i}}{1+e^{\theta - b_i}}$

where Pi (θ) represents the probability of a correct response given to the ith item and b*i* is the difficulty value of the ith item. In this, item difficulty, 'b' is used in selecting items. The parameter 'b' for an item is the point on the ability scale where the probability of a correct response is 0.5. This parameter is a location parameter, indicating the position of the ICC concerning the ability scale. The greater the value of the 'b' parameter, the higher the ability that is required by an examinee to have a 50% chance of getting the item right.

Difficult items are located to the right or the higher end of the ability scale while easy items are located to the left or the lower end of the ability scale (Bhaduri, & Singh, 2011). The simplest IRT model for a dichotomous item has only one item parameter. The item response function (i.e. the probability of a correct response given the single item parameter bi and the individual ability level θ) is shown in Figure 6. The function shown in the graph is known as the one-parameter logistic function.

Its values remain between 0 and 1 for any argument between −∞ and +∞ and this makes it appropriate for predicting probabilities, which are always numbers between 0 and 1. Besides, it is not at all a complicated function. The one-parameter logistic (1PL) model predicts the probability of a correct response from the interaction between the individual ability θj and the item parameter bi. The parameter bi is called the location parameter or, more aptly, the difficulty parameter.

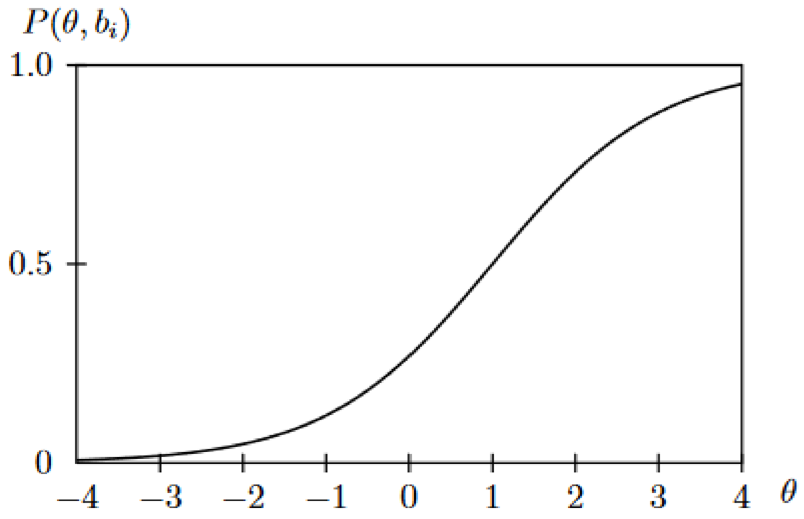

Figure 6: The item response function of the one-parameter logistic (1PL) model

IRT essentially equates the ability of the person with the difficulty of the test problem. One can find the position of *bi* on the common ability or difficulty axis at the point for which the predicted probability *Pij*(*θj-bi*) equals 0.5. This is illustrated in Figure 7. The item whose item response function is shown on the figure happens to have a difficulty of 1.

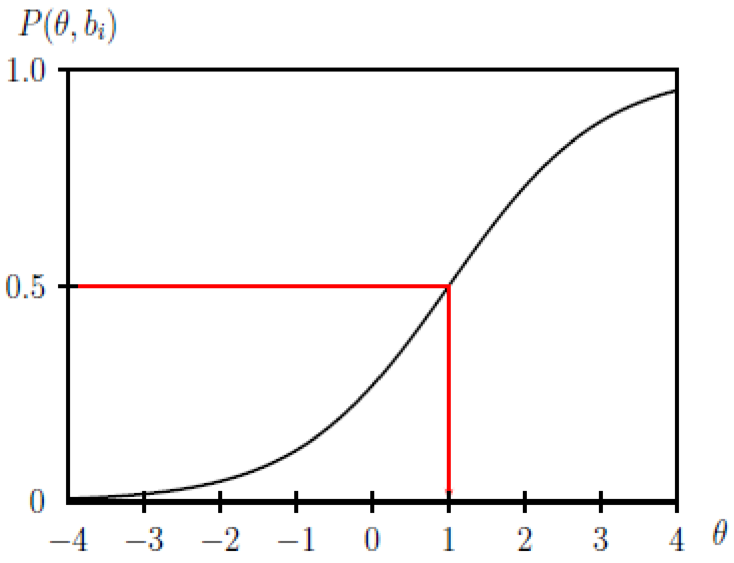

Figure 7: Locating the difficulty of an item on the ability or difficulty axis

The item information function of the 1PL model

Information functions have a prominent role in IRT. The *test information function* is related to the accuracy with which one can estimate ability. In other words, it measures the success to which one can do business as a psychometrician. For the time being, one item is considered, and its *item information function* examined. Any item in a test provides some information about the ability of the examinee, but the amount of this information depends on how closely the difficulty of the item matches the ability of the person. In the case of the 1PL model, this is the only factor affecting item information, while in other models it combines with other factors. The item information function of the 1PL model is

Ii(θ; *bi*) = *Pi*(θ; *bi*)*Qi*(θ; *bi*) (Ramp et al, 2009).

It is easy to see that the maximum value of the item information function is 0.25. It occurs at the point where the probabilities of a correct and an incorrect response are both equal to 0.5. In other words, items in the 1PL model are most informative for examinees whose abilities was equal to the difficulty of the item. As ability becomes either smaller or greater than the item difficulty, item information decreases. This is visible in Figure 8. The most important practical implication of all this is that items need to have different difficulty levels if one is to achieve good measurement for people having all sorts of different abilities.

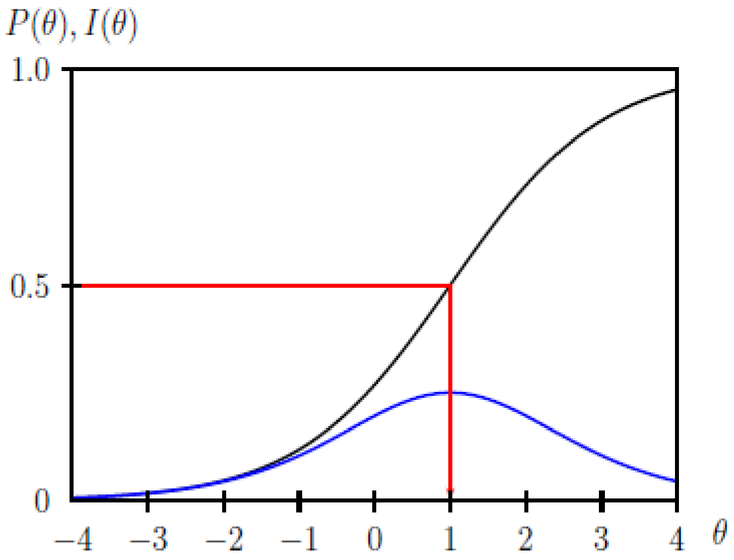

Figure 8: Item response function and item information function of the 1PL model

The two-parameter logistic (2PL) model

The *two-parameter logistic (2PL) model* predicts the probability of a correct response to any test item from ability and two item parameters. The item response function of the 2PL model is defined as

$P(Y_{is} = 1|\theta_s) = \frac{\exp(1.7a_i(\theta_s - b_i)}{1+exp[1.7a_i(\theta_s - b_i)]}$) (Ramp et al, 2009).

In this model, b_i is the difficulty parameter. The parameter, a_i, is called the *discrimination parameter*.

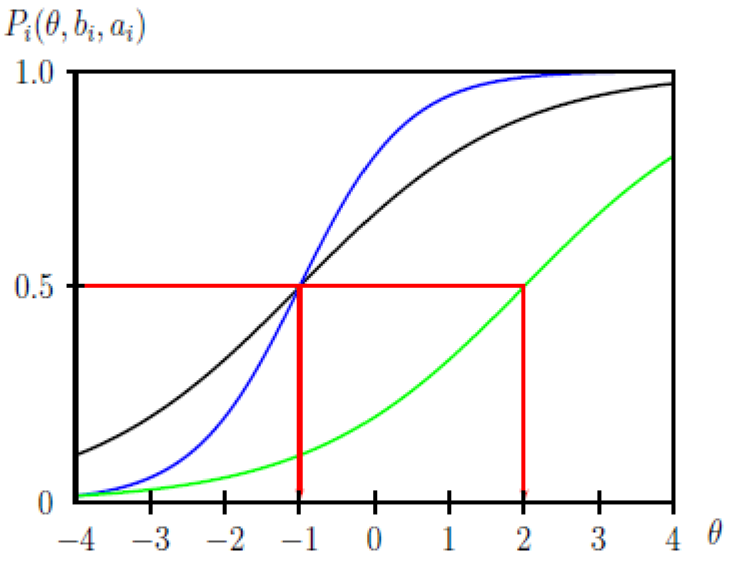

Figure 9: The item response functions of two 2PL items

Figure 9 depicts two items of a test having the same difficulty of -1.0. The difficulty parameter is found at the ability level that yields a probability of getting an item correct as 0.5. However, the blue curve is much *steeper* than the black one. This is because the item with the blue curve has a higher discrimination parameter than the item with the black curve. The discrimination parameters a_i are sometimes called *slope parameters*. The item difficulties are also known as *location parameters*. The slope of the 2PL item response function at b is equal to a = 4. The green curve has the same slope as the black one, but it is shifted to the right, hence the item with the green curve has the same discrimination parameter as the item with the black curve but a higher difficulty. The blue curve and the black curve cross. It means that the item with the black curves is the more difficult one for examinees of low ability, while the item with the blue curve is the more difficult one for examinees of higher ability. (Swaminathan, Hambleton, & Rogers, 2006).

The three-parameter logistic (3PL) model

The 3PL model is not a logistic model. Rather, it is a 2PL model whose item response function has been refashioned such that its lower asymptote is larger than zero. In other words, the probability of a correct response no longer approaches zero as true ability goes to -∞ (Ramp et al, 2009).

Instead, it approaches some positive value of 1/k; where *k* is the number of response categories in the multi-choice item. The argument is that examinees of very low ability will very likely switch to random guessing, and random guessing would enable them to choose the correct response with a probability of 1/*k* (Fox, 2010). The item response function of the 3PL model is

$$P(\theta_s, a_i, b_i, c_i) = c_i + (1 - c_i)\frac{\exp(a_i(\theta_s - b_i)}{1 + exp[a_i(\theta_s - b_i)]} \text{ (Ramp et al, 2009).}$$

The a_i-the parameter is the discrimination, b_i is the difficulty parameter and the third parameter, c_i , sets the lower asymptote, i.e., the probability of a correct response when true ability approaches -∞. The part multiplied with (1 - c_i) is the IRF of the 2PL model (with different numeric values for a_i and b_i).

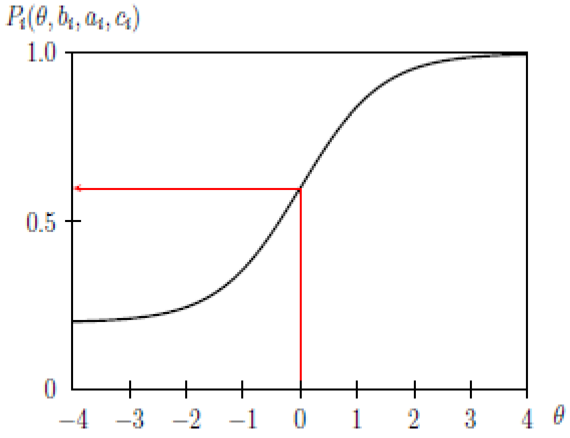

Figure 10: Item response function of a 3PL item

Figure 10 shows the IRF for a 3PL item with a_i = 1.4, b_i = 0, and c_i = 0.2. The lowest ability on the graph is only -4 and -∞ is a bit farther off than that. As in the1PL and the 2PL models, the curve turns from convex to concave at θ= b, but the probability of a correct response at θ = $b_i$is no longer 0.5. It is equal to c_i+(1-c_i)=2 = 0.2+0.4 = 0.6, instead. Furthermore, the slope at b_i is (1 - c_i) a_i/4 rather than a_i/4.

The four-parameter logistic (4PL) model

Barton and Lord (1981) introduced an upper asymptote parameter, expressed by d, into the 3PL model, resulting in the 4PL model:

$P_{4PL}(\theta) = C + (d_i - c_i)\frac{1}{1+exp[-1.702a_i(\theta-b_i)]}$ in which $P_{4PL}(\theta)$ ranges from the lower asymptote c_i to 1, and $P_{4PL}(\theta)$ ranges from c_i to the upper asymptote parameter d_i. The d_i-parameter is described as the item upper asymptote of carelessness.

Uses of IRT

The reason for the change of emphasis by the measurement community from classical to item response models is because of the benefits obtained through the application of the later to measurement problems. When IRT is used appropriately, it can increase the efficiency, accuracy or usefulness of a wide variety of measurement processes. The following advantages can be obtained by one or another of the classical procedures, but the IRT models provide a unified framework and system that facilities their accomplishments. According to Lee, Palazzo, Warnakulasooriya and Pritchard (2008), IRT has many strengths as compared to its limitations. These strengths include:

Test Construction**:** An IRT model can be used to create a pool of items that have known statistical characteristics, including descriptions of how well each item is measuring students at each ability level. The psychometric properties of any test created from the pool can be readily predicted for different groups of students, even when those students have not taken that test. These properties include number correct score means, standard deviations, and distributions, as well as reliabilities, standard errors of measurement, and item p-values. Thus, the IRT models are ideal for computer-assisted item selection systems which give the test constructor suggestions for items that meet various needs and instant feedback on the effects of alternative item selections.

Matrix Sampling: In matrix sampling procedure different groups of the sample from the same population are subjected to different sets of test items thereby decreasing the time taken for completion of the test by the examinee and also not subjecting the examinee to a huge examination load bringing in the factor of fatigue. The use of IRT is well suited to matrix sampling, in which multiple test forms are created and administered to different students using a random sampling procedure. Matrix sampling is useful for obtaining group-level data on a broad sample of items in a specified content domain while limiting the testing time required of each student.

It helps to improve the quality of the tests and scales produced: IRT helps in extracting information about the person from the item responses, although it is more intricate than the simple correct number count. The person's scale score is little affected by adding or deleting items from the test. In addition to this, it helps to develop a test that is tailored to proficiency with easy questions for low ability students and difficult questions for high ability students. The item-person map provides feedback to the system.

Test Equating and Administration: The test equating model helps to obtain comparable scores when more than one test form is used in a test administration. The process of equating is used to ensure that scores resulting from the administration of the multiple forms can be used interchangeably. Traditional equating procedures show how to translate a score on one test to a comparable score on another test. With traditional procedures, the intact tests are administered to students to collect the equating information.

However, IRT procedures are more flexible, because the item, rather than the whole test, can be the unit of scaling of equating. Once item scaling has been accomplished, the items can be selected for a variety of test configurations. IRT model shows how to aggregate the item information to get test information and how to produce equated ability scores for different tests. Equating test scores is a statistical procedure. IRT equating has some important advantages. It offers tremendous flexibility in choosing a plan for linking test forms. It is especially useful for adaptive testing and other situations where each test-taker gets a custom-built test form.

It handles a wider range of response modes*:* In case of polytomous item response, IRT helps to draw the curve on the probability of success with the ability of the student.

It helps to develop an understanding of Differential Item Functioning: A study of Differential Item Functioning (DIF) has been a point of interest in the interpretation of statistics resulting from the IRT. Distinct groups within a population may have higher or lower estimated scores on a construct that the scale is attempting to measure. This means that the scale can discriminate between these groups based on their estimated level of the underlying trait.

However, sometimes items are found to behave differently in distinct groups such as gender or language. In other words, two examinees with the same latent trait value but differing in other characteristics may have different probabilities of response. This tendency is referred to as DIF and IRT helps to analyze this differential item functioning.

Test Scoring and Interpretation: The user of any test score should know the amount of measurement error it is likely to contain. Classical test theory produces a single standard error of measurement (SEM) that applies to all scores obtained from a test. However, IRT goes beyond the classical approach to provide a different SEM for each score. For example, if a test emphasizes easy items, scores for low ability students will be more accurate than those for high ability students. Indices are available that reflect the appropriateness of a test in measuring a given student. For example, if a student's score appears to have been influenced by substantial guessing or non-completion of the test, the score can be flagged. In some cases, adjusted scores can be provided. Testers can use IRT in test scoring to increase accuracy by considering the statistical characteristics of the particular items that the student answered correctly.

Such scoring methods can be particularly helpful in increasing score accuracy for low-scoring students who have taken multiple-choice tests. From a student's score on a subset of the items in an item pool, IRT can yield that student's probability of passing any of the other items in the pool. Thus, scores can be referenced extensively to content, enhancing interpretation and instructional decisions.

Limitations

It is frequently argued that the benefits of the IRT model are only realized to the degree that the data meet the appropriate assumptions and the degree of the model to data fit. The evaluation of fit in IRT modelling is most challenging. Item response models also have technical and practical limitations (Kline, 2005). These include:

1. These models are complex, and the model parameter estimation problem does arise in practice.
2. The model fit can also be a problem and difficult to address at times.
3. The model fit needs more sophisticated software and expertise to handle.

Differential Item Functioning (DIF)

DIF refers to differences in the functioning of items across groups, oftentimes demographic, which are matched on the latent trait or more generally the attribute being measured by the items or test (Cho & Cohen, 2010). It is important to note that when examining items for DIF, the groups must be matched on the measured attribute, otherwise, this may result in inaccurate detection of DIF (Camilli,2006). To create a general understanding of DIF or measurement bias, Osterlind and Everson (2009) offered an example. In this case, Y refers to a response to a particular test item which is determined by the latent construct being measured. The latent construct of interest is referred to as theta (θ) where Y is an indicator of θ which can be arranged in terms of the probability distribution of Y on θ by the expression $f(Y)|\theta$ (Osterlind & Everson, 2009).

Response Y is conditional on the latent trait (θ). Because DIF examines differences in the conditional probabilities of Y between groups, these groups are usually labelled as the "reference" and "focal" groups. A typical practice in the literature is to designate the reference group as the group who is suspected to have an advantage while the focal group refers to the group anticipated to be disadvantaged by the test (Holland & Wainer,1993).

Given the functional relationship f(Y)|θ and under the assumption that there are identical measurement error distributions for the reference and focal groups it can be concluded that under the null hypothesis: $f\ (Y = 1 \mid \theta,\ G = r) = f\ (Y = 1 \mid \theta,\ G = f)$ with G corresponding to the grouping variable, "r" the reference group, and "f" the focal group represents an instance where DIF is not present.

In this case, the absence of DIF is determined by the fact that the conditional probability distribution of Y is not dependent on group membership. To illustrate, consider an item with response options 0 and 1, where Y = 0 indicates an incorrect response, and Y = 1 indicates a correct response. The probability of correctly responding to an item is the same for members of either group. This indicates that there is no DIF or item bias because members of the reference and focal group with the same underlying ability or attribute have the same probability of responding correctly. Therefore, there is no bias or disadvantage for one group over the other (Lord, 1980).

Ackerman (1992), considered the instance where the conditional probability of Y is not the same for the reference and focal groups. In other words, members of different groups with the same trait or ability level have unequal probability distributions on Y. Once controlling for θ, there is a clear dependency between group membership and performance on an item. For dichotomous items, this suggests that when the focal and reference groups are at the same location on θ, there is a different probability of getting a correct response or endorsing an item. Therefore, the group with the higher conditional probability of correctly responding to an item is the group advantaged by the test item. This suggests that the test item is biased and functions differently for the groups, therefore exhibits DIF.

Swaminathan and Rogers (1990) distinguished two types of non-uniform DIF as crossing and non- crossing DIF. Ability typically falls in the range -3 to +3 on the ability level scale in item response theory. When the IRFs cross in the middle of this range, a type of non-uniform DIF occurs that is analogous to a disordinal interaction in the analysis of variance (ANOVA) models (Swaminathan & Rogers, 1990). When the IRFs cross outside this range or when the IRFs are not parallel but do not cross (a situation that may occur with the three-parameter IRT model), a type of nonuniform DIP analogous to ordinal interaction occurs. Li and Stout (1993) termed these two types of DIF ‘nondirectional’ and ‘unidirectional’ respectively.

Test bias, impact, DIF, and item bias

Gierl (2005) stated that ‘bias occurs when tests yield scores or promote score interpretations that result in different meanings for members of different groups’ (p. 52). The group membership often refers to demographic information such as ethnicity, gender, language, or socioeconomic status. Test bias occurs when there is a systematic difference in the meaning of the test scores for different groups (Kim, 2003). According to AERA et al., (1999), test bias is attributed to the construct-irrelevant variances that affect the test scores of different group memberships. In other words, this means that there is a threat to the validity of inferences made from the test scores. There are instances where court decisions have been made to ban the use of certain tests for admission purposes because there is evidence that these tests are biased against female and /or minority examinees (Linn & Drasgow, 1987). It should be noted that test bias is a characteristic of the test as a whole and it is analysed at the level of the total score. It is not a compound concept because the characteristics of a sum of items can be different from the characteristics of items. For example, if the number of items that favour a group in a test is approximately equal to the number of items that are against the group, the bias might cancel out when the scores are summed indicating no bias in the test (Clauser & Mazor,1998).

There are terms related to testing bias such as impact, item bias and DIF that need to be clarified. According to Dorans and Holland (1993), a distinction can be made between impact and DIF. Impact refers to a difference in performance between groups on a test or an item. When there is a difference in the mean total score between groups, then this difference is called the 'impact', but only if the differences in test performance reflect the *true* differences in the overall ability in the distributions. On the other hand, DIF refers to differences in item functioning after groups have been matched concerning the ability or attribute that the item purportedly measures. DIF is an unexpected difference among groups of examinees who are supposed to be comparable concerning the attribute measured" (Dorans & Holland, 1993, pp.125-128). In short, the phrases 'true difference' and 'unexpected difference' can be used to distinguish between impact and DIF.

In the test development process, it is recommended that items flagged as having DIF undergo a review process by content experts to find out the source of unexpected group differences. If the source of the unexpected group differences is due to an unintended construct that is irrelevant to the construct being measured, then the item is considered to be biased (Camilli & Shepard, 1994). In another way, one can say that for item bias to occur, DIF is required but not enough. Furthermore, DIF is a statistical term and item bias is a judgmental term.

Zumbo (2007) when reflecting on the advancement of DIF research pointed out that the first generation and the transition to the second generation of DIF research focused on the terminology of DIF, bias, and impact as the framework for concept formation. In the first generation of DIF, the most commonly used term was item bias. It was only in the transition to the second generation of DIF research that there was a widespread acceptance of the term DIF rather than item bias and the distinction between impact and bias. For this reason, (Zumbo, 2007, p. 224) distinguished the three terms as follows:

1. DIF was the statistical term that was used to simply describe the situation in which persons from one group answered an item correctly more often than equally knowledgeable persons from another group.
2. Item impact described the situation in which DIF exists because there were true differences between the groups in the underlying ability of interest being measured by the item.
3. Item bias described the situations in which there is DIF because of some characteristic of the test item that is not relevant to the underlying ability of interest.

As noted above, DIF is the central concept to distinguish two terms, which are ‘impact’ and ‘bias’ in which item impact and item bias will only occur if DIF exists in the item. The definition of DIF also depends on the statistical DIF methods used to detect DIF.

Methods Used in Detecting DIF

A variety of statistical procedures have been developed for detecting DIF (Berk, 1982; Millsap & Everson, 1993). Current methods for DIF detection can be classified along two dimensions (Potenza & Dorans, 1995). The first of these dimensions is the nature of the ability estimate used for the matching or conditioning variable. The matching variable can use either an obtained or observed score, such as a total score, or a latent variable score, such as an estimate of the trait or ability level. The second dimension refers to the method used to estimate item performance at each level of the trait or ability. Because of `this, researchers have developed parametric and nonparametric methods to identify DIF that are effective and, at the same time, easy to implement in practice. Parametric procedures utilize a model, or function, to specify the relationship between the item score and ability level for each of the subgroups.

In nonparametric procedures, no model is required because item performance is observed at each level of the trait or ability for each of the subgroups. Parametric procedures generally require larger datasets and have the risk of model misspecification. Nonparametric methods for detecting DIF are the Mantel-Haenszel (MH) procedure (Holland & Thayer, 1988), the standardization procedure (Dorans & Kulick, 1986), and the simultaneous item bias procedure (SIBTEST, henceforth referred to as SIB; Shealy & Stout, 1993). MH and SIB share a common framework. They are computationally simple, inexpensive, easy to implement in practice, and do not require large sample sizes. Also, both procedures provide statistics that have associated tests of significance. Swaminathan and Rogers (1990) presented a logistic regression (LR) procedure and demonstrated that it can be implemented easily in practice. A major advantage of the LR procedure is that it is a model-based procedure with the ability variable treated as continuous. It also allows for testing the hypothesis of no interaction between the ability variable and the group variable.

Numerous statistical procedures have been developed to evaluate DIF. These procedures essentially fall into five categories: (a) descriptive statistical approaches (e.g., conditional proportion correct), (b) graphical display, (c) contingency tables, (d) regression models, and (e) methods based on item response theory (IRT). Camilli (2006) differentiated among these methods using only two classifications, namely: those based on observed scores and those based on IRT.

The MH procedure can be conceptualized as being based on the LR model in which the ability variable is treated as discrete and no interaction between the ability variable and group membership is permitted. The LR procedure is, therefore, expected to improve on the MH procedure for detecting non-uniform DIF. Previous research has shown that the MH, SIB, and LR procedures are equally effective in the identification of uniform DIF (Ackerman, 1992; Narayanan & Swaminathan, 1994; Rogers & Swaminathan, 1993a; Roussos & Stout, 2004; Swaminathan & Rogers, 1990**).**

Mantel-Haenszel

The M-H procedure is a chi-squared contingency table-based approach which examines differences between the reference group and focal group on all items of the test, one by one (Mantel & Haenszel, 1959). The ability continuum, defined by total test scores, is divided into *k* intervals which then serves as the basis for matching members of both groups (Marasculio & Slaughter, 1981). A 2 x 2 contingency table is used at each interval of *k* comparing both groups on an individual item. The rows of the contingency table correspond to group membership (reference or focal) while the columns correspond to correct or incorrect responses. Table 6 presents the general form for a single item at the *k*th ability interval.

Table 6: Contingency Table for a Dichotomous Item with Total Score k

	Item		
Source	Correct =1	Incorrect=0	Total
Reference Group	A_k	B_k	n_{RK}
Focal Group	C_k	D_k	n_{FK}
Total	n_{1k}	n_{0k}	T_k

The next step in the calculation of the M-H statistic is to use data from the contingency table to obtain an odds ratio for the two groups on the item of interest at a particular *k* interval. This is expressed in terms of *p* and *q* where *p* represents the proportion correct and *q* the proportion incorrect for both the reference (R) and focal (F) groups. For the M-H procedure, the obtained odds ratio is represented by $\boldsymbol{\alpha}$ with the possible value ranging from 0 to ∞. An $\boldsymbol{\alpha}$-value of 1.0 indicates an absence of DIF and thus similar performance by both groups. Values greater than 1.0 suggest that the reference group outperformed or found the item less difficult than the focal group. On the other hand, if the obtained value is less than 1.0, this is an indication that the item was less difficult for the focal group (Holland & Thayer,1988). Using variables from Table 6, the calculation is as follows:

$$\alpha = (p_{Rk}/q_{Rk})/(p_{Fk}/p_{Fk})$$

$$= (A_k/(A_k + B_k)/B_k(A_k + B_k))/(C_k/(C_k + D_k))/(D_k/C_k + D_k))$$

$$= (A_k/B_k)/C_k/D_k = A_kD_k/B_kC_k$$

The computation pertains to an individual item at a single ability interval. The population estimate **α** can be extended to reflect a common odds ratio across all ability intervals *k* for a specific item. The common odds ratio estimator is denoted $\alpha_{\text{M-H}}$ and can be computed by the following equation:

$$\alpha_{MH} = \varepsilon(\frac{A_kD_k}{N_k})/\varepsilon(B_kC_k/N_k \text{ (Marasculio \& Slaughter, 1981).}$$

for all values of *k* and where N_k represents the total sample size at the *kth* interval. The obtained α_{MH} is often standardized through log transformation, centering the value around 0 (Dorans & Holland, 1993). The new transformed estimator $\text{MH}_{\text{D-DIF}}$ is computed as follows:

$$MH_{D-DIF} = -2.35In(\alpha_{MH})$$

Thus, an obtained value of 0 would indicate no DIF. In examining the equation, it is important to note that the minus sign changes the interpretation of values less than or greater than 0. Values less than 0 indicate a reference group advantage whereas values greater than 0 indicate an advantage for the focal group. There are three types of effect sizes used in describing DIF size.

Type A items are negligible DIF: items with |ΔMHi | < 1. Type B items are moderate DIF: items with 1 ≤ |ΔMHi | <1,5 and Type C items are large DIF: items with |ΔMHi | ≥ 1,5 (Zieky, 1993). Longford, Holland and Thayer (1993) comment that if an item is classified as A one can still include the item in the test. If the item is classified as B, one should examine if there are other items one can choose to include in the test instead, i.e., an item with a smaller absolute value of $MH - DIF$.

Finally, an item classified as C should only be chosen if it meets essential specifications but documentation and corroboration by a reviewer are required. It should also be noted that the number of test-takers in the focal group can have a strong influence on the DIF categorization, i.e., more items are classified as categories B and C with larger focal and reference group sizes. A limitation of MH is that it may lack the power to detect non-uniform DIF (Hambleton & Roger, 1989; Swaminathan & Rogers, 1990; Uttaro & Millsap, 1994).

Item response theory

Item response theory (IRT) is another widely used method for assessing DIF. IRT allows for a critical examination of responses to particular items from a test or measure. As noted earlier, DIF examines the probability of correctly responding to or endorsing an item conditioned on the latent trait or ability. Because IRT examines the monotonic relationship between responses and the latent trait or ability, it is a fitting approach for examining DIF (Steinberg & Thissen, 2006). Three major strengths of using IRT in DIF detection, according to Camilli and Shepard (1994), are:

1. Compared to classical test theory, IRT parameter estimates are not as confounded by sample characteristics.
2. Statistical properties of items can be expressed with greater precision which increases the interpretation accuracy of DIF between two groups.

3. The statistical properties of items can be expressed graphically, improving interpretability and understanding of how items function differently between groups.

Measurement of DIF using ICCs

IRT has brought about significant changes in psychometric theory and test development. In its most basic form, it postulates that a single ability underlies examinee performance on a test and that the probability of a correct response on an item is a monotonically increasing curve (Hambleton & Slater, 1997). The ICC of an item plots the probability of the "correct" response against the magnitude or level of the underlying (latent) trait being measured. Osterlind (1983) describes ICCs as the most elegant of all the models to tease out DIF. IRT models assume unidimensionality, local independence of items and the fact that the probability that an examinee will respond correctly to an item depends upon the shape of the curve and the individual's level regarding the underlying construct being measured.

However, it is not dependent upon the individual's performance relative to any group (Osterlind, 1983). One of the most useful features of IRT is that the examinee's estimated ability level and item difficulty level are put on the same scale. This allows for the illustration of item difficulty and item discrimination simultaneously using ICC graphs to depict the characteristics of each item. This method provides a powerful base for assessing differential item functioning by also using visual inspection. In IRT terms, the "overall notion is that the item characteristic curves generated for each of the two contrasting groups should be alike for an item to be considered unbiased" (Osterlind, 1983, p. 16).

The use of ICCs for DIF detection concerns the comparison of differences in the ICCs for different subgroups. Only two groups can be compared at a time, but a sample can be divided into various subgroups for such comparisons. “The area between the equated ICCs is an indication of the degree of bias present in a considered test item” (Osterlind, 1983, p. 61). Although both groups are on essentially the same scale, they need to be equated employing a linear transformation. The difference in scales is caused by the fact that theta is arbitrarily defined as having a mean of 0 and a standard deviation of 1 in each separate group (Owen, 1992b). Even though the *a* and *b*-parameters are invariant from group to group, they are not invariant when the origin of theta changes arbitrarily in each new parameterisation. The scales of two different groups then must be equated before the respective parameters can be compared (Owen, 1992b). Once the theta scales have been equated, a meaningful comparison of the ICCs of the two groups is possible (Osterlind, 1983). Procedures for decision making may include simply inspecting the graphs visually or calculating the actual differences. Limits or cut-off criteria are arbitrary because no specific significance test is available to test differences between estimates of area (Osterlind, 1983).

The information can also be provided for two subgroups at a time, which results in two plotted graphs indicating subgroup performance over the spectrum of levels of the underlying trait. This allows for a comparison of subgroup performance at various levels of the underlying trait. If the graphs of two subgroups fall on top of each other, the level of performance (probability of “correct” response) at that level of the latent trait is the same. When the graphs differ, this indicates differences in performance despite the similarity in the latent trait being measured, thus giving evidence of test item bias or DIF. The extent of bias or DIF can be measured by the magnitude of the area between the two graphs. Since the ICC graph values for the respective comparison groups are known, the area between the two ICCs is calculated by dividing the area into small rectangular areas over the entire ability range and calculating and adding the areas formed between the ICCs. Figure 11 depicts ICC showing DIF free item

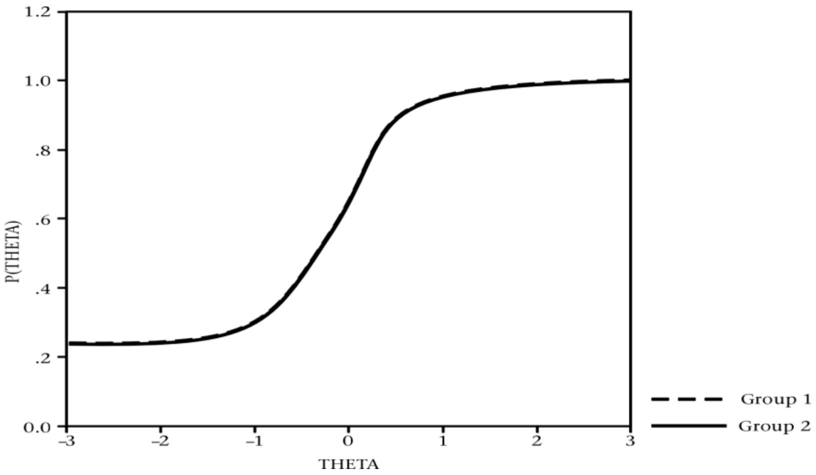

Figure 11: ICC showing no DIF item

Figure 11 shows the ICCs of an item showing no DIF since there is no difference between the ICCs of groups 1 and 2.

Assessing Item DIF

How IRT-based ICCs are used to evaluate DIF is to compare the ICCs of two groups (Osterlind, 1983). Various considerations make it extremely difficult to give one fixed magnitude at which an item should be considered biased or DIF. Visual inspection of the form of DIF, together with the magnitude of the area between the graphs of the two groups compared, is usually combined to determine whether an item should be flagged as biased. A distinction is made between uniform DIF and non-uniform DIF. In uniform DIF, the probability of answering an item correctly for one group is consistently lower than that of the other group. This results in the ICC for one group being below that of the other group over the entire ability range (see Figure 12). In non-uniform DIF, the curves cross at a certain point. Whereas for one range of ability the one group has a lower probability of answering the item correctly, the reverse is true of another range of ability. Figure 13 illustrates an item that shows non-uniform DIF. Of course, the ideal situation is that there should be little difference between the ICCs of the two groups being compared as illustrated in Figure 13.

Logistic Regression

Logistic regression approaches to DIF detection involve running a separate analysis for each item. The independent variables included in the analysis are group membership, an ability matching variable typically a total score, and an interaction term between the two. The dependent variable of interest is the probability or likelihood of getting a correct response or endorsing an item. Because the outcome of interest is expressed in terms of probabilities, maximum likelihood estimation is the appropriate procedure (Bock,1975). This set of variables is then expressed by the following regression equation:

$$Y = \beta_0 + \beta_1 M + \beta_2 G + \beta_3 MG \ ,$$

where β_0 corresponds to the probability of a response when M and G are equal to 0 with remaining β_s corresponding to weight coefficients for each independent variable. The first independent variable, M, is the matching variable used to link individuals on ability (a total test score). The group membership variable is denoted G and in the case of regression is represented through dummy coded variables. In the final term, MG corresponds to the interaction between the matching and grouping variables.

For this procedure, variables are entered hierarchically. Following the structure of the regression equation provided, variables are entered by the following sequence: matching variable M, grouping variable G, and the interaction variable MG. The determination of DIF is made by evaluating the obtained chi-square statistic with 2 degrees of freedom. Additionally, parameter MG estimate significance is tested.

From the results of the logistic regression, DIF would be indicated if individuals matched on ability have significantly different probabilities of responding to an item and thus differing logistic regression curves. Conversely, if the curves for both groups are the same, then the item is unbiased and therefore DIF is not present. In terms of uniform and non-uniform DIF, if the intercepts and matching variable parameters for both groups are not equal, then there is evidence of uniform DIF (Swaminathan, & Rogers, 1990). However, if there is a nonzero interaction parameter, this is an indication of nonuniform DIF as shown in Figures 12 and 13.

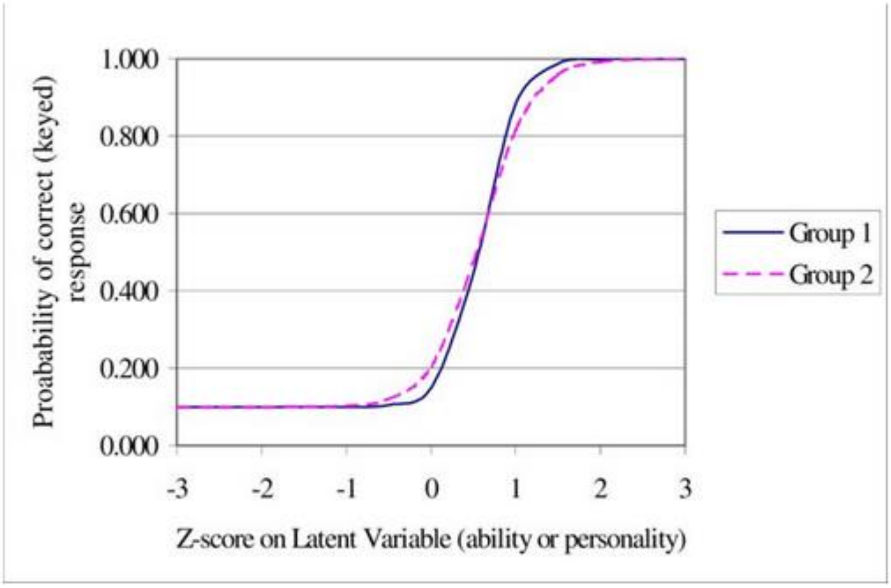

Figure 12: An example of an item that does not display DIF.

Figure 13 shows an item that does not display DIF since both groups increase and decrease across the same ability level.

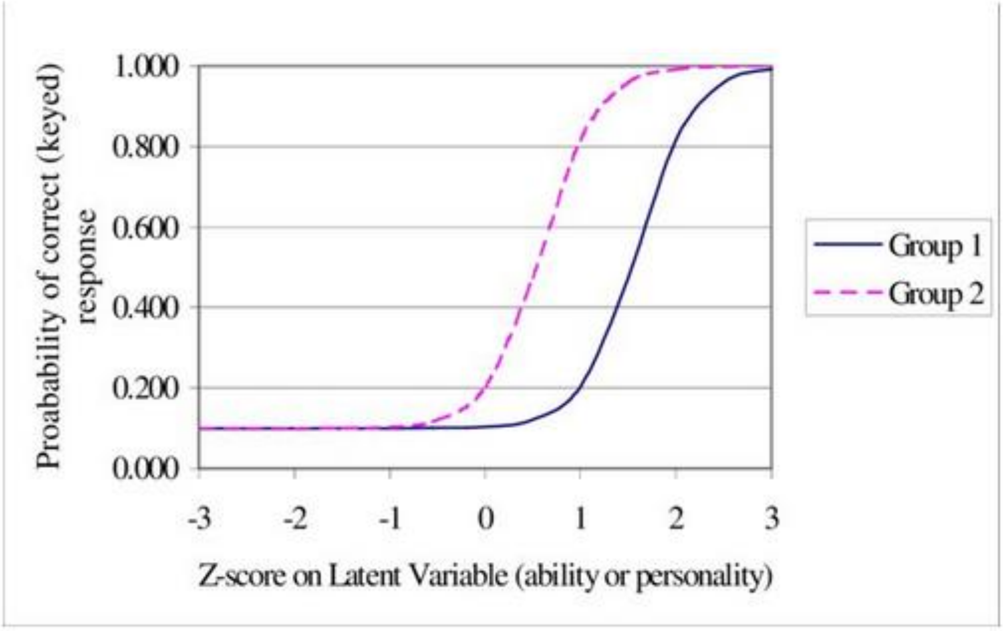

Figure 13: An example of an item that displays substantial uniform DIF

Figure 13 depicts an example of an item displaying Uniform DIF. Thus, examinees in both groups had equal chances of answering the item correctly irrespective of ability levels.

Measurement Issues

Measurement is the assignment of a number to a characteristic of an object or event, which can be compared with other objects or events (Pedhazur & Schmelkin, 2013). Validity and reliability relate to the interpretation of scores from psychometric instruments (eg, symptom scales, questionnaires, education tests, and observer ratings). Educationa tests include WASSCE which should provide scores that are free from DIF in order for the results to be reliable and vaild. This is because if the results from such examination are not free from DIF, thus if some of the items which constitute the examinations do not give equal chance for examinees of the same ability level to get the item(s) correctly. This (item that exhibit DIF) will be a treat to both validity and reliability.

Reliability

Reliability is the degree to which students' results remain consistent (are the same) over replications of an assessment procedure, that is when (a) they complete the same task(s) on two or more different occasions, (b) they complete two or more different but equivalent tasks on the same or different occasions or (c) two or more persons/teachers mark their performance on the same tasks (Nitko, 2004). Reliability is a major concern when a psychological test is used to measure some attributes or behaviour (Rosenthal & Rosnow, 1991). For instance, to understand the functioning of a test, the test which is used consistently must discriminate individuals at one time or over a course of time. In other words, reliability is the extent to which measurements are repeatable or when different persons perform the measurements, on different occasions, under different conditions, with supposedly alternative instruments that measure the same thing. In sum, reliability is the consistency of measurement (Bollen, 1989), or stability of measurement over a variety of conditions in which the same results should be obtained (Nunnally as cited in Rubio, Berg-Weger, Tebb, Lee, & Rauch, 2003). In terms of DIF, test items need to accurately tap into the construct of interest to derive meaningful ability level groups. One does not want to inflate reliability coefficients by simply adding redundant items. The key is to have a valid and reliable measure with enough items to develop meaningful matching groups in other to avoid DIF (Gadermann, Guhn, & Zumbo, 2012; Revelle, & Zinbarg, 2009; John, & Soto, 2007).

Validity

Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests (Linn, 2011). Although classical models divided the concept into various "validities" (such as content validity, criterion validity, and construct validity), according to Highhouse, Doverspike and Guion (2015), the currently dominant view is that validity is a single unitary construct (Messick, 2013).

Validity is, generally, considered the most important issue in psychological and educational testing (Popham, 2008) because it concerns the meaning placed on test results (Messick, 2013). Though many textbooks present validity as a static construct (Nitko & Brookhart, 2004), various models of validity have evolved since the first published recommendations for constructing psychological and education tests (Linn, 2011). These models can be categorized into two primary groups: classical models, which include several types of validity, and modern models, which present validity as a single construct. The modern models reorganize classical "validities" into either "aspects" of validity or "types" of validity-supporting evidence (Messick, 2013).

According to the 1999 Standards for educational and psychological testing, validity refers to the degree to which evidence and theory support the interpretations of the test. These shreds of evidence have been categorized in five namely,(1) Content: do instrument items completely represent the construct?. (2) Response process: the relationship between the intended construct and the thought processes of subjects or observers. (3) Internal structure: acceptable reliability and factor structure. (4) Relations to other variables: correlation with scores from another instrument assessing the same construct. (5) Consequences: do scores really make a difference?

Evidence-based on content

The content according to standards for educational and psychological testing 1999, refers to the themes, wording, and format of the items, tasks, or questions on the test, as well as the guidelines for procedures regarding administration and scoring. Evidence-based content can also come from expert judgments of the relationship between parts of the test and construct.

Evidence-based content can be used, in part to address questions about differences in the meaning and interpretations of test scores across relevant subgroups of examinees. Of concern is the extent to which content underrepresentation or content irrelevant components may give an undue advantage or disadvantage to one group or more subgroups of examinees. A careful review of the content domain by a different panel of experts may point to potential sources of irrelevant difficulty (or easiness) which can be described as DIF.

Evidence-based on the internal structure

The Standards for Educational and Psychological Testing (1999), refer to this type of evidence as to the internal structure and defines as the degree to which the relationship among test items and test components conform to the construct on which the proposed test score interpretations are based. A theory that posited unidimensionality would call for evidence of item homogeneity. In this case, the item interrelationships also provide an estimate of score reliability, but such an index would be inappropriate for a test with a more complex internal structure. Studies of internal struct of tests are designed to show whether item or items function differently for identifiable subgroups of examinees and that is an indication of DIF.

Evidence-based on criterion

Evidence of the relation of test scores to a relevant criterion may be expressed in various ways, but the fundamental question is always: How accurately do test scores predict a criterion performance? The degree of accuracy deemed necessary depends on the purpose for which the test is used. The criterion variable is a measure of some attribute or outcome that is of primary interest, as determined by test users, who may be administrators in the school system, management or clients.

Historically, two designs often called predictive and concurrent have been distinguished for evaluating test-criterion related evidence. A predictive study indicates how accurately test data can predict criterion scores that are obtained later whiles concurrent study obtains spredictor and criterion information at about the same time.

Evidence about relations to other variables can also be used to investigate questions of differential prediction for groups. However, the difference may also imply that the criterion has a different meaning for different subgroups.

One of the central issues in DIF analysis is the examination of the impact of DIF on test validity. Several research studies (e.g., Roznowski & Reith, 1999; Zumbo, 2003) have attempted to statistically model the impact of DIF on test performance. The results, however, have been mixed. While some researchers such as Roznowski and Reith (1999) and Zumbo (2003) have reported that DIF has little if any, impact, Pae and Park (2006) reported that

DIF may affect the performance on the test. The issue is of much significance because, as Pae and Park (2006) state, “it can provide new insights into how DIF items in the item bank should be dealt with, and because decisions with a test are made not by the result of an individual item score but by the result of a whole test score” (p. 476).

Test Validation and Testing Fairness

The term f*airness* is used in many ways and has no single technical meaning. This study relates fairness to the absence of DIF and equitable treatment of all examinees. There is a broad consensus that test items should be free from bias and that all examinees should be treated fairly in the testing process itself (e.g., afforded the same or comparable procedures in testing, test scoring, and use of scores). Another characteristic of test fairness addresses the equality of testing outcomes for the examinee subgroups defined by race, ethnicity, gender, disability, or other characteristics. The idea that fairness requires equality in the overall passing rate for different groups has been almost entirely repudiated in the professional testing literature.

A widely accepted view would hold that examinees of equal standing concerning the ability the test are intends to measure should on average earn the same test score, irrespective of a group membership.

Measurement Invariance

Measurement invariance is a desirable property of fundamental scientific requirements of good measurement. At the item-level, measurement invariance is defined as the independence of group membership and item response conditional on the intended construct being measured (Millsap & Meredith, 1992). According to Engelhard (1994), the invariant measurement can be viewed from person measurement or/and item calibration. In the person measurement case, the development of person measures must be independent of the subset of the test items or tests that are used to define person abilities on the latent trait. This is known as ‘item invariant **or** person measurement’.

In the item calibration case, the calibration of item locations does not depend on personal use and is known as ‘sample-invariant item calibration’. In sample-invariance item calibration, ‘ability’ is the term used for the quantity of the trait being measured. Violation of invariance reveals the presence of unintended item-level multidimensionality (Ackerman, 1992; Camilli, 1992), which is vital because it reflects whether the scores have the same meaning for different subgroups (Penfield, 2010a).

Stevens (1951, p.40) pointed out that “the scientist seeks measures that will stay put while his back is turned” (as cited in Engelhard, 1994). According to Engelhard (1994), measurement problems in human sciences mostly relate to the concept of invariance of measurement. For this reason, Engelhard (2009) stated five requirements for invariant measurement as follows:

1. The measurement of personability does not depend on the particular items that happen to be used for the measuring;
2. The calibration of the items does not depend on the particular persons used for calibration;
3. Items are measuring a unidimensional latent variable;
4. More ‘able’ persons have a better chance of success on an item than less ‘able’ persons;
5. Any person has a better chance of success on an easy item than on a more difficult item.

In practice, measurement invariance is ideal but is never achieved in constructing measures. However, test developers and researchers often aim to fulfil to a certain extent (satisfactorily) the requirement of measurement invariance to make valid inferences on the measures (Stevens,1951). As pointed out by Stevens (1951, p.40), “the scientist is usually looking for invariance whether he knows it or not" (cited in Engelhard, 2008). Whenever he discovers a functional relationship, his next question follows naturally: under what conditions does it hold? The same line of insight into measurement invariance has been noted by first, calibrating the measuring instruments which must be independent of those that happen to be used for calibration (the idea of sample-invariant item calibration).

Second, the measurement of objects must be independent of the instrument that happens to be used for measuring (item-invariant person measurement). DIF is related to the lack of measurement invariance, which often addresses issues on the validity, item bias, and test fairness, which constitute the focus of this study.

The Importance of Detecting DIF

As mentioned earlier, differential item functioning (DIF) is an analysis of performance across groups on specific test items. DIF occurs when examinees from different groups show different probabilities of success (i.e., a correct response) on a dichotomously scored item or a difference in probabilities of selecting a certain level of response on a polytomous item after matching on the underlying ability that the item is intended to measure (Zumbo, 1999).

In the field of education, the absence of DIF is regarded as an important aspect of test fairness by educational researchers (Rudas & Zwick, 1997) in that, the presence of DIF in an item indicates that the item is measuring some dimensions unrelated to the remainder of the test (Ackerman, 1992). Detecting such items that are identified as having DIF is critical to maintaining the tests' fairness and validity. Also, detecting DIF seems to provide some potential ways for improving test quality as stated by the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

Empirical Review of Related Studies

DIF detection methods

In terms of DIF detection methods, it is significant that all DIF detection methods available are designed to match the groups, either directly or indirectly, on the proficiency measured by the items (Angoff, 1993), and all DIF measurements investigate how different groups perform on individual test items to determine whether the test items are creating problems for a particular group (Zumbo, 1999). They are all based on such principles that if different groups of test-takers (e.g., males vs. females, Caucasian vs. African American, English Language as first language learner vs. English Language as second language learner) have approximately the same level of ability, they should perform similarly on individual test items regardless of group membership (Zumbo, 1999).

Mapuranga, Dorans, and Middleton (2008) stated that DIF analysis is an important step in evaluating tests for fairness and equity. DIF occurs when different groups of examinees with the same level of proficiency in a domain have a different expected performance on an item. In a DIF analysis, the sample is usually divided into two subgroups. The reference group typically provides a baseline for performance (e.g., White or male) and the focal group is typically the focus of fairness concerns (e.g., Black or female).

This section has focused on studies done using the most popular DIF detecting methods based on the two classifications by Camilli (2006). Specifically, this study reviewed the literature on studies done using the Mantel-Haenszel method (Holland & Thayer, 1988), logistic regression (Swaminathan & Rogers, 1990; Zumbo, 1999), and the IRT likelihood ratio test (Thissen, Steinberg, & Wainer, 1988).

DIF studies based on observed scores

Awulor (2008) conducted a simulation study that focused on determining the effect of unequal sample sizes on the statistical power of SIBTEST and Mantel-Haenszel procedures for detection of DIF of moderate and large magnitudes. The results indicated that not only the ratios but the magnitude of DIF influenced the behaviour of SIBTEST and M-H regarding their error rate behaviour. With moderate DIF magnitude, Type II errors were committed by both M-H and SIBTEST when the reference to the focal group sample size ratio was 1:10 due to low observed statistical power and inflated Type I error rates.

Gierl and Cui (as cited in Zhang, 2009) also investigated the consistencies of DIF detection and effect size measurements among M-H, SIBTEST, and LR procedures using gender DIF data from 2000 examinees across grade 3, 6, and 9 from University of Texas at Austin. They found out that in terms of DIF magnitude, the matching percentage between M-H and LR, M-H and SIBTEST, and SIBTEST and LR are 83.30%, 75.86%, and 90.00%, respectively, and in terms of effect size from each three measurements, the correlation between each other ranges from .79 to .92.

There is also evidence showing that the DIF indexes generated by M-H procedures and those generated by Rasch measurement, an IRT method, are equivalent (Engelhard., Anderson, & Gabrielson,1990; Schulz, Perlman, Rice, & Wright,1996). However, there is no common method for placing effect sizes from different DIF approaches on the same scale; consequently, only combined effect sizes under every single approach were calculated in this study.

Kabasakal, Arsan, Gök, and Kelecioglu (2014) conducted a study on comparing performances (Type I Error and Power) of IRT Likelihood Ratio SIBTEST and Mantel-Haenszel Methods in the determination of differential item functioning. This simulation study compared the performances (Type I error and power) of Mantel-Haenszel (MH), SIBTEST, and item response theory-likelihood ratio (IRT-LR) methods under certain conditions. Manipulated factors were sample size, ability differences between groups, test length, the percentage of differential item functioning (DIF), and the underlying model. Results suggest that SIBTEST had the highest Type I error in the detection of uniform DIF, but MH had the highest power under all conditions.

Also, the percentage of DIF and the underlying model appear to have influenced the Type I error rate of IRT-LR. Ability differences between groups, test length, the percentage of DIF, model, and the interactions between ability differences and percentage of DIF, ability differences and test length, test length and percentage of DIF, test length and model affected the SIBTEST methods' Type I error rate. In the MH procedure, effective factors for Type I error rate were, sample size, test length, the percentage of DIF, ability differences and percentage of DIF, ability differences and model, and ability differences and percentage of DIF and model. No factors were effective on the power of SIBTEST and MH, but the underlying model had a significant effect on the IRT-LR power rate.

All main effects were significant, while some of the interaction effects were significant for models. The sample size provided significant results for the MH method. In the case of unequal sample sizes for focal and reference groups, Type I error was found to be decreasing in the MH method. Ability distribution had a significant effect on the SIBTEST methods. Type I error was significantly lower under the condition in which the reference and focal groups' population standard deviations differed. In the SIBTEST and MH methods, test length caused significant differences. In the SIBTEST method, test length caused a decrease in the Type I error. In the MH method, there were significant differences between tests with 20 items and those with 40–80 items. Tests with 20 items had lower Type I error, but non-significant differences were found between tests with 40 and 80 items. The ratio of DIF items in the test affected all methods' Type I error.

Pei and Li (2010) conducted a study on the effects of unequal ability variances on the performance of logistic regression, Mantel-Haenszel, SIBTEST IRT, and IRT likelihood ratio for DIF detection using 1300 respondents. The results showed that the mean difference in ability alone has some impact on the validity of the logistic regression method, but it did not seem to affect the MH test, SIBTEST and IRT. The robustness of the SIBTEST to the mean difference inability is not surprising, because the SIBTEST was developed to separate true bias/DIF from this difference. Except for the IRT, the difference in ability variance inflates the Type I errors for all the DIF detection methods. The impact is the most significant on the logistic regression method, and the least significant on the SIBTEST. As the difference in ability variance increased, the inflation of the Type I error for all three methods became more and more severe. As expected, the only method that was not affected by both mean difference in ability and difference in ability variance is the IRT. In the IRT, the difference between the focal and reference group ability means and the ratio of the focal and reference group ability variances are properly formulated in the likelihood function and can be simultaneously estimated with the item parameters. Therefore, using IRT can separate the true DIF effect from those ability differences. This explains the well-controlled Type I errors in the IRT across different simulation settings

Swaminathan and Rogers (1990) conducted a simulation study to compare the LR procedure and the M-H under different conditions such as sample size, test length and the nature of the DIF. Results indicated that for the items with uniform DIF, the two procedures had very similar detection rates but with the Mantel-Haenszel procedure having a very slight advantage. Both were able to detect uniform DIF with about 75% accuracy in samples of 250 per group and with 100% accuracy in samples of 500. For non-uniform DIF, the picture was very different. The Mantel-Haenszel procedure was completely unable to detect non-uniform DIF under any condition.

In the Swaminathan and Rogers' (1990) study, the logistic regression procedure detected nonuniform DIF with about 50% accuracy in small samples and short tests and 75% accuracy in large samples and long tests. In terms of false positives, the Mantel-Haenszel procedure performed somewhat better than the logistic regression procedure. With a significance level of .01, the Mantel-Haenszel procedure consistently produced around 1% false positives under all conditions, while the logistic regression procedure produced between 1% and 6% false positives. There were 8 items with uniform DIF, 8 with nonuniform DIF, and 20 replications. The percentages were obtained by dividing the number of detections by 160 (8 x 20). Similarly, there were 64 items with no DIF, hence, the percentage of false positives was obtained by dividing the number of false positives by 1,280 (i.e., 64 x 20).

Swaminathan and Rogers (1990) reported that both procedures are equally and highly powerful to identify uniform DIF, but only the LR procedure can identify nonuniform DIF with consistency. In contrast, the false-positive detection (Type 1 error rate) for M-H method was 1% as expected. The Type 1 error rate for the LR procedure was 4%, which was higher than expected. To overcome the limitation of the M-H method in detecting non-uniform DIF, Mazor, Clauser and Hambleton (1994) proposed a modification of the M-H statistic, which involves splitting a sample into two groups, a group with high ability and the other at low ability level.

Hidalgo and Lopez-Pina’s (2004) simulation study compares the LR procedure, the M-H method and the modified M-H method in their efficacy for detecting DIF. The results of this study suggest that LR and the modified MH procedures are highly comparable. LR, generally, detected more DIF items than the standard MH procedure but detected a similar percentage when the modified MH procedure was used.

In asymmetrical nonuniform DIF identification, for items in which large differences existed between the parameters of the focal and reference group, LR and the modified MH procedure showed very similar results (87.92% overall for each of the procedures). In these situations, the standard MH procedure also showed similar confidence interval rates (77.50%). On the other hand, for identifying symmetrical nonuniform DIF, the LR analysis proved to be the most effective approach, with 68.75% of DIF items correctly identified compared to 61.25% for the modified MH procedure and 50% for the standard MH procedure. For identifying uniform DIF, the standard MH procedure appeared to be the most powerful (55% DIF items correctly identified), followed closely by LR (53.33%) and the modified MH procedure (50%). The effect size measures had the expected pattern, regardless of the procedure on which effect size measure was based. The effect size value was larger when the magnitude of DIF was greater and was insignificant for items with-out DIF.

However, the effect size measure based on LR appeared to be insensitive to the specified DIF conditions, with the DIF items classified as having moderate or large DIF ranging from 1% to 20%. Thus, when the criteria proposed by Zumbo and Thomas (1997) were used, only 1% of the items were classified with moderate DIF. When the criteria adopted by Jodoin and Gierl (2001) were used, a slightly larger percentage (5%) was classified as having large DIF, and 15% were classified as having moderate DIF. The effect size measures based on MH procedures (both the standard and modified procedure) were more sensitive to the specified DIF conditions, classifying a larger percentage of items as having moderate to the large DIF. Although the study by Jodoin and Gierl (2001) offered guidelines for interpreting DIF effect size, the performance of the statistic ΔR^2 as an effect size measure should first be studied in a wider range of experimental conditions, and interpretation criteria need to be established to optimize decision making, that is, to control for both false-positive rate and to offer greater statistical power.

Linacre and Wright (1987) compared the theoretical properties of the M-H method and the Rasch Model (RM). They argued that the M-H method in detecting DIF is an attempt to ascertain implicitly what the RM presents explicitly. In their arguments, it was made known that the basic requirement of the M-H method is that the probabilities of success for the reference and focus groups have the same relationships across all intervals. It was shown that the calculation of the $\hat{\alpha}$ estimate requires an arbitrary segmentation on the matching score on two groups to be compared. Hence the distribution of abilities, the selection of interval boundaries and the sample size of the reference and focal groups affect the magnitude of $\hat{\alpha}$. It was concluded that the M-H method involves theoretical uncertainties and depends on the arbitrary decisions made by users on the matching scores and their intervals. Linacre and Wright (1987) pointed out that the RM has been developed on the same assumptions that the M-H method implies and requires, but the RM used all relevant information available from every response by the reference and focus groups. Therefore, they claimed that the RM can provide an odds ratio estimate of smaller but a more accurate, and standard error that is independent of both ability distributions.

Abedalaziz (2010) conducted a study to explore gender-related differential item functioning in Mathematics using three methods (i.e. M-H, Transformed Item Difficulty (TID), and b-parameter difference) to find out the agreement among these methods. The samples used in this study were drawn from a dataset containing the responses of approximately 3390 (1600 males and

1790 females) eleventh-grade students from Malaysia to achievement test comprising of 45 items. In summary, the percentage of agreement among the three approaches in detecting DIF is relatively low. The range is from 43% to 65% for detecting DIF. The highest agreement was among M-H and TID methods, the lowest agreement was among TID and b-parameter difference. This study also provided evidence that there are gender differences in performance on test items in Mathematics that vary according to content even when content is closely tied to the curriculum.

DIF Studies based on IRT

Özdemir (2015) conducted a study to determine items which have differential item functioning (DIF) in TIMSS 2011 Mathematics subtest with three different item response theory (IRT)-based DIF methods (Lord’s Chi-Square, Raju’s Area and Likelihood-Ratio Test methods). Results indicated that two items were identified as DIF items by all three methods, whereas 12 other items were never identified as such. For four items, the Lord's Chi-square and Raju’s Area methods identified them as DIF, but the Lord’s Chi-Square did not. On the other hand, one item was detected as DIF item by only IRT methods. Although almost all items detected as DIF with three different methods were in favour of male students, Raju’s signed area method with item purification indicated that Item 8 and Item 21 were in favour of female students rather than male students with respect to Mathematics.

Performing item purification with Lord's Chi-square and Raju’s Area methods regarding Likelihood-Ratio test affected both the number of DIF items and DIF items themselves. However, performing item purification with LRT method did not affect the number of items detected as DIF. According to the results, the Lord's Chi-square method tended to be more sensitive than the other two methods concerning detecting DIF items. On the other hand, even when item purification was performed, LRT method failed to detect many items detected as DIF items by other methods. These three IRT-based techniques showed substantial agreement in the detection of DIF among the same set of Mathematics subtest items but vary in the number of items flagged with DIF due to different assumptions and criteria used.

The results also indicated that the 2PL IRT model fitted best to the data for both Lord’s Chi-Square method and Raju’s Signed Area method. Although several items detected as DIF differed for each of the methods, 2 items out of 22 dichotomous items in the test showed DIF consistently across all methods. These two items were more likely to be answered correctly by males after controlling for overall ability. Finally, results indicate that no single method can be guaranteed to identify all the DIF items in a test. Not only IRT-based methods but also Non-IRT-based methods should be used to address the instability problem which undermines the utility of current methods and results of both IRT-based and Non-IRT-based methods.

DIF across Gender

Research conducted nationally and internationally have revealed that test items contained differential item functioning whether the test is meant for certification, admission, recruitment or placement purposes. Doolittle and Cleary’s (1987) in their study, utilized a procedure for the detection of differential item performance (DIP) to examine the relationships between characteristics of mathematics achievement items and gender differences in performance in 8 samples (1,300–2,400 students each) of high school seniors.

The results indicated a relationship between item characteristics and gender-based DIP. Geometry and mathematics reasoning items were more difficult for female examinees, and the more algorithmic, computation-oriented items were easier. Predictions, based on previous research about the categories of items that would contribute to gender-based DIF were supported. For example, geometry and Mathematics reasoning items were relatively more difficult for female examinees and the more algorithmic, computation-oriented items were relatively easier.

Abedalaziz (2010) conducted a study to find out the agreement between two approaches (logistic regression model and M-H) in detecting a gender-related differential item functioning of a mathematical ability scale items. The scale was developed and administered to samples of 800 students (380 males and 420 females) in Jordan. The study pointed out that (1) the percentage of agreement between the two approaches in detecting DIF was 80% and (2) males outperformed females in spatial and deductive abilities, whereas females outperformed males in numerical ability.

Driana (2007) conducted studies on gender differential item functioning on a ninth-grade Mathematics proficiency test in Appalachian, Ohio, USA. The study was done on 40 multiple-choice items. Eighteen thousand, one hundred and ninety-eight (18,198) examinees participated in the study. This study was conducted to find out whether there was differential item functioning between male and female students and between Appalachian and non-Appalachian students. Female and Appalachian students were used as a focal group, whereas male and non-Appalachian students served as the reference groups. The M-H procedure was utilized to detect the existence of items showing uniform DIF and Breslow- Day test of homogeneity of the odds ratios was employed to identify items showing non-uniform DIF. The two-stage approach of M-H identifying the presence of items showing DIF was conducted at .05 significance level to answer four research questions. The first research question investigated the presence of items that function differently between males and females after matching the total score. The result of M-H analysis of the first and second stage showed no item was flagged at C-effect size item exhibiting large DIF. Four items were classified as moderate DIF (B-effect size) items showing medium DIF.

Driana's (2007) study also revealed that among the four items showing medium DIF, three favoured females. Two of these items that showed medium DIF were in the content area of algebra that tested knowledge and skills and conceptual understanding respectively while the other item that showed medium DIF was in the content area of data analysis that tested knowledge and skills. Hence only one item favoured males and it was also a medium DIF. This item was a measurement item that tested knowledge and skills (Driana, 2007).

The second research question examined items that function differently between groups of Appalachian and Non-Appalachian students. The results of the M-H analysis showed no DIF. The third research question investigated the presence of gender DIF among Appalachian and non- Appalachian students and the results showed that, at the first stage, two items were flagged as large DIF(C) items, while only an item was flagged as medium DIF (B). Even after purification, the three items continued to show DIF. This result was consistent with the direction when the analysis was conducted. Dodeen and Johanson (2003) analyzed and classified items that display sex-related differential item functioning (DIF) in attitude assessment. A total of 982 items that measured attitudes from 23 datasets were used in the analysis. Results showed that sex DIF is common in attitude scales. More than 27% of items showed DIF related to sex, 15% of the items exhibited moderate to large DIF, and the magnitudes of DIF against males and females were not equal.

Ryan and Chiu (2001) examined whether the patterns of gender differential item functioning (DIF) present in parcels of items were influenced by changes in item position. Items were studied collectively to detect differential bundle functioning (DBF) on 2 forms of a test of Mathematics for college freshmen using Simultaneous Item Bias Test (SIBTEST). The test included items in algebra (18), trigonometry (12), geometry (5), and analytic geometry (5). To investigate order effects, two forms of the test were assembled: Form 1 (random) and Form 2 (easy to difficult within a content area) with sample sizes of 3,932 and 1,074 for Form 1 and 2, respectively. Findings suggest that the amount of gender DIF and DIF present in item parcels tends not to be influenced by changes in item position. The gender DIF findings for the word problem category were an issue. The beta value was .46 for Form 1 and .39 for Form 2. In the case of DIF, if β is .46, the reference group (male students) has a probability of getting the item correct, that is, .46 greater than that of matched focal group members (female students).

Siamisang and Nenty (2012) researched gender differential item functioning on 2007 TIMSS among students from Botswana, Singapore and USA. A quantitative approach comprising the Scheunemann Modified Chi-Square (SSχ^2) and Mantel Haenszel (M-H) differential item functioning (DIF) analysis was used. Results from SSχ^2 the analysis showed that only four

Mathematics items and four Integrated Science items flagged DIF between male and female students among the three countries. However, the DIF detected did not significantly function differently among males and females from the three countries. The M-H analysis indicated that all items tested for gender DIF were flagging negligible DIF or no DIF.

The conclusion of Siamisang and Nenty (2012) study was based on the reflection of the objectives and research questions of the study. The findings of the study showed that while several gender differences were found, the effect size for these differences across the three countries was small. However, results of DIF analysis across nations showed that differences between Singapore and USA students were statistically significant but small while the differences between Botswana and Singapore students, Botswana and USA students were both statistically significant and quite large. The benchmarking that was done by the USA in Singapore during their last review and development of the Mathematics and Science curriculum could have contributed to bridging the differences between Singapore and the USA.

The findings of the studies reviewed further indicates that differential item functioning remains an important concept which cannot be ignored in the field of teaching and learning, conducting examinations and analysing of examination results.

In the study of gender-DIF in performance on sixty (60) multiple-choice tests items and 8 constructed-response type test items in Mathematics by Garner and Engelhard (1999), the results from a random sample of 3,952 eleventh graders who took the 1994 Georgia High School Graduation Testin the USA using many facet Rasch (1980) measurement models, showed that women consistently had an advantage over men on multiple-choice items involving algebra. Also, men showed less consistent advantage on items involving geometry and measurement, number and computation, data analysis and proportional reasoning in both the mean scores and DIF indexes. The mean scores were significantly higher for men than for women on 2 out of 8 constructed response items. However, when men and women were statistically matched according to ability, the only significant difference in performance on constructed-response items was in favour of women. It can, therefore, be concluded that gender DIF in Mathematics may be linked to content and item format.

Madu (2012) investigated differential item functioning (DIF) for male and female students in Mathematics examination conducted by the West African Examination Council (WAEC) in 2011 in Nigeria. The study was carried out in Nsukka Local Government Area using the responses of secondary school students who sat for June/July 2009 examination in Mathematics conducted by WAEC. Data were obtained from responses of 1671 students in 50 multiple-choice test items. The students (examinees) were drawn from 12 senior secondary schools and randomly sampled from 20 coeducation schools. DIF was investigated using Scheuneuman Modified Chi-square Statistics (SSχ2). The results indicated that male and female examinees functioned differently in 39 items and no difference in two items.

Taylor and Lee (2012) conducted a study of differential item functioning (DIF) for grades 4, 7, and 10 Reading and Mathematics items from state criterion-referenced test which was composed of multiple-choice and constructed-response items. The Reading items included 30 multiple-choice and 16 constructed-response for Mathematics while Reading had 28 multiple-choice,12 constructed-response tests items. Gender DIF was investigated using POLYSIBTEST and a Rasch procedure. The Rasch procedure flagged more items for DIF than did the poly simultaneous item bias procedure, particularly multiple-choice items.

Results indicated that for both Reading and Mathematics tests items, multiple-choice items, generally, favoured males while constructed-response items generally favoured females. Content analyses showed that flagged reading items typically measured text interpretations or implied meanings. Males tended to benefit from items that asked them to identify reasonable interpretations and analyses of informational text. Most items that favoured females asked students to make their interpretations and analyses, of both literary and informational text, supported by text-based evidence. Content analysis of Mathematics items showed that items favouring males measured geometry, probability, and algebra. Mathematics items favouring females measured statistical interpretations, multistep problem solving, and mathematical reasoning.

Abiam (1996) investigated gender DIF in mathematics examination conducted by the West African Examination Council (WAEC) in 2011 in Nigeria. The study was carried out in Nsukka Local Government Area using the responses of secondary school students who sat for June/July 2009 examination in Mathematics conducted by WAEC. Data were obtained from responses of 1671 students in 50 multiple-choice test items. The students (examinees) were sampled from 12 senior secondary schools and randomly sampled from 20 coeducation schools. DIF was investigated using

Scheuneuman Modified Chi-square Statistics (SSχ2). The results of the analysis indicated that male and female examinees functioned differently in 39 items and similarily in 2 items.

Turkan and Cetin (2017) conducted a study on the bias in 2012 Placement Test items in terms of gender variable using the Rasch Model in Turkey. The sample of the study was 216, 363 participants that were randomly selected out of 1,075, 546 students who took the Placement Exam in 2012. Stratified sampling was used to select 527,978 females and 547,568 males. Items in Turkish, Mathematics, Science and Technology and Social Sciences subtests in 2012 SBS that 8th-grade students were taken into consideration. The Turkish test includes 23 items. Science and Technology test and Social Sciences test each include 20 items. Results showed that two items in Mathematics subtest and one item in Social Sciences subtest, a total of three items, have Differential Item Functioning (DIF). In terms of gender DIF, it was detected that an item that demands geometry skills was in favour of females. Thus, female students were better at Mathematics and courses based on verbal language skills in primary school. However, the same study indicated that male students were better at geometry, beginning from high school (Amrein & Berliner, 2002). This finding contradicts the study carried out by Zenisky, Hambleton and Robin (2003).

Zenisky et al (2003) conducted a study using two-stage DIF analyses procedure known as weighted two-stage conditional p-value comparison procedure. Item-level responses from approximately 60,000 students participating in a large-scale state assessment programme in each of three subject areas: language arts (LA), mathematics (MA), and science (SCI). Tests in each subject area were administered at the elementary school (ES), middle school (MS), and high school (HS) education levels. At each education level, data were obtained from two forms (1 and 2). No items were common across forms or levels, but each test was built from the same test specifications regarding content and cognitive skills. In total, 18 datasets were evaluated for evidence of gender DIF: 3 Subject Areas (LA, MA, and SCI) ×3Education Levels (ES, MS, and HS) ×2 Test Forms (1 and 2). Findings showed that items demanding visual-spatial intelligence, that is, the ones with tables, figures, graphics, are in favour of males. A verbal item showed DIF in favour of males in this study. This finding was consistent with other studies (Yurdugul & Askar, 2004). In their study, Yurdugul and Askar used the latent growth analysis to investigate the development of students' programming learning in Tukey. The data were gathered from students who attended a computer programming course in a department of Computer Education and Instructional Technologies. The sample group consists of 86 students, 48 males and 38 females. The achievement measurement tool, designed to measure programming knowledge and skills of students in conceptual, syntactic and strategic learning domains, was administered three times periodically. Results indicated that when Mathematics problems are expressed verbally, they are in favour of males. This finding is compatible with a finding from a study by Kalaycıoglu and Kelecioglu (2011). The study investigated the gender-related differential item functioning (DIF) of 2005 University Entrance Examination (UEE) to decide whether a DIF item is biased. In this study, measurement specialists' opinions were gathered and MH and LR DIF detection methods were used. The analysis was based on the responses of 599,330 (273,419 females and 325,911 males) high school seniors to the Turkish, social sciences, mathematics and natural science subtests of the 2005 UEE. It was found that no items were flagged for gender DIF in the Turkish subtest. However, seven social sciences items and three mathematics and natural sciences items displayed DIF. Among these items, only one natural science item was identified as biased.

Researchers emphasized that questions about politics and war in History test may become advantageous for males (see Kaaycıoglu & Kelecioglu, 2011). Similarly, Zwick and Ercikan (1989) concluded in their study that some items in the History test may have DIF in favour of males. Zwick and Ercikan study used the Mantel-Haenszel approach to investigate differential item functioning on 141 U.S. history items that were administered as part of the National Assessment of Educational Progress. Results indicated that the male-female comparison yielded 67 items for which MH D-DIF was negative, indicating that males performed better on the item, conditional on the score. Of these items, 51 were A-level effect size items and thus not of concern, 12 were B-level effect size items, and 4 were C-level effect size items. On 74 items, the conditional performance of females was better. These items included 60 A-level effect size, 13 B-level effect size, and 1 C-level effect size item. In the white-black analysis, there were no C items that were conditionally easier for whites. The 15 B-level effect size items that were conditionally easier for whites included 7 items involving map reading and 4 items on World War II. The three C items on which blacks performed better than whites, conditional on a score, were about Martin Luther King, Harriet Tubman, and the Underground Railroad. The eight B items that were conditionally easier for blacks included two on slavery, three on the civil rights movement, and one on women's rights.

Le (1999) examined gender-based DIF on the 10th-grade history achievement test administered as part of the National Education Longitudinal Study of 1988(NELS:88) on a sample of 24,599 8th graders into the 10th and 12th grades. The participants were assessed in four achievement areas namely mathematics, science, reading and history/ geography/ citizenship. The test consisted of 30 dichotomously scored multiple-choice items. Results showed that two items in Mathematics subtest and one item in Social Sciences subtest, (three items in total) have Differential Item Functioning (DIF).

Karakaya (2012) conducted a study to determine whether 20 items in Science and Technology and 20 items in Mathematics subtests of 6th, 7th and 8th grades in the 2009 Level Determination Examination (LDE) on students who lived in Ankara, Turkey has gender DIF. In the study he employed the Mantel-Haenszel (MH) method and used 22,624 students in total, 6,913 of whom were 6th -year students (3,614 males and 3,299 females), 6,333 of whom were 7th -year students (3,277 males and 3,066 females) and 9,374 of whom were 8th-year students (4,290 males and 5,084 females). It was revealed in his study that 2 (items number 6 and 16) in 6th grade, 3 in 8th grade (item number 14, 15, and 17) tests have B-level DIF in the Science and technology subset items. Among those five items, 3 favoured girls and 2 favoured boys. Results in the Mathematics subtests indicated that one item each in 6th grade and 7th grade, and two items in 8th grade have DIF at the B level. While 3 of those items work in favour of males, the remaining 1 item works in favour of females.

Abedalaziz's (2010) study was conducted to explore a gender-related differential item functioning in mathematics. Three methods (i.e., M-H, TID, and b-parameter difference) were used to detect DIF and find out the agreement among these methods. The samples used in this study were drawn from a dataset containing the responses of approximately 3,390 (1,600 males and 1,790 females) eleventh-grade students to achievement test comprised 45 items. In summary, the percentage of agreement among the three approaches in detecting DIF is relatively low. The range is from 43% to 65% for detecting DIF. The highest agreement was among M-H and TID methods, the lowest agreement was among TID and b-parameter difference. This study provides evidence that there are gender differences in performance on test items in mathematics that vary according to content even when content is closely tied to the curriculum.

A critical look at the items that showed DIF in Mathematics subtest, revealed that the item that required the algorithmic calculations worked in favour of boys. Kalaycıoğlu and Berberoğlus' (2011) study was aimed to detect differential item functioning (DIF) items across gender groups, analysed item content for the possible sources of DIF, and eventually investigated the effect of DIF items on the criterion-related validity of the test scores in the quantitative section of the university entrance examination (UEE) in Turkey. The sample used included 35,372 students with 11,368 females and 24,004 males. They used 30 multiple-choice items each for mathematics and science. There are three phases in the analyses. These are (a) detecting DIF items both in the Mathematics-1 and Science-1 subtests, (b) disentangling the possible sources of DIF in the item content, and (c) studying the effect of DIF items on the validity of the test scores. In the first phase, four DIF detection methods were used. These were MH, LR, restricted factor analysis (RFA) and item response theory log-likelihood ratio (IRT-LR) methods.

In the second step of the analysis, content-wise evaluation of the DIF items was carried out by the subject matter experts to find out the reason why an item was flagged as DIF. In this evaluation, three sources of DIF were used: subject matter related factors, cognitive skill measured by item, and item format characteristics. As all the items are multiple choices with five alternatives in the tests, item format characteristics refer to the structure of the item stem, which is categorized into four major groups: (a) items presenting verbal materials only, (b) items using graphs or tables, (c) items using numerical or symbolic representations, and (d) items using figures or pictorial representations. It seems that higher-order cognitive skills and figural or graphical representations used in item content are the two sources of DIF for favouring male students, whereas routine algorithmic calculations could produce DIF against males. Among the factors considered, cognitive skills assessed by items seem the most effective factor in producing gender DIF. However, DIF items do not create a threat to the criterion-related validity of the quantitative section of the UEE.

DIF and Region/Nation/Location

Miller, Doolittle and Ackerman (as cited in Lee & Randall, 2011) study investigated the differential performance at the item level of Mexican American students who spoke English Language as a second language (ESL) versus White native English Language speakers using MH. It was hypothesized that (1) items that emphasize mechanics in the English Language Usage Test, such as grammar and punctuation, tend to favour ESL examinees; (2) items that focus upon style and structure in the English Language Usage Test tend to favour non-ESL students; and (3) mathematical items with the greatest verbal load tend to favour non-ESL examinees. Respondents included 471 Mexican American

ESL students and 1,000 White native speakers (non-ESL). Results indicated that none of the three hypotheses was supported. Although the mean score for the ESL students was almost a full standard deviation below that of the native speakers for both tests, it appeared that the group difference in performance was reflected throughout most of the test items.

Rogers and Kulick (1987) conducted a study using Mantel-Haenszel Delta-Difference and Differential Percentage Omitting. The study was based on the secondary analysis of data gathered from nine recent administrations of the SAT from June 1986 through December 1987. This pool of information included item statistics on 765 verbal and 540 mathematical items computed for subgroups of white, Hispanic, black, Asian American, male, and female examinees. The results indicated three out of 85 SAT- verbal items showed differential in favour of city (White) candidates. For SAT Mathematics, Rogers and Kulick reported 7, 7 and 4 out of 60 items on the three forms of SAT respectively as exhibiting DIF.

Van der Flier (1980) applied the logit model approach to a word exclusion and word analogies test administered to 500 Tanzanian and 500 Kenyan students. The logit model was able to exclude 8 out of the 29 items in the test as non DIF. Van der Flier, Mellenbergh, Ader and Wijn (1984) applied the iterative logit procedure to the same group of 500 Tanzanian and 500 Kenyan on 29 items used by Van der Flier (1980). The total scores were split into five categories and used 15 iterations to test the χ^2 statistics at 0.05 and 0.01 levels of significance. Without iteration, no item was identified as exhibiting DIF at 0.01 level of significance, but only two items were identified as exhibiting DIF when the significant level was 0.05. With iteration, 15 out of 29 items were flagged biased. They thus concluded that the iterative procedure was a more effective means of detecting biased items even though it takes a lot of computer time.

Nenty (1986) used Scheuneman's modified chi-square procedure (SSX2); Rudner and Convey's TID-45 degrees item difficulty p-value; one-parameter latent trait Rasch model; and Cochran's Test Method (CTX2) using chi-square. Detection methods to determine cross-cultural validity of the scale 2, Form A Cattell Culture Fair Intelligence Test (CCFIT) items on three mutually remote and culturally disparate groups. The sample was made up of 600 Americans, 231 Indians and 803 Nigerians. The results from the four detection procedures revealed that 23 out of 46 items detected as being biased by the Schueneman's chi-squares technique, 28 items were identified by the transformed item difficulty major axis (TID-45^0), 35 were identified by the 1-parameter item characteristic curve method (Rasch) and 34 were identified by Cochran's chi-square method (CTχ^2). The agreement between the detection methods was shown by the high correlation indices between the detection methods which ranged from .81 to .96. Also, their high agreement ratios which ranged from .68 to .83 are indications of significant regional bias in CCFIT.

An indigenous study of DIF was conducted by Inyan (1991) on location bias on 522 rural and 512 urban Akwa Ibom and the Cross-River States examinees in Nigeria at the 1986 Common Entrance Examination in Mathematics. She used three detection procedures; the modified Scheuerman chi-square (SSχ^2) procedure, transformed item difficulties (TID-45^0) and item discrimination methods. The SSχ^2 method identified 13 items out of the 33 multiple-choice test items as being biased. Five items were flagged biased by the transformed item difficulty, while nine (9) out of the 33 items in the test were identified as exhibiting DIF by the discrimination procedure. Her finding supported the hypothesis that there is location bias in Mathematics achievement tests.

A study was made by Umoinyang (2002) to determine the presence of DIF in the 50 West African Examination Council (WAEC) General Certificate in Education (GCE) Ordinary Level Objective test items using the Mantel-Haenszel (M-H) statistic. He used a sample of 1,458 (586 northern and 902 southern), Nigerian candidates who took GCE O/L Mathematics in November/December 1990. Results showed that eleven (11) items out of 50 exhibited significant DIF at 0.05 level of significance with four (4) and seven (7) items favouring northern and southerners, respectively. Based on the results of the study, he concluded that the WAEC Mathematics achievement test designed and used for certification in 1990 for GCE was not free from the regional DIF.

Ndifon, Umoinyang and Idiku (n.d.) conducted a research aimed at finding out whether the 2010 junior secondary certificate examination (JSSCE) in Mathematics exhibits gender, school location and school ownership differential item functioning (DIF) in the Southern Educational zone of Cross River State. A sample of 1,833 candidates was selected from a population of 11,811 candidates who sat for the examination in 2010. The instrument for the study was the 60 multiple-choice JSS three. Mantel-Haenszel statistics and Scheuneman chi-square (SSX^2) detection methods were used to identify items that exhibited DIF in 2010 JSSCE in Mathematics. The findings showed that there was no significant gender differential item functioning as none of the detection method identified items that function differentially between males and females but a significant school location differential item functioning as the Mantel-Haenszel Statistics detected two items that function differentially against urban students while the Scheunemann chi-square ($SS\chi^2$) detected one item that functions differentially against urban students. The findings were not in agreement with the findings of Inyang (1991), Umoinyang (1991), Amuche and Fan (2014) and Mokabi and Adedoyin, (2014) who have reported on the existence of differential item functioning between urban and rural students. However, the findings of the study agreed with the findings of Inyang's (2004) who reported that rural students performed better than their urban counterparts. Based on these findings it was concluded that the 2010 junior secondary school certificate Mathematics examination significantly exhibited location DIF in arithmetic and algebraic processes.

Summary of Literature review

In almost all studies reviewed on basis of gender DIF and location DIF, the literature showed that there was an incidence of DIF items concerning gender, location, and type of school, content area, ethnicity and language are spoken. Studies also provide evidence that there are gender differences in performance on test items in mathematics, social studies, English Language, geography, economics and science that vary according to content even when content is closely tied to the curriculum.

When the criteria values of different methods are strictly taken into consideration, no consistency is observed among the methods in flagging items as DIF. Some methods produce liberal, whereas some others produce conservative results. However, the magnitude of the index values is comparable. The large sample sizes or the lack of model-data fit in IRT scaling might be the reasons for liberal results in the IRT-LR statistic.

However, LR is rather a very conservative technique in detecting DIF items. Among the DIF methods, the MH produces quite consistent results when the magnitude of other index values across the methods are considered. The superiority of the MH procedure is evident in the related literature if there is no nonuniform DIF in the test.

In terms of the location within which an examinee finds himself/ herself goes a long way to determine one's academic achievement than those from the poor location may tend to perform poor. That is, areas that are well equipped with learning facilities, qualified teachers, good roads and good communication networks which put examinees located there an advantageous position when compared to other examines from other areas where opportunities are inadequate or somehow lacking although examinees have the same ability.

In terms of attributing gender DIF to item characteristics, existing studies have produced inconsistent results. Compared to male students, high school female students, tend to perform poorer on geometry in terms of mean scores (Doolittle & Cleary, 1987). Likewise, Harris and Carlton (1993) reported males performed better on geometry compared to a matched group of females on the SAT. However, a study conducted in 2006 did not find geometry as a source of gender DIF.

Conclusion: The gap in the literature, purpose and implication of this study

According to Penfield (2010) "the condition of invariance is desirable because it ensures that the responses to the item reflect only content-relevant variance" (p.151). Comparison between persons and items can only be made when both the instruments and persons possess the property of invariance. This study is applying the measurement theory to test for invariance in empirical data. The lack of invariance of item parameters across groups indicates DIF. The DIF serves as validity evidence in the validation process of a test, which is an issue of fairness and bias. To understand DIF, performance differences by gender and location in Mathematics, English Language, Integrated Science and Social Studies WASSCE items were investigated. The literature review of DIF suggests a gap in the literature concerning investigating DIF in subject areas such as Mathematics, English Language, Social Studies and Integrated Science specifically in Ghana. It also suggests a gap in the use of not more than a year's data for analysis and the use of MH, LR and 3PL IRT in one study. This study is therefore done to fill these gaps by using MH, LR and 3PL IRT DIF detecting methods to analysed 2012-2016 (5 years) WASSCE core subjects data in Ghana.

In short, this study aimed to advance the understanding of DIF in WASSCE examinations Core subjects in Ghana.

CHAPTER THREE

RESEARCH METHODS

Introduction

This chapter discusses the methodology employed to answer the research questions. The chapter begins with a philosophical discussion of two main worldviews, namely objectivism/ positivism and constructionism/ interpretivism. To achieve the aims of this study a post-positivist view of the world that underlies the quantitative methods was adopted. The section that follows then presents the research design, population, sampling procedure, data collection instruments and procedures and data processing and analysis.

Research Paradigm

Proctor (1998) states that when planning a research study, clarification of the basic beliefs (worldviews/paradigms) can assist the researcher in understanding the relationships between ontology (what is reality?), epistemology (what can be known) and methodology (how the researcher can discover what he/she believes can be discovered). According to Proctor (1998), the early stages of a research study involve much thought, reflection and planning.

A research philosophy consists of beliefs about the process of *how* data about a phenomenon should be gathered, analysed, interpreted and used (Proctor, 1998. According to Hewege and Perera (2013), the philosophical assumptions underlying social Science research rest on four core assumptions. These are the ontology, epistemology, human nature and methodology, and these assumptions are consequential to each other, no matter the persuasion of the researcher. Thus, the researcher’s view of ontology affects his/her epistemological persuasion which, in turn, affects his/her view of human nature, and consequently, the choice of methodology logically follows from the assumptions the researcher has already made (Holden & Lynch, 2004).

Ontology relates to the researcher's basic assumption about the nature of reality in the world, and this is the cornerstone of all other assumptions. Arbnor and Bjerke (as cited in Hewege & Perera, 2013) argued that researchers might have different assumptions about the form and nature of reality. Thus, researchers have different assumptions about what things, if any, have existence or whether reality is the product of one’s mind (Burrell & Morgan, 1979).

The second assumption, epistemology, concerns the study of the nature of knowledge. That is, how is it possible, if it is, for us to gain knowledge of the world? Epistemology (what is known to be true), encompasses the various philosophies of research approach and is concerned with the nature, validity, and limits of inquiry by Rosenau (as cited in Holden & Lynch, 2004). Most researches in psychological and organisational science have assumed that reality is objective and out there waiting to be discovered and that this knowledge can be identified and communicated to others.

The third assumption, concerning human nature, involves whether the researcher perceives “man” as the controller or as the controlled (Burrell & Morgan, 1979). Finally, the methodology represents all the means available to social scientists (researchers) to investigate the phenomena of interest (Burrell & Morgan, 1979).

There are two major epistemologies/theoretical perspectives in Social Sciences: These are objectivism/positivism and constructionism/interpretivism. Positivism is related to natural sciences and what research does is to uncover an existing 'reality'. 'The truth is out there' and positivists try to develop a natural Sciences approach towards social sciences by largely employing methods developed for and in natural sciences (e.g., biology, physics or chemistry) where the researcher needs to be as detached as possible from the research to be objective (Muijs, 2004). Confidence in Science stems from the belief that scientific knowledge, being objective, is both accurate and certain.

People ascribe subjective meanings to objects, but science does not. In other words, Science discovers meaning inherent in objects. Hence, this is reflected in the positivist view that objects in the world have meanings and it is for researchers to discover these meanings. Positivists view the world as a world of regularities, a highly systematic and well-organised world (Crotty, 1998). Therefore, according to positivists, the world works according to fixed laws of cause and effect and to understand the truth about how the world works one should test the theories about these laws by developing reliable measurement instruments (Muijs, 2004).

However, the positivists' view of the world is not the everyday world we experience but arguably an abstraction of a 'lived' world. This led others (i.e., Comte, Popper, Kuhn) to consider alternative worldviews to represent this paradigm shift. This alternative worldview is called 'post-positivism'. Post-positivists accept the fact that natural Sciences do not provide a model for all social research. Post-positivists reject the absolute truth, total objectivity and certainty of findings, and focus rather on a certain level of confidence, a certain level of objectivity and approximate the truth of findings as best as they can (Crotty, 1998).

In a way, post-positivism is a humbler version of the scientific approach and hence is a less arrogant form of positivism. Post-positivists agree that reality can be influenced by the view of a researcher, a theory, a hypothesis and a view of how knowledge is created. Researchers are part of the world of observation and very rarely can detach ourselves from what is observed. This adds the influence of the researcher's values in understanding the reality of the world (Crotty, 1998). Postpositivism emerged as a reaction to the critiques of positivism presented by subjectivists/constructivists.

According to Muijs (2004), to be completely objective is problematic. Historical research has shown that research studies and findings are influenced by the beliefs of people conducting the research and the political or the social climate when the research is conducted. This led to the increasing popularity of other epistemologies and theoretical perspectives to view and model the reality of the world. These perspectives are referred to by different names (symbolic interactionism, interpretivism, phenomenology, constructivism, etc.). Constructivism seems to be the most popular among these perspectives. It rejects the positivist view that 'the truth is out there' to be discovered but believes 'the world is out there', whose meaning can be constructed through social interaction between human beings and the world they live in. Therefore, from a constructivist viewpoint, meaning (or truth) cannot be described simply as objective or subjective. They argue that meaning is constructed and not given. Objectivity and subjectivity are two extreme views of the reality of the world.

According to Heidegger and Merleau-Ponty (as cited in Crotty, 1998), objectivity and subjectivity need to be brought together and held together indissolubly and constructionism does precisely that. The positivist/post-positivist theoretical perspectives usually underlie quantitative methods, whereas the constructivist theoretical perspective often underlies qualitative methods. Regardless of which design has been selected, both methods strive to answer the research questions through a rigorous systematic enquiry and are concerned with a contribution to knowledge.

To conclude, this thesis adopted a post-positivist theoretical perspective to answer the research questions and the methods selected are those that best serve the aims of this study. By taking this position, this study's knowledge claims focus on the probability of a certain level of objectivity and approximate truth rather than certainty and absolute truth. By taking this stance the research design of this study consisted of mostly quantitative methods.

Research Design

Research designs are procedures for collecting, analyzing, interpreting, and reporting data in research studies which guide the methods decisions that researchers must make during their studies and set the logic by which they make interpretations at the end of their studies (Morse, 2016). The design refers to the overall structure or plan of the study (Singleton & Straits, 2010). The research design that was used for the study is cross-sectional.

According to Sedgwick (2014), a cross-sectional study is defined as an observational research type that analyzes data of variables collected at one given point of time across a sample population. Population or a pre-defined subset. This study type is also known as cross-sectional analysis, transverse study or prevalence study.

Cross-sectional studies are generally quick, easy, and cheap to perform. In this study, data was collected across WASSCE examinees on 2012-2016 to estimate the presence of gender and location DIF in the 2012-2016 WASSCE core subjects in Ghana.

Crotty (1998) suggested that the four fundamental elements in the research process are epistemology, theoretical perspective, methodology, and methods. Crotty (1998) further suggested a framework that aims to provide a sense of stability and direction as a researcher proceeds on the research process to serve his/her research purposes. Figure 14 provides a visual framework for the research processes of this study.

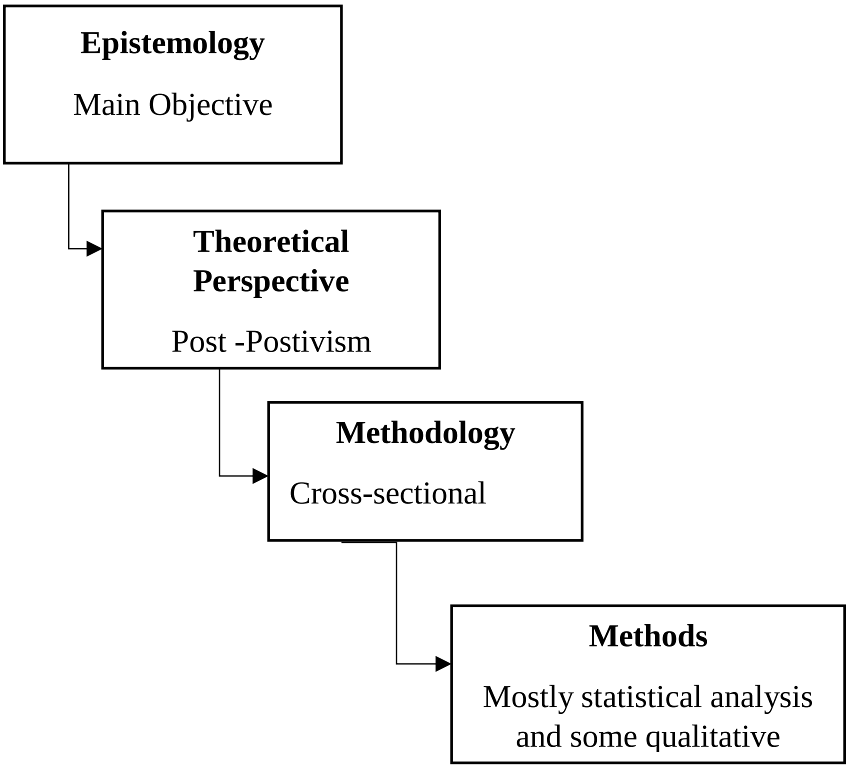

Figure 14: Interconnections of four Basic Elements in Research Design

As Figure 14 shows, this study lies within the objectivist epistemology views on how things exist, if truth and meaning reside in objects as well as believing that the world is out there to be discovered. This epistemology underlies the post-positivist stance. Research in this theoretical perspective selects a survey research study, and mostly employs quantitative methods of statistical analysis but with some qualitative methods. The purpose of this study was to examine gender differences and the validity of test scores from the perspective of DIF. Test scores are used to make inferences of pupils' ability on a construct being measured and these scores should reflect that construct and no other irrelevant constructs. This study employed a quantitative research design to fulfil the purpose of this study.

Population

The target population of this study comprised of all public Senior High School Form 3 (SHS3) candidates from five regions who sat for the 2012 to 2016 WASSCE. The accessible population comprised of students from the 20 out of 235 senior high schools who sat for English Language, Mathematics, Integrated Science and Social Studies WASSCE (i.e., core subjects from 2012 to 2016; an average of 289,210 candidates). These students have passed through nine years of basic education and three years of senior high school education before sitting for the WASSCE.

Sampling Procedure

Simple random technique was used to select a sample of twenty (20) senior high schools in Ghana from 235 schools. These 20 schools were made up of 16 single-sex schools and four mixed schools. Purposive sampling was used to select 16 single-sex schools for the gender DIF analysis whiles all 20 schools were used for the location DIF analysis. The corresponding regions where these selected schools were located was used for the locational DIF study respectively. In each of the sampled schools, the scores all the students who wrote the 2012-2016 WASSCE multiple-choice results were used. The distribution of the sample by gender and location (region) are shown in Tables 6 and 7.

Table 7: Sample Distribution of Core Subjects by Gender (2012-2016)

Year	Gender	English Language	Core Maths	Int. Science	Social Studies
2012	Male	3329	3352	3133	3357
	Female	2791	2790	2796	2793
2013	Male	7640	7477	7664	7285
	Female	6572	6505	6656	6357
2014	Male	3482	3236	3907	3902
	Female	3348	3363	3486	3350
2015	Male	4292	4285	4287	4287
	Female	3700	3712	3707	3705
2016	Male	5197	5018	5125	4934
	Female	4615	4926	4686	5113

Source: WAEC (2017)

Table 8: Sample Distribution of Core Subjects by Location (2012-2016)

Year	Location	English Language	Core Maths	Int. Science	Social Studies
2012	GAR	2051	1495	2075	2069
	ER	2426	2435	2422	2430
	CR	1556	1495	1286	1494
	WR	946	964	960	966
	VR	785	782	783	784
2013	GAR	4451	4602	4601	4608
	ER	5180	5325	5483	5096
	CR	3937	3589	3886	3592
	WR	2217	2232	2222	2224
	VR	1791	1692	1697	1792
2014	GAR	2102	2180	2664	3076
	ER	2644	2661	2654	2653
	CR	1854	1538	1859	1858
	WR	1294	1295	1293	1293
	VR	895	890	889	884
2015	GAR	2630	4285	2622	2631
	ER	2998	3005	3000	3005
	CR	1957	1968	1971	1959
	WR	1600	1594	1597	1594
	VR	948	946	946	945
2016	GAR	3311	3115	3254	3059
	ER	3796	3790	3623	3801
	CR	2257	2566	2487	2071
	WR	1837	1849	1833	2519
	VR	1277	1252	1243	1244

Source: WAEC (2017)

Data Collection Instrument

The West African Senior School Certificate Examination (WASSCE) is an achievement test administered to school candidates in the third year of the Senior High School. The examination measures the extent to which the candidates have understood the content of the teaching curriculum approved by the country. The West African Examinations Council (WAEC) as an international examining body periodically reviews its test administration procedures intending to minimize distortions and come out with very reliable scores. No specific instrument was developed for this study because the study analysed WASSCE data that had been collected. An introductory letter from the department signed by Head and both supervisors was sent to WAEC to obtain permission for 2012-2016 WASSCE (Core Subjects) dataset. The data was given in the form of codes and no names were given. This, in turn, ensured the confidentiality and anonymity of the respondents. An ethical clearance letter was collected from the Institutional Review Board (IRB) from the University of Cape Coast on submission of my proposal. The study did not include names of students or schools, and the data (students' scores) were kept safe and confidential in line with the recognition of the protection of human rights.

The West African Examinations Council (WAEC) procedures for conducting their examination have to be classified into three main sections namely,

1. Pre-examination procedure
2. During examination procedure
3. Post examination procedure

1. Pre-examination procedure

A. Data Capturing and Validation of Entries

Schools are expected to ensure that all students' information are validated before uploading them on the internet. When registration is completed, the entry information is printed to enable the candidates to confirm their entries and that errors are corrected. Rules and regulations together with the consequences of violating them are made available to candidates. The inspection of schools for recognition and approval is done by the WAEC and GES every year before examinations are conducted.

B. Selection of Supervisors

Proficiency and dedication of personnel used in the administration of tests influence the quality of the assessment. Though both public and private schools register candidates for the examinations, only personnel from the public schools are used as supervisors. This personnel are usually senior staff, nominated by the various Education Districts and sometimes universities. The final selection is done by the education offices in the various regions. This way, accountability for actions or inactions is achieved. The supervisors are trained on how to conduct standardized tests after which a manual of testing procedures is given to each of them.

C. Selection of Item Writers

The West African Examinations Council has a formal test development department that uses formal processes to review and to scrutinize some items selected from the item bank during the test development process. The items are reviewed before field testing by content specialists (i.e., subject officers at WAEC) as well as after field testing.

At WAEC, practising teachers and lecturers are involved throughout the test development process. Participating classroom teachers, lecturers and administrators from qualified content (subject) specialists have their names in the approved list which is revised every two years by the final appointment WAEC committee. This approved list consists of WAEC assistant examiners recommended by Heads of Department of participating schools through subject officers at WAEC. The teachers are provided with basic item-writing principles using content-specific manuals (i.e., item specification table) which contain explicit guidelines with examples and the examination syllabus which is a subset of the teaching syllabus.

D. Preparing the test

The test development process contains three general steps: Item writing, field testing, and creating the final form of the test. Item writing begins when teachers, lecturers and other subject experts (these people are neither teachers nor lecturers) are nominated to serve on the item writing committees.

WAEC uses a rotating selection procedure to ensure that teachers do not become too familiar with the items, hence the change of item writers every two years. The item writers meet between three to ten days once in a year to develop new items. At times, item writers from the Gambia, Nigeria, Sierra Leone, Liberia and Ghana are commissioned to provide test items within some specified number of days. If possible, the items are developed using a realistic context that would be familiar or topical for students in the regions.

2. Procedure during the examination procedure

A. Try Out of the Test:

Once the test is prepared, it is time to confirm the validity, reliability and usability of the test. Items to be tried out are reviewed to identify defective and ambiguous ones, determine the difficulty level of the items and therefore the test and to determine the discriminating power of the items. The items are then compiled into sets (series) by the subject officers and field-tested. The field tests undergo an internal review by both subject officers and heads of department in WAEC before they are administered to students from some selected schools. Members of the internal review committee are from the test development department unit at WAEC, Ghana. The purpose of the internal review is to examine the field tests for content validity, curricular validity, item appropriateness (e.g., wording, length, and interest), bias (e.g., gender, cultural, disability), balance to the test blueprint, and tone.

The field-tested items (i.e., only multiple-choice items) are then evaluated using classical item analysis. Three sources of information (i.e., the internal review committee's comments, the teachers'/students' comments from field testing, and the statistical results) are considered by the item developers during the selection of the final test items.

The final stage involves creating the final form of the test which includes the constructed response type items. This stage involves a second review by the moderation panel which includes the subject heads of departments, chief examiners and other subject-related experts. The purpose of this review is to finalize the selection of the test items and to ensure that the test meets the curriculum, assessment, and achievement standards. The moderation panel scrutinizes each test item looking for content and curricular match, item appropriateness, and possible biases based on the item analysis results that were reported after the field test (i.e., checking for validity and reliability of the test results). The test is then edited by subject officers and heads of department. The final version of the test is then sent to Security Printing Department (SPD) for printing and packaging. They examine the test for the accuracy of information presented in each item, wording, and possible biases. Once this step is completed, the subject officer and heads of department sign off the test and the extensive development process is finished. These items are kept at the item bank and used at least two years after it has been written.

B. Evaluating the tried-out test:

Evaluation is necessary to determine the quality of the test and its responses. Quality of the test implies that how good and dependable the test result is? That is the validity and reliability of test scores. The quality of the responses helps to check for misfits in the test. It also enables WAEC to evaluate the usability of the test scores in the general classroom situation.

2. Post examination procedure

A. Checking irregularities

At the end of each examination, all acts of malpractices are collated. Centres where collusion was prevalent, are listed region by region for thorough scrutiny by examiners. Such a list is circulated to all marking venues.

B. Marking of scripts

Marking of scripts commences four to five weeks after the examination. In preparing scripts for marking, control cards allocating the specified number of scripts to examiners by numbers rather than by names are produced to ensure that scripts are not marked within the zones where they were generated.

Selected qualified zonal examiners and chief examiners meet for 5 days to further standardize answers to questions in their various subjects and thereafter coordinate other assistant examiners for three days. Consistency is ensured through the standardized marking of dummy scripts before *live* scripts are issued to the examiners. During marking the activities of assistant examiners are monitored and information is provided to the examiners on how to detect malpractice in the scripts. The team leaders select randomly, one out of every ten scripts scored by an assistant examiner for vetting. The chief team leaders also select randomly, one to two of every ten scripts of team leaders' scripts and vet them. Where there is a difference of more than + 2 marks in a question, the assistant examiner is made to remark the scripts of all the candidates who answered that question number.

In addition to the vetting of marked scripts by team leaders and senior examiners, checkers are recruited to go through the marked scripts for possible errors of addition, omission and transfer of scores into the Optical Mark Reader which in turn is read against the mark sheets by qualified staff. These checks are to reduce possible human errors. Examiners whose marking were questionable are barred from further participation in marking. A list of such examiners is usually circulated to all marking venues across the country during subsequent marking exercises.

D. Movement of marked scripts (end documents) to scanning zones

From the marking venues, mark sheets, as well as the optical mark readers, are serially arranged by subject and transported in separate vehicles to offices from where they are sent to designated scanning zones. Scripts are evacuated for storage in designated offices and kept for three to five years. This period of storage is to ensure that all queries are resolved and that results are consolidated.

E. Release of Results

Once the grades are fixed, by converting score into standard scores through standard-setting, the Computer Services Division releases results except for those candidates whose results are to be withheld for irregular offences during the examinations. Results of examination cheats are cancelled, and, in some cases, the cheats are barred from sitting for the WASSCE for a specified period after investigations and due consideration by the appropriate Committee of Council. Chief examiners for various subjects also come out with reports every year after results have been released to educate people especially heads teachers and parents on the strengths and weaknesses identified from the examinations.

F. Review of marked scripts if requested

Schools and individual candidates may request a review of their scripts after the release of results. The request for review arises from dissatisfaction with a candidate's results and the desire to know what might have accounted for such a performance. The review of scripts is done to identify the weaknesses of the candidates and correct any defects in teaching and the way candidates tackle questions. For school-based examinations, requests for review of their scripts are entertained from heads of schools. Private candidates also may request for review of their scripts. Given this, all requests for review should be done within the stipulated period. The review of scripts attracts a fee that is reviewed periodically. Request for the review must be made formally to the Senior Deputy Registrar/Head of Test Administration Division.

Data Processing and Analysis

The data for this study were gathered from responses of candidates in 50 multiple- choice questions in Core Mathematics, Integrated Science and Social Studies and 80 to 100 multiple-choice items in English Language and administered by WAEC, for 2012-2016 WASSCE. Person-by-item response matrix obtained from WAEC office was used to map out the ability groups for each subgroup for the analysis of DIF. Data from the sixteen (16) single-sex schools was used to test research hypotheses 1, 3 and 4 (gender DIF) while data on all the twenty (20) schools were used to test research hypotheses 2, 5 and 6 (Location DIF). For gender DIF analysis, male students were used as the reference group because it is believed male students have an advantage over their female students' counterparts in terms of core subjects' performance. Candidates whose schools were located in Central (CR) and Greater Accra (GAR) regions were used as the reference groups because it is assumed that these two regions are very resourceful in terms of educational facilities and human resources. Female candidates and candidates who schooled in Eastern (ER), Western (WR) and Volta (VR) regions of Ghana constituted the focal groups for the study.

The multiple-choice items were scored 1 for the correct option and 0 for the wrong option with a maximum score of 50 each for Social Studies, Core Mathematics and Integrated Science, and 78 (2015), 80 (2016) and 100 for 2012-2014 English Language and a minimum of 0 for all subjects. To test the research hypotheses and answer the research question on understanding DIF by gender and location in the English language, Core Mathematics, Social Studies and Integrated Science assessments, the data on the research question was analysed at item-level. The IRT, MH and the LR procedures were employed to detect DIF for item-level analyses.

In this study, all tests of hypotheses were carried out at the 0.05 level of significance. Statistical bias or DIF is inferred if the probability associated with the obtained chi-square value is less than the set alpha level of 0.05 with one degree of freedom. Gender and location-based DIF refer to the differing probabilities of success on an item between the male and the female examinees and among examinees who schooled in Greater Accra Region (GAR), Central Region (CR), Western Region (WR), Eastern Region (ER) and Volta Region (VR) in the English Language, Mathematics, Integrated Science and Social Studies . Research hypotheses 1 to 6 were tested using MH, LR and 3PL IRT whiles research question was answered using frequencies, bar graphs and percentages.

The MH was used because of its associated test of significance and efficiency regarding computer time in detecting uniform DIF. The LR procedure was selected because of its power in detecting uniform and nonuniform DIF simultaneously. In contrast, the IRT is more sensitive in detecting both uniform and non-uniform DIF because of its equal discrimination assumption of items across different ability levels of examinees.

MH yields a chi-square test with one degree of freedom to test the null hypothesis that there is no significant relationship between group membership and test performance on the test items between the reference and the focal group. MH uses an internal matching variable (total test score) when evaluating the suspect item, to ensure that the examinees at each score level are comparable.

In the LR, the analysis was done between the matched examinees (scores of male and female for gender and scores for students who schooled in CR, VR, WR, ER and GAR for location) and the independent variables (gender and location). The dependent variable or logit for each of the matched examinees is the odds or likelihood of getting the item right. A significant score in each of the matched examinees indicates that examinees with higher total score tend to score better in the examination. A significant location DIF and gender DIF indicate that the odds of getting an item right are different between the male/female and students who schooled in CR, WR, ER, VR and GAR.

For the three-parameter model IRT, stable and accurate estimation of the item parameters requires large numbers of examinees over a broad range of ability. It is generally recommended that samples of at least 1000 be used for the three-parameter model (Baker, 1987; Hambleton, 1994; Hambleton & Swaminathan, 1985). The accurate estimation of the c-parameter also requires large numbers of examinees at (very) low ability levels.

Detecting DIF using DIFAS Mantel-Haenszel method (M-H)

The M-H method works by first dividing subgroups into the reference group (e.g., males) and the focal group (e.g., females). The focal group is of primary interest in the analysis and is compared to the reference group after being matched on θ (Uttaro & Millsap, 1994). The total test score usually serves as the θ estimate, and the performance (i.e. item endorsement rates) of the reference and focal groups is compared at unit intervals of θ weighted by the number of examinees at each level (Scheuneman & Gerritz, 2005). From this comparison, an odds-ratio estimator is calculated, and a χ^2 test of significance is carried out to assess the presence of DIF.

To assess the degree of DIF present, DIFAS as a programme analyses item characteristic and provides information about descriptive statistics (mean, standard deviation, minimum and maximum score) and the frequencies of choice for each category of item. It also analyses DIF for both dichotomous and polytomous items. Non-uniform DIF is analysed according to the description of the empirical item characteristic curves.

The results offered by this programme are displayed in two tables. The first of these shows the DIF statistics, while the second presents the conditional differences in the mean item scores between the reference and focal groups at ten intervals across the matching variable continuum. In the DIF analysis, the programme includes the following statistics: the Mantel-Haenszel chi-square statistic, the Mantel-Haenszel common log-odds ratio, the standard error of the Mantel-Haenszel common log-odds ratio, the Mantel-Haenszel log-odds ratio divided by the estimated standard error, the Breslow-Day chi-square test of trend in odds ratio heterogeneity and the combined decision rule and the Educational Testing Service (ETS) categorisation scheme. Each of the DIF statistics conducted by DIFAS for dichotomous items is briefly described as follows:

Mantel-Haenszel Chi-Square (MH CHI) – The Mantel-Haenszel chi-square statistic (Holland & Thayer, 1988; Mantel & Haenszel, 1959) is distributed as chi-square with one degree of freedom. Critical values of this statistic are 3.84 for a Type I error rate of 0.05 and 6.63 for a Type I error rate of 0.01.

Mantel-Haenszel Common Log-Odds Ratio (MH LOR) – The Mantel-Haenszel common log-odds ratio (Camilli & Shepard, 1994; Mantel & Haenszel, 1959) is asymptotically normally distributed. Positive values indicate

DIF in favour of the reference group, and negative values indicate DIF in favour of the focal groups.

The Educational Testing Services Categorization Scheme (ETS) – The ETS categorization scheme (Zieky, 1993) categorizes items as having small (A), moderate (B), and large (C) levels of DIF.

Longford, Holland and Thayer (1993) comment that if an item is classified as A, one can still include the item. If the item is classified as B, one should examine if there are other items one can choose to include in the test instead, i.e., an item with a smaller absolute value of $MH\ D - DIF$. Finally, an item classified as C should only be chosen if it meets essential specifications but documentation and corroboration by a reviewer are required.

It should also be noted that the number of test-takers in the focal group can have a strong influence on the DIF categorization, i.e., more items are classified as category B and C with larger focal and reference group sizes.

In addition to the guidelines for interpretation for MH according to DIFAS, the Educational Testing Service has proposed values of $MH\ \Delta$ for classifying the magnitude of the DIF as negligible, moderate or large (Zwick & Ericikan, 1989). Roussos and Stout (1996a, 1996b) modified the values and gave the following guidelines to aid in the interpretation of DIF:

1. Type A Items - negligible DIF: $|\ MH\ \Delta\ | < 1$,
2. Type B Items - moderate DIF: MH test is significant and $1.0 < |\ MH\ \Delta\ | < 1.5$,
3. Type C Items - large DIF: MH test is statistically significant and $|\ MH\ \Delta\ | > 1.5$.

The Mantel-Haenszel procedure is considered by some to be the most powerful test for uniform DIF for dichotomous items (Holland & Thayer, 1988). The Mantel-Haenszel procedure is easy to conduct, has an effect size measure and test of significance, and works well for small sample sizes. However, the Mantel-Haenszel procedure detects uniform DIF only (Narayanan & Swaminathan, 1994; Swaminathan & Rogers, 1990).

Research also indicates that the Mantel-Haenszel can indicate the presence of DIF when none is present in the data are generated by item response theory models (Meredith & Millsap, 1992; Millsap & Meredith, 1992; Zwick, 1990). Other factors that influence the performance of the Mantel-Haenszel include the amount of DIF, length of the test, sample size, and ability distributions of the focal and reference groups (Fidalgo, Mellenbergh, & Muniz, (2000); French & Maller, 2007; Jodoin & Gierl, 2001).

Analysis Using Logistic Regression (LR)

Swaminathan and Rogers (1990) applied the LR procedure to DIF detection. This was a response, in part, to the belief that the identification of both uniform and non-uniform DIF was important. The strengths of this procedure are well documented. It is a flexible model-based approach designed specifically to detect uniform and non-uniform DIF with the capability to accommodate continuous and multiple ability estimates.

Furthermore, simulation studies have demonstrated comparable power in the detection of uniform and superior power in the detection of non-uniform DIF compared to the MH (Mantel-Haenszel) and SIB (Simultaneous Item Bias) test procedures (Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990). These studies also identified two major weaknesses in the LR DIF procedure:

(1) the Type I error or false-positive rate was higher than expected; and (2) the lack of an effect size measure. LR has a formal mathematical equivalence to the log-linear model approach of Mellenbergh (1982). Coefficients for the group, total score and interaction terms are estimated and tested for significance with a model comparison strategy.

However, LR is highly like the standard ordinary least squares regression. It can be conceptualized as an equation that uses group, ability and group-by-ability terms to predict whether an item response is right (1) or wrong (0). This property is desirable for didactic purposes.

LR uses the examinee as the unit of analysis and has the following form:

$P(u|x,g) = \frac{e^{(1-u)[-\beta_0-\beta_1 x-\beta_2 g-\beta_3(xg)]}}{1+e^{[-\beta_0-\beta_1 x-\beta_2 g-\beta_3(xg)]}}$(Swaminathan & Rogers, 1990).

Where:

g: represents group membership (0 for focal group (female) and 1for reference group (male)).

x: the matching group (the observed total test score).

u represents the item response value (0 for an incorrect answer and 1 for a correct answer).

xg: represents the interaction between the matching variable and the group variable.

β0: β1, β2 and β3: parameters to be estimated.

The above equation is used for predicting the probabilities of correct and incorrect responses to each dichotomously scored item, given an observed total test score and its associated group membership. Once the estimates of the four coefficient parameters, $\beta_{0,}$ β1, β2 and β3, for an item are obtained from a sample of test responses, the usual likelihood ratio chi-square tests of significance of the estimates of β_2 and β_3 are conducted to examine if DIF exists. The null hypothesis is that $\beta2 = \beta3 = 0$. An item shows uniform DIF if $\beta_2 \neq 0$ and $\beta_3 = 0$ with one degree of freedom and non-uniform DIF if $\beta_3 \neq 0$ (whether $\beta2 =0$) with 1 degree of freedom (Swaminathan & Rogers, 1990).

CHAPTER FOUR

RESULTS AND DISCUSSION

Introduction

This study aimed at examining differential item functioning (DIF) of a test using different detection methods. The DIF detection methods selected are mathematical models, whose results and discussions are reported in this chapter. The unit of analysis is at the item-level, where the Mantel Haenszel (MH), the Logistic Regression (LR) and 3PL IRT measurement model (IRT) procedure were used to detect DIF. The dataset used consisted of the scores on multiple-choice test items of the 2012-2016 WASSCE core subjects.

The verifications of the models' assumptions were based on this dataset. The assumptions of these models need to be sufficiently fulfilled with empirical data before valid inferences could be made from the results. Therefore, this chapter answers the following questions on the models' assumptions.

For the MH Procedure: Two basic assumptions are considered.

1. Observations are independent of each other. In practice, this means that each observation comes from a different examinee, that the examinees were randomly selected from the population of interest, and that no specific group of examinees is purposefully omitted.
2. All observations are normally distributed

The Mantel-Haenszel analysis provides two closely related pieces of information. First, it provides statistical tests of whether the odds ratios are equal (homogeneous) or unequal (heterogeneous) across strata. Second, it provides an estimate of the odds ratio of the exposure variable, adjusted for the strata variable.

Considering the second assumption which states that all observations must be normally distributed, the data collected for this study satisfied this assumption . Plots to show this assumption are shown in figures 15 to 34 for all the four subjects across the five years under study.

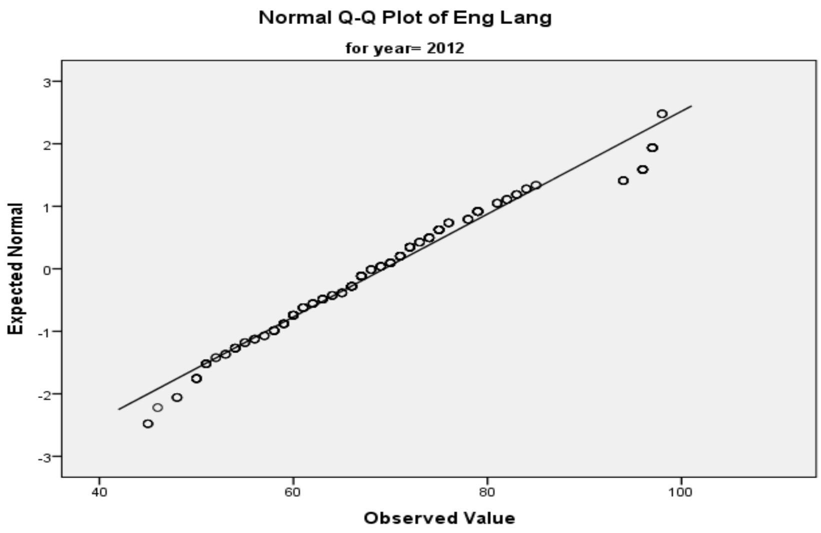

Figure 15: Plot of English Language scores for 2012.

The plot in Figure 15 show that the observations are normally distributed leaving few of them skewed.

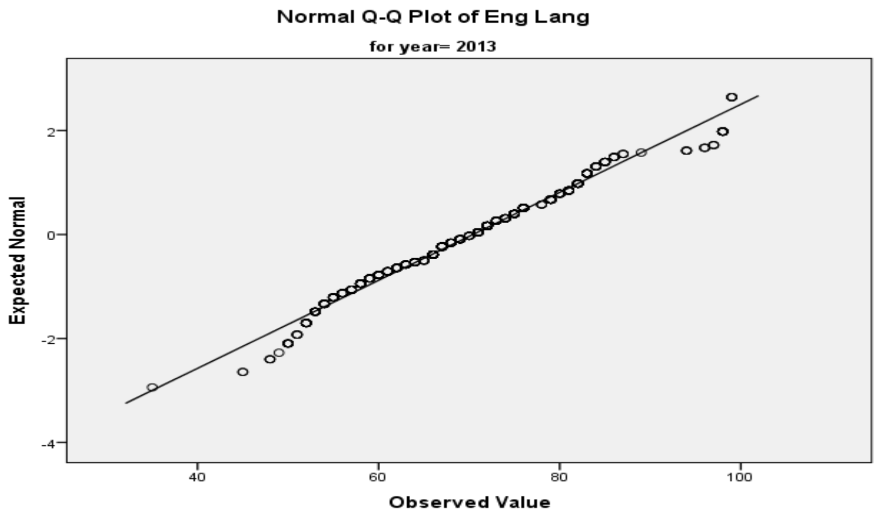

Figure 16: Plot for English Language scores for 2013.

From Figure 16, observations seem to be normally distributed with few observations slightly deviated from the line.

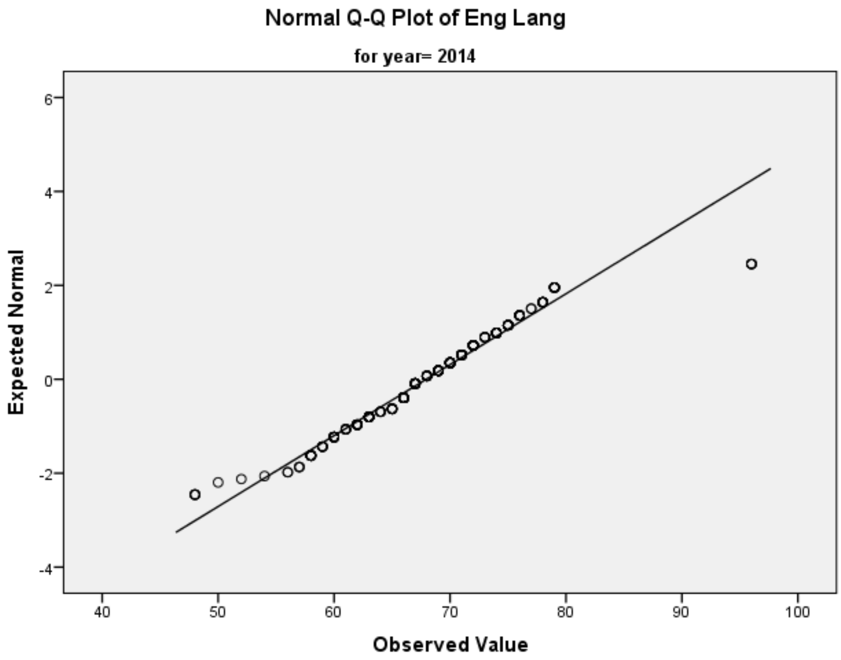

Figure 17: Plot of English Language scores for 2014

Figure 17 shows that the observations are normally distributed except for one observation that deviates from the line of fit.

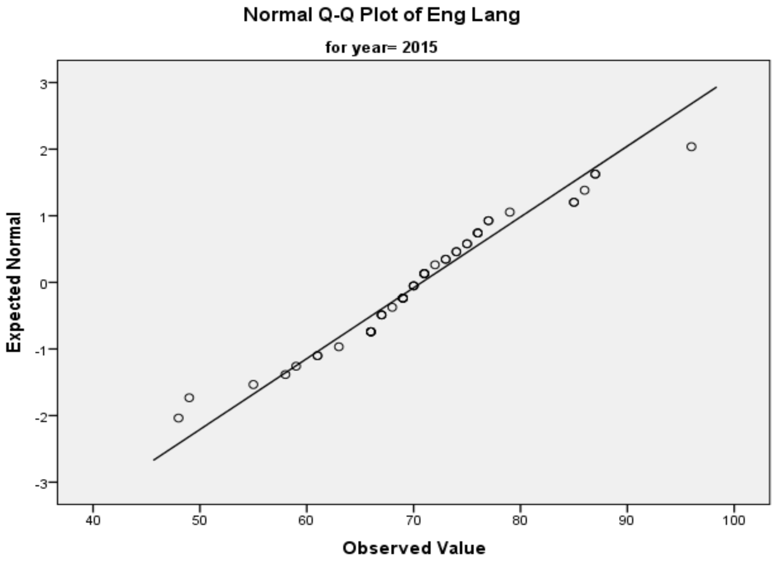

Figure 18: Plot of English Language scores for 2015

Plots in Figure 18 are quite close to the line of fit, indicating that the observations are not strongly normally distributed.

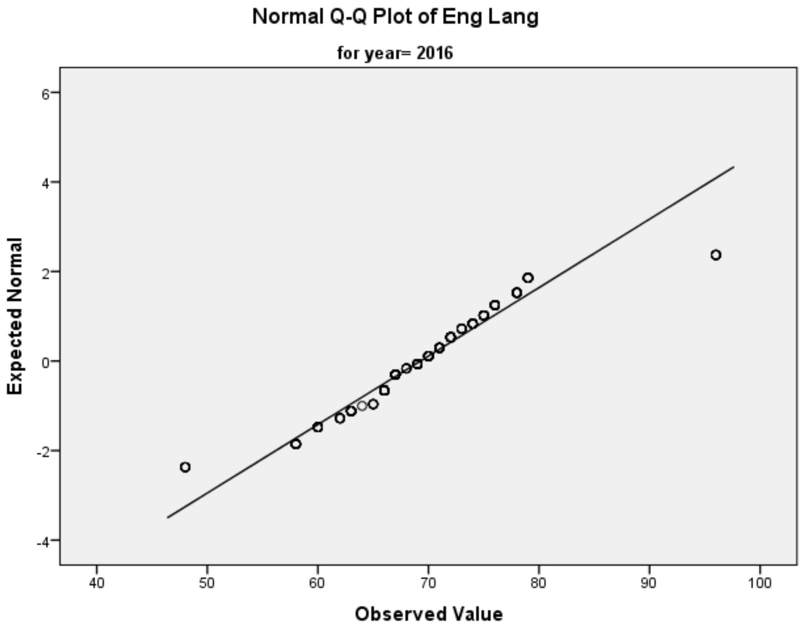

Figure 19: Plot for English Language scores for 2016

In Figure 19, data quite normally distributed. The observations of the English language scores seem quite normally distributed as shown in Figures 15 to 19. Figure 20 depicts the normality test of mathematics data.

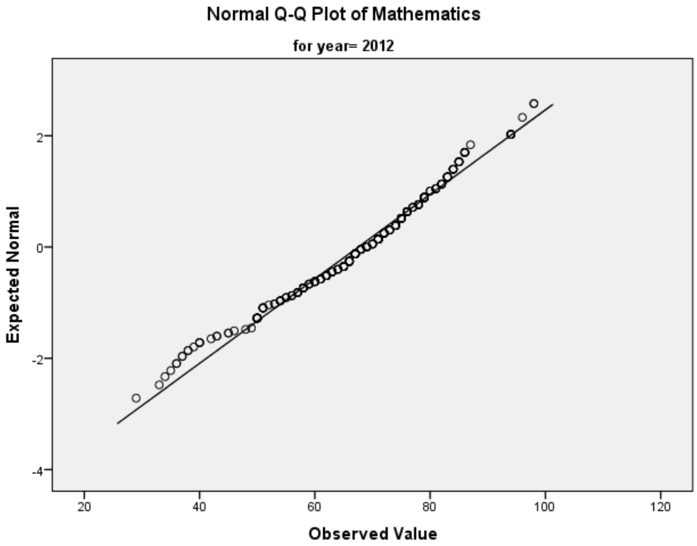

Figure 20: Plot of Mathematics scores for 2012

Observations in Figure 20 are quite normally distributed. Figure 21 depicts a graph showing the 2013 mathematics dataset.

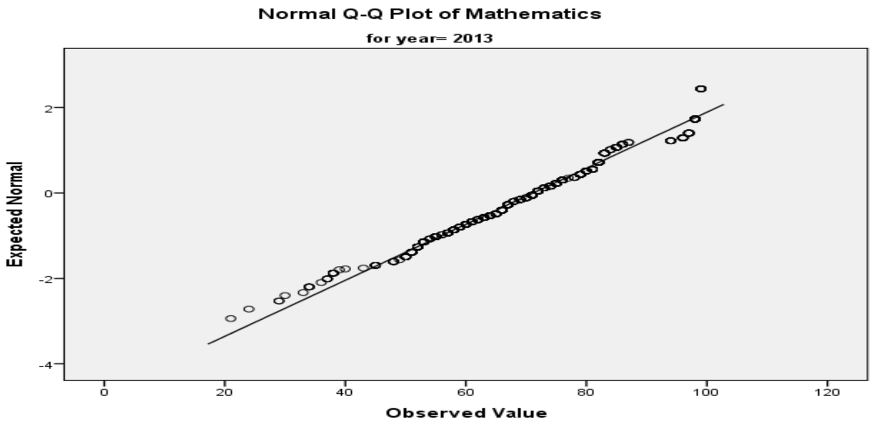

Figure 21: Plot for Mathematics scores for 2013

Observations in Figure 21 are quite normally distributed.

Figure 22 shows a normality test plot for 2014 mathematics dataset.

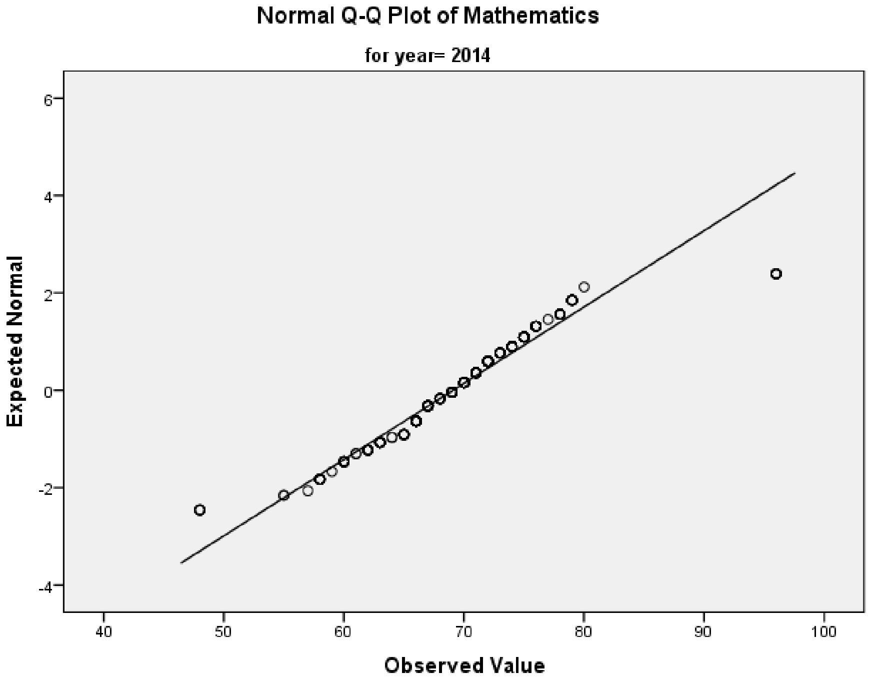

Figure 22: Plot of Mathematics for 2014.

Observations in Figure 22 are quite normally distributed.

Figure 23 shows the normality test plot for 2015 mathematics dataset.

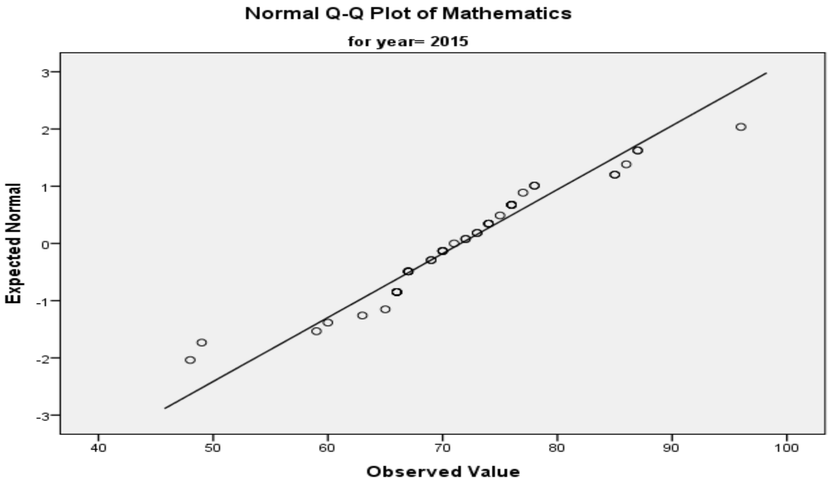

Figure 23: Plot of Mathematics for 2015

Figure 23 depicts a normally distributed graph for 2015 Mathematics dataset. The observations of the 2015 mathematics scores seem quite normally distributed 2016 mathematics dataset was tested for normality as shown in Figure 24.

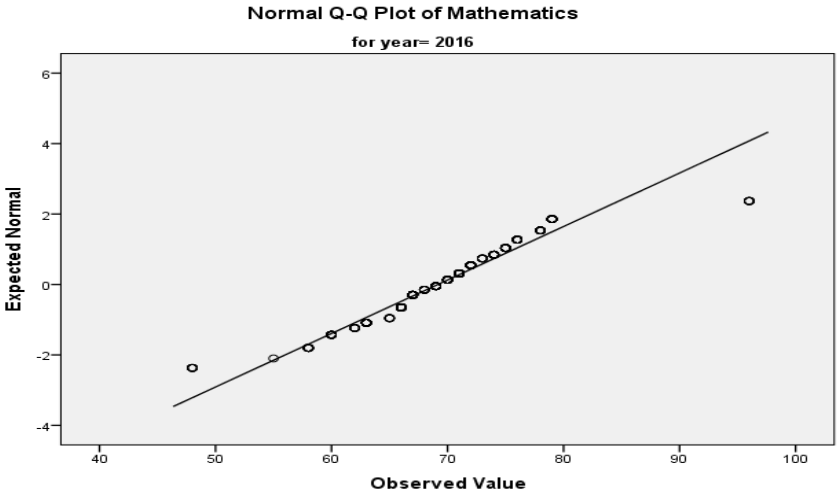

Figure 24: Plot of Mathematics scores for 2016

Plots in Figure 24 show that observations of 2016 mathematics test items are normally distributed. Figure 25 depicts a normality test plot for the 2012 Integrated Science dataset.

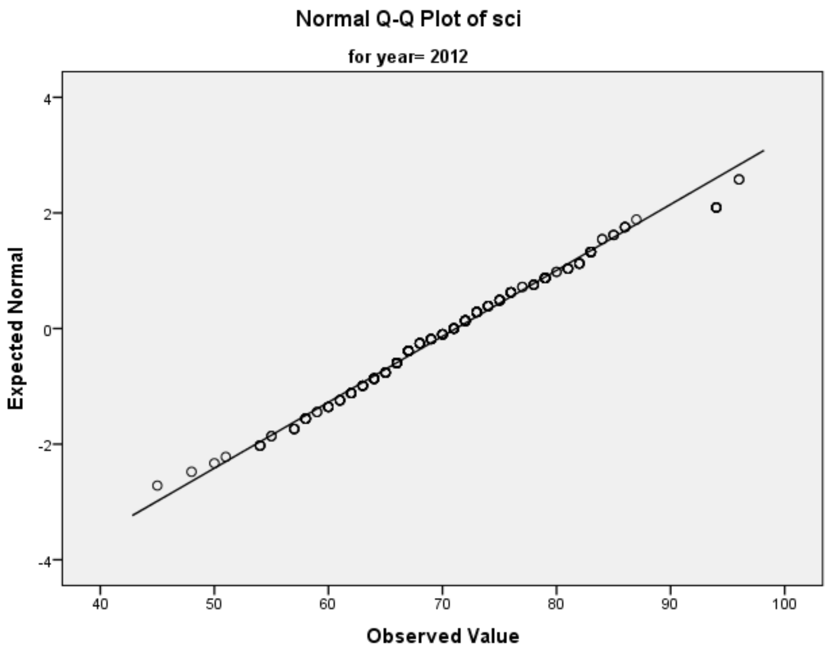

Figure 25: Plot for Integrated Science scores for 2012

For the 2012 Integrated Science scores, the data showed a normal distribution as shown in Figure 25. Figure 26 depicts a normality test plot for the 2013 Integrated Science data.

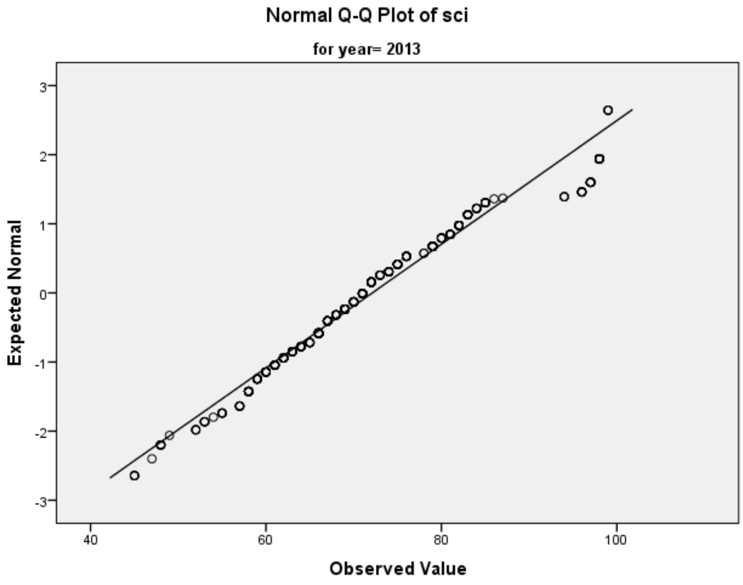

Figure 26: Plot of Integrated Science scores for 2013

The data were normally distributed as shown in Figure 26 for the 2013 Integrated Science scores. Figure 27 depicts a normality test plot for the 2014 Integrated Science data.

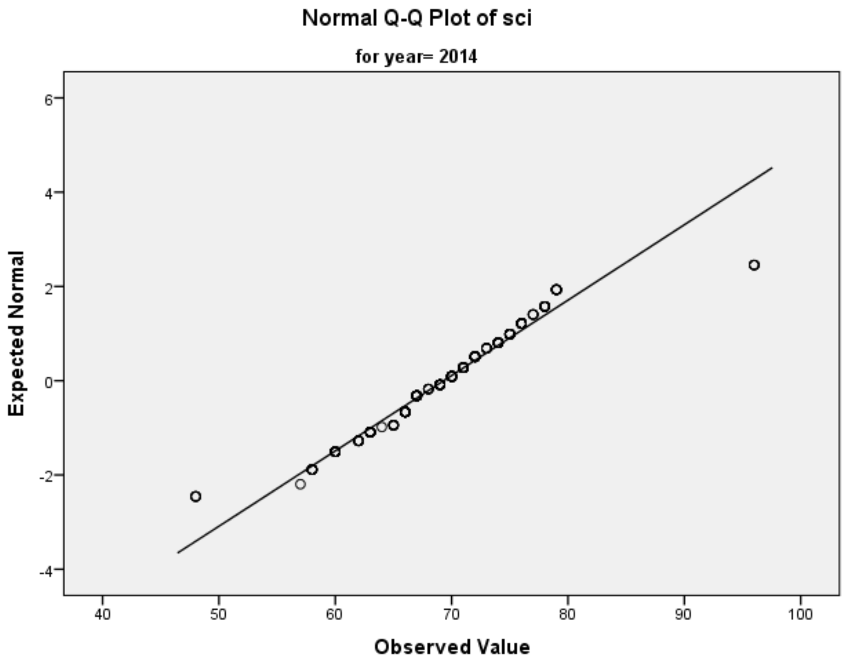

Figure 27: Plot of Integrated Science scores for 2014

For the 2014 Integrated Science scores were normally distributed as shown in Figure 27.

Figure 28 depicts a normality test plot for 2015 Integrated Science data.

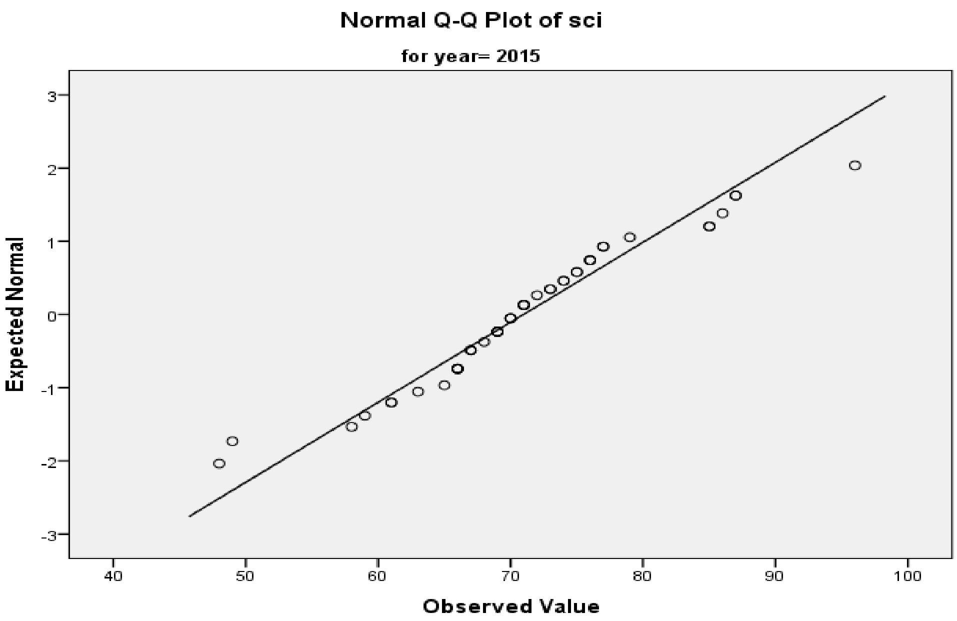

Figure 28: Plot of Integrated Science scores for 2015

2015 Integrated Science scores were normally distributed as shown in Figure 28.

Figure 29 depicts a normality test plot for the 2016 Integrated Science data.

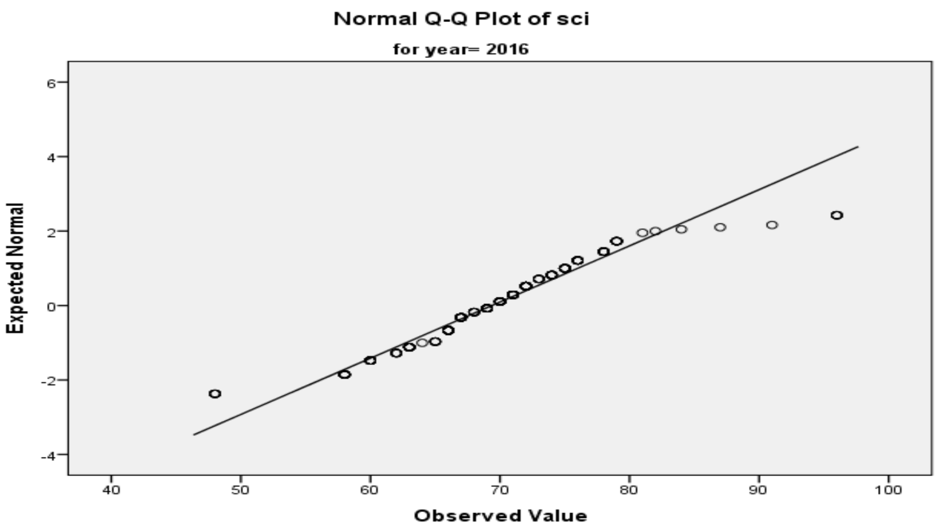

Figure 29: Plot of Integrated Science scores for 2016.

The 2016 Integrated Science scores were normally distributed as shown in Figure 29.

Figure 30 depicts the normality of 2012 social studies test scores.

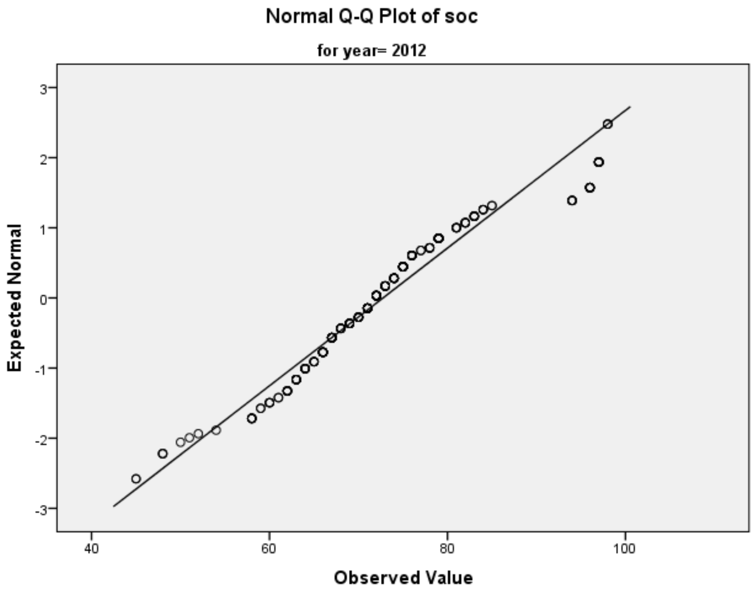

Figure 30: Plot of Social Studies scores for 2012

The 2012 social studies scores were normally distributed as shown in Figure 30. Figure 31 depicts a normality test plot for the 2013 social studies data.

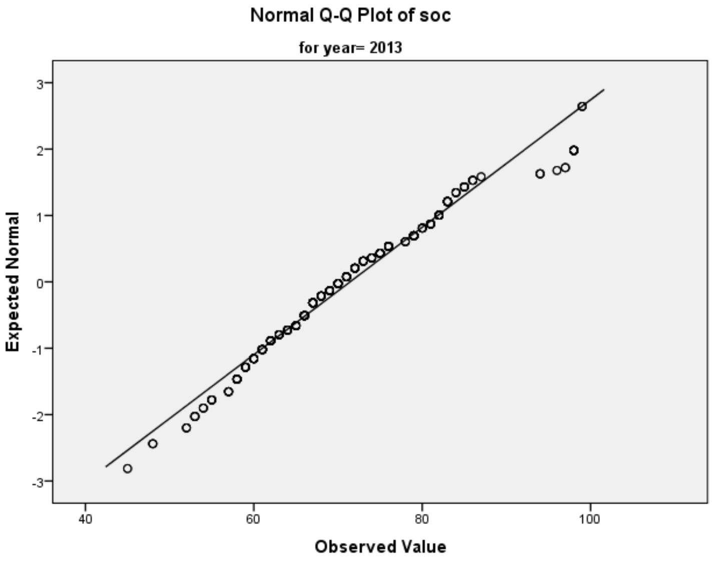

Figure 31: Plot of Social Studies scores for 2013

The 2013 social studies scores were normally distributed as shown in Figure 31. Figure 32 depicts a normality test plot for 2014 social studies data.

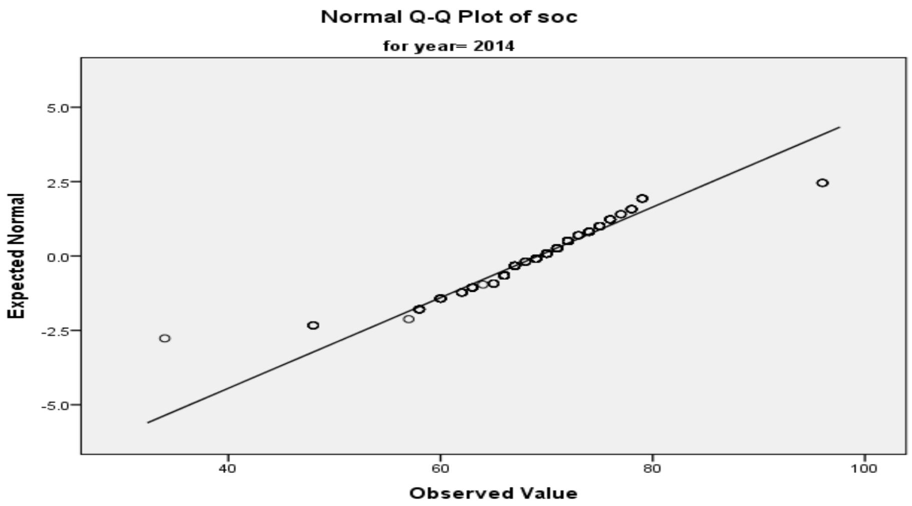

Figure 32: Plot of Social Studies scores for 2014

2014 social studies scores were normally distributed as shown in Figure 32. Figure 33 depicts a normality test plot for the 2013 social studies data.

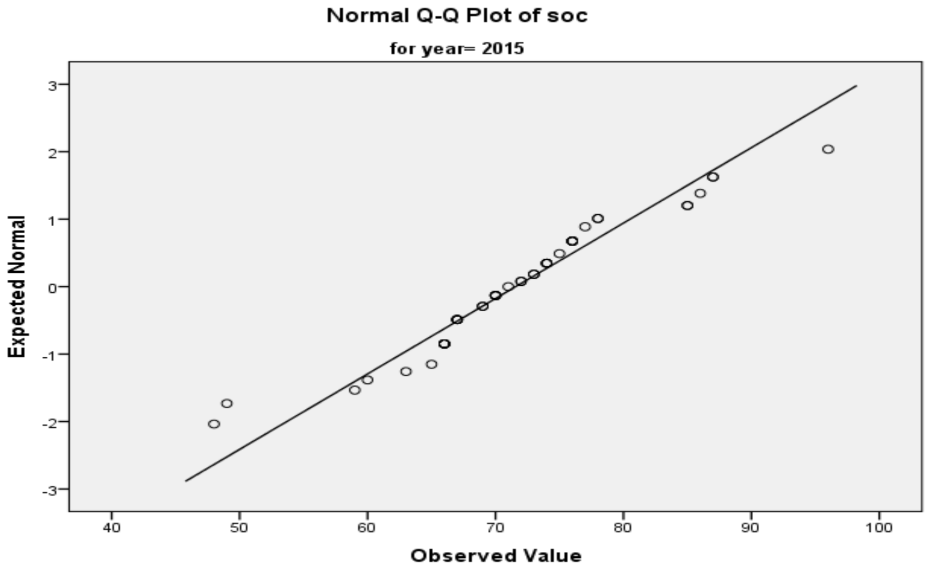

Figure 33: Plot of Social Studies scores for 2015

2015 Social studies scores were normally distributed as shown in Figure 33. Figure 34 depicts a normality test plot for 2016 social studies data.

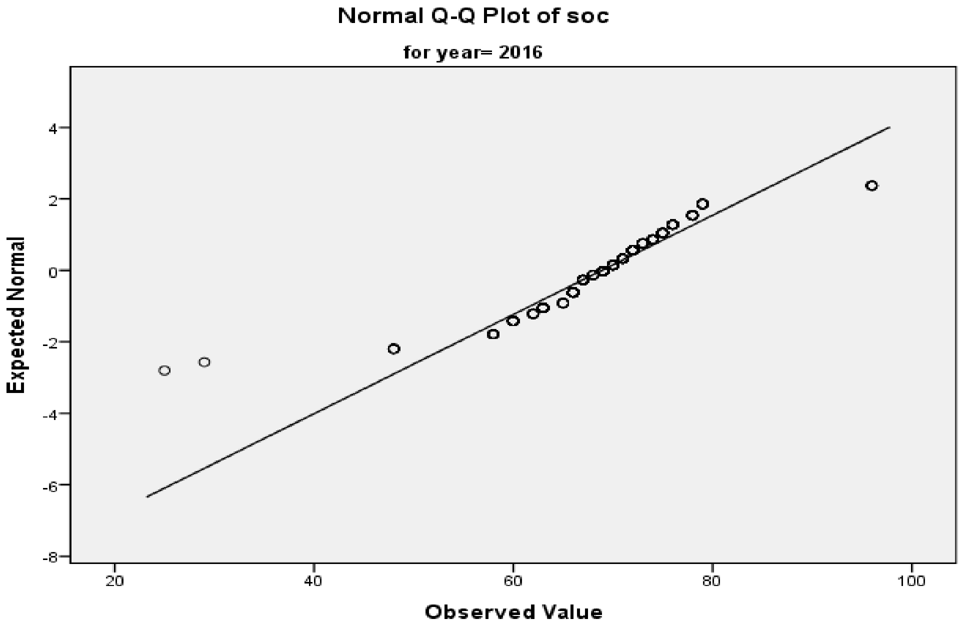

Figure 34: Plot of Social Studies for 2016

The 2016 Social studies scores were normally distributed as shown in Figure 34.

For the LR procedure:

One of the assumptions of binary logistic regression requires the study variable to be binary. The dataset for this study satisfied this assumption because the study variables are the gender and location. Gender is binary (male/female) and location(Central Region-CR / Eastern Region-ER, Western Region-WR, Volta Region-VR and Greater Accra Region-GAR (i.e., FG) and Greater Accra Region-GAR/ Central Region-CR / Eastern Region-ER, Western Region-WR and Volta Region-VR (i.e., FG))

Secondly, since logistic regression assumes that P(Y=1) is the probability of the event occurring, the dependent variable must be coded accordingly. That is, for a binary regression, the factor level 1 of the dependent variable should represent the desired outcome. This assumption was met because the dependent variable used in the study was the performance of students in English language, Mathematics, Integrated Science and Social Studies from 2012-2016 WASSCE multiple-choice items. The items were scored correct or wrong with 1 for correct response and 0 for the wrong response.

Lastly, it requires quite large sample sizes. The dataset for the study has a sample size that is quite large as shown in Table 9.

Table 9: Sample Size Distribution by Subject and Year

Subject Year	English Language	Mathematics	Integrated Science	Social Studies
2012	7711	7743	7526	7743
2013	17576	17440	17889	17312
2014	8795	8564	9359	9764
2015	10,133	10137	10130	10134
2016	12,478	12,602	12,440	12694

For the IRT:

(1) Do the data conform to the 3PL model unidimensionality assumptions?

This assumption implies the dataset for the study should measure a single latent trait. For example, the mathematics data for the study should measure only mathematical ability and nothing else.

(2) Do the data sufficiently fulfil the local independence assumptions?

The local independence assumption indicates that the observed items are conditionally *independent* of each other given an individual score on the latent variable(s).

(3) Do the data satisfy the invariance assumption?

Invariance is a property of real or formal systems for which types of transformations do not alter the relationships between the elements of a system. Thus, if a system is comprised of measurements of several objects using a single instrument, measurement invariance shall be defined as observing the same relationships between measurements when a second measurement instrument is used in the assessment. It can be defined as the equality of item and examinee parameters from different examinee populations or measurement conditions.

Checking unidimensionality assumption

To assess the unidimensionality of data used in this study, two approaches were employed, namely the Exploratory Factor Analysis and a principal component 3PL analysis of the residual via the SPSS and WINSTEPS software.

Assessing unidimensionality using principal component 3PL analysis

Reckase (1979) suggested that a measure should account for at least 20% of the variance to produce an acceptable unidimensional construct. The 2012-2016 dataset was analysed for unidimensionality using 3PL analysis. The interpretation of the terms in Tables 9 to 28 are as follows:

Empirical: Constitute the eigenvalue and observed value in percentages respectively.

In Eigenvalue units: Variance components are rescaled so that the total unexplained variance has its expected summed eigenvalue.

Observed: variance components for the observed data

Expected (Modeled): Variance components expected for these data if they exactly fit the Rasch model.

If **observed** and **expected** differ noticeably, then there is a problem in the estimation. This is not a symptom of multidimensionality.

Total variance in observations: Total raw-score variance in the observations

Variance explained by measures: Raw-score variance in the observations explained by the Rasch item difficulties, personal abilities and polytomous scale structures.

Unexplained variance (total): raw-score variance in the observations not explained by the Rasch measures

Unexplained variance in 1st, 2nd ... contrast: Variance that is not explained by the model measures is decomposed into Principal Component Analysis (PAC) components = Contrasts. The size of the first, second ... contrast (component) in the PCA decomposition of standardized residuals (i.e., the variance that is not explained by the Rasch measures, but that is explained by the contrast.

Table 10 depicts the results of the test of unidimensionality of 2012 English Language dataset.

Table 10: English Language 2012 Principal Components 3PL results

Standardized Residual variance (in Eigenvalue units)			
	Eigenvalue	Observed variance	Expected variance
Total variance in observations	20493.4	100.0%	100.0%
Variance explained by measures	20393.4	99.5%	99.9%
Unexplained variance (total)	100.0	.5%	.1%
Unexplained variance in 1st contrast	15.7	.1%	15.7%
Unexplained variance in 2nd contrast	6.4	.0%	6.4%
Unexplained variance in 3rd contrast	4.2	.0%	4.2%

Results according to Table 10 indicated that the empirical measure explained was above 95% of the variance, which suggests that the dataset is an acceptable one-dimensional construct. The next step is to check whether, after taking account of the 3PL measure, there are still large eigenvalue residual contrasts that may arguably alert the researcher to a likelihood of multidimensionality.

Notwithstanding, Linacre (2009a) argued that an eigenvalue of 2 units indicates 2-items strength and from Table 10, the variance explained by the 1st contrasts across the 2012 English Language dataset which is assumed to be the secondary dimensions have eigenvalues greater than 2 but variances explained less than 2%. A strength of fewer than 2 items for a second dimension is very weak and hence, this provides evidence that the measured 2012 English Language data is unidimensional.

Table 11 depicts the unidimensionality of the 2012 English Language data.

Table 11: Mathematics 2012 Principal Components 3PL results

Standardized Residual variance (in Eigenvalue units)			
	Eigenvalue	Observed variance	Expected variance
Total variance in observations	81790.6	100.0%	100.0%
Variance explained by measures	81741.6	99.9%	100.0%
Unexplained variance (total)	49.0	.1%	.0%
Unexplained variance in 1st contrast	4.3	.0%	8.7%
Unexplained variance in 2nd contrast	3.2	.0%	6.5%
Unexplained variance in 3rd contrast	2.5	.0%	5.1%
Unexplained variance in 4th contrast	2.2	.0%	4.5%
Unexplained variance in 5th contrast	2.1	.0%	4.2%

Results according to Table 11 indicated that the empirical measure explained was above 95% of the variance, which suggests that the dataset is an acceptable one-dimensional construct. The next step is to check whether, after taking account of the 3PL measure, there are still large eigenvalue residual contrasts that may arguably alert the researcher to a likelihood of multidimensionality.

From Table 11, the variance explained by the 1st contrasts across the 2012 mathematics dataset, which is assumed to be the secondary dimensions have eigenvalues greater than 2 but variances explained less than 2%. A strength of fewer than 2 items for a second dimension is very weak and hence, this provides evidence that the 2012 mathematics data is unidimensional.

Table 12 depicts the unidimensionality of 2012 Integrated Science data.

Table 12: Integrated Science 2012 Principal Components 3PL results

Standardized Residual variance (in Eigenvalue units)			
	Eigenvalue	Observed variance	Expected variance
Total variance in observations	21807.2	100.0%	100.0%
Variance explained by measures	21757.2	99.8%	100.0%
Unexplained variance (total)	50.0	.2%	.0%
Unexplained variance in 1st contrast	3.5	.0%	7.1%
Unexplained variance in 2nd contrast	3.3	.0%	6.7%
Unexplained variance in 3rd contrast	2.9	.0%	5.9%
Unexplained variance in 4th contrast	2.3	.0%	4.6%
Unexplained variance in 5th contrast	2.0	.0%	4.1%

Results according to Table 12 indicated that the empirical measure explained was above 95% of the variance, which suggests that the dataset is an acceptable one-dimensional construct.

From Table 12, the variance explained by the 1st contrasts across the 2012 Integrated Science dataset, which is assumed to be the secondary dimensions have eigenvalues greater than 2 but variances explained less than 2%. A strength of fewer than 2 items for a second dimension is very weak and hence, this provides evidence that the 2012 Integrated Science data is unidimensional.

Table 13 depicts the unidimensionality of 2012 Social Studies data.

Table 13: Social Studies 2012 Principal Components 3PL results

Standardized Residual variance (in Eigenvalue units)			
	Eigenvalue	Observed variance	Expected variance
Total variance in observations	41688.6	100.0%	100.0%
Variance explained by measures	41638.6	99.9%	99.9%
Unexplained variance (total)	50.0	.1%	1%
Unexplained variance in 1st contrast	3.6	.0%	7.2%
Unexplained variance in 2nd contrast	2.6	.0%	5.1%
Unexplained variance in 3rd contrast	2.1	.0%	4.3%
Unexplained variance in 4th contrast	2.0	.0%	3.9%

Results according to Table 13 indicated that the empirical measure explained was above 95% of the variance, which suggests that the dataset is an acceptable one-dimensional construct.

From Table 13, the variance explained by the 1st contrasts across the 2012 Social Studies dataset, which is assumed to be the secondary dimensions have eigenvalues greater than 2 but variances explained less than 2%. A strength of fewer than 2 items for a second dimension is very weak and hence, this provides evidence that the 2012 social studies data is unidimensional.

Table 14 depicts the unidimensionality of the 2013 English Language data.

Table 14: English Language 2013 Pricipal Componets 3PL results

Standardized Residual variance (in Eigenvalue units)	Eigenvalue	Observed variance	Expected variance
Total variance in observations	24076.5	100.0%	100.0%
Variance explained by measures	23976.5	99.6%	99.9%
Unexplained variance (total)	100.0	.4%	.1%
Unexplained variance in 1st contrast	6.2	.0%	6.2%

Results according to Table 14 indicated that the empirical measure explained was above 95% of the variance, which suggests that the dataset is an acceptable one-dimensional construct.

From Table 14, the variance explained by the 1st contrasts across the 2013 English Language dataset, which is assumed to be the secondary dimensions have eigenvalues greater than 2 but variances explained less than 2%. A strength of fewer than 2 items for a second dimension is very weak and hence, this provides evidence that the 2013 English Language data is unidimensional.

Table 15 depicts the unidimensionality of the 2013 mathematics data.

Table 15: Mathematics 2013 Principal Components 3PL results

Standardized Residual variance (in Eigenvalue units)	Eigenvalue	Observed variance	Expected variance
Total variance in observations	4672.1	100.0%	100.0%
Variance explained by measures	4622.1	98.9%	99.9%
Unexplained variance (total)	50.0	1.1%	.1%
Unexplained variance in 1st contrast	5.9	.1%	11.8%
Unexplained variance in 2nd contrast	3.5	.1%	7.0%
Unexplained variance in 3rd contrast	3.2	.1%	6.4%
Unexplained variance in 4th contrast	2.6	.1%	5.3%
Unexplained variance in 5th contrast	2.4	.1%	4.7%

Results according to Table 15 indicated that the empirical measure explained was above 95% of the variance, which suggests that the dataset is an acceptable one-dimensional construct.

From Table 15, the variance explained by the 1st contrasts across the 2013 mathematics dataset, which is assumed to be the secondary dimensions have eigenvalues greater than 2 but variances explained less than 2%. A strength of fewer than 2 items for a second dimension is very weak and hence, this provides evidence that the 2013 mathematics data is unidimensional.

Table 16 depicts the unidimensionality of 2013 Integrated Science data.

Table 16: Integrated Science 2013 Principal Components 3PL results

Standardized Residual variance (in Eigenvalue units)			
	Eigenvalue	Observed variance	Expected Variance
Total variance in observations	542.9	100.0%	100.0%
Variance explained by measures	492.9	90.8%	100.0%
Unexplained variance (total)	50.0	9.2%	.0%
Unexplained variance in 1st contrast	4.1	.8%	8.2%
Unexplained variance in 2nd contrast	3.9	.7%	7.8%
Unexplained variance in 3rd contrast	3.3	.6%	6.6%

Results according to Table 16 indicated that the empirical measure explained was above 95% of the variance, which suggests that the dataset is an acceptable one-dimensional construct.

From Table 16, the variance explained by the 1st contrasts across the 2013 Integrated Science dataset, which is assumed to be the secondary dimensions have eigenvalues greater than 2 but variances explained less than 2%. A strength of fewer than 2 items for a second dimension is very weak and hence, this provides evidence that the 2013 Integrated Science data measure is unidimensional.

Table 17 depicts the unidimensionality of 2013 social studies data.

Table 17: Social Studies 2013 Principal Components 3PL results

Standardized Residual variance (in Eigenvalue units)	Eigenvalue	Observed variance	Expected Varaiance
Total variance in observations	31009.8	100.0%	100.0%
Variance explained by measures	30959.8	99.8%	99.8%
Unexplained variance (total)	50.0	.2%	.2%
Unexplained variance in 1st contrast	5.8	.0%	11.5%
Unexplained variance in 2nd contrast	4.5	.0%	9.0%
Unexplained variance in 3rd contrast	3.1	.0%	6.1%
Unexplained variance in 4th contrast	2.8	.0%	5.6%
Unexplained variance in 5th contrast	2.5	.0%	5.0%

Results according to Table 17 indicated that the empirical measure explained was above 95% of the variance, which suggests that the dataset is an acceptable one-dimensional construct.

From Table 17, the variance explained by the 1st contrasts across the 2013 social studies dataset, which is assumed to be the secondary dimensions have eigenvalues greater than 2 but variances explained less than 2%. A strength of fewer than 2 items for a second dimension is very weak and hence, this provides evidence that the 2013 social studies data is unidimensional.

Table 18 depicts the unidimensionality of the 2014 English Language data.

Table 18: English Language, 2014 Principal Components 3PL results

Standardized Residual variance (in Eigenvalue units)	Eigenvalue	Observed variance	Expected variance
Total variance in observations	27705.1	100.0%	100.0%
Variance explained by measures	27605.1	99.6 %	99.9%
Unexplained variance (total)	100.0	.4%	.1%
Unexplained variance in 1st contrast	9.0	.0%	9.0%
Unexplained variance in 2nd contrast	6.8	.0%	6.8%
Unexplained variance in 3rd contrast	5.4	.0%	5.4%

Results according to Table 18 indicated that the empirical measure explained was above 95% of the variance, which suggests that the dataset is an acceptable one-dimensional construct.

From Table 18, the variance explained by the 1st contrasts across the 2012 English Language dataset, which is assumed to be the secondary dimensions have eigenvalues greater than 2 but variances explained less than 2%. A strength of fewer than 2 items for a second dimension is very weak and hence, this provides evidence that the 2014 English Language data measure is unidimensional.

Table 19 depicts the unidimensionality of 2014 mathematics data.

Table 19: Mathematics, 2014 Principla Components 3PL results

Standardized Residual variance (in Eigenvalue units)			
	Eigenvalue	Observed variance	Expected variance
Total variance in observations	13845.0	100.0%	100.0%
Variance explained by measures	13796.0	99.6%	100.0%
Unexplained variance (total)	49.0	.4%	.0%
Unexplained variance in 1st contrast	3.0	.0%	6.2%
Unexplained variance in 2nd contrast	2.6	.0%	5.3%
Unexplained variance in 3rd contrast	2.4	.0%	4.8%
Unexplained variance in 4th contrast	2.2	.0%	4.5%
Unexplained variance in 5th contrast	1.8	.0%	3.7%

Results according to Table 19 indicated that the empirical measure explained was above 95% of the variance, which suggests that the dataset is an acceptable one-dimensional construct.

From Table 19, the variance explained by the 1st contrasts across the 2014 mathematics dataset, which is assumed to be the secondary dimensions have eigenvalues greater than 2 but variances explained less than 2%. A strength of fewer than 2 items for a second dimension is very weak and hence, this provides evidence that the 2014 mathematics data is unidimensional.

Table 20 depicts the unidimensionality of 2014 Integrated Science data.

Table 20: Integrated Science 2014 Principla Components 3PL results

Standardized Residual variance (in Eigenvalue units)			
	Eigenvalue	Observed variance	Expected Variance
Total variance in observations	1779.5	100.0%	100.0%
Variance explained by measures	1729.5	97.2%	99.9%
Unexplained variance (total)	50.0	2.8%	.1%
Unexplained variance in 1st contrast	6.9	.4%	13.9%
Unexplained variance in 2nd contrast	3.7	.2%	7.5%
Unexplained variance in 3rd contrast	3.3	.2%	6.7%
Unexplained variance in 4th contrast	2.7	.2%	5.4%
Unexplained variance in 5th contrast	2.4	.1%	4.9%

Results according to Table 20 indicated that the empirical measure explained was above 95% of the variance, which suggests that the dataset is an acceptable one-dimensional construct.

From Table 20, the variance explained by the 1st contrasts across the 2014 Integrated Science dataset, which is assumed to be the secondary dimensions have eigenvalues greater than 2 but variances explained less than 2%. A strength of fewer than 2 items for a second dimension is very weak and hence, this provides evidence that the 2014 Integrated Science data is unidimensional.

Table 21 depicts the unidimensionality of 2014 social studies data.

Table 21: Social Studies 2014 Principla Components 3PL results

Standardized Residual variance (in Eigenvalue units)			
	Eigenvalue	Observed variance	Expected variance
Total variance in observations	8080.9	100.0%	100.0%
Variance explained by measures	8030.9	99.4%	99.7%
Unexplained variance (total)	50.0	.6%	.3%
Unexplained variance in 1st contrast	5.3	.1%	10.6%
Unexplained variance in 2nd contrast	2.9	.0%	5.9%
Unexplained variance in 3rd contrast	2.9	.0%	5.8%
Unexplained variance in 4th contrast	2.9	.0%	5.2%
Unexplained variance in 5th contrast	2.5	.0%	4.9%

Results according to Table 21 indicated that the empirical measure explained was above 95% of the variance, which suggests that the dataset is an acceptable one-dimensional construct.

From Table 21, the variance explained by the 1st contrasts across the 2014 social studies dataset, which is assumed to be the secondary dimensions have eigenvalues greater than 2 but variances explained less than 2%. A strength of fewer than 2 items for a second dimension is very weak and hence, this provides evidence that the 2014 social studies data is unidimensional.

Table 22 depicts the unidimensionality of the 2015 English Language data.

Table 22: English Language 2015 Principla Components 3PL results

Standardized Residual variance (in Eigenvalue units)			
	Eigenvalue	Observed variance	Expected Variance
Total variance in observations	8996.9	100.0%	100.0%
Variance explained by measures	8918.9	99.9%	100.0%
Unexplained variance (total)	78.0	.9%	.0%
Unexplained variance in 1st contrast	8.1	.1%	10.4%
Unexplained variance in 2nd contrast	4.0	.0%	6.5%
Unexplained variance in 3rd contrast	3.9	.0%	5.0%
Unexplained variance in 4th contrast	3.1	.0%	4.0%
Unexplained variance in 5th contrast	2.8	.0%	3.6%

Results according to Table 22 indicated that the empirical measure explained was above 95% of the variance, which suggests that the dataset is an acceptable one-dimensional construct.

From Table 22, the variance explained by the 1st contrasts across the 2015 English Language dataset, which is assumed to be the secondary dimensions have eigenvalues greater than 2 but variances explained less than 2%. A strength of fewer than 2 items for a second dimension is very weak and hence, this provides evidence that the 2015 English Language data is unidimensional.

Table 23 depicts the unidimensionality of 2015 mathematics data.

Table 23: Mathematics 2015 Principla Components 3PL results

Standardized Residual variance (in Eigenvalue units)			
	Eigenvalue	Observed variance	Expected Variance
Total variance in observations	28817.7	100.0%	100.0%
Variance explained by measures	28767.7	99.8%	99.8%
Unexplained variance (total)	50.0	.2%	.2%
Unexplained variance in 1st contrast	6.7	.0%	13.3%
Unexplained variance in 2nd contrast	4.8	.0%	9.7%
Unexplained variance in 3rd contrast	3.4	.0%	6.7%
Unexplained variance in 4th contrast	2.5	.0%	5.1%
Unexplained variance in 5th contrast	2.1	.0%	4.2%

Results according to Table 23 indicated that the empirical measure explained was above 95% of the variance, which suggests that the dataset is an acceptable one-dimensional construct.

From Table 23, the variance explained by the 1st contrasts across the 2015 mathematics dataset, which is assumed to be the secondary dimensions have eigenvalues greater than 2 but variances explained less than 2%. A strength of fewer than 2 items for a second dimension is very weak and hence, this provides evidence that the 2015 mathematics data is unidimensional.

Table 24 depicts the unidimensionality of 2015 Integrated Science data.

Table 24: Integrated Science 2015 Principla Components 3PL results

Standardized Residual variance (in Eigenvalue units)			
	Eigenvalue	Observed variance	Expected variance
Total variance in observations	18445.7	100.0%	100.0%
Variance explained by measures	18395.7	99.7%	100.0%
Unexplained variance (total)	50.0	.3%	.0%
Unexplained variance in 1st contrast	4.0	.0%	8.0%
Unexplained variance in 2nd contrast	3.1	.0%	6.3%
Unexplained variance in 3rd contrast	2.5	.0%	5.0%
Unexplained variance in 4th contrast	2.2	.0%	4.3%
Unexplained variance in 5th contrast	2.0	.0%	4.0%

Results according to Table 24 indicated that the empirical measure explained was above 95% of the variance, which suggests that the dataset is an acceptable one-dimensional construct.

From Table 24, the variance explained by the 1st contrasts across the 2015 Integrated Science dataset, which is assumed to be the secondary dimensions have eigenvalues greater than 2 but variances explained less than 2%. A strength of fewer than 2 items for a second dimension is very weak and hence, this provides evidence that the 2015 Integrated Science data is unidimensional.

Table 25 depicts the unidimensionality of 2015 social studies data.

Table 25: Social Studies 2015 Principal Components 3PL results

Standardized Residual variance (in Eigenvalue units)			
	Eigenvalue	Observed variance	Expected Variance
Total variance in observations	7873. 7	100.0%	100.0%
Variance explained by measures	7823.7	99.4%	100.0%
Unexplained variance (total)	50.0	.6%	.0%
Unexplained variance in 1st contrast	3.5	.0%	7.0%
Unexplained variance in 2nd contrast	2.8	.0%	5.6%
Unexplained variance in 3rd contrast	2.6	.0%	5.3%
Unexplained variance in 4th contrast	2.4	.0%	4.8%
Unexplained variance in 5th contrast	2.0	.0%	4.0%

Results according to Table 25 indicated that the empirical measure explained was above 95% of the variance, which suggests that the dataset is an acceptable one-dimensional construct.

From Table 25, the variance explained by the 1st contrasts across the 2015 social studies dataset, which is assumed to be the secondary dimensions have eigenvalues greater than 2 but variances explained less than 2%. A strength of fewer than 2 items for a second dimension is very weak and hence, this provides evidence that the 2015 social studies data is unidimensional.

Table 26 depicts the unidimensionality of 2016 English Language data.

Table 26: English Language 2016 Principal Components 3PL results

Standardized Residual variance (in Eigenvalue units)			
	Eigenvalue	Observed variance	Expected variance
Total variance in observations	40371.1	100.0%	100.0%
Variance explained by measures	40291.1	99.8%	100.0%
Unexplained variance (total)	80.0	.2%	.0%
Unexplained variance in 1st contrast	5.8	.0%	7.2%
Unexplained variance in 2nd contrast	4.6	.0%	5.8%
Unexplained variance in 3rd contrast	4.0	.0%	5.0%
Unexplained variance in 4th contrast	3.0	.0%	3.8%
Unexplained variance in 5th contrast	2.6	.0%	3.2%

Results according to Table 26 indicated that the empirical measure explained was above 95% of the variance, which suggests that the dataset is an acceptable one-dimensional construct.

From Table 26, the variance explained by the 1st contrasts across the 2016 English Language dataset, which is assumed to be the secondary dimensions have eigenvalues greater than 2 but variances explained less than 2%. A strength of fewer than 2 items for a second dimension is very weak and hence, this provides evidence that the 2016 English Language data is unidimensional.

Table 27 depicts the unidimensionality of 2016 mathematics data.

Table 27: Mathematics 2016 Principal Components 3PL results

Standardized Residual variance (in Eigenvalue units)			
	Eigenvalue	Observed variance	Expected variance
Total variance in observations	132333.2	100.0%	100.0%
Variance explained by measures	132283.2	100.0%	100.0%
Unexplained variance (total)	50.0	.6%	.0%
Unexplained variance in 1st contrast	8.5	.0%	16.9%
Unexplained variance in 2nd contrast	3.2	.0%	6.4%
Unexplained variance in 3rd contrast	2.5	.0%	5.1%
Unexplained variance in 4th contrast	2.0	.0%	4.0%
Unexplained variance in 5th contrast	1.8	.0%	3.7%

Results according to Tables 27 indicated that the empirical measure explained was above 95% of the variance, which suggests that the dataset is an acceptable one-dimensional construct. The next step is to check whether, after taking account of the 3PL measure, there are still large eigenvalue residual contrasts that may arguably alert the researcher to a likelihood of multidimensionality.

From Table 27, the variance explained by the 1st contrasts across the 2016 mathematics dataset, which is assumed to be the secondary dimensions have eigenvalues greater than 2 but variances explained less than 2%. A strength of fewer than 2 items for a second dimension is very weak and hence, this provides evidence that the 2016 mathematics data is unidimensional.

Table 28 depicts the unidimensionality of 2016 Integrated Science data.

Table 28: Integrated Science 2016 Principal Components 3PL results

Standardized Residual variance (in Eigenvalue units)			
	Eigenvalue	Observed variance	Expected Variance
Total variance in observations	15783.5	100.0%	100.0%
Variance explained by measures	15733.5	99.7%	99.8%
Unexplained variance (total)	50.0	.3%	.2%
Unexplained variance in 1st contrast	6.3	.0%	12.6%
Unexplained variance in 2nd contrast	3.9	.0%	7.8%
Unexplained variance in 3rd contrast	3.0	.0%	5.9%
Unexplained variance in 4th contrast	2.6	.0%	5.2%
Unexplained variance in 5th contrast	2.2	.0%	4.3%

Results according to Table 28 indicated that the empirical measure explained was above 95% of the variance, which suggests that the dataset is an acceptable one-dimensional construct.

From Table 28, the variance explained by the 1st contrasts across the 2016 Integrated Science dataset, which is assumed to be the secondary dimensions have eigenvalues greater than 2 but variances explained less than 2%. A strength of fewer than 2 items for a second dimension is very weak and hence, this provides evidence that the 2016 social studies data is unidimensional.

Table 29 depicts the unidimensionality of 2012 Integrated Science data.

Table 29: Social Studies 2016 Pricipal Components 3PL results

Standardized Residual variance (in Eigenvalue units)			
	Eigenvalue	Observed variance	Expected Variance
Total variance in observations	7985.2	100.0%	100.0%
Variance explained by measures	7935.2	99.4%	99.9%
Unexplained variance (total)	50.0	.6%	.1%
Unexplained variance in 1st contrast	8.7	.1%	17.4%
Unexplained variance in 2nd contrast	6.6	.1%	13.2%
Unexplained variance in 3rd contrast	3.2	.0%	6.4%
Unexplained variance in 4th contrast	3.2	.0%	6.4%
Unexplained variance in 5th contrast	2.7	.0%	5.5%

Results according to Table 29 indicated that the empirical measure explained was above 95% of the variance, which suggests that the dataset is an acceptable one-dimensional construct. The next step is to check whether, after taking account of the 3PL measure, there are still large eigenvalue residual contrasts that may arguably alert the researcher to a likelihood of multidimensionality.

From Table 29, the variance explained by the 1st contrasts across the 2016 social studies dataset, which is assumed to be the secondary dimensions have eigenvalues greater than 2 but variances explained less than 2%. A strength of fewer than 2 items for a second dimension is very weak and hence, this provides evidence that the 2016 social studies data is unidimensional.

Checking of Local Independence Assumption of the 3PL model

The local independence assumption is that the true score or the latent trait gives all relevant information about an examinee's performance and that the contribution of each item in the test can be assessed independently among other items.

According to Smith (2005), mathematically, local independence means that for a given value of θ, the joint probability of correct responses to an item pair is the product of the probabilities of correct responses to the two items that are,

$$\Pr\{X_1, X_2, \ldots X_k|\beta_n\} = \Pi_{i=1}^{k}\Pr\{X_i|\beta_n\}$$

Where *βn* is the latent ability of each person *n*, *{Xi =xi}* is the answer of a randomly selected person to item *i* (where *i=1,2,…,k*), and $\Pr\{X_i|\beta_n\}$ represents the probability of a person responding to item *i*. Local independence is a requirement of the IRT.

In practice, local independence will normally be violated when responses to items are related in some way (Yen, 1993). For example, if a correct response to an item is necessary to answer the subsequent item correctly or if the content and knowledge of one item gives relevant information to answer another item correctly or if the scoring rubrics are used in the same way, or if a set of items all refer to a common stimulus such as a passage, a graph, a table, or a diagram, then local independence might be violated (Smith, 2005).

When investigating local dependence (LD) based on Yen's Q_3, residuals for any pair of items should be uncorrelated, and generally close to 0. Residual correlations that are high indicate a violation of the local independence assumption, and this suggests that the pair of items are highly related to each other than to the rest of the item set (Marais, 2013).

As noted by Yen (1984), a negative bias is built into Q3. This problem is since measures of association will be biased away from zero even though the assumption of local independence applies, due to the conditioning on a proxy variable instead of the latent variable (Rosenbaum, 1984). A second problem is that the way the residuals are computed induces a bias (Kreiner & Christensen, 2011). Marais (2013) recognized that the sampling properties among residuals are unknown, therefore these statistics cannot be used for formal tests of LD.

A third, and perhaps the most important, problem in applications, is that there are currently no well-documented suggestions of the critical values which should be used to indicate LD, and for this reason, arbitrary rules of thumb are used when evaluating whether an observed correlation is such that it can be reasonably supposed to have arisen from random sampling.

Standards often reported in the literature include looking at fit residuals over the critical value of 0.2, as proposed by Chen and Thissen (1997) and can be seen in studies like Elden and Reeve (2007), Hissbach, Klusmann and Hampe (2011); Makransky and Bilenberg (2014) and Makransky, Rogers and Creed (2014).

However, other critical values are also used, and there seems to be a wide variation in what is seen as indicative of dependence. Marais and Andrich (2008b) investigated dependence at a critical residual correlation value of 0.1, but a value of 0.3 has also often been used (La Porta, Maselli, & Petrioli, 2011; Das Nair, Moreton, & Lincoln, 2011; Ramp et al. 2009), and critical values of

0.5 (ten Klooster, Taal, &Van De Laar,.(2008), Davidson & MacKinnon (2004) and even 0.7 (González-de Paz et al., 2014) can be found in use.

Yen (1984) proposed a Q_3 statistic as an index to flag items with local dependence. The Q_3 index is the correlation of residuals for a pair of items after the primary measure is partially out through the ability estimates. To calculate the value of Q_3 statistic, a proficiency estimate is calculated for each examinee. Then the expected score (denoted E_{ni} where n denotes an examinee and i denotes an item) is computed for each examinee for each item. The residual (denoted d_{ni}), which is the deviation of an examinee's observed score (denoted O_{ni}) from the expected score can be written as:

$$d_{ni} = O_{ni} – E_{ni}$$

Thus, for item i and item j, the statistic Q_3 is the correlation of residuals taken over all examinees ($Q_{3ij} = r_{didj}$).

Yen's Q_3 statistic was computed using the WINSTEPS software (Linacre, 2009b). It should be noted that the Q_3 statistic exists for diagnostic purposes rather than for hypothesis testing, therefore caution was be taken in the interpretation of the statistics (Chen & Thissen, 1997). According to Linacre (2009a), local dependence items are likely to have a large positive correlation. Highly locally dependent items are any pairs of items with a correlation value greater than 0.70 or correlation value less than -.07. Tables 35 to 53 present results of standardized residual correlations of pair items that are used to identify dependent items of the 2012-2016 dataset for English Language, Mathematics Integrated Science and Social Studies. Paired items with correlations greater than 0.7 are the ones which have violated the assumption of local independence in the dataset used for the analysis.

Table 30 depicts the standardized residual correlations of paired items that are used to identify dependent items for 2012 English Language data.

Table 30: Standardized Residual Correlations for English Language 2012

Residual Correlation	.92	.91	.87	.86	.85	.84	.83	-.84
Item Pair	13/14	58/84	45/62	42/71	28/60, 22/74, 28/69	28/69	60/69	28/62

Using a critical value of 0.7 led to the conclusion of 10 out of 4,950 English Language paired items identified violate the 3PL model as seen in Table 30.

Table 31 depicts the standardized residual correlations of paired items that are used to identify dependent items for 2012 mathematics data.

Table 31: Standardized Residual Correlations for Mathematics 2012

Residual Correlation	.73	.56	.41	.38	.-44	-.39	-.38
Item Pair	34/45	9/46	23/49	4/12	21/45 8/34	23/32	4/39 20/27

Using a critical value of 0.7 led to the conclusion of one out of 1,225 mathematics paired items identified violates the 3PL model as seen in Table 31.

Table 32 depicts the standardized residual correlations of paired items that are used to identify dependent items for 2012 Integrated Science data.

Table 32: Standardized Residual Correlations for Integrated Science 2012

Residual Correlation	.83	.78	.65	.43	-.55	-.52	-.49	-.46	-.45
Item Pair	48/49	49/50	48/50	20/33	46/49	46/48	10/47	10/33	27/48 17/49

Using a critical value of 0.7 led to the conclusion of 2 out of 1,225 Integrated Science paired items identified violate the 3PL model as seen in Table 32.

Table 33 depicts the standardized residual correlations of paired items that are used to identify dependent items for 2012 social studies data.

Table 33: Standardized Residual Correlations for Social Studies, 2012

Residual Correlation	.54	.53	.47	.46	.40	.39	-.61	-.57	-.49
Item Pair	10/27	6/15	10/28	7/8	13/25	27/28	10/14	14/27	14/28

Using a critical value of 0.7 led to the conclusion of none out of 1,225 social studies paired items identified violate the 3PL model as seen in Table 33.

Table 34 depicts the standardized residual correlations of paired items that are used to identify dependent items for 2013 English Language data.

Table 34: Standardized Residual Correlations for English Language 2013

Residual Correlation	.92	.91	.83	.80	.77	.75	.73	.72	-.74
Item Pair	34/88	64/73	3/71	31/41	3/74	39/41	73/74	40/46 29/45	40/87

Using a critical value of 0.7 led to the conclusion of 9 out of 4,950 English Language paired items identified violate the 3PL model as seen in Table 34.

Table 35 depicts the standardized residual correlations of paired items that are used to identify dependent items for 2013 mathematics data.

Table 35: Standardized Residual correlations for Mathematics, 2013

Residual Correlation	.70	.67	.66	.63	.57	.54	.52	-.69	-.58	-.54
Item Pair	27/44	21/44	26/36	19/31	31/42	39/40	21/27	26/43	39/43	36/43

Using a critical value of 0.7 led to the conclusion of none of the 1,125 mathematics paired items identified violate the 3PL model as seen in Table 35.

Table 36 depicts the standardized residual correlations of paired items that are used to identify dependent items for 2013 Integrated Science data.

Table 36: Standardized Residual Correlations for Integrated Science, 2013

Residual Correlation	.64	.57	.56	.55	.51	.48	.47	.46
Item Pair	11/18	42/47	6/8	26/42	5/34 5/29	5/29	29/40	14/23 10/20

Using a critical value of 0.7 led to the conclusion of none of the 1,225 Integrated Science paired items identified violate the 3PL model as seen in Table 36.

Table 37 depicts the standardized residual correlations of paired items that are used to identify dependent items for 2013 social studies data.

Table 37: Standardized Residual Correlations for Social Studies, 2013

Residual Correlation	.64	.63	.57	.55	.52	.51	-.54	-.53
Item Pair	23/24	26/30	8/17	7/16	30/39 25/36	15/35	23/39	36/47 8/43

Using a critical value of 0.7 led to the conclusion of none of the 4,950 English Language paired items identified violate the 3PL model as seen in Table 37.

Table 38 depicts the standardized residual correlations of paired items that are used to identify dependent items for 2013 mathematics data.

Table 38: Standardized Residual Correlations for English Language 2014

Residual correlation	1.0	1.0	.88	.84	.83	.82	.76	.75	-.81
Item Pair	87/88	96/97	3/11	80/86	82/84	84/85	95/99	83/86	89/96 89/97

Using a critical value of 0.7 led to the conclusion of 10 out of 4,950 English Language paired items identified are in violation of the 3PL model as seen in Table 38.

Table 39 depicts the standardized residual correlations of paired items that are used to identify dependent items for 2014 mathematics data.

Table 39: Standardized Residual Correlations for Mathematics 2014

Residual correlation	.55	.40	.33	.31	.60	-.49	-.32	-.30
Item Pair	11/48	17/18	31/35	46/47 18/19	48/49	2/46	11/49 9/26	21/36

Using a critical value of 0.7 led to the conclusion of none of the 1,225 mathematics paired items identified violate the 3PL model as seen in Table 39.

Table 40 depicts the standardized residual correlations of paired items that are used to identify dependent items for 2014 Integrated Science data.

Table 40: Standardized Residual Correlations for Integrated Science 2014

Residual correlation	.81	.74	.73	.72	.69	.64	.62	-.78	-.64
Item Pair	36/37	7/14	9/40	40/41	23/25 17/40	10/25	9/41	26/42	7/10

Using a critical value of 0.7 led to the conclusion of 4 out of 1,225 Integrated Science paired items identified violate the 3PL model as seen in Table 40. Table 41 depicts the standardized residual correlations of paired items that are used to identify dependent items for 2014 social studies data.

Table 41: Standardized Residual Correlations for Social Studies, 2014

Residual correlation	.64	.57	.52	.49	-.69	-.64	-.55	-.47
Item Pair	13/44	19/31	19/34	28/35 26/33 16/25	14/44	13/14	16/21	16/39

Using a critical value of 0.7 led to the conclusion of none of the 1,225 social studies paired items identified violate the 3PL model as seen in Table 41.

Table 42 depicts the standardized residual correlations of paired items that are used to identify dependent items for 2015 English Language data.

Table 42: Standardized Residual Correlations for English Language 2015

Residual correlation	.80	.79	.71	.70	.69	-.79	-.76	-.74	-.73
Item Pair	25/34 25/65	35/56	25/63	34/63	34/65	34/75	56/75	35/75	25/75

Using a critical value of 0.7 led to the conclusion of 3 out of 3,008 English Language paired items identified violate the 3PL model as seen in Table 42. Table 43 depicts the standardized residual correlations of paired items that are used to identify dependent items for 2015 mathematics data.

Table 43: Standardized Residual Correlations for Mathematics 2015

Residual correlation	.69	.63	.60	.56	.55	-.68	-.60	-.57
Item Pair	11/26	10/11	10/23 3/11	16/18 21/23	39/47	26/50	11/50	10/50

Using a critical value of 0.7 led to the conclusion that none of the 1,225 mathematics paired items identified violates the 3PL model as seen in Table 43.

Table 44 depicts the standardized residual correlations of paired items that are used to identify dependent items for 2015 Integrated Science data.

Table 44: Standardized Residual for Integrated Science 2015

Residual correlation	.69	.54	.52	.43	-.64	-.63	-.61	-.58	-.54	-.47
Item Pair	22/28	5/12	14/42	38/47	45/47	22/47	28/47	24/50	7/38	14/50

Using a critical value of 0.7 led to the conclusion that none out of the 1,225 Integrated Science paired items identified violate the 3PL model as seen in Table 44.

Table 45 depicts the standardized residual correlations of paired items that are used to identify dependent items for 2015 social studies data.

Table 45: Standardized Residual Correlations for Social Studies 2015

Residual correlation	.62	.57	.52	.48	-.56	-.51	-.50	-.46	-.44
Item Pair	16/18	12/15	10/40	6/12	4/10 9/31	22/31	34/50	12/43	15/43

Using a critical value of 0.7 led to the conclusion of none of the 1,225 social studies paired items identified violate the 3PL model as seen in Table 45.

Table 46 depicts the standardized residual correlations of paired items that are used to identify dependent items for 2016 English Language data.

Table 46: Standardized Residual Correlations for English Language 2016

Residual correlation	.87	.74	.73	.70	-.75	-.74	-.69	-.67	-.62	-.61
Item Pair	4/29	4/5	5/29	21/29	76/77	29/50	62/72	4/50	29/42	5/50

Using a critical value of 0.7 led to the conclusion of 5 out of 3,160 English Language paired items identified violate the 3PL model as seen in Table 46.

Table 47 depicts the standardized residual correlations of paired items that are used to identify dependent items for 2016 mathematics data.

Table 47: Standardized Residual Correlations for Mathematics 2016

Residual correlation	.65	.61	.54	-.78	-.60	-.59	-.57	-.55
Item Pair	21/29	6/21	19/23 6/29 43/47	11/28	21/47	21/43	21/32	6/32

Using a critical value of 0.7 led to the conclusion of none of the 1,225 mathematics paired items identified violate the 3PL model as seen in Table 47.

Table 48 depicts the standardized residual correlations of paired items that are used to identify dependent items for 2016 Integrated Science data.

Table 48: Standardized Residual Correlations for Integrated Science 2016

Residual correlation	.79	.66	.65	.63	.62	-.66	-.60	-.58
Item Pair	31/37	14/43	37/48 16/47	20/42	27/31	13/16	13/47	13/50 16/44

Using a critical value of 0.7 led to the conclusion of 1 out of 1,225 Integrated Science paired items identified violate the 3PL model as seen in Table 48.

Table 49 depicts the standardized residual correlations of paired items that are used to identify dependent items for 2016 social studies data.

Table 49: Standardized Residual Correlations for Social Studies, 2016

Residual correlation	.94	.93	.92	.90	.89	.80	.78	.73	-.70
Item Pair	8/14	32/40	12/14	28/36 10/30	8/12	28/30	10/28	10/16	9/40

Using a critical value of 0.7 led to the conclusion of 8 out of 1,225 social studies paired items identified violate the 3PL model as seen in Table 49.

Checking measurement invariance assumption

Measurement invariance or measurement equivalence is a statistical property of measurement that indicates that the same construct is being measured across some specified groups. For example, measurement invariance can be used to test whether a given measure is interpreted in a conceptually similar manner by respondents representing different genders or cultural backgrounds. Violations of measurement invariance may preclude the meaningful interpretation of measurement data.

Measurement invariance is often tested in the framework of multiple-group confirmatory factor analysis (CFA) according to Chen, Sousa and West (2005). In the context of structural equation models, including CFA, measurement invariance is often termed *factorial invariance* (Widaman, Ferrer, & Conger, 2010).

Tests for invariance

Although there is a need for further research on the application of various invariance tests and their respective criteria across diverse testing conditions, two approaches are common among applied researchers. For each model being compared (e.g., Equal form, Equal intercepts), a χ^2 fit statistic is iteratively estimated from the minimization of the difference between the model implied mean and covariance matrices and the observed mean and covariance matrices (Loehlin, 2004). As long as the models under comparison are nested, the difference between the χ^2 values and the respective degrees of freedom of any two CFA models of varying levels of invariance follows a χ^2 distribution (DIF χ^2) and as such, can be inspected for significance as an indication of whether increasingly restrictive models produce appreciable changes in model-data fit (Loehlin, 2004). However, there is some evidence the DIF χ^2 is sensitive to factors unrelated to changes in invariance targeted constraints (e.g., sample size) according to Cheung and Rensvold (2002).

As a result, researchers also recommend the use of the difference between the comparative fit indexes (ΔCFI) of two models specified to investigate measurement invariance. When the difference between the CFIs of two models of varying levels of measurement invariance (e.g., equal forms versus equal loadings) is greater than 0.01, then invariance in likely untenable (Cheung and Rensvold, 2002). It is important to note that the CFI values being subtracted are expected to come from nested models as in the case of DIF χ^2 testing (Widaman & Thompson, 2003). However, there is an indication that applied researchers rarely consider this when applying the CFI test (Kline, 2011).

According to Drasgow and Kanfer (1985), a test or a subscale is said to have measurement invariance/equivalence across groups or populations if persons with identical scores on the underlying/latent construct have the same expected raw score or true score at the item level, the subscale total score level, or both. Without measurement equivalence, it is difficult to interpret observed mean score differences meaningfully. That is, observed mean score differences may reflect the true mean difference between the groups as well as a difference in the relationship between the latent variable and the observed score that is not identical across groups. When measurement invariance is present, the relationship between the latent variable and the observed variable remains invariant across populations. In this case, the observed mean difference may be viewed as reflecting only the true difference between the populations.

Dataset on 2012-2016 of English Language, Mathematics, Integrated Science and Social Studies was tested for measurement invariance and results are presented in Tables 50 -53. The results are presented for each subject across the five years under study for each dataset.

Table 50: Test of Measurement Invariance for 2012-2016 English Language Dataset

	Chi-square	Df	p-value	Invariance?
Overall Model (2012)				
Unconstrained	537557.597	4851	.000	
Fully constrained	1075115.193	9801	.000	
Number of groups		2		
Difference	537557.596	4950	.000	NO
2013				
Unconstrained	1243869.722	4850	.000	

Fully constrained	2463616236	9800	.000	
Number of groups		2		
Difference	2462372366	4950	.000	NO
2014				
Overall Model				
Unconstrained	503325.123	5732	.000	
Fully constrained	1234566.213	98203	.000	
Number of groups		2		
Difference	731241.09	92471	.000	NO
2015				
Unconstrained	533019.577	2925	.000	
Fully constrained	879842.274	5928	.000	
Number of groups		2		
Difference	346822.697	3003	.000	NO
2016				
Unconstrained	660499.113	3081	.000	
Fully constrained	1320998.227	6241	.000	
Number of groups		2		
Difference	660499.114	3160	.000	NO

From Table 50, results show that the 2012-2016 English Language dataset has no measurement invariance. In this case, the observed mean difference may be viewed as reflecting only the abituary difference between the populations and not true difference. Hence the assumption of measurement invariance is violated meaning there is variation within the dataset.

Table 51 displays the results of the 2012-2016 Mathematics dataset to test for measurement invariance.

Table 51: Test of Measurement Invariance for 2012-2016 Mathematics Dataset

2012	**Chi-square**	**df**	**p-value**	**Invariance?**
Overall Model				
Unconstrained	187997.114	1175	.000	
Fully constrained	375994.228	2400	.000	
Number of groups		2		
Difference	187997.114	1225	.000	NO
2013				
Unconstrained	495294.02	1175	.000	
Fully constrained	990588.04	2400	.000	
Number of groups		2		
Difference	495294.02	1225	.000	NO
2014				
Overall Model				
Unconstrained	197949.75	1127	.000	
Fully constrained	395899.499	2303	.000	
Number of groups		2		
Difference	197949.749	1176	.000	NO
2015				
Unconstrained	294531.235	1175	.000	
Fully constrained	491783.776	2400	.000	
Number of groups		2		
Difference	197252.541	1225	.000	NO
2016				
Unconstrained	374430.832	1175	.000	
Fully constrained	748861.664	2400	.000	
Number of groups		2		
Difference	374430.832	1225	.000	NO

From Table 51, results indicate that measurement invariance is absent, thus the relationship between the latent variable (mathematical ability) and the observed variable remains non-invariant across the populations. In this case, the observed mean difference may be viewed as not reflecting the true difference between the populations.

Table 52 depicts the results of measurement invariance for 2012-2016 Integrated Science dataset. The results indicate whether measurement invariance assumption is violated or satisfied for Integrated Science dataset for 2012-2016.

Table 52: Test of Measurement Invariance for 2012-2016 Integrated Science Dataset

	Chi-square	**df**	**p-value**	**Invariance?**
Overall Model (2012)				
Unconstrained	175469.127	1175	.000	
Fully constrained	350938.255	2400	.000	
Number of groups		2		
Difference	175469.128	1225	.000	NO
2013				
Unconstrained	383109.85	1175	.000	
Fully constrained	766219.699	2400	.000	
Number of groups		2		
Difference	383109.849	1225	.000	NO
2014				
Overall Model				
Unconstrained	228797.976	1175	.000	
Fully constrained	457595.953	2400	.000	
Number of groups		2		
Difference	228797.977	1225	.000	NO

2015				
Unconstrained	287494.586	1175	.000	
Fully constrained	574989.172	2400	.000	
Number of groups		2		
Difference	287494.586	1225	.000	NO
2016				
Unconstrained	366723.081	1175	.000	
Fully constrained	5777544.02	2400	.000	
Number of groups		2		
Difference	5410820.939	1225	.000	NO

Results from Table 52 showed that the 2012-2016 Integrated Science dataset violated the assumption of measurement invariance which means that there is group difference at the model level. In this case, the observed mean difference may be viewed as not reflecting the true difference within the population of 2012-2016 Integrated Science WASSCE examinees.

Table 53 shows measurement invariance results on 2012-2016 Social Studies dataset.

Table 53: Test of Measurement Invariance for 2012-2016 Social Studies Dataset

	Chi-square	**df**	**p-value**	**Invariance?**
Overall Model (2012)				
Unconstrained	215907.692	1175	.000	
Fully constrained	431815.383	2400	.000	
Number of groups		2		
Difference	215907.691	1225	.000	NO

2013				
Unconstrained	394285.105	1175	.000	
Fully constrained	7888570.21	2350	.000	
Number of groups		2		
Difference	7494285.105	1175	.000	NO
2014				
Overall Model				
Unconstrained	216457.856	1175	.000	
Fully constrained	432915.712	2400	.000	
Number of groups		2		
Difference	216457.856	1225	0.000	NO
2015				
Unconstrained	256012.309	1175	.000	
Fully constrained	512024.618	2400	.000	
Number of groups		2		
Difference	256012.309	1225	.000	NO
2016				
Unconstrained	418135.272	1175		
Fully constrained	648967.614	2400	.000	
Number of groups		2	.000	
Difference	230832.342	1225	.000	NO

From Table 53, results depict that Social Studies dataset violates the measurement invariance assumption. This means that the subpopulations within the 2012-2016 WASSCE examinees are different at the model level, thus the relationship between the latent variable (Social Studies skill) and the observed variable remains non- invariant across populations. In this case, the observed mean difference may be viewed as not reflecting the true difference between the populations.

IRT Model-Data Fit

Item fit is a question of the utility of the data for analysis by the measurement model, whereas person fit is a question of the interpretation and inference (i.e., validity) of the measure of an examinee. According to Smith (1990), item fit is concerned with whether the data fit the model and this question must be answered before further analyses of the data are useful. This study's focus is DIF, and one needs to verify that the data adequately fit the IRT model before 3PL DIF analyses can be conducted.

The fit of the data to the 3PL IRT model was verified by examining the infit mean square (infit MNSQ) and outfit mean square (outfit MNSQ) statistics for the calibration of items and the estimation of persons' ability (Linacre, 2002). The former is standardized information weighted mean square statistic, which is more sensitive to responses near the person's ability. The latter is a standardized outlier-sensitive mean square statistic, which is more sensitive to responses far away from the person's ability (Linacre, 2009a). Real data depart from the 3PL model to some extent as no data could ever perfectly fit a model. An outfit MNSQ statistic and an infit MNSQ statistic value of 1 and 1.1 respectively is the ideal of IRT model specification. This is when the data fit the model, and then only the advantages of the model can be used in constructing a measure. Hambleton (1993) indicated that:

> The potential of item response theory for solving many problems in testing and measurement is high; however, the success of particular IRT applications is not assured simply by processing test results through one of the available computer programmes. The advantages claimed for item response models can be realized only when the fit between the model and the test dataset of interest is satisfactory. A poorly fitting model cannot yield invariant item and ability parameter estimates (p.172).

Linacre and Wright (1994) wondered how much noise is tolerable. How close to 1 is good enough? According to Bond and Fox (2007), there is no standard rule for this, and it depends on the testing context. An acceptable fit range of 0.80 to 1.20 seemed tolerable for a high-stakes examination. Values greater than 1.20 indicate noise in the data and values lower than 0.80 may indicate item redundancy. For this reason, fit statistic values greater than 1.20 are further investigated, to ascertain whether there should be any reason to be concerned that the data are behaving in the same way as the construct being measured.

Item fit

The fit to the 3PL model was examined by inspecting summary fit statistics (i.e., overall items or persons) as well as fit of individual items. Table 59 shows a summary of fit statistics of measured items for 2012 dataset, which indicates that the mean of infit MNSQ statistic and the mean of outfit MNSQ statistic of English Language, are 0.94 and 1.67 respectively.

Tables 54 to 58 depict means and standard deviations of the infit and outfit of 2012-2016 English Language, Mathematics, Integrated Science and Social Studies items.

Table 54: Summary of Data Fit Statistics of 2012 Items

	INFIT		OUTFIT	
Subject	Mean	Standard deviation	Mean	Standard deviation
English Language	.94	.43	1.67	1.29
Mathematics	.94	.47	1.56	2.67
Integrated Science	.95	.36	1.02	.68
Social Studies	.92	.39	.73	.95

All values are close to the expected value of 1.00 as seen in Table 54. The fact that the means of the infit MNSQ and the means of the outfit MNSQ statistics are close to 1 and the relatively small standard deviations provide evidence that test items fit the 3PL model and are behaving in the same way as the construct being measured.

Table 55: Summary of Data Fit Statistics of 2013 Items

	INFIT		OUTFIT	
Subject	Mean	Standard deviation	Mean	Standard deviation
English Language	.96	.45	1.89	1.39

Mathematics	.95	.49	1.65	2.80
Integrated Science	.98	.37	1.05	.78
Social Studies	.93	.40	.74	.97

Table 55 depicts that results are close to the expected value of 1.1 and 1.0 of infit and outfit respectively. The fact that the means of infit MNSQ and the means of outfit MNSQ statistics fit the 3PL model and are behaving the same way as the construct being measured.

Table 56: Summary of Data Fit Statistics of 2014 Items

	INFIT		OUTFIT	
Subject	Mean	Standard deviation	Mean	Standard deviation
English Language	.96	.56	1.42	2.50
Mathematics	.96	.37	1.12	2.18
Integrated Science	.91	.44	2.51	3.59
Social Studies	.94	.35	1.61	2.45

From Table 56, the means of the infit MNSQ and outfit MNSQ statistics of 2014 are close to 1 except for Integrated Science and Social Studies which have MNSQ statistics a little higher than 1. The relatively small standard deviations provide evidence that test items fit the 3PL model and are behaving in the same way as the construct being measured.

Table 57: Summary of Data Fit Statistics of 2015 Items

	INFIT		OUTFIT	
Subject	Mean	Standard deviation	Mean	Standard deviation
English Language	.93	.46	1.07	2.41
Mathematics	1.00	.39	.90	1.42
Integrated Science	.92	.40	1.35	2.71
Social Studies	.94	.27	1.72	2.52

Table 57 showed that values of the INFIT MNSQ statistics of the 2015 dataset were higher the established rule of 1.0 and 1.1 for mathematics and OUTFIT MNSQ for all except mathematics. From the analysis, the 2015 dataset is considered reliable for analysis.

Table 58: Summary of Data Fit Statistics of 2016 Items

	INFIT		OUTFIT	
Subject	Mean	Standard deviation	Mean	Standard deviation
English Language	.94	.36	1.15	2.52
Mathematics	.94	.47	1.34	2.62
Integrated Science	.98	.37	1.05	.78
Social Studies	.91	.38	2.15	3.36

The MNSQ statistics of the dataset of 2016 (Table 58) had values that were a little higher than one. Since the infit MNSQ statistics of the items do not depart from the satisfactory range (according to the predetermined range of 0.8 and 1.20), these items are not considered to be unreliable enough to vitiate the measurement system. Though the statistics of the outfit seems to be large, it is worth noting that it poses less threat to measurement (Linacre, 2009a).

This study concludes that the model-data fit statistics provide evidence that the data conformed adequately to the 3PL model and the items are measuring single English Language, Mathematics, Integrated Science and Social Studies ability constructs.

In summation, it is concluded that the 2012-2016 dataset contended sufficiently the 3PL model requirements namely, the local independence, the unidimensionality and model-data fit satisfactorily for further analyses on DIF.

Results

The results of the DIF analyses for the 2012-2016 WASSCE dataset are presented according to the DIF procedures MH, LR and 3PL IRT for both gender and region. The results are discussed based on subjects and DIF procedures used. The corresponding items identified as DIF were denoted by the square (□) for MH), circle (○) for LR and triangle (△) for 3PL IRT. The criteria for items identified as DIF are classified based on the effect size, thus only B (medium) and C (large) categories are counted as DIF items through the data analysis for all DIF procedures. Items that exhibit DIF in favour of reference group were indicated by a plus sign (+) and minus (-) for DIF in favour of the focal group in the analysis. For gender-DIF, male are classified as reference group and females as focal group whiles for location-DIF, CR and GAR served as a reference against FG (WR, VR, ER and CR/ GAR) as focal group.

Effect Size

The importance of utilizing an ES with a statistical significance finding has been demonstrated in the DIF literature (DeMars, 2009; Meade, 2010; Jodoin & Gierl, 2001). Zwick and Ercikan (1989) proposed the following interpretation guidelines to evaluate the DIF effect size as:

I. Negligible or A-level DIF: the item is not statistically significant,

II. Medium or B-level DIF: the item is statistically significant,

III. Large or C-level DIF: the item is statistically significant

Results of Research hypotheses one and two

1. H_0: There is no statistically significant gender differential item functioning in 2012-2016 WASSCE core subjects' examinations using MH DIF detecting procedure.

   H_1: There is a statistically significant gender differential item functioning in 2012-2016 WASSCE core subjects' examinations using MH DIF detecting procedure.

2. H_0: There is no statistically significant location differential item functioning in 2012-2016 WASSCE core subjects' examinations using MH DIF detecting procedure.

   H_1: There is a statistically significant location differential item functioning in 2012-2016 WASSCE core subjects' examinations using MH DIF detecting procedure.

The results of gender and location-DIF analysis by MH procedure for English Language, Mathematics, Social Studies and Integrated Science are presented in Tables 60-71.

Table 59 depicts results of gender-DIF by MH procedure for 2012-2016 English Language dataset. The + values indicate items that showed DIF in favour of males and – values in favour of females.

Table 59: Distribution of DIF Items in 2012-2016 English Language Exams Using Gender

Year	MH LOR		USING ETS CRITERIA		
	Male	Female	A	B	C
2012	39	60	+15 -10	+7 -8	+17 -42
2013	45	52	+10 -9	+10 -5	+25 -38
2014	45	39	+10 -9	+10	+25 -30
2015	31	45	+8 -8	-5	+5 -50
2016	36	39	+10 -7	+8	+18 -32

From Table 59, it can be reported that for the 78-100 English Language items examined, items that indicated DIF showed a decreasing rate. Thirty-one (31) to forty-five (45) items indicated DIF in favour of males (positive values) whiles 39 to 60 items indicated DIF in favour of females (negative values). The items that exhibited DIF were 24 (using B & C levels as rule) items having actual DIF in favour of male and 50 (using B & C levels as rule) items in favour of females in 2012. There is 35 actual DIF in favour of male as against 43 items in favour of females in 2013. The years 2014, 2015 and 2016 showed similar results. Table 58 shows that using the MH procedure, English Language items had more DIF items in favour of males than females. Items identified as DIF has been indicated on the instrument (question paper for English Language) as Appendix A.

Table 60 showed data analysis of location-DIF, where CR was used as the reference group for English Language 2012-2016 dataset. Items that showed DIF were out of 100 for 2012, 2013 and 2014, 78 for 2015 and 80 for 2016. The + values indicate items that showed DIF in favour of reference group (RF) and – values in favour of focal group (FG).

Table 60: Distribution of DIF Items in 2012-2016 English Language Exams Using Location with CR as Reference Group (RF)

Year	MH LOR		USING ETS CRITERIA		
	CR	FG	A	B	C
2012	44	55	+14	+7	+23
			-10	-9	-36
2013	49	49	+12	+2	+35
			-9	-20	-20
2014	43	43	+11	+7	+25
			-12	-5	+26
2015	37	40	+10	+6	+21
			-16	-2	-22
2016	25	54	+7	+3	+15
			-10	-4	-40

Table 60 indicates the results of location DIF with CR as the reference group and WR, GAR, VR, ER as the focal group. The results indicate that more items showed DIF in favour of the examinees from the other four regions as compared to the examinees from CR in 2012 and 2014. Thus, 45 (using items exhibiting B and C level of effect size) items in 2012 and 44 items in 2013 all exhibited DIF in favour of the focal group. 2013 test items had 37 DIF items in favour of examinees from CR whiles 40 items exhibited DIF in favour of the focal group. In 2014, 2015 and 2016, 32, 27and 18 items were identified as DIF respectively based on ETS in favour of students who schooled in CR. In all English Language test, items exhibited more DIF items in favour of the focal group than the reference group.

Table 61 shows the results of English Language 2012-2016 location-DIF analysis using the MH procedure. Items that showed DIF were out of 100 for 2012, 2013 and 2014, 78 for 2015 and 80 for 2016.

Table 61: Distribution of DIF Items in 2012-2016 English Language Exams Using Location with GAR as RG

Year	MH LOR		USING ETS CRITERIA		
	GAR	FG	A	B	C
2012	50	50	+9	+3	+38
			-10	-10	-30
2013	49	51	+10	+7	+22
			-18	-10	-23
2014	34	62	+30	+4	-4
			-48	-10	
2015	35	42	+10	+7	+18
			-14	-10	-18
2016	27	53	+18	+5	+4
			-10	-7	-36

According to Table 61, 41 items as against 40 items exhibited DIF in favour of examinees who wrote the test in GAR as compared to their counterpart in who wrote at the other regions respectively in 2012 based on the ETS. Also, 14 items were identified as DIF in favour of the focal group whiles only 4 items were identified as DIF in favour of the reference group in 2014. In addition to these 9 items as against 43 items were identified as DIF in favour of the focal group in 2016 based on the ETS (Thus, using the B and C levels as criteria). In 2014(14 items), 2015(28 items) and 2016(43 items) exhibited DIF in favour of students who schooled in other regions.

Table 62 presents the result of gender-DIF for 2012-2016 Mathematics test items using the MH procedure. Items that showed DIF were out of 50 for all years.

Table 62: Distribution of DIF Items in 2012-2016 Mathematics Exams Using Gender

Year	MH LOR		USING ETS CRITERIA		
	Male	Female	A	B	C
2012	15	34	+9	+3	+3
				-5	-29
2013	18	32	+7	+5	+6
			+5		-27
2014	20	29	+8	+3	+9
			-8	-5	-16
2015	21	29	+8	+2	+11
			-10	-3	-16
2016	18	29	+11	-4	+7
					-25

The results in Table 62 based on the ETS (B and C levels) indicated that 34 items showed DIF in favour of focal group (females) whiles only 6 items were identified as DIF in favour of the reference group (males) in 2012. In 2013, the results are no different from 2012 since 27 items exhibited DIF in favour of females as compared to 11 items identified as DIF in favour of males. In general, Mathematics test items showed more items identified as DIF in favour of females than males. Items identified as DIF are indicated in Appendix B.

Table 63 indicate results of location-DIF for 2012-2016 Mathematics test items by MH procedure with CR serving as the reference group. The positive values indicate DIF in favour of the reference group whiles the negatives values indicate DIF in favour of the focal group. Items that showed DIF were out of 50 for all years under study.

Table 63: Distribution of DIF Items in 2012-2016 English Language Exams Using Location with CR as the RG

Year	MH LOR		USING ETS CRITERIA		
	CR	FG	A	B	C
2012	13	36	+11 -10	+2 -5	-11
2013	22	27	+6 -6	+4	+12 -21
2014	21	24	+10 -4	+1 -1	+10 -19
2015	18	32	+9 -5	-4	+5 -27
2016	20	28	+5 -5	+1 -3	+14 -20

Table 63 results based on ETS indicated 16 DIF items in favour of the focal group (students who schooled in other regions) and 2 DIF items in favour of the reference group in 2012. It can be seen from the result that the year 2012 candidates who schooled in CR had the least number of items indicating DIF in their favour. The results once again showed that most items (34 out of 50) answered by 2016 candidates were at the large (C) level of DIF of which 20 were in favour of the focal group. In 2015, out of 50 items, 31 were identified as DIF in favour of the focal group whiles 5 items were in favour of the reference group.

Table 64 depicts location-DIF of 2012-2016 Mathematics test items with GAR as reference group using the MH procedure. Items that showed DIF were out of 50 each for each year under study.

Table 64: Distribution of DIF Items in 2012-2016 Mathematics Exams Using Location with GAR as the RG

Year	MH LOR		USING ETS CRITERIA		
	GAR	FG	A	B	C
2012	16	32	+10	+3	+3
			-5	-6	-21
2013	19	31	+9	-4	+10
			-6		-21
2014	15	33	+10	+4	+1
			-4	-2	-27
2015	21	29	+15	+6	-25
			-4		
2016	18	30	+9	+2	+8
			-5	-2	-20

Table 64 shows that throughout the five years, over 40 items were identified as DIF using the MH procedure. Items that were not flagged as DIF were few(i.e.,2- 6 items). The year 2012 showed the highest number of items (27 based on ETS) DIF in favour of the focal group. In 2012, 6 items were flagged in favour of the reference group whiles 27 items flagged in favour of the focal group. The results for 2012 were not quite different from the subsequent years 2013, 2014, 2015 and 2016 because, in 2014, 29 items were flagged ad DIF(based on ETS) in favour of the focal group whiles in 2015, 25 items were identified as DIF in favour of the examinees who wrote the examination in FG (CR, VR, ER and WR). Thus, for Mathematics, most of the examinees who wrote in GAR had an advantage over the examinees in other regions.

Table 65 provides results on gender-DIF for 2012-2016 Integrated Science dataset using the MH procedure. Items that showed DIF in the Table were out of 50 each for each year under study.

Table 65: Distribution of DIF Items in 2012-2016 Integrated Science Exams Using Gender

Year	MH LOR		USING ETS CRITERIA		
	Male	Female	A	B	C
2012	16	32	+6	+3	+7
			-4		-28
2013	20	29	+9	-3	+8
			-4		-25
2014	21	29	+9	+3	+9
				-5	-24
2015	15	33	+8	+4	+3
			-7	-4	-22
2016	16	34	+7	+2	+7
			-12	-4	-18

Results from Table 65 results based on MH LOR shows that 2014 flagged the highest number (50) of DIF items even though 2012, 2013, 2015 and 2016 also flagged more than 40 items each. The years, 2012 and 2013 showed an equal number (29) DIF items. It can be noticed from Table 70 that, most of the items (based on the ETS) which were identified as DIF were in favour of the focal group. Notwithstanding, 10 (in favour of reference group) items as against 28 (in favour of the focal group) items were considered as actual DIF(based on B and C levels). In 2012 items exhibited negligible DIF(A) and these were not included in this study. A similar trend occurred in 2013, 2014, 2015 and 2016. Items identified as is DIF in 2012-2016 Integrated Science data have been identified as such in Appendix C.

Table 66 demonstrates the results of Integrated Science items for 2012-2016 DIF analysis for the location with CR as the reference group using the DIFAS. Thus, examinees that wrote the examinations in the Central region were assumed to have an advantage over examinees from other regions in Ghana. Items that showed DIF in the Table were out of 50 each for each year under study.

Table 66: Distribution of DIF Items in 2012-2016 Integrated Science Exams Using Location with CR as the RG

Year	MH LOR		USING ETS CRITERIA		
	CR	FG	A	B	C
2012	19	28	+10	+9 -3	-28
2013	22	28	+6 -7	+2 -1	+14 -20
2014	19	30	+3 -7	-2	+16 -21
2015	22	26	+8 -2	+4 -2	+10 -22
2016	24	25	+5 -5	+5 -3	+14 -17

The results show that items classified as DIF in favour of the focal group (GAR, WR, ER and VR) were more than that of the reference group. In 2016, the items identified as DIF based on MH LOR were 19 and 20 in the reference and focal groups respectively. The MH LOR 2014 results were quite different from the rest of the years because even though 30 items exhibited DIF in favour of the focal group, only 23 (based on ETS) were considered as DIF in this study.

Information in Table 67 indicates the results of the MH procedure using DIFAS in the analysis of location-DIF with GAR as the reference group. Items that showed DIF in the Table were out of 50 each for each year under study.

Table 67: Distribution of DIF Items in 2012-2016 Integrated Science Exams Using Location with GAR as the RG

Year	MH LOR		USING ETS CRITERIA		
	GAR	FG	A	B	C
2012	20	29	+8	+6	+6
			-4	-5	-20
2013	24	20	+3	+3	+18
			-1		-19
2014	16	32	+6	-5	+10
			-4		-23
2015	15	34	+10	-3	+5
			-4		-27
2016	19	30	+5	+2	+12
			-5	-4	-21

Results from Table 68 indicates that all years but 2013 flagged 49 items as DIF out of 50 items. In 2012, twenty-four items out of 29 items were considered as DIF in favour of focal group whiles 12 items out of 20 were considered as DIF in favour of the reference group. The result was not too different from the dataset for 2013 (19 items), 2014 (28 items), 2015 (27 items) and 2016 (25 items) of the Integrated Science data based on ETS criteria exhibited DIF in favour of the focal group.

Table 68 reported on the gender-DIF analysis of Social Studies items for 2012-2016 WASSCE.

Table 68: Distribution of DIF Items in 2012-2016 Social Studies Exams Using Gender

Year	MH LOR		USING ETS CRITERIA		
	Male	Female	A	B	C
2012	14	32	+7	+5	-27
			-5	-2	
2013	15	33	+8	+2	+5
			-10	-5	-18
2014	21	29	+6	+6	+9
			-5		-24
2015	15	34	+10	+3	+2
			-4	-3	-27
2016	19	29	+7	+9	+2
				-1	-29

From Table 68, ETS criteria analysis indicated that 2015 recorded the highest number (30) of DIF items which were in favour of the females as compared to the rest of the years under study. Even though 2016 had the highest number of items considered as DIF, 30 out of 41 items were identified as DIF in favour of the focal group. In 2015 out of 35 items considered as DIF, only 5 items were in favour of the reference group, thus the females were at a disadvantage in answering most of the Social Studies items.

Table 69 presents results on 2012-2016 WASSCE dataset for Social Studies location-DIF analysis using MH procedure by DIFAS. CR was used as the reference group for the analysis.

Table 69: Distribution of DIF Items in 2012-2016 Social Studies Exams Using Location with CR as the RG

Year	MH LOR		USING ETS CRITERIA		
	CR	FG	A	B	C
2012	19	27	+5	+9	+5
			-10	-1	-16
2013	18	31	+8	+2	+8
			-6	-4	-21
2014	17	33	+10	+4	+3
			-6		-27
2015	17	31	+5	+5	+7
			-8	-3	-20
2016	21	28	+6	+2	+13
			-4	-5	-19

CR as a reference group in location DIF analysis results, show that 2013 and 2015 had an equal number (31 items) that showed DIF in favour of the focal group whiles 2014 and 2015 show equal number (17 items) of DIF in favour of candidates. In 2013 (25 out of 31 items) and 2015 (23 out of the 31 items) were considered as DIF in favour of the focal group based on the ETS criteria. In 2014, out of the items identified as DIF, 7 were considered to exhibit DIF in favour of the reference group as shown in Table 70.

Table 70 indicates the results of data analysis using the MH procedure for 2012-2016 Social Studies using GAR as the reference group.

Table 70: Distribution of DIF Items in 2012-2016 Social Studies Exams Using Location with GAR as the RG

	MH LOR		USING ETS CRITERIA		
Year	GAR	OTHERS	A	B	C
2012	22	26	+10	+3	+9
			-2	-3	-21
2013	15	34	+5	-7	+10
			-6		-21
2014	19	31	+8	+2	+9
			-5		-26
2015	19	30	+6	+5	+8
			-6	-6	-18
2016	18	32	+6	+3	+9
			-3	-2	-27

The least number of items (26) that showed DIF in favour of the focal group was recorded in 2012 as demonstrated in Table 70 based on ETS criteria. Thus, out of the 26 items, 24 items were considered as DIF. The largest number (36 items) that exhibit DIF at level C was in 2016 of which 27 were in favour of the focal group. Items that were identified as DIF are indicated in Appendix D.

Summary

The data analysis of 2012-2016 English Language, Mathematics, Integrated Science and Social Studies WASSCE using MH procedure through DIFAS indicated that the items were not free from both gender and location-DIF. It was also revealed that items identified as DIF were in favour of both subgroups understudy even though more items were identified to exhibit DIF in favour of the focal group in all four subjects. English Language, Integrated Science and Mathematics dataset exhibited more items in favour of females as compared to their male counterparts. Social Studies dataset identified an almost equal number of items that exhibited gender-DIF in favour of both groups. Location –DIF results were quite interesting because the analysis revealed that even though the number of items identified as DIF was different from year to year for the four subjects, the difference between the focal and reference groups was not large.

Research hypotheses three and four

3. H_0: There is no statistically significant gender differential item functioning in 2012-2016 WASSCE core subjects Examinations using LR DIF detecting procedure.

   H_1: There is a statistically significant gender differential item functioning in 2012-2016 WASSCE core subjects Examinations using LR DIF detecting procedure.

4. H_0: There is no statistically significant location differential item functioning in 2012-2016 WASSCE core subjects Examinations using LR DIF detecting procedure.

   H_1: There is a statistically significant location differential item functioning in 2012-2016 WASSCE core subjects Examinations using LR DIF detecting procedure.

In the present study, an item reveals uniform DIF when the significant odds ratio is for the group, whereas the item reveals non-uniform DIF when the significant odds ratio is for the interaction between the group and total score. The item reveals DIF in favour of the reference group when the significant odds ratio is greater than one, whereas the item reveals DIF in favour of the focal group when the significant odd ratio is less than one (α = 0.05).

Table 71 shows the summary results of the LR method to identify DIF on the English language ability scale for each of 100 items for 2012, 2013 and 2014, 78 items for 2015 and 80 items for 2016.

Table 71: Distribution of DIF Items in 2012-2016 English Language Exams Using Gender

	Number of DIF Items Flagged		Type of DIF	
Year	Male	Female	Uniform	Non Uniform
2012	43	38	1(M)	80

2013	31	37	2(M)
			6(F)
				60

2014	32	27	14(M)
			4(F)
				41

2015	35	22	4(M)	53

2016	28	27	3(M)
			1(F)
				55

Eighty-one items or 81% of the items revealed DIF of which only one exhibited uniform DIF in favour of males in 2012. Out of 80 nonuniform DIF, 43 of them were in favour of males in 2012 as shown in Table 71. In 2013, there were 60 nonuniform DIF items as compared to 2 uniform DIF items. Items identified as DIF seemed to decrease as the years go by. Observation for the five years showed that most of the English Language items revealed non-uniform DIF.

Table 72 shows the summary results of the LR method to identify location-DIF on the English language ability scale for 2012-2016 WASSCE items.

Table 72: Distribution of DIF Items in 2012-2016 English Language Exams Using Location with CR as the RG

Year	Number of DIF Items Flagged		Type of DIF	
	CR	FG	Uniform	Non Uniform
2012	31	33	4 (CR)	60
2013	32	36	5 (CR)	63
2014	33	27	2 (CR) 2 (FG)	56
2015	27	21	3 (CR) 4 (FG)	41
2016	14	23	10 (CR) 4 (FG)	27

From Table 72, results revealed that fourteen (28%) to thirty-three (66%) of the items exhibited DIF in favour of CR whereas 42% to 72% of the items revealed DIF in favour of the other regions (i.e., GAR, WR, VR and ER). Four to fourteen items exhibited uniform DIF, whereas thirty-seven to sixty-eight items exhibited non-uniform DIF. The uniform DIF identified in the year 2016 consist of 10 items that favoured examinees who wrote in CR whiles only four of these items exhibited uniform DIF in favour of examinees from other regions.

Table 73 shows the results of DIF analysis using LR procedure for 2012-2016 English language data with GAR as a reference group.

Table 73: Distribution of DIF Items in 2012-2016 English Language Exams Using Location with GAR as the RG

Year	Number of DIF Items Flagged		Type of DIF	
	GAR	FG	Uniform	Non Uniform
2012	34	40	2 (GAR)	72
2013	32	43	1 (GAR)	74
2014	40	44	0	84
2015	27	32	1 (GAR)	58
2016	23	30	3 (GAR) 3 (FG)	47

Table 73 shows that 47 to 84 nonuniform DIF items or 47 to 84% of the items revealed DIF (i.e., 1-6 items exhibited uniform DIF, whereas the 47 to 84 items exhibited non-uniform DIF). Twenty-three to forty items in favour of students who schooled in GAR, whereas 30 to 44 items were in favour of the FG. All items that exhibit DIF in 2014 were all non-uniform DIF. Six items (3 each in favour of both subgroups) were identified as uniform DIF in 2016. Items identified as DIF are indicated by a circle symbol in Appendix A.

Table 74 shows the summary results of the LR method to identify DIF on the mathematics ability scale for each of 50 items.

Table 74: Distribution of DIF Items in 2012-2016 Mathematics Exams Using Gender

Year	Number of DIF Items Flagged		Type of DIF	
	Male	Female	Uniform	Non Uniform
2012	22	20	2 (M)	40
2013	21	20	1 (M)	40
2014	17	19	1 (M) 1 (F)	34
2015	14	19	2 (F)	33
2016	15	22	1 (M) 2 (F)	37

From Table 74, 42 items exhibited DIF consisting of 40 uniform DIF and 2 non-uniform DIF in 2012. The results from Table 78 indicate 22 out 42 items were in favour of males, whereas the 20 items were in favour of females for the 2012 dataset.

In 2013, 21 items were identified as DIF in favour of males whereas 20 items exhibited DIF in favoured females. From 2014 -2016 items that were identified as DIF in favour of the males are 17 and 15 respectively. Items have been indicating in appendix B.

Table 75 shows the summary results of the LR method to identify DIF on the mathematics ability scale for the location with CR as the reference group.

Table 75: Distribution of DIF Items in 2012-2016 Mathematics Exams Using Location with CR as the RG

Year	Number of DIF Items Flagged		Type of DIF	
	CR	FG	Uniform	Non Uniform
2012	18	15	1 (CR)	32
2013	16	17	2 (CR)	31
2014	17	19	1 (FG)	35
2015	13	18	1 (CR)	30
2016	17	13	3 (CR) 2 (FG)	25

From Table 75, results indicated that out of 50 items, 18 items were identified as DIF in favour of examinees from CR in 2012 while in 2013 16 items exhibited DIF in favour of examinees who wrote the examination in CR. In 2016, five uniform DIF items were identified of which three were in favour of candidates who wrote the examination at CR whiles two items exhibited DIF in favour of the FG (GAR, VR, WR and ER).

Results from Table 76 present location-DIF of 2012-2016 Mathematics analysis using LR procedure with GAR as the reference group.

Table 76: Distribution of DIF Items in 2012-2016 Mathematics Exams Using Location with GAR as the RG

Year	Number of DIF Items Flagged		Type of DIF	
	GAR	FG	Uniform	Non Uniform
2012	21	20	0	41
2013	22	17	1 (GAR)	38
2014	18	21	0	39
2015	19	17	1 (GAR)	35
2016	14	16	2 (GAR)	28

Table 76 shows that there were more non-uniform DIF items than uniform DIF items through the five years under study. The non-uniform DIF items decreased as the years go by, thus from 41 to 28 items. The 2013 and 2015 items showed an equal number (17) of DIF items in favour of the focal group. The year 2016 identified two items as uniform DIF and were all in favour of examinees who wrote in GAR. Items that showed DIF has been indicated in Appendix B.

The results in Table 77 show the summary results of the LR method to identify DIF on the mathematics ability scale for each of 50 items across 2012-2016.

Table 77: Distribution of DIF Items in 2012-2016 Integrated Science Exams Using Gender

Year	Number of DIF Items Flagged		Type of DIF	
	Male	Female	Uniform	Non Uniform
2012	11	18	4 (M) 1 (F)	24
2013	21	17	3 (M) 2 (F)	33
2014	16	21	0	37
2015	23	19	0	42
2016	22	16	1 (M)	37

Results from Table77 show that out of the 50 items, 37 exhibited nonuniformed DIF in 2013 while in 2015 42 out of 50 items exhibited non-uniform DIF. The highest number of uniform DIF identified in the dataset were five items in 2012 of which three of them were in favour of males while two were in favour of females. None of the items in 2014, 2015 exhibited uniform DIF.

Table 78 presents results of 2012-2016 Integrated Science dataset using LR procedure for location-DIF analysis.

Table 78: Distribution of DIF Items in 2012-2016 Integrated Science Exams Using Location with CR as the RG

Year	Number of DIF Items Flagged		Type of DIF	
	CR	FG	Uniform	Non Uniform
2012	17	19	0	36
2013	19	16	3 (CR)	32
2014	18	16	1(FG)	33
2015	19	15	1 (CR)	33
2016	11	13	4 (CR) 2(FG)	18

Table 78 indicated that 2016 showed the highest number (6) of uniform DIF whiles 2012 had the highest number (36) of non-uniform DIF. The 2013 and 2015 items exhibit nineteen items DIF each in favour of CR. Uniform DIF identified in 2016 consists of four items uniform DIF in favour of CR examinees and two items uniform DIF in favour of examinees who wrote the examination in GAR, VR, ER and WR.

In general, a high number of non-uniform DIF were identified in all the five years under study in favour of examinees who sat in CR. The results indicate that examinees who sat and wrote the examination in CR were at a disadvantage in answering the examination items as compared with candidates who wrote the examination in GAR, WR, VR and ER.

Table 79 provides results of location-DIF of 2012-2016 Integrated Science analysis using LR with GAR as the reference group.

Table 79: Distribution of DIF Items in 2012-2016 Integrated Science Exams Using Location with GAR as the RG

Year	Number of DIF Items Flagged		Type of DIF	
	GAR	FG	Uniform	Non Uniform

2012	15	14	2(GAR)
			1(FG)
				26

2013	21	17	3(GAR)	35
2014	21	20	1(GAR)	40
2015	20	21	2(GAR)	38

2016	14	16	4(GAR)
			3(FG)
				23

From Table 79, items that showed DIF in favour of the reference group were from 14 to 21 from 2012-2016. 2013 and 2014 had an equal number (21) of DIF items in favour of GAR. The result also indicated that in 2016, seven items exhibited uniform DIF with three in favour of examinees who wrote the exams in CR, WR, VR and ER whiles four items were in favour of examinees who wrote the examination in GAR.

Table 80 shows for the DIF results of Social Studies items from 2012-2016 WASSCE using LR procedure.

Table 80: Distribution of DIF Items in 2012-2016 Social Studies Exams Using Gender

Year	Number of DIF Items Flagged		Type of DIF	
	Male	Female	Uniform	Non Uniform
2012	16	18	2(M)	32
2013	18	23	0	41
2014	20	16	2(M)	34
2015	17	20	2(M)	35
2016	18	17	0	35

The results in Table 80 show that the majority of the items exhibit nonuniform DIF with several items ranging from 32 to 41. In the year 2013 most DIF items in favour of females whereas 2014 showed the most DIF items in favour of males. All items in 2013 and 2016 were revealed as non-uniform DIF. Only two (2) items showed uniform DIF in 2012, 2014 and 2015 and were in favour of males.

Table 81 revealed results on DIF analysis on the location with CR as the reference group for 2012-2016 WASSCE Social Studies using LR.

Table 81: Distribution of DIF Items in 2012-2016 Social Studies Exams Using Location with CR as the RG

Year	Number of DIF Items Flagged		Type of DIF	
	CR	FG	Uniform	Non Uniform
2012	18	21	1(CR)	38

2013	19	14	3(CR)
			3(FG)
				27

2014	20	22	1(CR)	41

2015	19	18	1(CR)
			1(FG)
				35

2016	14	15	3(CR)
			1(FG)
				25

Twenty items showed DIF in favour of CR whiles 22 items revealed in favour of the focal group in 2014. Six items were revealed as uniform DIF, three items in favour of CR. A total of 38 nonuniform DIF items were identified in 2012 and out of these 18 items were in favour of CR as shown in Table 82. There were 19 nonuniform DIF items identified in favour of CR in 2013 and 2015.

Table 82 provides location-DIF analysis results using LR procedure for 2012-2016 WASSCE Social Studies with GAR serving as the reference group.

Table 82: Distribution of DIF Items in 2012-2016 Social Studies Exams Using Location with GAR as the RG

	Number of DIF Items Flagged		Type of DIF	
Year	GAR	FG	Uniform	Non Uniform
2012	17	17	2(GAR)	32
2013	19	22	0	41
2014	19	21	1(GAR)	39
2015	17	17	1(GAR)	33
2016	16	17	2(GAR)	31

Table 82 shows equal (17) nonuniform DIF items in favour of both the reference and focal groups in 2012. An equal number (2) of items were revealed as uniform DIF in favour of the reference group in 2012 and 2016. All forty-one items identified as DIF in 2013 were all non-uniform with 21 of them being favour of examinees who wrote the examination in CR, VR, ER and WR.

Summary

The gender and location-DIF analysis for 2012-2016 WASSCE English Language language, Mathematics, Integrated Science and Social Studies items using LR procedure showed that concerning gender, there were more items identified as DIF in favour of males than females in English Language and Integrated Science whiles in mathematics more items showed DIF in favour of females. In Social Studies there was a balance in the number of DIF items identified.

In terms of location-DIF with CR as the reference group, DIF identified items were more in favour of the focal group in English Language and Mathematics whiles in Integrated Science and Social Studies, there was a balance of several items that exhibited DIF.

Also, location-DIF with GAR as reference had more items exhibiting DIF in favour of the focal group in the English Language while in Mathematics, Integrated Science and Social Studies the items that exhibited DIF were balanced. More items exhibited nonuniform DIF in all subjects than uniform DIF. The items that were identified as uniform DIF were in favour of the reference groups than focal groups in both gender and locationDIF.

Research hypotheses five and six

5. H_0: The 2012-2016 WASSCE core subjects' examinations do not statistically significantly exhibit gender differential item functioning using 3PL IRT model.

   H_1: The 2012-2016 WASSCE core subjects' examinations statistically significantly exhibit gender differential item functioning using 3PL IRT model.

6. H_0: The 2012-2016 WASSCE core subjects' examinations do not statistically significantly exhibit location differential item functioning using 3PL IRT model.

   H_1: The 2012-2016 WASSCE core subjects' examinations statistically significantly exhibit location differential item functioning using 3PL IRT model.

In this study, the interest is in the difference between a- and b-parameters and their effect size. This study did not investigate differences in the parameter *c* as they were not readily interpretable and no accepted criteria to classify the size of the differences were found in the literature.

However, DIF in the 3PL model context in this study used the criteria employed by Santelices and Wilson (2012) to analyse the data. It states that the DIF estimate is obtained using the standardization procedure by finding the differences between all three parameters estimated in the focal and reference groups. That is, $D_a = a_r - a_f$, $Db = b_r - b_f$ *and* $D_c = c_r - c_f$. Santelices and Wilson (2012) established the cutoff scores (effect size) for the difference between parameters *b* (D*b*) as follows:

|DIF| < 0.426: Negligible

0.426 ≤ |DIF| < 0.638: Intermediate

0.638 ≤ |DIF|: Large.

and that of *a (Da)* was classified using the cutoff scores as:

|DIF| < 0.213: Negligible

0.213 ≤ |DIF| < 0.319: Intermediate

0.319 ≤ |DIF|: Large.

According to Freedle (2003), the more difficult items would exhibit larger positive DIF estimates, indicating items that benefit the focal group, and easier items would exhibit small positive or negative DIF estimates, indicating items that benefit the reference group.

Table 83 presents results of number of items that exhibited DIF in 2012-2016 WASSCE in English Language, Mathematics, Integrated Science and Social Studies dataset by 3PL IRT model and their effect sizes (ES) of A, B and C levels. Since an effect size of A level is seen as negligible or not statistically significant, DIF analysis of this study considered items that exhibited DIF at B and C levels as actual DIF.

Table 83: Distribution of DIF Items in 2012-2016 English Language Exams Using Gender

	Number of DIF Items Flagged		Type of DIF	
Year	Male	Female	Uniform	Non Uniform
2012	43	38	1(M)	80

2013	31	37	2(M)
			6(F)
				60

2014	32	27	14(M)
			4(F)
				41

2015	35	22	4(M)	53

2016	28	27	3(M)
			1(F)
				55

It is shown in Table 83 that 2012 exhibited 17 DIF items in favour of males and 82 DIF items in favour of females according to item difficulty parameter (b_i). Thus, easier items exhibit DIF in favour of males whereas harder items showed DIF favouring the females as proposed by Wilson (2010b).

Out of 17 items that exhibited DIF in favour of males, 8 of them showed DIF in favour of males according to the effect size whiles out of the 80 items, 42 were truly showing DIF in favour of females. This information is not different for 2013 but from 2014-2016 items exhibited DIF in favour of males than females according to the corresponding effect sizes.

In terms of discrimination (a_i), 2012 displayed the highest (42) number of DIF items in favour of females whiles in 2014, 10 items exhibited DIF in favour of males.

The estimate of the pseudoguessing parameter for males and females was .034 and .041 respectively, which suggests a modest degree of guessing on the test. The pseudoguessing parameter represents the smallest probability of a correct response during exams. Thus, according to this model, even the least able male and female student has, at minimum, a .1% and 0% for 2012, 3.8% and 4.1% for 2013, 0.4% and 1.3% for 2014, 0.1% and 0.1% for 2015 and 0.3% and 1.7% chance of responding correctly on any given item in 2016.

Table 84 provides results on gender-DIF for 2012-2016 Mathematics using 3PL IRT model. It provides results on the a_i and b_i-parameters for both male and female examinees.

Table 84: Distribution of DIF Items in 2012-2016 Mathematics Exams Using Gender

Year	Number of Items Flagged (a)		Number of Items Flagged (b)		Effect Size of DIF Items (a)			Effect Size of DIF Items (b)		
	M	F	M	F	A	B	C	A	B	C

2012	48	1	2	48	48(M)	0	1(F)	2(M)
								10(F)
									12(F)	26(F)

2013	35	15	18	32	30(M)
					7(F)
						0	5(M)
							8(F)
								18(M)
								32(F)
									0	0

2014	26	23	34	15	14(M)
					12(F)
						0	12(M)
							11(F)
								34(M)
								15(15)
									0	0

2015	24	24	3	45	14(M)
					10(F)
						1(F)	10(M)
							13(F)
								3(M)
								38(F)
									7(F)	0

2016	1	49	46	4	1(M)	0	49(F)	46(M)
								4(F)
									0	0

Results from Table 84 depicts that in 2014, 12 items exhibited DIF in favour of males whiles in 2016, 49 items exhibited DIF in favour of females regarding the difficulty parameter. In 2012, 26 items were indicted as DIF in favour of females based on the discrimination parameter. In 2015,7 items exhibited DIF in favour of females based on the discrimination parameter.

Table 85 present 3PL IRT analysis on 2012-2016 Integrated Science data for gender DIF.

Table 85: Distribution of DIF Items in 2012-2016 Integrated Science Exams Using Gender

Year	Number of Items Flagged (a)		Number of Items Flagged (b)		Effect Size of DIF Items (a)			Effect Size of DIF Items (b)		
	M	F	M	F	A	B	C	A	B	C

2012	9	40	26	24	9 (M)
					3 (F)
						3(F)	34(F)	26(M)
								20(F)
									3(F)	1(F)

2013	28	22	24	26	20(M)
					10(F)
						1(M)	7(M)
							12(F)
								20(M)
								23(F)
									3(M)	1(M)
										3(F)

2014	30	20	8	42	27(M)
					5(F)
						1(M)	2(M)
							15(F)
								0	4(M)
									30(F)
										4(M)
										12(F)

2015	5	45	25	25	5(M)	0	45(F)	25(M)
								22(F)
									3(F)	0

2016	17	33	39	11	10(M)
					9(F)
						0	7(M)
							24(F)
								39(M)
								10(F)
									1(F)	0

From the results of Integrated Science items for 2012-2016, forty items showed DIF in favour of females (i.e., the discrimination parameter-a_i) in 2012 but 34 out of the 40 items discriminated well among female examinees. The data of 2014 identified 20 items exhibiting DIF but 15 of them discriminated among the female examines whiles only 3 out of thirty items identified discriminated well among the male examinees. Besides, in 2014, all items identified to exhibit DIF concerning the difficulty parameter (b_i) were very difficult for both male and female examinees. The female examinees had more items being difficult than male examinees as shown in Table 85. The years, 2014 (0.013) and 2016 (0.013) showed an equal chance of guessing for male students. It indicates the probability that very low ability individuals have 1.3% of getting an item correct by chance. This implies , items that exhibited DIF in terms of the guessing parameter were equal with respect to gender.

Table 86 display the data analysis of 2012-2016 WASSCE Social Studies using 3PL IRT model.

Table 86: Distribution of DIF Items in 2012-2016 Social Studies Exams Using Gender

Year	Number of Items Flagged (a)		Number of Items Flagged (b)		Effect Size of DIF Items (a)			Effect Size of DIF Items (b)		
	M	F	M	F	A	B	C	A	B	C

2012	31	19	0	50	30(M)
					5(5)
						1(M)	14(F)	13(F)	15(F)	22(F)

2013	32	17	9	39	32(M)	0	17(F)	9(M)	37(F)	2(F)

2014	24	26	19	31	24(M)
					2(F)
						1(F)	23(F)	17(M)
								20(F)
									2(M)	11(F)

2015	25	25	30	20	25(M)
					1(F)
						0	24(F)	30(M)
								17(F)
									2(F)	1(F)

2016	26	24	37	13	24(M)
					4(F)
						0	2(M)
							20(F)
								37(M)
								11(F)
									2(F)	0

In Table 86, 2012-2016 Social Studies results for gender show that 39 items exhibit DIF in favour of females in terms of the difficulty parameter whiles in the same year more items (32) also showed DIF in favour of males by the discrimination parameter. According to a 2013 analysis, even the least able male and female student has, at minimum, a 6.1% and 5.2% chance of responding correctly on any given item respectively. According to Table 91, 2014 Social Studies items exhibited 24 and 12 items in favour of females based on the difficulty and discrimination parameters respectively. In 2015, the least able student has virtually no chance of responding correctly to any given item.

Table 87 depicts results of 2012-2016 English Language location-DIF where CR was used as the reference group and other locations as the focal group (FG).

Table 87: Distribution of DIF Items in 2012-2016 English Language Exams Using Location with CR as the RG

Year	Number of Items Flagged (a)		Number of Items Flagged (b)		Effect Size of DIF Items (a))			Effect Size of DIF Items (b)		
	CR	FG	CR	FG	A	B	C	A	B	C

2012	78	22	31	69	78(CR)
					2(FG)
						20(FG)	0	30(CR)
								8(FG)
									1(CR)
									3(FG)
										58(FG)

2013	18	81	26	74	18(CR)	1(FG)	80(FG)	20(CR)
								28(FG)
									6(CR)
									2(FG)
										38(FG)

2014	82	16	30	70	80(CR)
					5(FG)
						2(CR)	11(O)	30(CR)
								8(FG)
									55(FG)	7(FG)

2015	32	46	75	2	30(CR)
					2(FG)
						2(CR)
						2(FG)
							36(FG)	75(CR)	0	2(FG)

2016	72	8	19	71	72(CR)	0	8(FG)	17(CR)
								44(FG)
									2(CR)
									10(FG)
										7(FG)

Table 87 shows that 2012 English Language had 20 items identified as DIF in terms of the difficulty parameter in favour of students who schooled in the four regions (namely, ER, WR, VR and GAR). In terms of the discrimination parameter, 58 (2012), 44 (2013), 62 (2015) exhibited DIF in favour of students who schooled in other regions. The pseudo guessing parameter of this analysis revealed that in 2014, even the least able student who schooled in CR has, at minimum, a 2.7% chance of responding correctly on any given item but no chance at all in 2016.

Table 88 show results of location-DIF for 2012-2016 English Language items where GAR was used as the reference group.

Table 88: Distribution of DIF Items in 2012-2016 English Language Exams Using Location with GAR as the RG

Year	Number of Items Flagged (a)		Number of Items Flagged (b)		Effect Size of DIF Items (a)			Effect Size of DIF Items (b)		
	GAR	FG	GAR	FG	A	B	C	A	B	C

2012	77	21	39	61	77(GAR)
					6(FG)
						2(FG)	13(O)	39(GAR)
								20(FG)
									8(FG)	3(FG)

2013	4	96	60	40	4(GAR)	1(FG)	95(FG)	50(GAR)
								46(O)
									3(FG)	1(FG)

2014	48	52	30	70	0	38(GAR)
						17(O)
							10(GAR)
							35(O)
								30(GAR)
								20(O)
									11(FG)	39(FG)

2015	66	12	32	46	66(GAR)	0	12(FG)	32(GAR)
								39(FG)
									4(FG)	3(FG)

2016	80	0	36	44	80(GAR)	0	0	36(GAR)
								38(FG)
									2(FG)	4(FG)

DIF analysis results for the location for the English language for 2012-2016 as shown in Table 88 indicated in 2012, 15 items showed DIF in favour of FG based on the difficulty parameter. In 2013, 95 items exhibited DIF in favour of FG based on difficulty parameter. In 2013 and 2014, most of the items showed DIF in favour of examinees who wrote the examination in other regions and the items discriminated well among the examinees. In 2012 and 2014, 11 items as against 50 items exhibited DIF in favour of FG concerning the discrimination parameter. The minimum chance a weak student has in responding to any item correctly is 1.8%, 1.3%, 1.2%, 0.8% and 1.5% for GAR and 0.3%, 7.7%, 2.7%,0.0% and 0.0% for CR through 2012 to 2016.

Table 89 presents results of the location-DIF analysis for 2012-2016 Mathematics using CR as a reference group.

Table 89: Distribution of DIF Items in 2012-2016 Mathematics Exams Using Location with CR as the RG

Year	Number of Items Flagged (a)		Number of Items Flagged (b)		Effect Size of DIF Items (a)			Effect Size of DIF Items (b)		
	CR	FG	CR	FG	A	B	C	A	B	C

2012	0	50	43	7	0	0	50(FG)	43(CR)
								6(FG)
									0	1(FG)

2013	23	27	48	2	23(CR)
					2(FG)
						0	25(FG)	48(CR)
								2(FG)
									0	0

2014	23	26	13	36	23(CR)
					2(FG)
						0	24(FG)	13(CR)
								29(FG)
									7(FG)	0

2015	39	11	33	17	39(CR)
					2(FG)
						1(FG)	8(FG)	50	0	0

2016	47	3	30	20	47(FG)
					1(CR)
						0	2(CR)	30(CR)	8(FG)	12(FG)

Data analysis for Mathematics using CR as a reference group indicated that in terms of the a_i-parameter all the 50 items exhibit DIF in favour of the examinees who wrote in other four regions in 2012 while in 2016, 94% of the items exhibit DIF in favour of the reference group even though the DIF was negligible as shown in Table 89. This means items did not discriminate well among examinees in 2016 but did good discrimination among examines from the ER, WR, GAR and VR in 2012. The b_i-parameter indicated that 98% and 66% of the items showed DIF in favour of the reference group in 2013 and 2015 respectively. The items were easy for reference group because the discrimination level was at A. For 2012, 2013, 2014 and 2015 items, the probability for even the least able student who schooled in CR to get the item correct was virtually 0% but 2.2% for 2016.

The results of 2012-2016 mathematics using GAR as a reference group for the location DIF analysis is shown in Table 90.

Table 90: Distribution of DIF Items in 2012-2016 Mathematics Exams Using Location with GAR as the RG

Year	Number of Items Flagged (a)		Number of Items Flagged (b)		Effect Size of DIF Items (a)			Effect Size of DIF Items (b)		
	GAR	FG	GAR	FG	A	B	C	A	B	C

2012	4	46	2	48	4(GAR)	0	46(FG)	40(FG)	7(FG)	2(GAR)
										1(FG)

2013	12	38	21	29	12(GAR)
					2(FG)
						1(FG)	35(FG)	21(GAR)
								29(FG)
									0	0

2014	37	13	19	31	37(GAR)	1(FG)	12(FG)	48	1	1

2015	38	12	28	22	38(GAR)
					3(CR)
						0	9(FG)	28(GAR)
								22(FG)
									0	0

2016	7	43	43	7	7(GAR)
					1(FG)
						1(FG)	41(FG)	43(GAR)
								7(FG)
									0	0

From Table 90, items exhibit 80-100% of DIF throughout the five years of examination under review concerning the difficulty index. This means the items were not difficult for examinees hence the items did not discriminate well among examinees. In 2014, the least able student has, at minimum, a 90.5% chance of responding correctly on any given item if schooled in GAR.

Table 91 presents the results of the 2012-2016 Integrated Science location DIF with CR as a reference group.

Table 91: Distribution of DIF Items in 2012-2016 Integrated Science Exams Using Location with CR as the RG

Year	Number of Items Flagged (a)		Number of Items Flagged (b)		Effect Size of DIF Items (a)			Effect Size of DIF Items (b)		
	CR	FG	CR	FG	A	B	C	A	B	C

2012	7	43	15	35	7(CR)	0	43(FG)	35(FG)
								7(CR)
									6(CR)	2(CR)

2013	36	14	28	22	36(CR)
					4(FG)
						1(FG)	9(FG)	28(CR)
								15(FG)
									2(FG)	5(FG)

2014	20	30	49	1	20(CR)	0	30(FG)	49(CR)	0	1(FG)

2015	38	12	41	9	38(CR)
					2(FG)
						0	10(FG)	41(CR)
								8(FG)
									0	1(FG)

2016	39	11	50	0	39(CR)
					1(FG)
						0	10(FG)	50(CR)	0	0

Results in Table 91 indicate that 43, 10, 30 items showed DIF in favour of CR in terms of the a_i-parameter in 2012, 2013 and 2014 respectively From Table 91, results indicate that the items were not difficulty for examining as well. In terms of the c_i-parameter, the highest chance for the least student who schooled in CR to get any item correct was 4.5% in 2013 and that of the focal group is 2.9% in 2015.

Table 92 indicates the results of 2012-2016 Integrated Science location-DIF using 3PL IRT model with GAR as the reference group.

Table 92: Distribution of DIF Items in 2012-2016 Integrated Science Exams Using Location with GAR as the RG

Year	Number of Items Flagged (a)		Number of Items Flagged (b)		Effect Size of DIF Items (a)			Effect Size of DIF Items (b)		
	GAR	FG	GAR	FG	A	B	C	A	B	C

2012	48	2	11	39	48(GAR)
					1(FG)
						0	1(FG)	11(GAR)	0	39(FG)

2013	14	36	22	28	14(GAR)
					6(FG)
						30(FG)	0	22(GAR)
								20(FG)
									2(FG)	6(FG)

2014	19	31	50	0	19(GAR)
					1(FG)
						2(FG)	28(FG)	50(GAR)	0	0

2015	33	17	50	0	0	33(GAR)
						1(FG)
							16(FG)	50(GAR)	0	0

2016	40	10	49	1	0	40(GAR)
						1(FG)
							9(FG)	49(GAR)
								1(FG)
									0	0

The 2012-2016 DIF analysis for the location with GAR as a reference group for Integrated Science test items in Table 92 revealed that, 30 items each for 2013, 2014 and 49 each for 2015 and 2016 exhibited DIF in favour of OTHERS in terms of difficulty parameter. Notwithstanding the items for difficulty parameter, 39 items exhibited DIF in favour of FG in 2012. According to this analysis, even the least able student has, at minimum, a .9% chance of responding correctly on any given item in 2012 and 2016 if he/she schooled in GAR.

Table 93 displays location-DIF of 2012-2016 Social Studies data analysis using 3PL IRT model with CR as the reference group.

Table 93: Distribution of DIF Items in 2012-2016 Social Studies Exams Using Location with CR as the RG

Year	Number of Items Flagged (a)		Number of Items Flagged (b)		Effect Size of DIF Items (a)			Effect Size of DIF Items (b)		
	CR	FG	CR	FG	A	B	C	A	B	C

2012	37	13	38	12	37(CR)
					1(FG)
						1(FG)	11(FG)	38(CR)
								12(FG)
									0	0

2013	34	14	40	10	34(CR)
					4(FG)
						0	10(O)	40(CR)
								10(FG)
									0	0

2014	19	31	19	31	19(CR)	0	31(O)	19(CR)	0	31(FG)

2015	10	40	38	12	10(CR)
					5(FG)
						0	35(0)	38(CR)
								12(FG)
									0	0

2016	13	37	46	4	13(CR)	0	37(O)	46(CR)
								4(FG)
									0	0

Table 93 shows the results DIF analysis on Social Studies items using CR as the reference group. Results showed that items that exhibit DIF in favour of students who schooled in CR increased as the years go by from 11 to 37, in terms of difficulty parameter. It is also clear in Table 93 that items did not exhibit DIF in terms of discrimination parameter for all years under study except in 2014 where 31 items were identified as showing DIF in favour of students who schooled in other regions. In terms of the guessing parameter, even the least able student has at minimum 8.4% chance of responding any item correct if he/she schooled in CR.

Table 94 demonstrates location-DIF on 2010-2016 Social Studies items using GAR as a reference group using 3PL IRT model.

Table 94: Distribution of DIF Items in 2012-2016 Social Studies Exams Using Location with CR as the RG

Year	Number of Items Flagged (a)		Number of Items Flagged (b)		Effect Size of DIF Items (a)			Effect Size of DIF Items (b)		
	GAR	FG	GAR	FG	A	B	C	A	B	C

2012	38	12	36	14	38(GAR)	0	12(FG)	36(GAR)
								14(FG)
									0	0

2013	25	25	8	42	25(GAR)
					2(FG)
						1(FG)	22(FG)	8(GAR)
								29(FG)
									8(FG)	5(FG)

2014	29	21	11	39	29(GAR)
					1(FG)
						1(FG)	19(FG)	10(GAR)
								18(FG)
									1(GAR)
									5(FG)
										16(FG)

2015	3	47	8	42	5(FG)	3(GAR)	42(FG)	8(GAR)
								34(FG)
									3(FG)	1(FG)

2016	19	31	32	18	19(GAR)
					2(FG)
						1(FG)	28(FG)	32(GAR)
								17(FG)
									1(FG)	0

Table 94 indicates that 42 test items exhibited DIF in favour of the focal group in 2015 in terms of the difficulty parameter. The guessing parameter indicates that the least student who schooled in either of the five regions has approximately 0.3 chance of responding to an item correctly.

Summary

The gender and location-DIF of 2012-2016 WASSCE in English Language, Mathematics, Integrated Science and Social Studies using 3PL IRT model provided results on discrimination index (a_i), difficulty index(b_i) and guessing parameter(c_i) of the test items. English Language, Mathematics and

Integrated Science items generally showed large DIF in favour of males concerning the discrimination parameter. There was a moderate number of items that exhibited DIF in terms of difficulty (b-parameter) in favour of males. Social Studies items exhibited about equal items as DIF based on both the discrimination and difficulty parameters.

For location-DIF analysis results, the 2012-2016 WASSCE English Language, Mathematics, Integrated Science and Social Studies items exhibited an almost equal number of DIF in terms of discrimination and difficulty parameters across the five years under study.

Research question

What is the level of agreement among the MH, LR and 3PL IRT DIF detection methods?

In this study, three different DIF procedures were used comparatively to ascertain their effectiveness and sensitivity to detect DIF for gender and location of schools in the test items in the context of the core subject multiple-choice test items (Paper 1) administered to Senior High School students who sat for the 2012 to 2016 WASSCE.

The three DIF procedures involve subgroups referring to the reference group and the focal group. The comparison focused on the performance on test items of the reference group (males in the gender-based DIF; CR and GAR in the location-based DIF) and focal group (females in gender-based DIF; FG {GAR, VR, ER, WR} or FG {CR, VR, ER, WR} in location-based DIF).

The detection rates of the three procedures are presented according to subjects for gender and location-based DIF in Tables 99 to 109. Figures 39-50 also depicts the level of agreement among the DIF detecting methods for 2012-

2016 English Language, Mathematics, Integrated Science and Social Studies for both gender and location DIF.

Table 95 and Figure 35 present level of agreement among the DIF detecting methods used for 2012-2016 English Language.

Table 95: Distribution of Number of Gender DIF Items Using MH, LR and 3PL IRT Detecting Methods for 2012-2016 English Language with Females as Reference Group

Year	MH	LR	IRT	
			a-parameter	b-parameter
2012	90	76	88	85
2013	91	70	89	85
2014	78	71	87	85
2015	71	55	64	62
2016	66	53	68	64

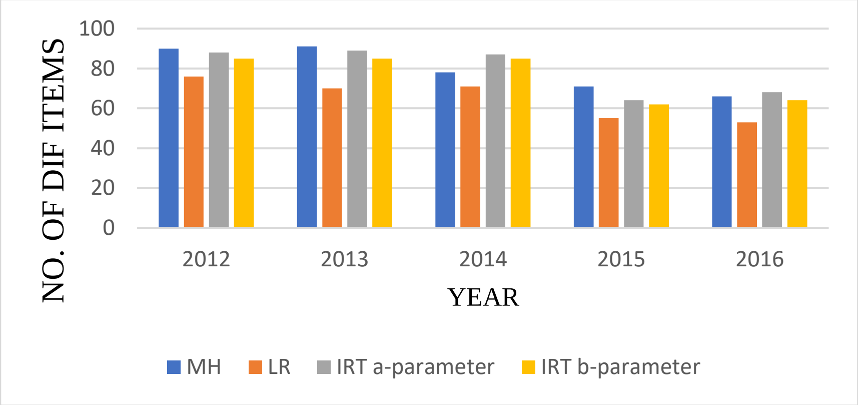

Figure 35: Level of agreement among MH, LR and 3PL IRT DIF detecting methods for 2012-2016 English Language.

Results from Table 95 and Figure 35 indicate the level of agreement among all the three DIF detecting methods in detecting gender DIF for 2012-2016 English Language. MH procedure detected the highest number of DIF items in 2012-2013 . MH detected more DIF items than LR whiles the 3PL IRT detected the highest number of DIF items in 2015-2016 than LR. The a-parameter of the IRT procedure detected more items than the b-parameter.

Table 96 and Figure 36 present level of agreement among the DIF detecting methods used for 2012-2016 Mathematics for gender.

Table 96: Distribution of Number of Gender DIF Items Using MH, LR and 3PL IRT Detecting Methods for 2012-2016 Mathematics with Females as Reference Group

Year	MH	LR	IRT	
			a-parameter	b-parameter
2012	45	40	43	42
2013	46	37	44	42
2014	42	34	43	41
2015	46	31	42	40
2016	43	36	44	42

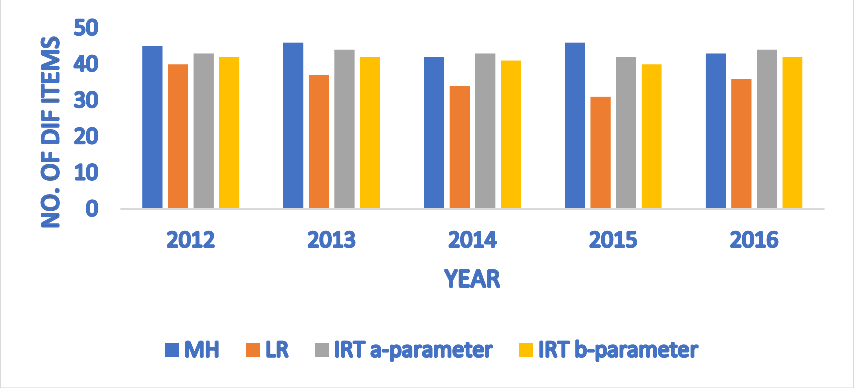

Figure 36: Level of agreement among MH, LR and 3PL DIF detecting methods for 2012-2016 Mathematics.

Results from Table 96 and Figure 36 indicate the level of agreement among all the three DIF detecting methods in detecting gender DIF for 2012-2016 Mathematics. LR procedure detected the least number of DIF items. MH detected a few more item as DIF as compared to LR whiles the 3PL IRT detected the highest number of DIF items. The a-parameter of the IRT procedure detected more items than the b-parameter.

Table 97 and Figure 37 present level of agreement among the DIF detecting methods used for 2012-2016 Integrated Science based on gender.

Table 97: Distribution of Number of Gender DIF Items Using MH, LR and 3PL IRT Detecting Methods for 2012-2016 Integrated Science with Females as Reference Group

Year	MH	LR	IRT	
			a-parameter	b-parameter
2012	42	34	44	43
2013	45	43	44	43
2014	47	37	46	43
2015	44	42	46	43
2016	43	39	46	43

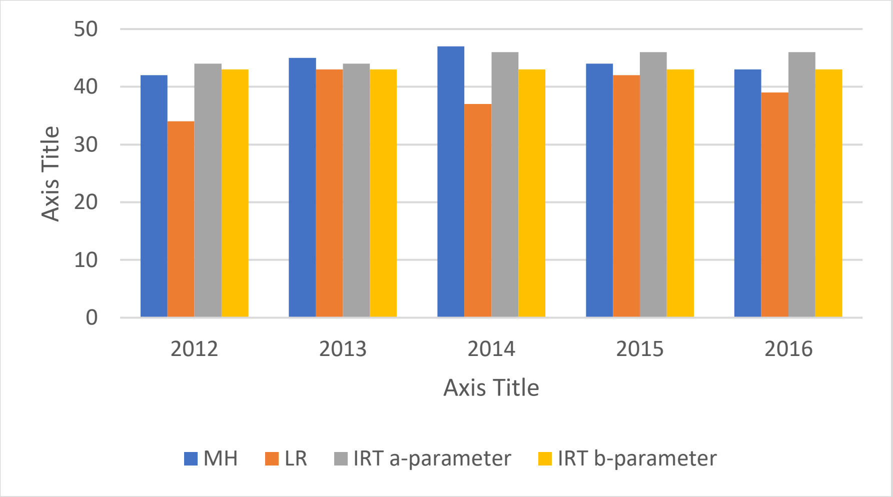

Figure 37: Level of agreement among MH, LR and 3PL IRT DIF detecting methods for Integrated Science.

Results from Table 97 and Figure 37 indicate the level of agreement among all the three DIF detecting methods in detecting gender DIF for 2012-

2016 Integrated Science. LR procedure detected the least number of DIF items. MH detected a few more DIF items than LR whiles the 3PL IRT detected the highest number of DIF items. The a-parameter of the IRT procedure detected more items than the b-parameter.

Table 98 and Figure 38 present level of agreement among the DIF detecting methods used for 2012-2016 Social Studies.

Table 98: Distribution of Number of Gender DIF Items Using MH, LR and 3PL IRT Detecting Methods for 2012-2016 Social Studies with Females as Reference Group.

Year	MH	LR	IRT	
			a-parameter	b-parameter
2012	42	36	44	45
2013	45	41	43	43
2014	47	38	44	45
2015	44	39	44	45
2016	43	35	44	45

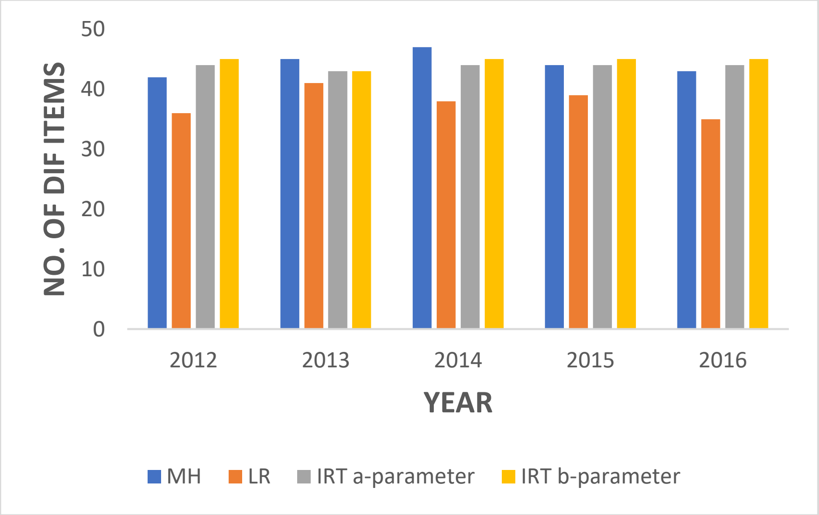

Figure 38: Level of agreement among MH, LR and 3PL IRT DIF detecting methods for Social Studies.

Results from Table 98 and Figure 38 indicates the level of agreement among all the three DIF detecting methods in detecting gender DIF for 2012-2016 Social Studies. LR procedure detected the least number of DIF items. MH detected a few more DIF items than LR whiles the 3PL IRT detected the highest number of DIF items. The b-parameter of the IRT procedure detected more items than the a-parameter.

Table 99 and Figure 39 present level of agreement among the DIF detecting methods used for 2012-2016 English Language for the location using CR as the RG.

Table 99: Distribution of Number of Location DIF Items Using MH, LR and 3PL IRT Detecting Methods for 2012-2016 English Language with CR as Reference Group

Year	MH	LR	IRT	
			a-parameter	b-parameter
2012	88	61	89	90
2013	92	66	88	90
2014	76	57	87	90
2015	70	48	67	67
2016	72	45	69	70

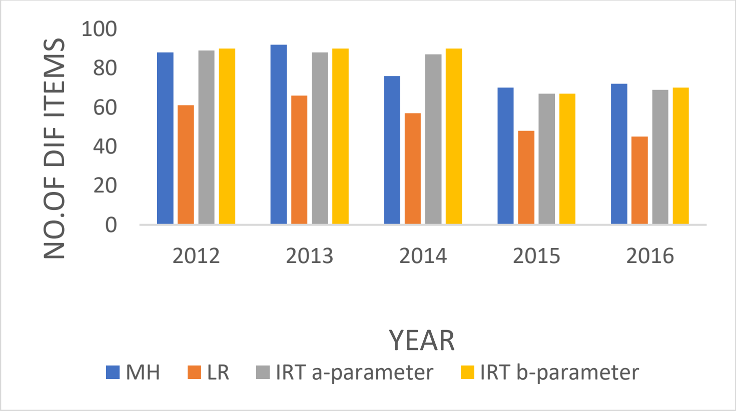

Figure 39: Level of agreement among MH, LR and 3PL IRT DIF detecting methods for 2012-2016 English Language with CR as RG.

Results from Table 99 and Figure 39 indicate a high level of agreement among all the three DIF detecting methods in detecting Location (CR as the reference group) DIF for 2012-2016 English Language. LR procedure detected the least number of DIF items. MH detected a few more DIF items than LR whiles the 3PL IRT detected the highest number of DIF items. The b-parameter of the IRT procedure detected more items than the a-parameter.

Table 100 and Figure 40 present level of agreement among the DIF detecting methods used for 2012-2016 English Language.

Table 100: Distribution of Number of Location DIF Items Using MH, LR and 3PL IRT Detecting Methods for 2012-2016 English Language with GAR as Reference Group

Year	MH	LR	IRT	
			a-parameter	b-parameter
2012	91	72	88	86
2013	90	72	91	86
2014	70	80	91	86
2015	65	56	69	64
2016	68	55	71	66

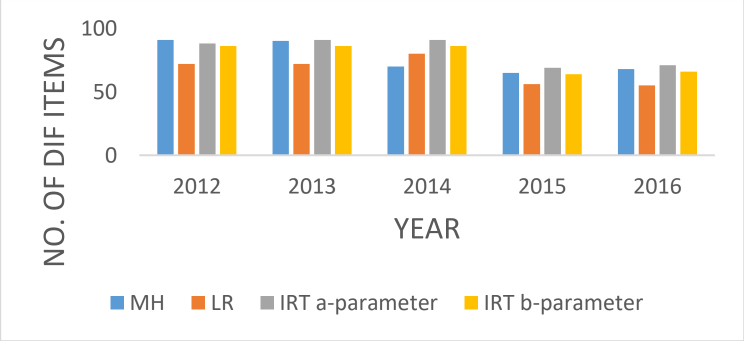

Figure 40: Level of agreement among MH, LR and 3PL IRT DIF detecting methods for the English Language with GAR as RG.

Results from Table 100 and Figure 40 indicate the level of agreement among all the three DIF detecting methods in detecting location DIF for 2012-

2016 English Language. LR procedure detected the least number of DIF itemsin all years except 2014. MH detected a few more DIF items than LR whiles the 3PL IRT detected the highest number of DIF items. The a-parameter of the IRT procedure detected more items than the b-parameter.

Table 101 and Figure 41 present level of agreement among the DIF detecting methods used for 2012-2016 Mathematics.

Table 101: Distribution of Number of Location DIF Items Using MH, LR and 3PL IRT Detecting Methods for 2012-2016 Mathematics with CR as Reference Group

Year	MH	LR	IRT	
			a-parameter	b-parameter
2012	39	30	44	41
2013	46	31	44	41
2014	40	33	43	41
2015	47	28	44	41
2016	45	31	44	41

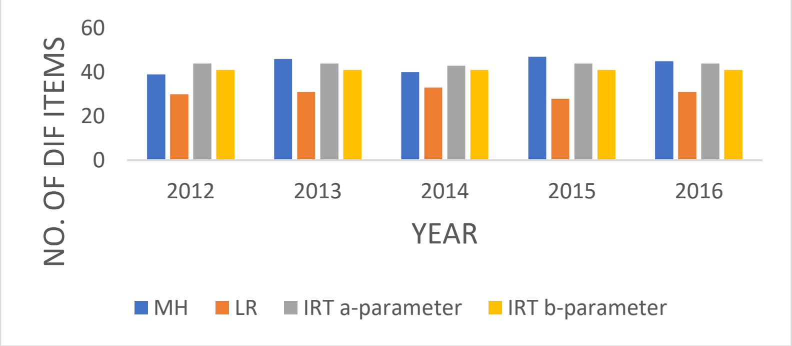

Figure 41: Level of agreement among MH, LR and 3PL IRT DIF detecting methods for 2012-2016 Mathematics with CR as RG.

Results from Table 101 and Figure 41 indicate the level of agreement among all the three DIF detecting methods in detecting location DIF for 2012-2016 Mathematics. LR procedure detected the least number of DIF items. MH detected a few more DIF items than LR whiles the 3PL IRT detected the highest number of DIF items. The a-parameter of the IRT procedure detected more items than the b-parameter.

Table 102 and Figure 42 present level of agreement among the DIF detecting methods used for 2012-2016 Mathematics.

Table 102: Distribution of Number of DIF Items Using MH, LR and 3PL IRT Detecting Methods for 2012-2016 Mathematics with GAR as Reference Group

Year	MH	LR	IRT	
			a-parameter	b-parameter
2012	43	40	46	40
2013	46	39	46	40
2014	45	38	45	39
2015	45	36	46	40
2016	42	31	46	40

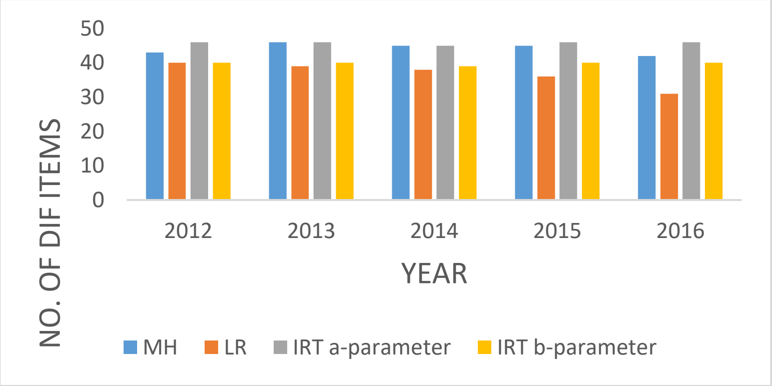

Figure 42: Level of agreement among MH, LR and 3PL IRT detecting methods for 2012-2016 Mathematics using GAR as RG.

Results from Table 102 and Figure 42 indicate the level of agreement among all the three DIF detecting methods in detecting location DIF for 2012-2016 Mathematics. LR procedure detected the least number of DIF items. MH detected a few more DIF items than LR whiles the 3PL IRT detected the highest number of DIF items. The a-parameter of the IRT procedure detected more items than the b-parameter.

Table 103 and Figure 43 present level of agreement among the DIF detecting methods used for 2012-2016 Integrated Science for the location with CR as RG.

Table 103: Distribution of Number of Location DIF Items Using MH, LR and 3PL IRT Detecting Methods for 2012-2016 Integrated Science with CR as Reference Group

Year	MH	LR	IRT	
			a-parameter	b-parameter
2012	39	36	44	41
2013	48	38	44	41
2014	47	35	44	41
2015	46	35	44	41
2016	46	30	44	41

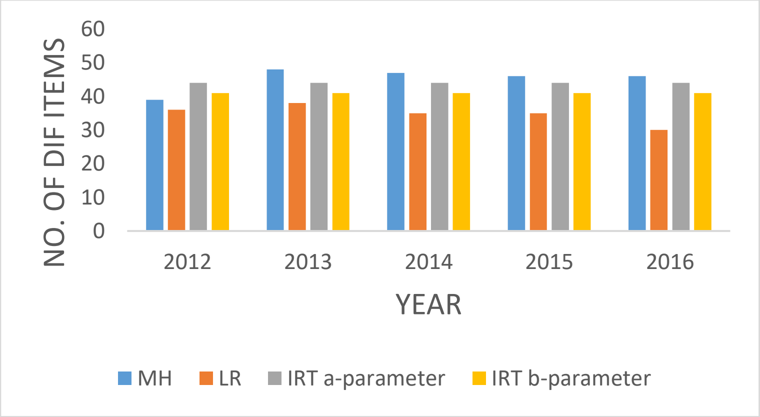

Figure 43: Level of agreement among MH, LR and 3PL IRT DIF detecting methods for 2012-2016 Integrated Science using CR as RG.

Results from Table 103 and Figure 43 indicate the level of agreement among all the three DIF detecting methods in detecting location DIF for 2012-2016 Integrated Science. LR procedure detected the least number of DIF items. MH detected a few more DIF items than LR whiles the 3PL IRT detected the highest number of DIF items. The a-parameter of the IRT procedure detected more items than the b-parameter.

Table 104 and Figure 44 present level of agreement among the DIF detecting methods used for 2012-2016 Integrated Science (Region-GAR).

Table 104: Distribution of Number of Location DIF Items Using MH, LR and 3PL IRT Detecting Methods for 2012-2016 Integrated Science with GAR as Reference Group

Year	MH	LR	IRT	
			a-parameter	b-parameter
2012	43	32	46	42
2013	42	41	46	42
2014	46	42	46	42
2015	46	42	46	42
2016	45	37	46	42

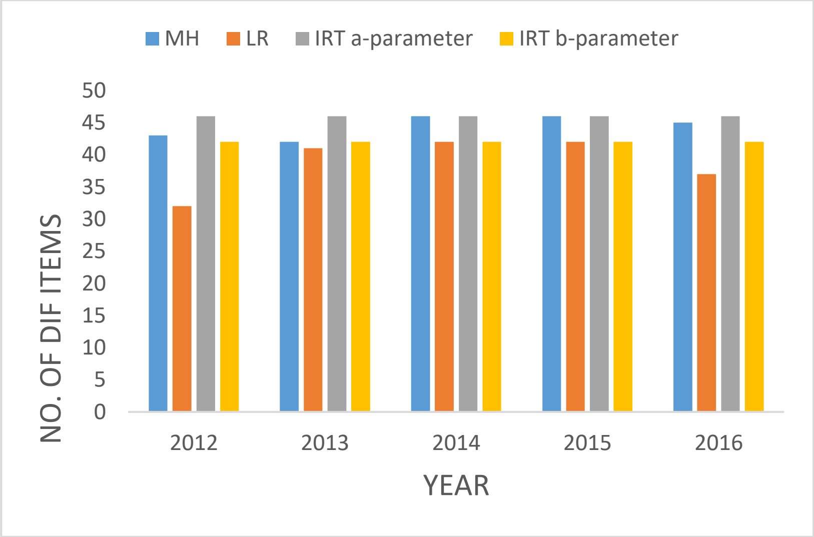

Figure 44: Level of agreement among MH, LR and 3PL IRT DIF detecting methods for 2012-2016 Integrated Science using GAR as RG.

Results from Table 104 and Figure 44 indicate the level of agreement among all the three DIF detecting methods in detecting location DIF for 2012-2016 Integrated Science using GAR as RG. LR procedure detected the least number of DIF items. MH detected a few more DIF items than LR whiles the 3PL IRT detected the highest number of DIF items. The a-parameter of the IRT procedure detected more items than the b-parameter.

Table 105 and Figure 45 present level of agreement among the DIF detecting methods used for 2012-2016 Social Studies using CR as RG.

Table 105: Distribution of Number of Location DIF Items Using MH, LR and 3PL IRT Detecting Methods for 2012-2016 Social Studies with CR as Reference Group

Year	MH	LR	IRT	
			a-parameter	b-parameter
2012	35	40	45	41
2013	41	39	45	41
2014	44	42	45	41
2015	42	39	45	41
2016	43	33	45	41

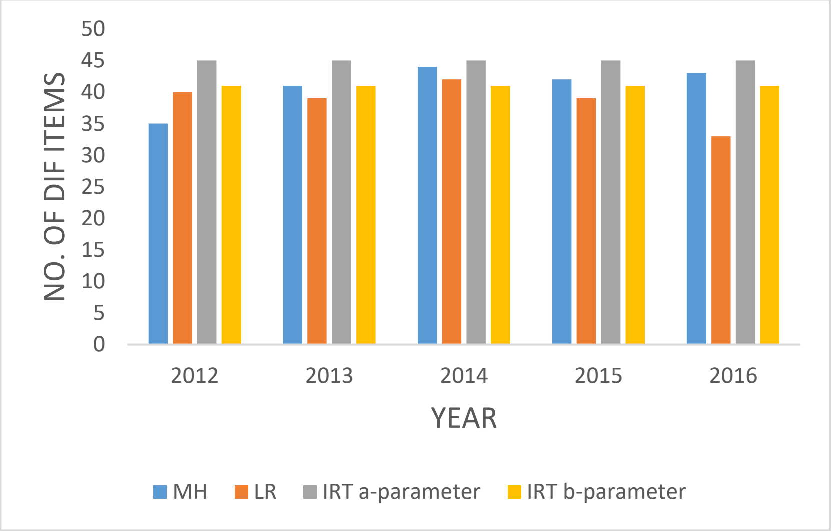

Figure 45: Level of agreement among MH, LR and 3PL IRT DIF detecting methods for 2012-2016 Social Studies using CR as RG.

Results from Table 105 and Figure 45 indicate the level of agreement among all the three DIF detecting methods in detecting location DIF for 2012-

2016 Social Studies using CR as RG. LR procedure detected the least number of DIF items. MH detected a few more DIF items than LR whiles the 3PL IRT detected the highest number of DIF items. The a-parameter of the IRT procedure detected more items than the b-parameter.

Table 106 and Figure 46 present level of agreement among the DIF detecting methods used for 2012-2016 Social Studies (Region-GAR).

Table 106: Distribution of Number of Location DIF Items Using MH, LR and 3PL IRT Detecting Methods for 2012-2016 Social Studies with GAR as Reference Group

Year	MH	LR	IRT	
			a-parameter	b-parameter
2012	36	35	45	42
2013	44	40	45	42
2014	38	40	45	42
2015	44	34	45	42
2016	45	34	45	42

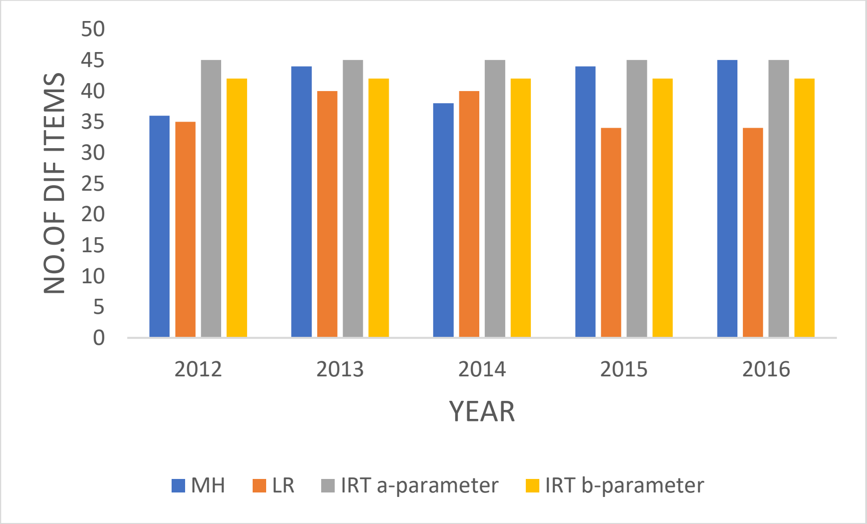

Figure 46: Level of agreement among MH, LR and 3PL IRT DIF detecting methods for 2012-2016 Social Studies using GAR as RG.

Results from Table 106 and Figure 46 indicate the level of agreement among all the three DIF detecting methods in detecting Gender DIF for 2012-2016 Integrated Science. LR procedure detected the least number of DIF items. MH detected a few more DIF items than LR whiles the 3PL IRT detected the highest number of DIF items. The a-parameter of the IRT procedure detected more items than the b-parameter.

Discussion

The purpose of this study was to examine DIF among the WASSCE English language, Mathematics, Integrated Science and Social Studies (Core Subjects) multiple-choice test items for 2012-2016 and the consistency among the three frequently used DIF detection procedures namely MH, LR and IRT.

Overall, the percentage of agreement between the two approaches (MH and LR) in detecting DIF is relatively high as compared to IRT. However, this may be because both methods are related to the classical theory of measurement. This finding seems to be consistent with the previous studies (e.g., Hambleton & Rogers, 1989; Baghi & Ferrara, 1989; Skaggs & Lists, 1992; Hakim & Cohen, 1995; Stage, 2000). The different procedures provided consistent estimates on the magnitude and direction of DIF and thus supports the recommendation that multiple DIF detection procedures should be used in real testing situations to reduce the uncertainty.

This confirms several studies done by other researchers (e.g., Rogers & Swaminathan, 1993; Narayanan & Swaminathan, 1996). The results also showed that the 3PL IRT identified more items as DIF as compared to MH and LR and that supports literature which indicates that IRT is sensitive in detecting

DIF. (Edelen & Reeve, 2009). On the other hand, LR detected the least number of items exhibiting DIF as compared to MH and IRT.

The outcomes of this study show that IRT identified more items displaying DIF than MH and LR. This sensitive nature of IRT perhaps maybe because, in the IRT model, an item shows DIF if people from different subgroups but at the same level on the underlying construct measured have unequal probabilities of responding symptomatically to a particular item (Teresi, Ramirez, Lai, & Silver, 2008. That is IRT, LR and MH have different assumptions and variance of measurement errors (Gruijter & Kamp, 2008). They also have different focuses. MH focuses on the test and is sample dependent (Hambleton, 1991). LR focuses on modelling the probability of answering an item correctly and a conditioning variable usually the observed total test score (Camilli & Shepherd, 1994). IRT focuses on the item by modelling the response of an examinee of given ability to each item in the test (Baker, 2004; DeMars, 2010; Embretson & Reise, 2000).

From the results of this study, the Mantel Haenszel had the greatest advantages (detected the least DIF items but at C -level of ETS criteria) as a DIF detection technique because it provided the best results (Navas-Ara & Gómez-Benito, 2002), and was conceptually uncomplicated, and did not require highly specialized software (Camili & Shepard, 1994).

Overall, this study indicated that the LR procedure provided as good or better uniform DIF detection than MH (Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990). Because it detected more DIF items than MH and at C-level uniform DIF items (Mazor, Kanjee, & Clauser, 1995).

Performances of the comparison groups are also different under the three DIF procedures. In MH and LR there is an almost equal set of items favouring either the focal group or the reference group. The LR detected both uniform and non-uniform DIF with most of the items showing non-uniform DIF. IRT-3PL detected DIF in terms of discrimination, and difficulty level for both gender and region. There are marked differences between the items disadvantaging either group.

Differential Item Functioning Across Gender

The DIF analysis between male and female students showed more items identified as DIF in favour of female students across the years and in subjects in MH analysis whereas in the LR analysis, more English Language items exhibit DIF in favour of males. The MH analysis demonstrated its efficiency when it indicated that most of the items tested for gender DIF and flagged large DIF (classified as Category C).

The LR analysis identified more of the Mathematics and Social Studies DIF items in favour of females while an equal number of Integrated Science items significantly functioned differently between male and female students across the years. The two procedures of MH and LR complimented each other by reflecting similar findings in English Language, Integrated Science and Social Studies but 3PL IRT reflected similar DIF items but at A-level of ETS.

The DIF indices of this study point to the conclusion that females had an advantage over males in English Language and Integrated Science whereas males had an advantage in Mathematics items. Consequently, the number of items that function significantly different for male and female students is not significantly different from the number that did not differentially function in

2012-2016 WASSCE Social Studies. This implies that WASSCE 2012-2016 core subjects' multiple-choice test items functioned differentially for male and female students. The finding of this study is in line with a research study by Abedalaziz (2010) who reported an incidence of gender DIF in mathematics. Also, Odili (2003) revealed that there was evidence of gender DIF in WAEC/SSCE Biology paper 2 for 1999, 2000 and 2001 where females performed better than males. The tendency for males to perform better than females in mathematics is consistent with previous findings (e.g., Willson, Fernandez and Hadaway, 1993; Gallagher, DeLisi, Holst, McGillicuddy-DeLisi, Morely and Cahalan, 2000)

Geary (1996) found that male students were superior in geometry and visualization. On the other hand, female students were superior in computation-based on the data. Gender differences in achievement in mathematics in favour of boys have been found in standardized tests and are most prominent at the very high levels of achievement (Leder, 1992). These differences are likely to be both content and ability dependent.

While males outperform females in scientific and mathematical tasks, females outperform males in tasks involving verbal abilities (Benbow, & Stanley, 1980; Becker, 1990). From the findings of this study males perform well than females in mathematics (Özdemir, 2015) while females perform better than males in English Language (Ahmadi & Jalili, 2014)) and Integrated Science. Social studies did not significantly function differently among male and female examinees in 2012-2016 WASSCE. Items were able to discriminate among examinees on 2012-2016 WASSCE core subjects, with most of the items exhibiting negligible to moderate difficulty indices (Osadebe & Agbure, 2018).).

Differential Item Functioning Across Regions

The findings of this study show that test items for 2012-2016 WASSCE Core subjects significantly functioned differently among students from schools in focal and reference groups.

In testing the region-DIF of the study, the findings revealed that the 2012-2016 WASSCE core subjects' examinations significantly exhibited location differential item functioning. This study is not in agreement with the findings of Inyang 1991, Umoinyang (1991), Eng and Hoe (2010), Amuche and Fan (2014), Mokabi and Adedoyin, (2014) who have reported on the existence of differential item functioning based on location. These findings of this study align with the result of the study carried out by Odili (2003) whose result agreed with Umoinyang (1991) who analysed Mathematics multiple-choice test used by West African Examination Council (WAEC) in the 1990 General Certificate Examination (GCE). Odili (2003) results revealed 29 items that differentially function in favour of candidates from educationally advantaged setting.

This study agrees with the findings of Inyang (2004) who reported that rural students performed better than their urban counterparts. The reason for CR or GAR students to out-perform other students could be due to their interpersonal ties with their community which provides a conducive learning environment. It is assumed that every individual one way or the other is influenced by the community he/she lives in. Another reason could be that the CR or GAR students had adequate coverage of their syllabus in those areas that the items were set. Based on the findings, it was concluded that the 2012-2016 WASSCE core subjects' examinations significantly exhibited location-DIF.

There were more locational differential item functioning items detected by LR and MH analysis at B and C levels as compared to IRT analysis DIF in this study. The results of this study in consistent with Siamisang and Nenty's (2012) study reported significant DIF items in Mathematics and Integrated Science among students from different geographical locations from Botswana, Singapore and the USA who participated in the 2007 TIMSS examinations. The findings of the study showed that mathematics and Integrated Science items significantly functioned differently among students from Botswana, Singapore and the USA who participated in the 2007 TIMSS examinations for both mathematics and Integrated Science tests.

The SSX^2 analysis showed that Singapore students and students from the USA had more test items favouring their students in both mathematics and Integrated Science as compared to Botswana who had few items favouring their students in both mathematics and Integrated Science. Mathematics item showed DIF in favour of USA whiles Integrated Science items exhibited DIF in favour of Singapore.

It was interesting to note that the findings of this current study showed that all the sampled regions have items that either favoured their students or did not favour their students and this result confirms the study by Nenty (2012). As Ndifon, Umoinyang, and Idiku (n.d) reported in their study that the school location of examinees can results in DIF items, seems to be consistent to this study, since English Language and Mathematics items exhibited more DIF in favour of Central and Greater Accra regions as compared to Integrated Science and Social Studies. The findings of this study establish the fact that geographical location can cause DIF especially when examinees come from educationally less endowed environment.

Comparison of MH, LR and 3PL IRT in detecting DIF

In general, these three main DIF analysis methods namely the Mantel-Haenszel (MH), logistic regression (LR) and item response theory (IRT) were used to detect uniform DIF for dichotomous items. Results indicate that the LR procedure is effective for detecting non-uniform DIF as compared to MH and confirms that of Acar, and Kelecioglu. (2010) results. The 3PL IRT is very sensitive in identifying DIF than MH and LR. The current study findings connote the finding of Baghi and Ferrara, (1990) on the fact that IRT detects more DIF items as compared to LR and MH. It was also shown in this study that it is shown that the logistic regression procedure is more powerful than the Mantel-Haenszel procedure for detecting nonuniform DIF and as powerful in detecting uniform DIF in confirmation to the findings of Güler and Penfield (2009).

CHAPTER FIVE

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

Summary

The purpose of this study was to examine gender and regional differential item functioning of 2012-2016 WASSCE in English language, Mathematics, Integrated Science and Social Studies (core subjects) using MH, LR and 3PL IRT DIFF detecting methods.

Research hypothesis one examined gender DIF in 2012-2016 WASSCE core subjects using MH. Results showed DIF in favour of females in English Language and DIF in favour of males in Mathematics and Integrated Science. In Social Studies, there was no statistically significant gender DIF.

Research hypothesis two showed statistically significant more location DIF in favour of examinees who schooled in the CR and GAR as compared to OTHERS in English Language, Mathematics, Integrated Science and Social Studies using MH DIF detecting method.

Research hypotheses three and four sought to examine gender and location DIF using LR. The results on gender and location DIF were not too different from the results when MH procedure was used in terms of identification of uniform DIF, only that LR identified more items as DIF than MH in some years. Nevertheless, LR identified few nonuniform gender and location DIF in all the four subjects under study.

Research hypotheses five and six examined gender and location DIF respectively using 3PL IRT DIF detecting procedure. Findings on research hypothesis five indicated DIF in favour of females in English Language whiles Mathematics items exhibited DIF in favour of males. There is an approximately equal number of DIF items in favour of both groups in terms of Integrated Science and Social Studies based on difficulty and discrimination parameters. In terms of guessing parameter, there was a higher probability of the least knowledgeable female (4.1%) student to get an item correct in English Language than the least male student (3.4%) while the least able male student had higher guessing parameter in mathematics (4.7%) than the female counterpart (1.3%). Concerning Integrated Science and Social Studies, there was an equal probability of the least able male (1.3%) and female (1.2%) getting an item correct. The findings on location DIF showed more DIF in favour of examinees who schooled in CR and GAR.

The research question examined the level of agreement among MH, LR and 3PL IRT detecting methods. Results showed that there is moderate level of agreement among MH, LR and 3PL IRT in detecting DIF based on gender and location. LR and MH were most effective in detecting gender and location DIF as compared to 3PL IRT. MH detected about 2% more items in gender and location DIF than LR. 3PL IRT detected the greatest number of gender and location DIF but at the A-level of ETS criteria.

Conclusions

This study was conducted to find out items that exhibited differential item functioning in 2012-2016 WASSCE in core subjects by gender and location (CR, GAR, WR, ER and VR) in Ghana using MH, LR and IRT DIF detecting methods. Based on the findings of this study, it is concluded that the

WASSCE examination in core subjects was not free from differential item functioning (DIF).

It is obvious that WASSCE results for examinees in senior secondary schools in CR, GAR, WR, ER and VR were not as valid as they should be

The results of the study imply that the results were not valid in terms of its uses and interpretation since groups of examinees were given an undue advantage or disadvantage over other groups. The location, in terms of regions, where examinees attended schools also detected DIF in which more items favoured the examinees who attended schools in CR and GAR. Students who attended school in CR and GAR had an undue advantage over those students who attended schools in WR, ER and VR. The examinees in CR and GAR may have had better infrastructure, adequate teaching and learning materials.

Lastly, the results of this study, showed that 2012-2016 WASSCE English Language, Mathematics, Integrated Science and Social Studies Examinations as national assessment tool (WASSCE), seems to be unfair to some students.

Recommendations

The findings and observation from the study showed that 2012-2016 WASSCE, as international examination, was not free from gender and location DIF in English Language, Mathematics, Integrated Science and Social Studies. Based on the above findings and conclusions, the following recommendations are made:

1. It is recommended that WAEC subject experts and people responsible for test development, validation and administeration need to carry out differential item functioning analysis for all items after administering and scoring the test.

2. Given the high-stake decisions in WASSCE examinations, test users should provide evidence that inferences made are valid for the test results.
3. WAEC should make it a yearly task to analyse responses to test items through DIF analysis to improve their item banks.

Contribution to knowledge

This study is making a significant contribution to knowledge and the current literature concerning three areas namely methodology, educational measurement and assessments in Mathematics, English Language, Integrated Science and Social Studies.

Methodology

This study employed a systematic framework from the measurement theory of invariance to examine DIF by using three different approaches to ascertain DIF. This study is the first in the literature that used three common DIF detection methods, namely the 3PL IRT, to calculate the a_i-parameter (difficulty), b_i-parameter (discriminating), c_i-parameter (pseudo guessing), the MH and the LR procedures, to differentiate two different types of non-uniform DIF.

This contributes to the psychometrics literature, because of often in psychometrics the aim is to measure persons as well as to generalize over a domain of items rather than to an instrument (Briggs & Wilson, 2007). Furthermore, the proposed 3PL IRT, which is a three-parameter logistic model, enables researchers to estimate the difficulty level, the discrimination index and guessing index problems by considering both responses and items simultaneously in the analysis (Heck & Thomas, 1999). In this context, a three-parameter logistic model provided more information to explain the nature of

DIF in the data, because it models the difficulty, discriminating and guessing parameters respectively.

To sum up, it is concluded that the contribution to knowledge with regards to the methodology can be seen from the proposed systematic framework to investigate DIF.

This study contributed by adding to literature with regards to understanding DIF; (a) explicitly differentiating and detecting two different types of DIF, namely uniform and nonuniform DIF; and (b) comparing the difficulty, discriminating and pseudo guessing parameters using 3PL IRT and (c) finding the agreement level of MH, LR and 3PL IRT within the Ghanaian Context.

Implications

The findings of this study have implications for researchers, test developers, teachers, teacher trainers, policymakers and those who use tests to inform high-stakes decisions. The extent to which test scores are used to make inferences on examinees' performance is an important issue of fairness and validity. Decisions that are based on invalid test scores may cause considerable harm to test takers and stakeholders affected by testing decisions.

Therefore, test developers and test users must justify the proposed uses and interpretations of test scores as well as taking responsibility for determining the trustworthiness of the test. Indeed evidence that supports the trustworthiness should be presented as well as the interpretations of the test scores. This study investigated the trustworthiness of test scores in measurement theory by testing the lack of measurement invariance in empirical data.

Suggestions for Further Research

1. A study should be conducted on the analysis of item distraction to assess DIF because the distribution of distractors has the potential for DIF which can influence the performance of a student.
2. Studies should be conducted on other types of DIF such as school type, school ownership and socioeconomic status of examinees using data from WASSCE and BECE.
3. Studies should also be done to give reasons why items exhibit gender and location DIF in English Language, Mathematics, Integrated Science and Social Studies.

REFERENCES

Abedalaziz, N. (2010). A gender-related differential item functioning of mathematics test items. *The International Journal of Education and Psychological Assessment, 5,*101-114.

Abedalaziz, N. (2010). Detecting gender-related DIF using logistic regression and Mantel-Haenszel approaches. *Procedia-Social and Behavioral Sciences*, *7*, 406-413

Acar, T., & Kelecioglu, H. (2010). Comparison of differential item functioning determination techniques: HGLM, LR and IRT-LR. *Educational Sciences: Theory and Practice*, *10*(2), 639-649.

Ackerman, T. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*, 674-691.

Adedoyin, O. O. (2010). Using IRT approach to detect gender-biased items in public examinations. *Educational Research and Reviews Academic Journals*, *5*(7), 385-399

AERA, A. P. A. (1999). NCME (American Educational Research Association,

American Psychological Association, & National Council on Measurement in Education). (1999). *Standards for educational and psychological testing and skills for life: First results from PISA 2000.* Paris, France: OECD.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks.

Allen, M. J., & Yen, W. M. (2001). *Introduction to measurement theory*. Florida, USA: Waveland Press.

Ahmadi, A., & Jalili, T. (2014). A confirmatory study of Differential Item Functioning on EFL reading comprehension. *Applied Research on the English Language*, *3*(6), 55-68.

Ahmadi, A., & Bazvand, A. D. (2016). Gender differential item functioning on a national field-specific test: The case of PhD entrance exam of TEFL in Iran. *Iranian Journal of Language Teaching Research*, *4*(1), 63-82.

Amuche, C.I & Fan, A. F. (2014) An assessment of item bias using differential item functioning technique in NECO biology conducted examinations in Taraba State, Nigeria. *American International Journal of Research in Humanities, Arts and Social Sciences*, *6*(1), 95-100

Angoff, W. (1993). Perspective on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-24). Hillsdale, NJ: Lawrence Erlbaum Associates.

Atalay Kabasakal, K., Arsan, N., Gök, B., & Kelecioglu, H. (2014). Comparing performances (Type I error and power) of IRT Likelihood Ratio SIBTEST and Mantel-Haenszel methods in the determination of differential item functioning. *Educational Sciences: Theory and Practice*, *14*(6), 2186-2193.

Amirian, S. M. R., Alavi, S. M., & Fidalgo, A. M. (2014). Detecting gender DIF with an English Language proficiency test in the EFL context. *Iranian Journal of Language Testing*, *4*(2), 187-203.

Ariffin, S. R., Ishak, N. M., Rahmad, R. A. O., Ahmad, A. G., Idris, R., Najmuddin, N. A., ... & Samsuri, S. (2008b). Assessing generic skills using Rasch model approach: A method for construct validity and reliability. In *International Conference on Education on Learning Diversity*, Bangkok, 7-10 April 2008.

Baghi, H., & Ferrara, S. (1990). Detecting differential item functioning using IRT and Mantel-Haenszel technique: *Implementing procedures and comparing results*. (ERIC Document Reproduction Service No. ED 325 479).

Baker, F. B. (2001). *The basics of item response theory*. ERIC Clearinghouse on Assessment and Evaluation.

Bar-Hillel, M., Budescu, D. & Attali, Y. (2005). Scoring and keying multiple-choice test: A case study in irrationality. *Mind & Society, 4*, 3–12.

Barton, M. A., & Lord, F. M. (1981). An upper asymptote for the three-parameter logistic item-response model. *Research Bulletin 8*, 1-20. Princeton, NJ: Educational Testing Service.

Baker, F. B. (1987). Methodology review: Item parameter estimation under the one-, two-, and three-parameter logistic models. *Applied Psychological Measurement*, *11*(2), 111-141.

Becker, B. J. (1989). Gender and science achievement: A reanalysis of studies from two meta-analyses. *Journal of Research in Science Teaching, 26*, 141–169.

Becker, B. J. (1990). Item characteristics and gender differences in the SAT-M for mathematically able youths. *American Educational Research Journal*, *27*(1), 65-87.

Berk, R. A. (Ed.). (1982). *Handbook of methods for detecting test bias*.Baltimore, MD: Johns Hopkins University Press.

Benbow, C. P., & Stanley, J. C. (1983). Differential course-taking hypothesis revisited. *American Educational Research Journal*, *20*(4), 469-473.

Benbow, C. P., & Stanley, J. C. (1980). Sex differences in mathematical ability: Fact or artefact. *Science, 210*(4475), 1262-1264.

Bhaduri, I., & Singh, A. (2011). *Learning levels of Grade V students in language*. In International Association for Educational Assessment Conference, Manila, Philippines.

Birenbaum, M., & Tatsuoka, K. K. (1993). Applying an IRT-based cognitive diagnostic model to diagnose students' knowledge states in multiplication and division with exponents. *Applied Measurement in Education*, *6*(4), 255-268.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores* (pp 397-479). Reading, MA: Addison-Wesley.

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. *Educational Measurement*, *5*(2) 1-16.

Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods & Research*, *17*(3), 303-316.

Bolger, N., & Kellaghan, T. (1990). Method of measurement and gender differences in scholastic achievement. *Journal of Educational Measurement*, *27*(2), 165–174.

Bock, R. D. (1975). *Multivariate statistical methods*. New York, NY: McGraw-Hill.

Bond, T. B., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum

Burkam, D. T., Lee, V. E., & Smerdon, B. A. (1997). Gender and science learning early in high school: Subject matter and laboratory experiences. *American Educational Research Journal, 34*(2), 297–331.

Burrell, G., & Morgan, G. (2017). *Sociological paradigms and organisational analysis: Elements of the sociology of corporate life*.Abingdon, UK: Routledge: Taylor & Francis Group.

Burton, R. F. (2002). Misinformation, partial knowledge and guessing in true/false tests. *Medical Education*, *36*(9), 805–811.

Bock, R.D.; Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika, 46* (4), 443–459.

Camilli, G. (1992). A conceptual analysis of differential item functioning in terms of multidimensional item response model. *Applied Psychological Measurement*, *16*(2), 129-147.

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Camilli, G. (2006). Test fairness. In R. L. (Ed.), *Educational measurement* (4th ed., pp. 220-256). Westport, CT: American Council on Education.

Cauley, K. M., & McMillan, J. H. (2010). Formative assessment techniques to support student motivation and achievement. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, *83*(1), 1-6.

Coulombe, H., & Canagarajah, S. (1997). Child labour and schooling in Ghana. *World Bank policy research paper*, (1844).

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265-289.

Cho, S. J., & Cohen, A. S. (2010). A multilevel mixture IRT model with an application to the DIF. *Journal of Educational and Behavioral Statistics, 35*(3), 336-370.

Chen, Fang Fang; Sousa, Karen H.; West, Stephen G. (2005). Testing measurement invariance of second-order factor models. *Structural Equation Modeling, 12*(3), 471– 492.

Cheung, G. W.; Rensvold, R. B. (2002). "Evaluating goodness-of-fit indexes for testing measurement invariance*". Structural Equation Modeling. 9* (2), 233–255.

Cizek, G. J., Rosenberg, S., & Koons, H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68,* 397-412.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*(1), 31-44.

Cole, N. S., & Zieky, M. J. (2001). The new faces of fairness. *Journal of Educational Measurement, 38*(4), 369-382.

Collins, K. (2010). Advanced sampling designs in mixed research: Current practices and emerging trends in the social and behavioural sciences.

Sage handbook of mixed methods in social and behavioural research, *2*, 353-377.

Cronbach, L. J. (1970). *Essentials of psychological testing* (3rd ed.). New York, NY: Harper & Row.

Crocker. L., & Angina, J. (1986). *Introduction to classical and modern test theory.*Toronto, Canada: Holt Rinehart & Winston.

Cole, N. S. (1997). *The ETS gender study: How males and females perform in educational settings.* Princeton, NJ: Educational Testing Service.

Cohen, A., Bolt, D. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement, 42*, 133-148.

Covic, T., Pallant, J. F., Conaghan, P. G., & Tennant, A. (2007). A longitudinal evaluation of the Center for Epidemiologic Studies-Depression scale (CES-D) in a rheumatoid arthritis population using Rasch analysis. *Health and Quality of Life Outcomes, 5*(1), 41.

Davidson, M., Keating, J. L., & Eyres, S. (2004). A low back-specific version of the SF-36 physical functioning scale. *Spine*, 29, 586-594.

Denga, D. I. (1998). *Educational measurement, continuous assessment and Psychological testing*. Calabar, Nigeria: Rapid Educational publishers.

DeMars, C. E. (1998). Gender differences in mathematics and science on a high school proficiency exam: The role of response format. *Applied Measurement in Education, 11*(39), 279-299.

DeMars, C. (2010). *Item response theory*. NewYork, NY: Oxford University Press.

DeVellis, R. F. (2016). *Scale development: Theory and applications* (26). Thousand Oaks, CA: Sage Publications.

Doolittle, A. E., & Cleary, T. A. (1987). Gender-based differential item performance in mathematics achievement items. *Journal of Educational Measurement*, *24*(2), 157-166.

Dorans, N. J., & Holland, P. W. (1992). DIF detection and description: Mantel-Haenzel and standardization.1, 2. *ETS Research Report Series*, *1992*(1), 1- 40.

Dorans, N.J., & Holland, P.W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Erlbaum.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, *23*(4), 355-368.

Driana, E. (2007). *Gender differential item functioning on a ninth-grade mathematics proficiency test in Appalachian Ohio*. (Unpublished doctoral dissertation). Ohio University.

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modelling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, *16*(1), 5.

Eleje, L. I., Onah, F. E., & Abanobi, C. C. (2018). Comparative study of classical test theory and item response theory using diagnostic quantitative economics skill test item analysis results. *European Journal of Educational & Social Sciences*, *3*(1), 57-75.

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis and the interpretation of research results*. Cambridge, UK: University Press.

Embretson, S. E., & Reise, S. P. (2000). Multivariate applications books series. *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. London, UK: Psychology Press.

Engelhard Jr, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*(2), 93-112.

Engelhard Jr, G. (2009). Using item response theory and model data fit to conceptualize differential item and person functioning for students with disabilities. *Educational and Psychological Measurement*, *69*(4), 585-602.

Engelhard Jr, G. (2008). Historical perspectives on invariant measurement: Guttman, Rasch, and Mokken. *Measurement*, *6*(3), 155-189.

Engelhard, G., Anderson, D., & Gabrielson, S. (1990). An empirical comparison of Mantel Haenszel and Rasch procedure for studying differential item functioning on teacher certification tests. *Journal of Research and Development in Education*, *23*, 172-179.

Ercikan, K., Simon, M., & Oliveri, M. E. (2013). Score comparability of multiple language versions of assessments within jurisdictions. In M. Simon, K. Ercikan, & M. Rousseau. (Eds.*), Improving large-scale* *assessment in education: Theory, issues and practice* (pp.110-124). New York. NY: Routledge/Taylor & Francis.

Erguven, M., & Erguven, C. (2014). An empirical study on the assessment of item-person statistics and reliability using classical test theory measurement methods. *Journal of Technical Science and Technologies*, *2*(2), 25-33.

Erdem, K. D. (2014). Comparison of Mantel-Haenszel and logistic regression techniques in detecting differential item functioning. *Journal of Measurement, Evaluation, Education and Psychology .5*(2), 12-25.

Fennema, E., & Leder, G. C. (1990). Gender differences in mathematics: A synthesis In Fennema & G. C. Leder (Eds.), *Mathematics and gender* (pp. 188-199). New York: Teachers College Press.

Fennema, E. (1974). Mathematics learning and the sexes: A review. *Journal for Research in Mathematics Education*, 126-139.

Fennema, E. (1980). Teachers and sex bias in mathematics. *The Mathematics Teacher*, *73*(3), 169-173.

Finch, W. H., & French, B. F. (2007). Detection of crossing a differential item functioning: A comparison of four DIF detecting methods. *Educational and Psychological Measurement*, *67*(4), 565-582.

Fidalgo, A. M., Mellenbergh, G. J., & Muñiz, J. (2000). Effects of amount of DIF, test length, and purification type on robustness and power of Mantel-Haenszel procedures. *Methods of Psychological Research Online*, *5*(3), 43-53.

Freedle, R. (2003). Correcting the SAT's ethnic and social-class bias: A method for reestimating SAT scores. *Harvard Educational Review*, *73*(1), 1-43.

Fox, L. H., Brody, L., & Tobin, D. (1985). The impact of early intervention programs upon course-taking and attitudes in high school. *Women and Mathematics: Balancing the Equation 2,* 249-274.

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size used with logistic regression for differential item functioning detection. *Educational and Psychological Measurement*, *67*(3), 373-393.

Frost, L. A., Hyde, J. S., & Fennema, E. (1994). Gender, mathematics performance, and mathematics-related attitudes and affect: A meta-analytic synthesis. *International Journal of Educational Research*, *21*(4), 373-385.

Gadermann, A., M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research, & Evaluation, 17(3),* 1-13.

Gamer, M., & Engelhard, G. (1999). Gender differences in performance on multiple-choice and constructed response mathematics items. *Applied Measurement in Education, 12,* 29-51.

Geary, D. C. (1996). Sexual selection and sex differences in mathematical abilities. *Behavioural and Brain Sciences*, *19*(2), 229-247.

Gessell, N. R. (2004). *Classroom management: Principles and practice*. London, UK: George Allen and Unwin.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioural sciences*. New York, NY: WH Freeman.

Gierl, M. J. (1999). *Test bias*. Thousand Oaks,CA: Sage Publications, Inc.

Gierl, M.J., & Cui, Y. (n.d.) Defining characteristics of diagnostic classification models and the problem of retrofitting in cognitive diagnostic. *Assessment Measurement Interdisciplinary Research and Perspectives 6*(4), 263-268.

Gierl, M. J. (2005). Using dimensionality-based DIF analyses to identify and interpret constructs that elicit group differences. *Educational Measurement: Issues and Practice*, *24*(1), 3-14.

Goswami, U. (1991). *Put to the Test: The Effects of External Testing on Teachers Educational Researcher* 20(4), 8-11. Thousand Oaks, CA: Sage Publications.

Gommez-Benito J, Navas-Ara MJ (2000) A Comparison of chi 2, RFA and IRT based procedures in the detection of DIF. *Quality Q, 34*(1)17-31.

González-de, L. P., Kostov, B., López-Pina, J. A., Solans-Julián, P., Navarro-Rubio, M. D., & Sisó-Almirall, A. (2015). A Rasch analysis of patients' opinions of primary health care professionals' ethical behaviour concerning communication issues. *Family practice*, *32*(2), 237-243.

Güler, N., & Penfield, R. D. (2009). A comparison of the logistic regression and contingency table methods for simultaneous detection of uniform and non-uniform DIF. *Journal of Educational Measurement*, *46*(3), 314-329.

Gulliksen, H. (2013). *Theory of mental tests*.NewYork, NY: Routledge, Taylor & Francis Group.

Gruijter, D. N. M. D., & Kamp, L. J. T. v. d. (2007). *Statistical test theory for the behavioral sciences*. Boca Raton, FL: Chapman & Hall/CRC.

Guion, R. M. (2011). *Assessment, measurement, and prediction for personnel decisions*. NewYork, NY: Routledge.

Halpern, D. (1992). *Sex differences in cognitive abilities*. Hillside, NJ: Lawrence Erlbaum.

Halpern, D. F. (2000). *Sex differences in cognitive abilities*. New York, NY: Psychology Press.

Hamilton, L. S. (1999). Detecting gender-based differential item functioning on a constructed-response science test. *Applied Measurement in Education, 12*, 211-235.

Ha-Kim, S., & Cohen, A. (1995). Comparison of Lord chi-square and rajus measures and the likelihood method on detecting of differential item functioning, *Applied Measurement in Education, 14*, 291-312.

Hambleton, R. K., & Swaminathan, H. (2013). *Item response theory: Principles and applications*. NewYork: Springer Science & Business Media.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: progress report. *European Journal of Psychological Assessment*, *10*, 229-244

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education, 2*(4), 313-334.

Hambleton, R. K., & Traub, R. E. (1973). Analysis of empirical data using two logistic latent trait models. *British Journal of Mathematical and Statistical Psychology*, *26*(2), 195-211.

Hambleton, R.K. & Rogers, H.J. (2000). *Developing an item bias review form*. Amherst, MA: The University of Massachusetts at Amherst.

Hambleton, R. K., & Slater, S. C. (1997). Item response theory models and testing practices: Current international status and future directions. *European Journal of Psychological Assessment*, *13*(1), 21-28.

Hand, D. J. (1996). Statistics and the theory of measurement. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *3*, 445-492.

Harnisch, D. L., Tatsuoka, K., & Wilkins, J. L. M. (1995). *Reporting math proficiencies based on the new SAT-M*. In Annual Meeting of the American Evaluation Association, British Columbia, Canada.

Haralambos, M., & Holborn, M. (2008). *Sociology: Themes and perspectives*. Hammer Smith: Harper Collins Limited.

Harris, A. M., & Carlton, S. T. (1993). Patterns of gender differences in mathematics items in the scholastic aptitude test. *Applied Measurement in Education*, *6*(2), 137-151.

Hedges, L. V., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science*, *269*, 41-45.

Hewege, C. R., & Perera, L. C. R. (2013). In search of alternative research methods in marketing: Insights from Layder's adaptive theory methodology. *Contemporary Management Research*, *9*(3).

Hidalgo, M. D., & Lopez-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement*, *64*(6), 903-915.

Higgins, G.E. 2007. Examining the Original Grasmick Scale: A Rasch model approach. *Journal of Criminal Justice and Behavior 34*, 157-158.

Hissbach, J. C., Klusmann, D., & Hampe, W. (2011). Dimensionality and predictive validity of the HAM-Nat, a test of natural sciences for medical school admission. *BMC Medical Education*, *11*(1), 83.

Highhouse, S., Doverspike, D., & Guion, R. M. (2015). *Essentials of personnel assessment and selection*. NewYork, NY: Routledge.

Hockemeyer, C. (2002). A comparison of non-deterministic procedures for the adaptive assessment of knowledge. *Psychologische Beiträge*, *44*, 495-503.

Holden, M. T., & Lynch, P. (2004). Choosing the appropriate methodology: Understanding research philosophy. *The Marketing Review*, *4*(4), 397-409.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Erlbaum.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum.

Hsin-Huang Li & Stout, W. (1995). *A new procedure for detection of crossing DIF*: *Research Report*. The University of Illinois at Urbana-Champaign, Illinois.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin 107*(2), 139-155.

Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin*, *104*(1), 53.

Hyde, J. S., Geiringer, E. R., & Yen, W. M. (1975). On the empirical relation between spatial ability and sex differences in other aspects of cognitive performance. *Multivariate Behavioral Research*, *10*(3), 289-309.

ISSER (2008). *The state of the Ghanaian economy*. Accra, Ghana: University of Ghana.

Inyang, S. N. (2004). *Analysis of item difficulty and students' performance in the 2002* junior *secondary school mathematics test.* (Unpublished M.Ed thesis). Faculty of Education, University of Calabar.

John, O. P., & Soto, C. J. (2007). The importance of being valid: Reliability and the process of construct validation. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 461-494). New York, NY: Cambridge University Press.

Jovanovic, J., Solano-Flores, G., & Shavelson, R. J. (1994). Performance-based assessments: Will gender differences in science achievement be eliminated? *Education and Urban Society, 26*, 352–366.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*(1), 79-93.

Kamata, A., & Vaughn, B. K. (2004). An Introduction to differential item functioning analysis. *Learning disabilities: A Contemporary Journal*, *2*(2), 49-69.

Kaplan, R. M., & Saccuzzo, D. P. (2009) *Psychological testing* Belmont, CA: Wadsworth.

Kalaycioglu, D. B., & Kelecioglu, H. (2011). Item bias analysis of the university entrance examination. *Egitim ve Bilim*, *36*(161), 3.

Kalaycioğlu, D.B., Berberoğlu, G. (2011), Differential Item Functioning analysis of the science and mathematics items in the university entrance examinations in Turkey, *Journal of Psychoeducational Assessment, 29* (5), 467-478.

Karakaya, I. (2012). An investigation of item bias in science and technology subtests and mathematics subtests in Level Determination Exam (LDE). *Educational Sciences: Theory and Practice*, *12*(1), 222-229.

Karami, H., & Salmani-Nodoushan, M. A. (2011). Differential Item Functioning (DIF): Current problems and future directions. *Online Submission*, *5*(3), 133-142.

Kline, P. (2014). *The new psychometrics: science, psychology and measurement*. New York, NY: Routledge.

Kline, T. (2005). *Psychological testing: A practical approach to design and evaluation*. Thousand Oaks, CA: Sage.

Kim, J.S.,& Bolt. D. M. (2007). Estimating item response theory models using Markov chain Monte Carlo methods. *Educational Measurement: Issues and Practice,* 26: 38–51.

Kim, W. (2003). *Development of a differential item functioning (DIF) procedure using the hierarchical generalized linear model: A comparison study with the logistic regression procedure.* (Unpublished PhD thesis). The Pennsylvania State University.

Kimball, M. M. (1989). A new perspective on women's math achievement. *Psychological Bulletin, 105*(2), 198.

Kreiner, S., & Christensen, K. B. (2011). Item screening in graphical log-linear Rasch models. *Psychometrika*, *76*(2), 228-256.

La Porta, F., Franceschini, M., Caselli, S., Cavallini, P., Susassi, S., & Tennant, A. (2011). Unified balance scale: an activity-based, bed to the community, and aetiology-independent measure of balance calibrated with Rasch analysis. *Journal of Rehabilitation Medicine, 43,* 435-444.

Langenfeld, T. E. (1997). Test fairness: Internal and external investigations of gender bias in mathematics testing. *Educational Measurement: Issues and Practice*, *16*(1), 20-26.

Lane, S., Wang, N., & Magone, M. (1996). Gender-Related Differential Item

Functioning on a Middle-School Mathematics Performance Assessment. *Educational Measurement: Issues and Practice*, *15*(4), 21-27.

Leder, G. C. (1992). Mathematics and gender: Changing perspectives. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 597-622). Oxford, UK: Maxwell Macmillan International.

Leder, G. (1985). Sex-related differences in mathematics: An overview. *Educational Studies in Mathematics*, *16*(3), 304-309.

Lee, Y. J., Palazzo, D. J., Warnakulasooriya, R., & Pritchard, D. E. (2008). Measuring student learning with item response theory. *Physical Review Special Topics-Physics Education Research*, *4*(1), 101-102.

Lee, Y. W. (2004). Examining passage-related local item dependence (LID) and measurement construct using Q3statistics in an EFL reading comprehension test. *Language Testing, 21*(1), 74-100.

Lee, M. K., & Randall, J. (2011). *Exploring language as a source of DIF in a math test for English language learners*. Paper presented at Northeast Educational Research Association Conference Proceedings, University of Connecticut, 21 October 2011.

Leland, H. E. (1999). Beyond Mean-Variance: Performance Measurement in a Nonsymmetrical World (corrected). *Financial Analysts Journal*, *55*(1), 27-36.

Lewin, K. M., & Sayed, Y. (2005). *Non-government secondary schooling in sub-Saharan Africa: Exploring the evidence in South Africa and Malawi Department for international development*, London, UK: Fuller Davies Limited.

Lewis, J. (2001) Language isn't needed: nonverbal assessments and gifted learners. (ERIC Document Reproduction Service No. ED453026)

Le, V. N. (1999). *Identifying differential item functioning on the NELS: 88 History achievement test*. CSE Technical Report.

Linn, R. L. (2011). The standards for educational and psychological testing: Guidance in test development. In *Handbook of test development* (pp. 41-52).New York, NY: Routledge.

Linn, M. C., & Hyde, J. S. (1989). Gender, mathematics, and science. *Educational Researcher*, *18*(8), 17-27.

Li, H., & Stout, W. F. (1993). *A new procedure for detection of crossing DIF/bias*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta.

Li, H., & Stout, W.F. (1996). A new procedure for detection of crossing DIF. *Psychometrika*, *61*(4), 647-677.

Linacre, J. M. (2009a). *A user's guide to Winsteps-Ministep: Rasch-model computer programs*. Program manual 3.68. 0. Chicago, IL: Winsteps.

Linacre, J. M. (2009b). Local independence and residual covariance: A study of Olympic figure skating ratings. *Journal of Applied Measurement 10*(2):157-69.

Linacre, J. M., & Wright, B. D. (1994). Chi-square fit statistics. *Rasch Measurement Transactions*, *8*(2), 350.

Linn, R. L., & Drasgow, F. (1987). Implications of the golden rule settlement for test construction. *Educational Measurement: Issues and Practice, 6(2), 13-17.*

Linne, M. C., & Hyde, J. S. (1989). Gender, mathematics, and science. *Educational Researcher, 18*(8), 17-27.

Linacre, J. M., & Wright, B. D. (1987). *Item bias: Mantel-Haenszel and the Rasch model*. Finnish Association of Mathematics and Science Education Research, Memorandum, 39.

Longford, N. T., Holland, P. W., & Thayer, D. T. (1993). Stability of the MH DIF across populations. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 171–196). Hillsdale, NJ: Erlbaum.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.

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

Lord, F. M., & Novick, M. R. (2008). *Statistical theories of mental test scores*. IAP.

Lord, F. M. (2012). *Applications of item response theory to practical testing problems*. New York, NY: Routledge.

Madu, B. C. (2012). Analysis of gender-related differential item functioning In Mathematics multiple-choice items administered by the West African Examination Council (WAEC). *Journal of Education and Practice, 3*(8), 71-79.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719-748.

Mann, V. A., Sasanuma, S., Sakuma, N., & Masaki, S. (1990). Sex differences in cognitive abilities: A cross-cultural perspective. *Neuropsychologia*, *28*(10), 1063-1077.

Mapuranga, R., Dorans, N. J., & Middleton, K. (2008). A review of recent developments in differential item functioning. *ETS Research Report Series*, *2008*(2), 1-32.

Makransky, G., Rogers, M. E., & Creed, P. A. (2015). Analysis of the construct validity and measurement invariance of the career decision self-efficacy scale: A Rasch model approach. *Journal of Career Assessment*, *23*(4), 645-660.

Makransky, G., & Bilenberg, N. (2014). Psychometric properties of the parent and Teacher ADHD Rating Scale (ADHD-RS) measurement invariance across gender, age, and informant. *Assessment, 21*(6), 694-705.

Marais, I.,& Andrich, D. (2008a). Formalising dimension and response violations of Local independence in the unidimensional Rasch model. *Journal of Applied Measurement*, *9*(3), 200-15.

Marais, I.,& Andrich, D. (2008b). Effects of varying magnitude and patterns of response dependence in the unidimensional rasch model. *Journal of Applied Measurement*, *9*(2), 1-20.

Marais, H. (2013). *South Africa Pushed to the limit: The political economy of change*. London, UK: Zed Books Ltd.

Marasculio, L. A., & Slaughter, R. E. (1981). Statistical procedures for identifying possible sources of item bias based on 2 x 2 statistics. *Journal of Educational Measurement*, *18*, 229-248.

Martin, M. O., Mullis, I. V., Gonzalez, E. J., Gregory, K. D., Smith, T. A., Chrostowski, S. J., & O'Connor, K. M. (2008). *TIMSS 1999. International Science Report*.

Martin, M.O., Mullis, I.V.S., Gonzalez, E.J., & Chrostowski, S.J. (2004). *TIMSS 2003 international science report: Findings from IEA's Trends in International mathematics and science study at the fourth and eighth grades*. Chestnut Hill, MA: Boston College.

Mantel Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, *17*(2), 105-116.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1994). Identification of nonuniform differential item functioning using a variation of the Mantel-Haenszel procedure. *Educational and Psychological Measurement*, *54*(2), 284 - 291.

Mazor, K. M., Kanjee, A., & Clauser, B. E. (1995). Using logistic regression and the Mantel-Haenszel with multiple abilities estimates to detect differential item functioning. *Journal of Educational Measurement*, *32*(2), 131-144.

McMillan, J. H. (2003). Understanding and improving teachers' classroom assessment decision making: Implications for theory and practice. *Educational Measurement: Issues and practice, 22*(4), 34-43.

Maccoby, E. E., & Jacddin, C. N.(1974) *The psychology of sex differences.* Stanford, CA: Stanford Univer. Press.

McTighe, J., & O'Connor, K. (2005). Seven practices for effective learning. *Educational Leadership*, *63*(1), 11-12.

Mendes-Barnett, S., & Ercikan, K. (2006). Examining sources of gender DIF in mathematics assessments using a confirmatory multidimensional model approach. *Applied Measurement in Education*, *19*(4), 289-304.

Mellenberg, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics*, *7*(2), 105-108.

Meredith, W., & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika*, *57*(2), 289-311.

Messick, S. J. (2013). *Assessment in higher education: Issues of access, quality, student development and public policy*. New York, NY: Routledge.

Michell, J. (2014). *An introduction to the logic of psychological measurement*. London, UK: Psychology Press.

Miles, M. B., Huberman, A. M., Huberman, M. A., & Huberman, M. (1994). *Qualitative data analysis: An expanded sourcebook*. Thousand Oaks, CA: Sage.

Millsap, R. E., & Meredith, W. (1992). Inferential conditions in the statistical detection of measurement bias. *Applied Psychological Measurement*, *16*(4), 389-402.

Millsap, R. E., & Everson, H. T. (1993). Methodological review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17(4),* 297-334.

Mislevy, R. J., & Verhelst, N. (1990). Modelling item responses when different subjects employ different solution strategies. *Psychometrika*, *55*(2), 195-215.

Morse, J. M. (2003). Principles of mixed methods and multimethod research design. *Handbook of mixed methods in social and behavioural research*, *1*, 189-208.

Morse, J. M. (2016). *Mixed method design: Principles and procedures*. New York, NY: Routledge.

Mullis, I. V., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 Assessment Frameworks*. International Association for the evaluation of educational achievement. Herengracht 487, Amsterdam, 1017 BT: Netherlands.

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Gregory, K. D., Garden, R. A., M., O. C. (2000). *TIMSS 1999 International Mathematics Report*. Chestnut Hill, MA: Boston College.

Mullis, I. V., Martin, M. O., Gonzalez, E. J., & Chrostowski, S. J. (2004). *TIMSS 2003 International Mathematics Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades*. TIMSS & PIRLS International Study Center. Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467.

Mullis, I., Martin, M., Gonzalez, E., Gregory, K., Garden, R., O'Connor, K., & Smith, T. (2000). *TIMSS 1999. Findings from IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade*. International Mathematics Report. Boston.

Muijs, D. (2004). *Doing qualitative research in education with SPSS.* London, UK: SAGE Publication.

Nair, R., Moreton, B. J., & Lincoln, N. B. (2011). Rasch analysis of the Nottingham extended activities of daily living scale. *Journal of Rehabilitation Medicine*, *43*(10), 944-950.

Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews*, *82*(4), 591-605.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement*, *18*(4), 315-328.

Navas-Ara, M. J., & Gómez-Benito, J. (2002). Effects of ability scale purification on the identification of dif. *European Journal of Psychological Assessment*, *18*(1), 9.

Ndifon, M. B. O., Umoinyang, I. E., & Idiku, F O.(n.d). *Differential item functioning of 2010 junior secondary school certificate mathematics examination in the southern educational zone of Cross River State, Nigeria.* Retrieved from https://iaea.info/documents/differential-item-functioning-of 2010-junior-secondary-school-certificate-mathematics-examination-in-southern-educational-zone-of-cross-river-state-Nigeria.

Nitko, A. J. (1996). *Educational assessment of students*. Des Moines, IA: Prentice-Hall.

Nitko, J.J., Brookhart, S. M. (2004). *Educational assessment of students*. Upper Saddle River, NJ: Merrill-Prentice Hall.

Nijenhuis, E. R. (2004). *Somatoform dissociation: Phenomena, measurement, and theoretical issues*. New York, NY: West Norton & Company.

Odili, J. N. (2010). Effect of language manipulation on differential item functioning of Test items in Biology in a multicultural setting. *Journal of Educational Assessment in Africa*, *4*(2), 6-8.

OECD Programme for International Student Assessment. (2007). *PISA 2006: Science competencies for tomorrow's world data*. Paris, France: OECD.

Osadebe, P. U., & Agbure, B. (2018). Assessment of Differential Item Functioning in Social Studies Multiple-Choice Questions in Basic Education Certificate Examination. *European Journal of Education Studies*, *4*(9) .236-257.

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (Vol. 161). Thousand Oaks, CA: Sage Publications.

Onyeneke, C., Olorunju, S., Eta, U., & Nwaonu, C. (2018). Weibull Transformation Approach to Formulation of Reliability Model for Analysis of Filth Formation Using Zenith Grinding Machine. *American Journal of Aerospace Engineering*, *5*(1), 30-35.

Organization for Economic Co-Operation and Development. (2004). *Problem-solving for tomorrow's world*. First measures of cross-curricular competencies from PISA 2003"

Organisation for Economic Co-operation and Development (OECD). (2001). *Knowledge Organisation for Economic Co-operation and Development. (2007), PISA 2006: Science competencies for tomorrow's world Volume 1: Analysis*. Paris, France: Author.

Osterlind, S. J. (1983). *Test item bias*. Thousand Oaks, CA: Sage.

Owen, A. B. (1992b). A central limit theorem for Latin hypercube sampling. *Journal of the Royal Statistical Society: Series B (Methodological), 54*(2), 541-551.

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

Pae, T. (2004b). Gender effect on reading comprehension with Korean EFL learners. *A system*, *32*(2), 265–281.

Pae, T. (2012). Causes of gender DIF on an EFL language test: A multiple-data analysis over nine years. *Language Testing*, *29*(4), 533–554.

Pallas, A. M., & Alexander, K. L. (1983). Sex differences in quantitative SAT performance: New evidence on the differential coursework hypothesis. *American Educational Research Journal*, *20* (2), 165-182.

Pedhazur, E. J., & Schmelkin, L. P. (2013). *Measurement, design, and analysis: An integrated approach*. London, UK: Psychology Press.

Pei, L. K., & Li, J. (2010). Effects of unequal ability variances on the performance of logistic regression, Mantel-Haenszel, SIBTEST IRT, and IRT likelihood ratio for DIF detection. *Applied Psychological Measurement*, *34*(6), 453-456.

Penfield, R. D. (2010). Test-based grade retention: Does it stand up to professional standards for fair and appropriate test use?. *Educational Researcher*, *39*(2), 110-119.

Penfield, R. D., & Camelli, G. (2007). Differential item functioning and bias. In C. R. Rao & S. Sinharay (Eds.), *Psychometrics* (Vol. 26). Amsterdam, Netherlands: Elsevier.

Penner, A. M., & Paret, M. (2008). Gender differences in mathematics achievement: Exploring the early grades and extremes. *Social Science Research*, *37*(1), 239-253.

Peterson, P. L., & Fennema, E. (1985). Effective teaching, student engagement in classroom activities, and sex-related differences in learning mathematics. *American Educational Research Journal*, *22*(3), 309-335.

Plake, B. S. (1980a). A comparison of a statistical and subjective procedure to ascertain item validity: One step in the test validation process. *Educational and Psychological Measurement*, *40*(2), 397–404.

Polikoff, M. S. (2010). Instructional sensitivity as a psychometric property of assessments. *Educational Measurement: Issues and Practice*, *29*(4), 3-14.

Popham, W. J. (2008). All about Assessment: A misunderstood grail. *Educational Leadership*, *66*(1), 82-83.

Popham, W. J.(2016). Standardized tests: Purpose is the Point. *Educational Leadership*. *73* (7), 44-49.

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomous scored items: A framework for classification and evaluation. *Applied Psychological Measurement*, *19*(1), 23-37.

Proctor, J. D. (1998). The social construction of nature: Relativist accusations, Pragmatist and critical realist responses. *Annals of the Association of American Geographers*, *88*(3), 352-376.

Programme for International Student Assessment. (2003). *PISA*. Technical Report. Paris, France: OECD Publication.

Rallis, S. F., & Rossman, G. B. (2003). Mixed methods in evaluation contexts: A pragmatic framework. *Handbook of mixed methods in social and behavioural research*, 491-512.

Ramp, M., Khan, F., Misajon, R. A., & Pallant, J. F. (2009). Rasch analysis of the multiple sclerosis impact scale (MSIS-29). *Health and Quality of Life Outcomes*, *7*(1), 58.

Randhawa, B. S. (1994). Self-efficacy in mathematics, attitudes, and achievement of Boys and girls from restricted samples in two countries. *Perceptual and Motor Skills*, *79*(2), 1011-1018.

Rao, C., & Sinharay, S. (Eds.). (2007). *Handbook of statistics*: *Psychometrics*. 26, 45-74. Amsterdam, Netherlands: Elsevier.

Reback, R., Rockoff, J., & Schwartz, H. L. (2014). Under pressure: Job security, resource allocation, and productivity in schools under No Child Left Behind. *American Economic Journal: Economic Policy*, *6*(3), 207-41.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, *4*(3), 207-230.

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the global: Comments on Sijtsma. *Psychometrika*, *74*(1), 145.

Reeve, B. B., Hays, R. D., Chang, C. H., & Perfetto, E. M. (2007). Applying item response theory to enhance health outcomes assessment. *Quality of Life Research*, *16*(1), 1-3.

Retnawati, H. (2008). *Identifying item bias using the simple volume indices and multidimensional item response theory likelihood ratio (IRTLR) test*. Paper on International Conference on Mathematics 3th on July 18th 2008.

Rogers, H. J., & Kulick, E. (1987). An investigation of unexpected differences in item performance between Blacks and Whites taking the SAT. *Differential item functioning on the Scholastic Aptitude Test*, 87-100.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioural research: Methods and data analysis* (Vol. 2). New York, NY: McGraw-Hill

Rosenbaum, P. R. (1984). Conditional permutation tests and the propensity score in observational studies. *Journal of the American Statistical Association*, *79*(387), 565-574.

Roussos, L., & Stout, W. (1996a). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20,* 355–371.

Roussos, L., & Stout, W. (1996b). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-

Haenszel Type I error performance. *Journal of Educational Measurement*, *33*, 215–230.

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance. *Journal of Educational Measurement*, *33*(2), 215-230.

Roussos, L. A., & Stout, W. (2004). Differential item functioning analysis. *The Sage handbook of quantitative methodology for the social sciences*, 107-116.

Ross, J. A., Xu, Y., & Ford, J. (2008). The Effects of a Teacher In-Service on Low-Achieving Grade 7 and 8 Mathematics Students. *School Science and Mathematics, 108*(8), 362-379.

Røe, C., Damsgård, E., Fors, T., & Anke, A. (2014). Psychometric properties of the pain stages of change questionnaire as evaluated by Rasch analysis in patients with chronic musculoskeletal pain. *BMC musculoskeletal disorders*, *15*(1), 95-97.

Rowntree, D. (2015). *Assessing students: How shall we know them?*. New York, NY: Routledge.

Roznowski, M., & Reith, J. (1999). Examining the measurement quality of tests containing differentially functioning items: Do biased items result in poor measurement? *Educational and Psychological Measurement*, *59*(2), 248– 70.

Rubio, D. M., Berg-Weger, M., Tebb, S. S., Lee, E. S., & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. *Social Work Research*, *27*(2), 94-104.

Rudas, T., & Zwick, R. (1997). Estimating the importance of differential item functioning. *Journal of Educational and Behavioral Statistics, 22*(1), 31-45.

Rulison, K., & Loken, E. (2009). I've fallen and I can't get up: Can high-ability students recover from early mistakes in CAT? *Applied Psychological Measurement*, *33*(2), 83-101.

Ryan, K., & Bachman, L.F. (1992). Differential item functioning on two tests of EFL proficiency. *Language Testing*, *9*(1), 12–29.

Ryan, K. E., & Chiu, S. (2001). An examination of item context effects, DIF, and gender DIF. *Applied Measurement in Education*, *14*(1), 73-90.

Ryan, K. E., & Fan, M. (1996). Examing Gender DIF on a Multiple-choice Test of Mathematics: A Confirmatory Approach. *Educational Measurement: Issues and Practice*, *15*(4), 15-20.

Kline, R. (2011). *Principles and practice of structural equation modeling*. London, UK: Guilford Press.

Santelices, M. V., & Wilson, M. (2012). On the relationship between differential item functioning and item difficulty: An issue of methods? Item response theory approach to differential item functioning. *Educational and Psychological Measurement*, *72*(1), 5-36.

Schulz, E. M., Perlman, C., Rice, W. K., & Wright, B. D. (1996). An empirical comparison of Rasch and Mantel-Haenszel procedures for assessing differential item functioning. In G. Engelhard & M. Wilson (Eds.), *Objective measurement: Theory into practice* (Vol.3, pp. 65-82). Norwood, NJ: Ablex.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*(2), 159-194.

Siamisang, F. T., & Nenty, H. J. (2012). Analysis of gender-based differential item functioning (DIF) in 2007 TIMSS examination among students from Botswana, Singapore and USA. *Journal of Educational Assessment in Africa*, *7*, 43-54.

Singleton, R., & Strait, B. (2010). Straits. *Approaches to Social Research* (5th ed.) New York, NY: Oxford University Press.

Simon, F., Malgorzata, K., & Beatriz, P. O. N. T. (2007). *Education and training policy no more failures ten steps to equity in education: Ten steps to equity in education*. Paris, France: OECD Publishing.

Smith, R.M. (1990). Theory and practice fit. *Rasch Measurement Transactions. 3*(4), 78-80.

Skaggs, G., & Lissits, R. (1992). The consistency of Detecting item Bias across different test Administration: Implications of anther Failure. *Journal of Educational Measurement*. *29*(3).

Stage, C. (2000). Predicting Gender Differences in word item. A comparison of item response theory and classical test theory. A study of SAT Subtest. ERIC. *Journal of Educational Measurement*, *29*(30), 5-15.

Stoneberg, Jr. B.D. 2004. A Study of Gender-Based and Ethnic-Based Differential Item Functioning (DIF) in the Spring 2003 Idaho Standards Achievement Tests Applying the Simultaneous Bias Test (SIBTEST) and the Mantel- Haenszel Chi-Square Test, Idaho Standards

Achievement Tests; Reading, Language Usage, and Mathematics, Grades 4, 8, and 10. *Idaho State Department of Education*.

Smith, T. W. (1990). *Ethnic images (No. 19).* Chicago, IL: National Opinion Research Center, The University of Chicago.

Smith, E. V. (2005). Effect of item redundancy on Rasch item and person estimates. *Journal of Applied Measurement*, *6*(2), 147-163.

Snider, P., & Styles, I. (2003). *Psychometric analysis of triads' instrument of collectivism and individualism using modern latent trait theory*. Murdoch, Australia: Murdoch University: Western Australia.

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. *Psychological Methods, 11(4),* 402-415.

Stobart, G. & Gipps, C. 1997. *Assessment: A Teacher's Guide to the Issues. Ed. 3.* London, UK: Hodder & Stoughton.

Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health measurement scales: A practical guide to their development and use*. Oxford University Press, USA.

Suskie, L. (2018). *Assessing student learning: A common sense guide*. Hoboken, NJ: John Wiley & Sons.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370.

Swaminathan, H., Hambleton, R. K., & Rogers, H. J. (2006). 21 Assessing the fit of item response theory models. *Handbook of Statistics*, *26*, 683-718.

Swanson, L. A. (2013). A strategic engagement framework for nonprofits. *Nonprofit Management and Leadership*, *23*(3), 303-323.

Tae, P. (2004). Gender effect on reading comprehension with Korean EFL learners. *System*, 32, 265-281.

Taylor, C. S., & Lee, Y. (2012). Gender DIF in reading and mathematics tests with mixed item formats. *Applied Measurement in Education*, *25*(3), 246-280.

Ten Klooster, P. M., Taal, E., & Van De Laar, M. A. (2008). Rasch analysis of the Dutch Health Assessment Questionnaire disability index and the Health Assessment Questionnaire II in patients with rheumatoid arthritis. *Arthritis Care & Research*, *59*(12), 1721-1728.

Thissen, D., & Wainer, H. (Eds.). (2001). *Test scoring*. New York, NY: Routledge.

Thissen, D., & Steinberg, L. (1988). Data analysis using item response theory. *Psychological Bulletin*, *104*(3), 385-387.

Thissen, D., Steinberg, L., Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-114). Hillsdale, NJ: Lawrence Erlbaum.

Thissen, D., Steinberg, L., Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H.Wainer & H. Braun (Eds.), *Test validity* (pp. 147-150). Hillsdale, NJ: Lawrence Erlbaum.

Traub, R. E. (1983). A priori considerations in choosing an item response model. *Applications of item response theory*, *57*, 70.

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement*, *16*, 8-13.

Turkan, A., & Cetin, B. (2017). Study of Bias in 2012-Placement Test through the Rasch Model in Terms of Gender Variable. *Journal of Education and Practice*, *8*(7), 196-204.

Umoinyang, I. E. (1991). *Differential item functioning (DIF) resulting from the level of states educational development in Nigeria.* Paper presented at the first regional conference of world council for curriculum and Instruction region. 2 (South of Sahara) Lagos.

Uttaro, T., & Millsap, R. E. (1994). Factors influencing the Mantel-Haenszel procedure the detection of differential item functioning. *Applied Psychological Measurement*, *18*(1), 15-25.

Vandenberg, R. J., & 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), 4–70.

Van Der Flier, H., Mellenbergh, G. J., Adèr, H. J., & Wijn, M. (1984). An iterative item bias detection method. *Journal of Educational Measurement*, *21*(2), 131-145.

Van der Flier, H. (1980). *Vergelijkbaarheid van individuele testprestaties*. Swets & Zeitlinger.

Victoriano, A. N. (2011). Factors Affecting the National Achievement Test Performance of Selected Second Year High School Students in Santa Maria, Bulacan. A. (Unpublished master's thesis). Open University of the Polytechnic University of the Philippines.

Volante, L., & Melahn, C. (2005). Promoting assessment literacy in teachers: Lessons from the Hawaii School Assessment Liaison Program. *Pacific Educational Research Journal*, *13*(1), 19-34.

Walker, C. M. (2011). What's the DIF? Why differential item functioning analyses are an important part of instrument development and validation. *Journal of Psychoeducational Assessment*, *29*(4), 364-376.

Walker, C. M., & Beretvas, S. N. (2003). Comparing multidimensional and unidimensional proficiency classifications: Multidimensional IRT as a diagnostic aid. *Journal of Educational Measurement*, *40*(3), 255-275.

Walker, C. M., & Beretvas, S. N. (2001). An empirical investigation demonstrating the multidimensional DIF paradigm: A cognitive explanation for DIF. *Journal of Educational Measurement*, *38*(2), 147-163.

Walker-Gleaves, C. (2009). *A study of 'caring' academics and their work within a UK university*. (Unpublished doctoral thesis). The University of Leicester.

Waller, W. I. (1989). Modelling guessing behaviour. *Applied Psychological Measurement,13*(3), 233-243.

Weiss, D. J., & Davison, M. L. (1981). Test theory and methods. *Annual Review of Psychology*, *32*(1), 629-658.

Wechsler, D. (1981). WAIS-R manual*: Wechsler adult intelligence scale-revised.* New York, NY: Psychological Corporation.

Whitmore, M. L., & Schumacker, R. E. (1999). A comparison of logistic regression and analysis of variance differential item functioning detection methods. *Educational and Psychological Measurement, 59*(6), 910-927.

Wiersma, W., & Jurs, S. G. (1985). *Educational measurement and testing*. Boston, MA: Allyn & Bacon.

Wilson, M., & Iventosch, L. (1988). Using the partial credit model to investigate responses to structured subtests. *Applied Measurement in education, 1*(4), 319-334.

Widaman, K. F.; Ferrer, E., & Conger, R. D. (2010). Factorial Invariance within Longitudinal Structural Equation Models: Measuring the Same Construct across Time. *Child Dev Perspect, 4*(1), 10-18

Widaman, K. F., & Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modelling. *Psychological Methods, 8* (1), 16–37.

World Bank. (2007). *The World Bank Annual Report 2007*. Washington, DC: The World Bank.

Wright, H. N. (1968). The effect of sensorineural hearing loss on threshold-duration functions. *Journal of Speech and Hearing Research*, *11*(4), 842-852.

Wright, B. D., & Stone, M. H. (1979). *Best Test Design: Rasch Measurement*. Chicago. IL: MESA.

Xie, Y. (2005). *Three studies of a person by item interactions in an international assessment of educational achievement* (Unpublished doctoral dissertation). University of California, Berkeley.

Yamamoto, K. (1987). *A model that combines IRT and latent class models* (Unpublished doctoral dissertation). University of Illinois, Champaign-Urbana.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence *Journal of Educational Measurement, 30*(3), 187-213.

Yen, Y.-C., Ho, R.-G., Chen, L.-J., Chou, K.-Y., & Chen, Y.-L. (2010). Development and evaluation of a confidence-weighting computerized adaptive testing. *Journal* of *Educational Technology & Society*, *13*(3), 163-176

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125-145.

Young, D. J., & Fraser, B. J. (1994). Gender differences in science achievement: Do school effects make a difference? *Journal of Research in Science Teaching*, *31*(8), 857-871.

Young, J. W., & Fisler, J. L. (2000). Sex differences on the SAT: An analysis of demographic and educational variables. *Research in Higher Education*, *41*(3), 401-416.

Zhang, M. (2009). *Gender-related differential item functioning in mathematics tests: A meta-analysis*. (Unpublished doctoral dissertation). Washington State University.

Zenisky, A. L., Hambleton, R. K., & Robin, F. (2004). DIF detection and interpretation in large scale science assessments: Informing item writing practices. *Educational Assessment*, *9*(1-2), 61-78.

Zenisky, A. L., Hambleton, R. K., & Robin, F. (2003). Detection of differential item functioning in large-scale state assessments: A study evaluating a two-stage approach. *Educational and Psychological Measurement*, *63*(1), 51-64

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2002). Identification and evaluation of local item dependencies in the Medical College Admissions Test. *Journal of Educational Measurement*, *39*(4), 291-309.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 337–347). Hillsdale, NJ: Erlbaum.

Zumbo, B. D., & Thomas, D. R. (1997).A measure of effect size for a model-based approach For studying DIF: *Working paper of the Edgeworth Laboratory for Quantitative Behavioral Science*. Prince George, Canada: University of Northern British Columbia.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modelling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, Canada: National Defense Headquarters.

Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language Testing*, *20*(2), 136-147.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*(2), 223-233.

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement, 26*(1), 55-66.

Zwick, R.; Thayer, D.T., & Wingersky, M. (1995). Effect of Rasch calibration on ability and DIF estimation in computer-adaptive tests. *Journal of Educational Measurement. 32* (4), 341–363.