THE EXAMINATION OF THE PSYCHOMETRIC QUALITY OF THE COMMON EDUCATIONAL PROFICIENCY ASSESSMENT (CEPA)-ENGLISH TEST

by

Salma A. Daiban

M.A., Quantitative Psychology, Middle Tennessee State University, 2003

Submitted to the Graduate Faculty of

School of Education in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2009

UNIVERSITY OF PITTSBURGH

SCHOOL OF EDUCATION

This dissertation was presented

by

Salma A. Daiban

It was defended on

November 16th, 2009

and approved by

Clement A. Stone, PhD, Professor, School of Education

Feifei Ye, PhD, Assistant Professor, School of Education

Lauren Terhorst, PhD, Assistant Professor, School of Nursing

Dissertation Advisor: Suzanne Lane, PhD, Professor, School of Education

THE EXAMINATION OF THE PSYCHOMETRIC QUALITY OF THE COMMON EDUCATIONAL PROFICIENCY ASSESSMENT (CEPA)-ENGLISH TEST

Salma A. Daiban, PhD

University of Pittsburgh, 2009

The CEPA-English test is used for achievement, selection, and placement purposes. Since this test has heavily influences student’s academic futures, it is imperative to ensure that the test functions as intended and provides meaningful results. Therefore, the purpose of this study was to examine the technical quality of the CEPA-English test in relation to Forms A and B. This study evaluated 1) the psychometric properties of the CEPA-English test, 2) the extent to which DIF occurs, 3) the comparability of Forms A and B, and 4) the amount of information provided at the cutoff score of 150, which is the mean of the test in the NAPO study.

The study sample included 9,496 students for Form A and 9,296 for Form B, taken from the 2007 administration. The results for both Forms A and B test data revealed that the unidimensional 3PL IRT model provided a better fit at both item and test levels than the 1PL or 2PL models and the assumptions of the 3PL model were met. However, the property of invariance of item parameters was not strictly met for Form A and to some extent for Form B.

Overall, the analyses revealed that the CEPA-English test demonstrated good psychometric properties, since in both forms, the majority of the items were of moderate difficulty. In addition, items moderately discriminated between high-performing and low-performing students, and both forms showed a high internal reliability. Yet, it was also found that the test could be improved by eliminating items with negative discrimination and adding easier items to gain more precise information at the cutoff score of 150. In addition, the test developer may want to evaluate items that misfit the 3PL model. Finally, while DIF items were detected between males and females, and between Arts and Sciences students, nevertheless a significant proportion of DIF items were flagged by school type, which may indicate curriculum differences across private, public, and home schools. Therefore, the test developer could evaluate items with a medium and large DIF to determine whether to revise or eliminate them from Forms A and B of the CEPA-English test.

DEDICATION

To the endless sources of love and giving

My parents Fatima and Ali

Who

Instilled in me the value of an education

Who

Taught me the work should be useful and skillful

Who

Are behind everything I have ever achieved

And to my sisters and brothers

Who

Have always been there with love and support

PREFACE

In the Name of Allah, the Most Gracious, the Most Merciful

Praise be to Allah (God), the Cherisher and Sustainer of the World, for giving me the strength, patience, and ability to accomplish my PhD degree. Thank you Allah for all the blessings you gave me and will give me. Thank you for always being with me.

To my advisor and dissertation chair, Dr. Suzanne Lane, I am indebted for your constant support, encouragement, and guidance on this research and throughout my PhD study. Thank you for your constructive suggestions that resulted in creating work of which I am proud.

To Drs. Clement Stone, Feifei Ye, and Lauren Terhorst, thank you for serving on my committee. I am grateful to you for your support and valuable suggestions, which enhanced the quality of this research.

Special thanks go to Dr. Annie Brown, the head of Educational Assessment in the National Admissions and Placement Office in the UAE, for providing me with the data and information needed to make this research possible.

My deepest gratitude and love go to my parents. I am so grateful for all your unconditional love, sacrifices, and prayers. Thank you for your unwavering trust and belief in me, and for supporting the pursuit of my academic dreams. I will be forever thankful for what you did and do for me. I am blessed to be your daughter.

Many thanks and much love go to my sisters and brothers. Your love, prayers, and support encouraged me all the way. I never dreamed that I would have sisters and brothers like you.

To my family and friends, thank you for always being caring, loving, supportive, and encouraging.

Finally, to my government represented by the United Arab Emirates University (UAEU), I am grateful for the opportunity and financial support you provided to pursue my graduate study in the United States. I am looking forward to serving my country.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

1.1 STATEMENT OF THE PROBLEM

The Common Educational Proficiency Assessment (CEPA)-English test, a standardized paper-and-pencil exam, is designed to measure the English proficiency of 12th grade students. The CEPA-English test is administered once a year and has multiple forms used to ensure test security. The 2007 CEPA-English test, which is the focus of this study, had four forms (A, B, C, and D), consisting of 120 multiple-choice items—90 items in the Grammar and Vocabulary Section and 30 items in the Reading Section. The Grammar and Vocabulary Section consisted of 90 items: 10 parts-of-speech items, 40 grammar items, and 40 vocabulary items.

The CEPA-English test is the first national high-stakes test and one of the most important exams students take in their academic career in the United Arab Emirates (UAE). This test is mandatory for all 12th grade students seeking undergraduate studies at the UAE's higher education institutions. The CEPA-English test score serves three purposes. First, it is used for assessing the English achievement of 12th grade students who follow the Ministry of Education English curriculum, and it accounts for 25% of the students' overall General

Secondary Certificate (GSC) English grade.[1] Second, it is used as a basis for admission into three major higher education institutions: the United Arab Emirates University (UAEU), Zayed University (ZU), and Higher Colleges of Technology (HCT). Students must achieve a minimum average of 70% on the GSC exam in addition to a minimum score of 150 on the CEPA-English test in order to be eligible for the Bachelor programs at the three institutions and for the Higher Diploma programs at the HCT. Students who score below 150 are admitted to the HCT Diploma program. Third, it is used for placing students into the appropriate levels of English proficiency in the remedial program, prior to starting their programs across the three institutions.[2]

Because the CEPA-English test has serious consequences for students, it is crucial that the CEPA-English test score inferences for the 12th grade are valid. Until now, there has been no study that extensively validates the CEPA-English exam. Focusing on Forms A and B, this study evaluates: 1) the psychometric properties of the CEPA-English test, 2) the extent to which differential item functioning (DIF) occurs, 3) the comparability of two forms (A and B) of the test, and 4) the amount of information provided at the cutoff score of 150, which is the mean of the test in the National Admission Placement Office (NAPO) study.

Chapter 1 is divided into five sections. The first section begins with a description of the purposes of English language testing in an educational setting. It also illustrates the need for English language testing in the UAE's major higher educational institutions. A general

[1] The GSC is a public examination taken by all 12th grade students; it is held at the end of the first and second semesters. The GSC exam consists of various subjects, depending on the student's stream (Arts or Science). To pass this exam and move to the next grade level, students must get at least a score of 60 % on each subject, including English. In fact, students must pass the GSC exam to undertake higher studies at the university or college level (MOE, 2009).

[2] The remedial program is an academic program designed to prepare students with a limited educational background to undertake study at university.

description of the CEPA-English test is also provided. The second section assesses procedures used to examine the test’s technical quality along with a rationale for using these procedures: item response theory (IRT), the Mantel-Haenszel (MH) DIF detection procedure, and equipercentile equating with the cubic spline postsmoothing method. Finally, the last three sections present the purpose of the study, the research questions, and the significance of the study.

1.1.1 Purposes of Standardized English Language Testing

Tests are defined as instruments or systematic procedures for observing or sampling behavior. They are typically standardized when administered and scored in a consistent manner (Nitko, 2004). Standardized tests are vital tools for assessing student performance. The primary purpose of standardized tests in educational settings is to measure and compare student performance, to make various decisions about students’ behaviors, and/or to predict students’ behaviors based on their performance.

English language tests, in particular, are commonly used to place or assign students into appropriate courses according to the students’ language ability levels. The tests are also used to assess achievement in order to determine how well a student has acquired knowledge of the skills addressed in the tests. Finally, they are used to determine whether a student has reached a certain level of language proficiency (that is, in listening, speaking, reading comprehension, and writing) needed to perform successfully in future academic courses. Each of these uses, therefore, implies a somewhat different interpretation of the English language test scores (Alderson et al., 1995; Hughes, 1989).

English tests are used for making both low-and-high-stakes decisions. Tests are low-stakes if their outcomes are less likely to affect the students' academic futures. For example, such tests are used for placing students into appropriate level of English classes. On the other hand, tests are high-stakes if their outcomes are used to make important decisions regarding grade –to-grade promotion or graduation from high school; thus, high-stakes tests may impose serious consequences on students. Two important high-stakes English language proficiency tests are the Test of English as a Foreign Language (TOEFL) and the International English Language Testing System (IELTS) (Jones & Hargrove, 2003).The TOEFL and IELTS tests are used to assess the English proficiency of students who are non-native English speakers (Alderson et al., 1995; Hughes, 1989). These tests are also used to predict students' future academic success at college.

1.1.2 The Need for English Language Testing in Higher Educational Institutions in the UAE

All subjects taught in the UAE's public schools (from primary to secondary levels) use Arabic as a medium of instruction, except for English language classes. The English language is taught from Grades 1 to 12 (MOE, 2009; uaeinteract, 2009). Despite having twelve years of English education in school, the majority of 12th grade students do not have the sufficient level of English required to succeed at those higher educational institutions that use English as the primary language for academic instruction. The three institutions found that even though a large majority of 12th grade students apply for acceptance, few of these students are proficient in English. They also found that the GSC English exam scores *alone* were not adequate for placing students into appropriate preparatory English levels. This finding reflected the gap between English language skills required to succeed at the university level and the English curriculum taught in secondary schools and assessed by the GSC English exam. Therefore, each institution has developed its own English language placement test, which is administered at the beginning of the academic year to place students into appropriate preparatory English levels (Brown, 2008). For example, the UAEU offers a first-year developmental program through its University General Requirements Unit. The length of time students spend in the program depends on both their English entry levels and their rate of progress. Similarly, ZU provides the Readiness Program, and HCT offers the Foundations Program. In fact, the majority of all preparatory programs incorporate the English language in all area of the study; it is estimated that more than 30% of higher education resources and curriculum time is devoted to English language usage as a means for preparing students to work effectively at the college or university level (Brown, 2008).

The three institutions decided to collaborate on the development of a common English placement instrument that provides the following four advantages (Brown, 2008):

- Economic advantage in terms of test development and administration, since representatives from the three institutions work together to develop and nationally administer the CEPA-English test to all 12th grade students
- Coordination advantage in terms of having the National Admission Placement Office (NAPO) which uses the NAPO database to compile candidate information and to distribute the CEPA-English test results to the three institutions
- Administrative advantage in terms of having a single national test for all 12th grade students, which offers institutions prior knowledge about the applicants' English language proficiency levels and allows institutions to have an advanced plan

- Compilation advantage in terms of collecting significant data on students' English language proficiency levels across the three institutions using the CEPA-English test results.

1.1.3 The CEPA-English Test

The CEPA-English test began as a joint venture between NAPO and the three higher education institutions—UAEU, ZU, and HCT. The CEPA-English test was developed because of the need for an accurate and reliable English selection and placement test, since the GSC English exam scores *alone* were not adequate for placing students into appropriate preparatory English levels across the three higher education institutions. The CEPA-English test was administered for the first time in March 2002 to over 13,000 12th grade students who applied through NAPO for admission into the three institutions (NAPO, 2009, Brown, 2008).

Since 2007, the CEPA-English test has been used as an important requirement for admission into Bachelor's and Higher Diploma programs at UAEU, ZU, and HCT. To be eligible for these programs, applicants must achieve a minimum average of 70% on the GSC or equivalent exam, and a minimum score of 150 on the CEPA-English exam. Students who score below 150 on the test are eligible for the HCT Diploma program.[3] A CEPA-English test score of 150 assumes that a student has attained the minimum level of English proficiency to study at a college or university (NAPO, 2009).

Since 2007, the CEPA-English test has also been used as the second semester English

[3] Diploma program (2 years), Higher Diploma (3 years), and Bachelor program (4 to 6 years, depending on the college).

exam for all 12th grade students who follow the Ministry of Education English curriculum. The CEPA-English test score accounts for 25% of the students' overall GSC English grade (NAPO, 2009; Brown, 2007). Thus, the CEPA-English test's purpose has changed from a low-stakes placement test to a high-stakes achievement, selection, and placement test (NAPO, 2009; Brown, 2007).

The CEPA-English test has multiple forms. The 2007 CEPA-English test has four forms (A, B, C, and D) that were randomly distributed to the examinees. Forms A to C were administered in the morning and Forms C to D were administered in the afternoon. A student receives only one form and is required to complete the form in two- and-half hours.

As earlier stated, the CEPA-English test consists of 120 multiple-choice items—90 items in the Grammar and Vocabulary Section and 30 items in the Reading Section (see Appendix A). Of these 120 items, 115 items are unique to each section and form, while one set of five common items are in Forms A and B and another set of five common items are in Forms C and D. These two sets of five items, which only represent grammar and vocabulary domains, were used for the purpose of equating the four forms.

In 2007, the CEPA was administered to a total of 32,500 students with 74% of students achieving a score of 150, compared to 69% in 2006. This indicates that there was an improvement in the students' performance across the country (Brown, 2007).

1.1.4 Examining the Psychometric Properties of the Test using IRT

As Standard 13.2 states in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999), "In educational settings, when a test is designed or used to serve multiple purposes, evidence of the test's technical quality should be provided for each purpose" (p. 145). The CEPA-English exam is used for achievement, selection, and placement purposes. It is imperative, then, to ensure the technical quality of the CEPA-English test for these purposes. Using IRT, this study provides evidence for the use of the CEPA-English test as an achievement measure. IRT was used to evaluate the quality of Forms A and B of the CEPA-English items and the test as a whole. This evaluation included examining the psychometric properties of the test such as the difficulty of individual items and their ability to discriminate among persons of different abilities.

IRT is commonly used in educational measurement to analyze test data at the item level. IRT is a powerful statistical test theory that links observable examinee performance to items in a test to an unobservable trait(s) of interest via statistical models. More specifically, IRT consists of a set of mathematical models that use a latent trait (θ) and item parameters (difficulty, b_i, discrimination, a_i, and guessing, c_i) to predict the probability of a correct response to an item. IRT offers several important advantages over classical test theory (CTT), such as the properties of invariance of ability and item parameters as well as the estimation of different standard error of measurement (SEM) for each of the examinees' ability levels. Because of the invariance property, IRT provides a useful framework for solving a variety of measurement problems, including selecting items, creating an item bank, equating scores from different test administrations, and evaluating DIF (Embretson & Reise, 2000; Hambleton, 1993; Hambleton, et al., 1991; Hambleton & Swaminathan, 1985).

The three most commonly used IRT models for dichotomous items are the 1-parameter logistic model (1PL), the 2-parameter logistic model (2PL), and the 3-parameter logistic model (3PL). The application of IRT models requires checking if the assumptions of the IRT model are met in the observed data. These assumptions are related to dimensionality, local independence, non-speededness, and the form of the IRT model. The application also requires assessing model-data-fit since the use of an IRT model is valid *only* when the model fits the data. This includes assessing the fit of the model at both the item and test levels. Finally, it is important to assess the degree to which the invariance of item and ability parameters holds in the test data.

1.1.5 Detecting DIF using Mantel-Haenszel Procedure

It is crucial that items in a test are fair to all examinees and are not biased against any particular group. Because the CEPA-English test imposes serious consequences on 12th grade students, it is important to examine whether the items on this test exhibit DIF. DIF refers to items that function *differently* for subgroups of examinees of approximately *equal* ability. In other words, DIF analyses, by examining the extent to which items may have differential validity for subgroups of examinees, can thus help monitor the validity and fairness of a test. In fact, DIF can be an important indicator of irrelevant constructs that pose particular difficulty for one subgroup.

There are two types of DIF: uniform and nonuniform. The former occurs when there is no interaction between the ability level and group membership; the latter occurs when there is an interaction between the ability level and group membership (Swaminathan & Rogers, 1990). However, it is commonly acknowledged that nonuniform DIF does occur but at a substantially lower rate than uniform DIF (e.g., Gierl et al., 1999; Camilli & Shepard, 1994; Mazor et al., 1994). The most commonly used non-IRT method for detecting DIF in dichotomous items is the Mantel-Haenszel (MH) procedure (Swaminathan & Rogers, 1990).

The MH procedure is a more powerful test for detecting uniform DIF items (Rogers & Swaminathan, 1993; Lopez-Pina, 2001; Swaminathan & Rogers, 1990). This method does not require a large sample size, and it is relatively easy to perform with computer software. In addition, MH procedure provides effect size measures to interpret the magnitude of DIF and to determine whether DIF items are negligible (Rogers & Swaminathan, 1993). In this study, the MH procedure was used to examine whether any items on Forms A and B of the CEPA-English test exhibit DIF.

1.1.6 Equating using Equipercentile Procedure

The CEPA-English test is administered repeatedly each year which increases threats to test security. To ensure test security, NAPO uses multiple forms for the CEPA-English test. These forms are constructed on the same specifications, so that they are enhanced to be similar to each other in content and statistical characteristics. Although multiple test forms are carefully constructed, the forms differ somewhat in difficulty; therefore, scores from forms are not interchangeable without some type of equating. Equating adjusts for differences in difficulty among forms that are built to be similar in difficulty and content, so that the forms can be used interchangeably. Once forms are successfully equated, it should not matter which test form an examinee is administered; examinees would have the same expected scores, regardless of which form they receive (Kolen & Brennan, 2004).

The first step in the equating process involves selecting an appropriate equating design. Four data collection designs are commonly used in equating: (a) single-group design; (b) single-group design with counterbalancing; (c) random-group design; and (d) anchor-item design. After choosing the appropriate design, the second step is to select the statistical equating methods. Various equating procedures, including procedures based on CTT and IRT, have been utilized to maintain comparable test scores. CTT has three equating methods: mean, linear, and equipercentile. IRT has two equating methods: IRT true-score equating and IRT observed-score equating. The final step examines the standard error of equating to evaluate the amount of random error in equating (Kolen & Brennan, 2004).

The current study did not replicate the equating method that was used by NAPO because the NAPO method only used five common items in Forms A and B (administered in the morning) and a different set of five common items in Forms C and D (administered in the afternoon).Thus, A and B are not linked with C and D. NAPO used different common-items in the morning and afternoon to prevent the morning anchors from becoming compromised. In this study, equipercentile equating method under the random-groups design was used to equate Forms A and B of the CEPA-English test.

1.1.7 Assessing Test Information Function Using IRT

The CEPA-English test should be designed to provide the most information at the cutoff score of 150, which is used as a basis for admission into three major higher educational institutions (UAEU, ZU, and HCT). Therefore, it is essential to examine whether the test provides the most precise information at the cutoff score of 150, which is the mean of the NAPO test distribution. To do so, this study examined the amount of information at the cutoff score of 150 for Forms A and B using the 3PL IRT model.

To obtain the test information function (TIF) for Forms A and B, the item information function for each item at each ability level was summed. The amount of information at a θ level is inversely related to the standard error (*SE*) of the estimate (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985; Hambleton, et al., 1991). The higher the information function, the lower the *SE* will be, and hence, the more precise the test. Generally, higher TIF's and, consequently smaller *SE*'s are associated with longer tests with highly discriminating items and with tests composed of items with *b* values close to the examinee's true ability (Hambleton, et al., 1991; Hambleton, 1993).

1.2 PURPOSE OF THE CURRENT STUDY

The present study has four main aims related to providing evidence for the test's technical quality. The first aim is to examine the psychometric properties of Forms A and B of the CEPA-English test using IRT. The second aim is to examine whether Forms A and B exhibits DIF using the MH detection method. The third aim is to examine the extent to which the CEPA-English test scores are equivalent across Forms A and B by using the equipercentile equating method under the random-groups design. This also involves evaluating the quality of equating Forms A and B through examining the error associated with this design. Finally, the fourth aim is to examine the amount of information provided at the cutoff score of 150 for equated Forms A and B using IRT.

1.3 RESEARCH QUESTIONS

This study addressed the following research questions:

1. Do Forms A and B of the CEPA-English test data meet the assumptions of IRT?

   Examining the assumptions of IRT include: a) determining which IRT model (1PL, 2Pl, or 3PL) is the preferred model for each form of the CEPA-English test data; b) assessing the internal structure and unidimensionality of each form of the CEPA-English test; c) investigating whether the items of each form of the CEPA-English test are locally independent; d) investigating whether each form of the CEPA-English is non-speeded; and e) evaluating the extent to which examinees are guessing on items.

2. Does the preferred IRT model fit each item on Forms A and B of the CEPA-English test data?

   It was expected that the unidimensional 3PL IRT model would provide a better fit for each form of the CEPA-English test data than the 1PL or 2PL IRT models. Therefore, the study examines the extent to which the preferred 3PL IRT model fits each item on each form of the CEPA-English test data.

3. Does the property of invariance of item parameters hold true for Forms A and B of the CEPA-English test data?

   This question involves examining the degree to which the item parameter estimates of the 3PL IRT model are invariant across different samples of examinees for each form of the CEPA-English test.

4. Are there any DIF items on Forms A and B of the CEPA-English test?

   This question requires examining whether the items on each form of the CEPA-English test exhibits DIF between males and females, between study types (i.e., Arts and Science), and between school types (i.e., public, private, and home-schooled).

5. To what extent are the CEPA-English test scores equivalent across Forms A and B?

   This question involves examining the error associated with the equipercentile equating for Forms A and B under the random-groups design.

6. To what extent is the test information function for Forms A and B of the CEPA-English test maximized at the cutoff score of 150?

1.4 SIGNIFICANCE OF THE STUDY

This study examined the psychometric quality of Forms A and B of the CEPA-English test. The results of this study will not only show the quality of the CEPA-English exam, but will also allow decision makers in the UAE to evaluate the trustworthiness of the CEPA-English test scores as a major indicator of candidates' performances or abilities. Firstly, this study will provide the developers of the CEPA-English test with evidence on the technical quality of the test which will lead to the improvement of test design and use. Secondly, examining DIF on Forms A and B of the CEPA-English test will provide important information as to whether some items may be measuring an irrelevant construct in addition to English language proficiency. This will not only have important implications for 12th grade English curriculum and instruction, but it will also provide information for further item design. This study will also contribute to DIF studies on language testing which in turn will lead to the improvement of language testing design and construction. Third, equating Forms A and B using equipercentile equating will help in obtaining meaningful comparison of students' scores, as well as ensuring that students are neither advantaged nor disadvantaged for taking either Forms A or B of the CEPA-English test. Finally, examining the amount of information provided by Forms A and B items at the cutoff score of 150 will indicate whether test was designed to provide maximum information at the cutoff score of 150. This consequently will provide some information on whether items on the test match the purpose of the CEPA-English test as a basis for admission into higher education institutions

1.5 ORGANIZATION OF THE STUDY

The material in this study is organized into five chapters. Chapter one provides the reader with background information about the study and introduces the purpose of the study, the research questions, and the significance of the study. Chapter two provides some background on UAE education as well as information on English tests, including the CEPA-English test. This chapter also provides an overview of IRT, the assessment of dimensionality of dichotomous data using EFA and DIMTEST methods, DIF analyses using the MH procedure, and equipercentile equating method. Chapter three describes the overall design of the study, the data source, the testing instrument, the sample, the data collection procedures, and the statistical analysis procedures employed in this study. Chapter four provides findings of the study corresponding to each research question. The final chapter provides a summary and interpretation of the research findings along with implications of the major findings, limitations of the present study and recommendations for further research.

2.0 LITERATURE REVIEW

The purpose of this study was to examine the technical quality of Forms A and B of the CEPA-English test. Using IRT, the study first examined the psychometric properties of Forms A and B of the CEPA-English test. Using the MH DIF detection method, the study then examined whether any items on Forms A and B exhibit DIF. Afterwards, using equipercentile equating method, the study examined the extent to which the CEPA-English test scores are equivalent across Forms A and B. Lastly, using IRT, the study examined how much information provided at the cutoff score of 150 for Forms A and B.

To cover the necessary background for this study, the first section of the chapter provided the essential information on UAE education. The second section gives a general description of the TOEFL and the IELTS tests as well as the CEPA-English test. The third section offers an overview of IRT and the assessment of dimensionality of dichotomous data using EFA and DIMTEST methods. The fourth part is an overview of detecting DIF using MH Procedure, while the last section provides an overview of equipercentile equating method.

2.1 BACKGROUND OF THE UAE EDUCATION

2.1.1 General Information Regarding UAE

The UAE is one of the most developed countries in the Middle East and currently has one of the fastest growing economies in the world. The UAE is a federation of seven independent emirates (states): Abu Dhabi, Dubai, Sharjah, Ajman, Umm al-Qaiwain, Ras al-Khaimah, and Fujairah (see Figure 1). The UAE was formed on December 2, 1971, after gaining independence from Britain. Each of the emirates has its own ruler, which together form the Supreme Council of Rulers, with the emirates of Abu Dhabi and Dubai holding the positions of president and vice-president respectively. The total area of the UAE is about 83,600 square kilometers (32,400 square miles) (UAE Yearbook, 2009).

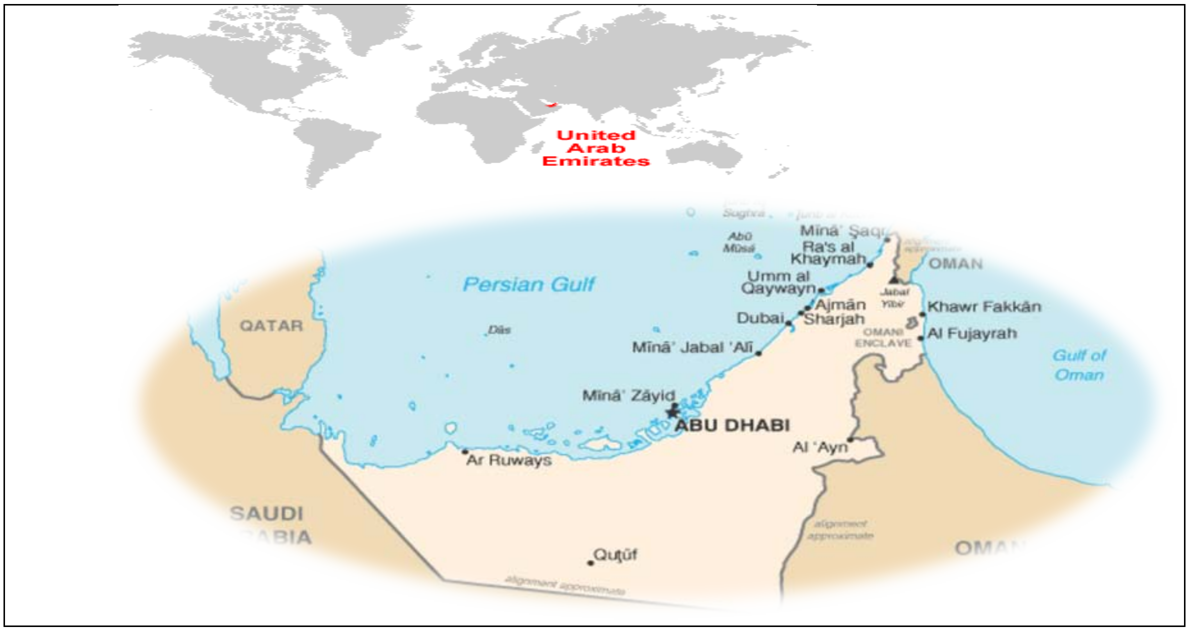


Figure 1. Map of the UAE (From: http://simple.wikipedia.org/wiki/United_Arab_Emirates)

The UAE has significant oil and gas industries, with reserves that are expected to last more than 150 years at present production rates; it is the world's fifth-largest oil producer and has nearly 9% of the world's proven oil reserves and almost 5% of the world's natural gas. However, the UAE government is committed to diversifying the economy and has also established strong manufacturing, agricultural, tourism, and service sectors (UAE Yearbook, 2009).

The three largest emirates are Abu Dhabi, Dubai and Sharjah. Abu Dhabi is the capital of the UAE and wealthiest of the seven emirates, owning the largest share of the oil and gas resources in the UAE—95% of the oil and 92% of the gas. Dubai is the second-largest emirate and the commercial center of the UAE, while Sharjah is the third-largest and declared as the "cultural capital" of the UAE (UAE Yearbook, 2009).

The population of the UAE is expected to be about 5.06 million at the end of 2009; UAE citizens account for less than 25% of the population, and the remaining come from other Middle Eastern countries, as well as from India, South East Asia, Europe, and America. Islam is the official religion of the country, but there are a significant number of Christians and Hindus. Arabic is the official language; however, English is widely spoken, particularly in government, businesses, and universities. Other languages include Hindi and Urdu, as well as Farsi, Pashto, and Malayalam (UAE Yearbook, 2009).

2.1.2 The UAE Education System

The UAE education system was initially established in 1971, and it has dramatically improved after the discovery of oil. The general objectives of the UAE education system are to offer equal educational opportunities to all students in all stages of education, to encourage a sense of self-worth in students, to pass on the heritage and goals of the nation, and to create an educated citizenry that will continue the development of the UAE civilization (MOE, 2009).

The current education system in the UAE is divided into three categories: public, private, and adult education. Education from kindergarten to the university level are free for all UAE citizens, including the primary and secondary adult education, but private education is not free (MOE, 2009; uaeinteract, 2009). All public schools are mandatory single-sex schools, whereas private schools are both coeducational and single-sex. About two thirds of private schools are single-sex, and less than one third are coeducational.

Education systems up to the secondary level are monitored by the Ministry of Education, which is the central authority holding the responsibility for administering all public and adult education programs as well as controlling the national curriculum for primary and secondary public education. Essentially, the curriculum of UAE schools is uniform across the nation, although school activities are somewhat different depending on the school level. In general, individual public schools do not have enough autonomy to decide which subjects are taught or even which teaching strategies are used; both the content as well as the number of periods for each subject is determined by the Ministry of Education (MOE, 2009; uaeinteract, 2009).

Private schools in the UAE operate under the licensing and supervision of the Ministry of Education. The Ministry of Education has a private education department to supervise private schools, providing the regulations, resolutions, and follow-up procedures for the implementation of national policy guidelines. For example, government policy stipulates that private schools must offer Islamic education, social studies, and Arabic as the core subjects for Arab students and as additional subjects for non-Arab students. Also, the curriculum of the private school must be approved by the Ministry of Education (MOE, 2009; uaeinteract, 2009). In fact, most private schools follow the curriculum of foreign countries such as India, France, Germany, United States (US), or United Kingdom (UK). Yet, there are some schools that follow the Ministry of Education with more of an emphasis on English and French language learning. In most private schools, the primary language of instruction is English, except for French and German schools (MOE, 2009; uaeinteract, 2009).

The current UAE public schooling system consists of four stages which cover 14 years of education: two years of kindergarten (4-5 year olds); nine years of primary school, which is divided into two levels—level 1 is a five-year program (6-10 year olds), and level 2 is a four-year program (11-14 year olds); and either three years of secondary school (15-17 year olds) or three years of vocational school (15-17 year olds) (see Figure 2). In addition to these schools, there is adult education program enabling individuals who do not complete the formal education requirement to demonstrate that they have acquired a level of learning comparable to high school graduates (MOE, 2009; uaeinteract, 2009).

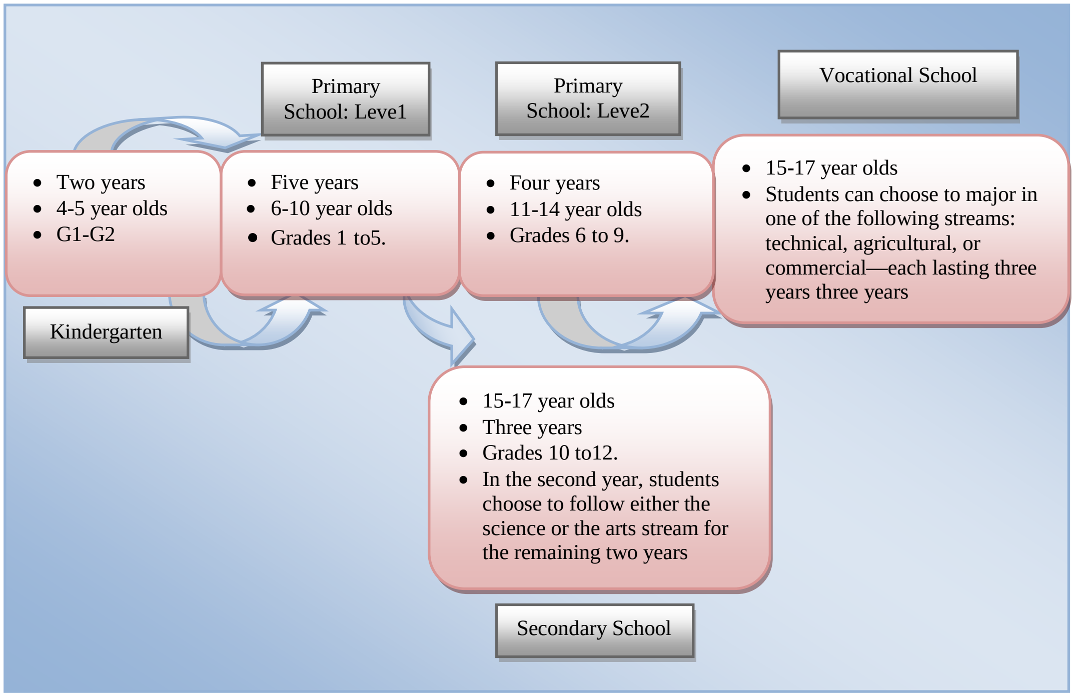

Primary School: Leve1
Primary School: Leve2
Vocational School
Two years
4-5 year olds
G1-G2
Kindergarten
Five years
6-10 year olds
Grades 1 to5.
Four years
11-14 year olds
Grades 6 to 9.
15-17 year olds
Students can choose to major in one of the following streams: technical, agricultural, or commercial—each lasting three years three years
15-17 year olds
Three years
Grades 10 to12.
In the second year, students choose to follow either the science or the arts stream for the remaining two years
Secondary School

Figure 2. Diagram shows the UAE public schooling system

At the first level of primary education (Grades 1 to 5), students typically spend five years learning basic skills and knowledge, which cover subjects such as Islamic education, Arabic, English, mathematics, science, social studies, computer science, and various activity subjects (art, physical education, and music).[4] At the second level of primary education (Grade 6 to 9), students typically spend four years learning the same subjects taught in the first primary level but with an increase in content and difficulty, as well as number of class periods. Additionally, music activity is dropped and social studies are divided into three separate areas: history, geography, and civics (uaeinteract, 2009).

[4] English and mathematics subjects are extended to the first three grades of all public schools in the UAE in 2003—before, these were taught in Grade 4 (MOE, 2009; uaeinteract, 2009).

The secondary education program lasts three years. In the first year, students follow the same curriculum: Islamic education, Arabic, English, history, geography, mathematics, physics, chemistry, biology, geology, computer science, and physical education. Once they complete the first year, they can choose to follow either the science or the arts stream for the remaining two years. Students in the arts stream take geology, history, geography, and economics; those in the science stream take physics, chemistry, and biology. In addition, all students in both streams take Islamic education, Arabic, English, mathematics, and physical education. However, if a student does not attend a secondary level program, vocational education is the other option. In vocational schools, students can choose to major in one of the following streams: technical, agricultural, or commercial—each lasting three years (uaeinteract, 2009). All subjects taught in public schools (from primary to secondary levels) use Arabic as a medium of instruction, except for the English language subject.

The school year starts in September and ends in early June, and it is divided into two terms. Student achievement is measured twice during each academic year, during the first and second terms. 50% mastery is required in each subject to pass the grade. A student who fails to attain 50% mastery in any subject matter is required to retake the test before the beginning of the next academic year. If the student fails to pass, s/he must repeat the same grade. A successful student is awarded a certificate and is promoted to the next grade. At the middle and end of the third year of the general and technical secondary schools, students must pass the General Secondary Certificate (GSC) exam to undertake higher studies at the university or college level. Students need to get 60% in each subject to pass on the GSC exam (MOE, 2009).

2.1.3 Changes in the UAE Education System

In order to bridge the gap between secondary school and university education, the UAE Ministry of Education has shifted its focus to computer literacy and academic skills required for successful performance in college. To this effect, the UAE Ministry of Education released a policy document "Education Vision 2020" outlining a strategy for education development in the UAE up to the year 2020 based on a five-year plan. The strategy aims to switch from "instruction-orientated education" to "self-education," creating a learning environment conducive to creativity and innovation, promoting computer literacy at high school, and promoting the learning of the English language from a primary school level (Rassekh & Thomas 2001).[5]

2.1.4 The UAE Higher Education Institutions

Public Higher education is free to all UAE citizens, and it is monitored by the Ministry of Higher Education and Scientific Research, which coordinates admissions to the higher education institutions, namely United Arab Emirates University (UAEU), Zayed University (ZU), and Higher Colleges of Technology (HCT)

Established in 1977, the UAEU is the first and largest degree-granting public university

[5] Specifically, the strategy aims to introduce the latest information technology at all school levels, including a computer for every 10 students in kindergarten, every five students in the first level of primary school, every two students in the second level of primary, and for every student in secondary school. Also, technology training program will be provided for teachers. Before 2006, courses in UAE University are taught in Arabic except in the faculties of Sciences, Engineering, and Medicine and Health Sciences, where they are taught in English (MOE, 2008; uaeinteract, 2008).

in the UAE. The UAEU is located in Al Ain, a city within the Abu Dhabi emirate. The university includes nine colleges: Food and Agriculture, Education, Humanities and Social Sciences, Medicine and Health Sciences, Science, Law, Business and Economics, Engineering, and Information Technology. UAEU offers a variety of specialist undergraduate and postgraduate programs, which are internationally accredited (UAEU, 2009).

ZU was established in 1998 and has two campuses in Abu Dhabi and Dubai. ZU has five colleges: Arts and Sciences, Business Sciences, Communication and Media Sciences, Education, and Information Technology. Furthermore, the university offers a number of undergraduate and graduates programs in business, health care, information technology, and education. ZU is fully accredited in the UAE as well as by the Middle States Commission on Higher Education in the US (ZU, 2009).

HCT was established in 1988 to provide post-secondary vocational education to UAE nationals. HCT is the largest institution of higher education in the UAE with 14 campuses across the country. HCT offers more than 90 programs in business, communication technology, education, engineering technology, health sciences, and information technology. HCT offers four different credential degrees—Diploma, Advanced Diploma, Higher Diploma, and Masters (HCT, 2009).

English is the medium of instruction in the three institutions (UAEU, HCT, and ZU), except in courses in Arabic and Islamic studies. Because of this, it is necessary to ensure that all students have achieved a certain level of English proficiency required for admission. Therefore, students who apply for a Bachelor's or Higher Diploma program at the three institutions must achieve a minimum score of 150 on the CEPA-English test in addition to an average of 70% on the GSC exam. Students who score below 150 on the test are eligible for the HCT Diploma program.

CEPA-English test scores are used by the three institutions to place admitted students in the appropriate level of English in the remedial program. The main purpose of this program is to boost students' English skills needed to perform successfully in future courses. Students may need to spend up to two years in the preparatory program before they are permitted to start with their coursework. At any time students obtain a score of 61 or above on the TOEFL internet-based test (iBT) or a score of 5.0 and above in the IELTS test, they will be exempt from the English remedial program.[6] In fact, these tests are used as an exit requirement for academic courses; both the UAEU and ZU require an IELTS or TOEFL score for students to proceed to undergraduate studies, while HCT requires an IELTS score for graduation.

2.2 AN OVERVIEW OF THE TOEFL AND IELTS TESTS

Standardized English proficiency tests are commonly used in the admission process to select qualified applicants (Angoff, 1971; Tracey & Sedlack, 1987). The TOEFL and the IELTS are the two most important standardized tests of the English language. These tests are used in university programs as a benchmark of English proficiency for entrance requirements. The TOEFL and IELTS test scores are also used internationally as assessment tools to reliably assess students' English proficiency (Alderson et al., 1995; Hughes, 1989).

[6] A score of 500 or above on the TOEFL paper-based test (PBT), or a score of 173 or above on the TOEFL computer-based test (CBT).

2.2.1 The TOEFL Test

The TOEFL test was developed by the Educational Testing Service (ETS) to measure the ability of nonnative speakers of English to use and understand North American English as it is used in college and university settings. The test is available in computer-based (CBT), paper-based (BT), and an internet-based test (iBT) format. The test includes a new “Speaking” section in addition to the “Listening”, “Reading” and “Writing” sections (ETS, 2009):

(PBT), and an internet-based test (iBT) format. The test includes a new “Speaking” section in addition to the “Listening”, “Reading” and “Writing” sections (ETS, 2009):

- The Speaking section includes six tasks that measure the ability to speak English in academic setting
- The Listening section measures ability to understand spoken English in colleges and universities
- The Reading section measures ability to understand academic reading material.
- The Writing section includes two tasks that measure the ability to write effectively for college and university course work.

2.2.2 The IELTS Test

The IELTS is a European and Australian English language test, jointly administered by the University of Cambridge ESOL Examinations, the British Council, and the International Development Program of Australia. The IELTS is available in two formats—Academic and General Training. The Academic exam is mainly used as an entry requirement to universities in an English-speaking country; on the other hand, the General exam is used as an entry requirement for immigration to Australia, Canada, or New Zealand (IELTS, 2009).

The IELTS test consists of four sections: Listening, Reading, Writing, and Speaking, which last around two hours and 45 minutes. There are 40 questions on the Listening section, which is divided into four parts, and it lasts, around half an hour to 40 minutes. The Reading section contains three reading passages, and each of the reading passages is approximately 700 to 800 words in length. There are 40 questions on the reading test, which last for 60 minutes. The Writing section consists of two essay tasks, each lasting 60 minutes. The writing tasks are on a variety of subjects. For one essay, students are asked to write a report describing information presented in the form of a graph, table or diagram. For the second essay, students are asked to write a response to a statement or question, with a minimum of 250 words. Finally, the IELTS Speaking section has three parts lasting around 15 minutes and is in the form of an interview—an interview between one candidate and one examiner (IELTS, 2009).

All candidates take the same Listening and speaking tests. However, the Reading and Writing tests are different in the Academic and General Training tests. The first three tests—Listening, Reading, and Writing—are administered via paper-and-pencil and must be completed in one day; the Speaking test, on the other hand, may be taken within a seven-day period before or after the other tests. Students receive two scores in the IELTS test; a band score of 1 to 9 for each individual section as well an overall band score of 1 to 9. Most universities and colleges in the UK, Australia, New Zealand, Canada and USA accept an overall band score of 6.0 or 6.5 for entry to academic programs of study (IELTS, 2009).

2.3 AN OVERVIEW OF THE CEPA-ENGLISH TEST

2.3.1 Test Development and Design

The CEPA-English test was developed by a group of English language specialists from three higher institutions (UAEU, ZU, and HCT) and NAPO, which is in charge of administering this test to all 12th grade students seeking higher education in the UAE. This test was initially developed because of the need for an English placement test, since the GSC English exam scores *alone* were not adequate for placing students into appropriate instructional levels. Essentially, the CEPA-English test began as an internal English placement exam for the HCT. The content of the test was tailored to the UAE region in an effort to avoid cultural biases inherent in international tests. After the modification, the CEPA-English test was then used as a common placement exam in all three institutions, and it was administered for the first time in March, 2002 to over 13,000 12th grade UAE students (NAPO, 2009, Brown, 2008).

The CEPA-English test is still used as a means for placing students into appropriate levels of English proficiency courses in the remedial programs. For example, there are three levels of English proficiency courses in the UAE University program (namely, the University General Requirements Unit or UGRU):

- Level 1: students at this level have a score between 150 and 164
- Level 2: students at this level have a score between 165 and 174
- Level 3: students at this level have a score between 175 and 184

Student who have a CEPA-English score of 185 and above are eligible to take the University IELTS exam, and if they pass it (5.0 or above), they do not need to take an English course in the

UGRU; otherwise, they will be placed in the Level 3 course. If students obtain a score of 61 or above on the TOEFL internet-based test (iBT) or a score of 5.0 and above in the IELTS test, they will be excused from the remedial program.

Since 2006 the CEPA-English test has been used as an important requirement for selecting applicants for bachelor and higher diploma programs. To be eligible for these programs, applicants must achieve a minimum average of 70% on the GSC exam and a minimum score of 150 on the CEPA-English exam. Students who score below 150 on the latter are eligible to enter diploma courses at HCT (NAPO, 2009).

At the beginning of the 2006-7 academic year, the content of the CEPA-English test was reviewed by the CEPA committee group. They reviewed test specifications, materials for writing and test construction, the existing bank of test materials and items, and item banking procedures. As a result of that review, the specifications were revised as well as the item writing and editing procedures, including the re-training of current item writers and training new writers recruited from the three institutions. In addition, new items were added, including a second writing task, which assessed functional writing in the form of a letter, and a fourth text in the reading section, which is an “authentic” text using non-prose layout. Furthermore, the process of revising the content of the CEPA-English test was amended to take into account the new use of this test as the 12th grade second semester English exam (NAPO, 2009).

Since 2007, the CEPA-English test has been used as the second semester English exam for all 12th grade students who follow the Ministry of Education English curriculum. The CEPA-English score counts for 25% of the student’s overall GSC English grade (NAPO, 2009).

The purpose of the CEPA-English test has changed from a low-stakes placement test to a high-stakes achievement, selection, and placement test. Raising the CEPA-English test to a high-stakes test means that it imposes serious consequences on students. This requires raising students' awareness about the importance of the test and providing them with opportunity to prepare for the test. Further, this requires the accumulation of validity evidence to support the uses of the test.

2.3.2 Preparing for the CEPA-English Test

To use the CEPA-English test as the 12th grade second semester English exam, NAPO and the Ministry of Higher Education established a "Professional Development Training Program," targeting secondary school English language teachers. The training program began in February 2006. This program aimed to help teachers improve their teaching skills and to raise students' English language skills and proficiency levels in preparation for the CEPA-English exam. Also, the program aimed to enable students to foster positive attitudes towards English as a means of communication, thus enhancing students' performance on the CEPA-English exam (Brown, 2007).

Furthermore, to help students prepare for the CEPA-English test, additional practice sample materials were distributed to students and teachers, including a mock exam. Practice materials for this exam were also available on the NAPO website (http://ws2.mohesr.ae/cepa/).

As of 2007, 12th grade students receive practice CEPA-English questions through the "CEPAlearn" Short Message System (SMS) program, which was developed by NAPO. The CEPAlearn program allows students to access the CEPA-English practice questions on grammar and vocabulary items via SMS on their mobile phones. The CEPAlearn program is available for the three months prior to the exam. Students need to register for the program, and they can take one practice test per day by sending an SMS to a specified number.[7] Each day, a practice test consisting of 10 multiple choice items (4 grammar, 1 word form, and 5 vocabulary) is available for downloading. Students receive immediate feedback on their answers, as well as summary statistics on their performance (their score out of 10, the average score, and the best score, time taken to answer the questions, average time) (NAPO, 2009).

Currently, NAPO provides an official CEPA-English test preparation book. The book includes ten units with exercises to improve student reading comprehension, writing, vocabulary and knowledge of grammar. It also includes four full practice tests with an answer key and a CD. Finally, the book provides helpful tips on improving student English skills, as well as on preparing for the CEPA-English exam (NAPO, 2009).

2.3.3 The Content of the CEPA-English Test

The CEPA-English exam is a paper-and-pencil test, lasts for two-and-half hours, and consists of three sections: Grammar and Vocabulary, Reading, and Writing (see Appendix A).[8] The Grammar and Vocabulary section, which lasts 45 minutes, consists of 90 multiple-choice items, including 40 multiple-choice grammar items, 10 parts of speech items, and 40 vocabulary items. The grammar items measure a student's ability to recognize common grammatical patterns in English; the parts of speech items measure knowledge of word forms in English; and the

[7] Registration is restricted to one mobile number per user, and the cost of using the service is only 18 fills per SMS (NAPO, 2009).

[8] NAPO is currently developing computer-based versions of the CEPA-English test. NAPO recently piloted the English version successfully at the UAE University, and subsequently will pilot the English versions in some public high schools. Once the program is working well in the pilot schools, it will be offered to other schools across the country (NAPO, 2009).

vocabulary items measure knowledge of common English vocabulary (NAPO, 2009).

The Reading section consists of three descriptive or narrative prose texts, and one non-prose text, and with a total of 30 multiple-choice items. This section measures the student's ability to understand academic reading material. The three prose texts are based on three general subjects: a simple descriptive passage on an everyday topic, a passage on social science or humanities, and one on science or technology. The Writing section consists of two tasks: Task 1 is an essay which requires expressing an opinion, and Task 2 requires writing a letter (known as a "functional writing"). Each task lasts 30 minutes. This section requires students to provide their point of view related to an assigned issue and then to employ reasoning and evidence to support their ideas using varied and accurate grammatical and lexical resources. The student's writing is evaluated on fluency and coherence, grammar, vocabulary, spelling, punctuation, and content (NAPO, 2009).

2.3.4 The CEPA-English Test Administration

The CEPA-English test is administered to all 12th grade students once a year in May 19; examinees take the test in the morning or afternoon. Students have *only one* opportunity to take the CEPA-English exam. The higher education institutions and NAPO collaborate with the Ministry of Education to administer the CEPA-English exam. NAPO distributes application forms to every school in September and October, and the school forwards them to NAPO. All 12th grade students who are applying to the three institutions complete the application. Each individual student receives a letter from NAPO through the school informing her/him of the date, time, and location of the test for which they have been scheduled (see Appendix A). In addition, all the dates, times and locations for all test sessions are announced in the local newspapers (NAPO, 2009).

Students receive their CEPA-English scores via SMS in early June. The scores are also reported to the Ministry of Education. In addition, the three institutions review the students' score records in NAPO. The scores will then be used by the three institutions to determine students' placement into diploma, higher diploma or bachelor's degrees, and then into appropriate English level courses (NAPO, 2009).

2.3.5 The CEPA-English Test Scoring

The CEPA-English test is administered once a year on May 19 and has multiple forms. Seven forms were developed for use in 2007—Forms A to D were used in the main administration on May 19. Each form consists of unique items and five common (or anchor) items that are used across the test forms to support equating of tests. The CEPA-English test is statistically equated to adjust raw scores for differences in difficulty among forms by placing scores from multiple test forms on the same IRT scale. By doing so, students with the same ability level should receive the same scaled score, regardless of which form of the test was taken. In this way, performance on the CEPA-English test can be compared across forms for different cohorts of students. Thus, the students' raw scores on the Grammar and Vocabulary section and Reading section are analyzed using the 3PL IRT model and are converted to a scaled scored ranging from 90 to 210 (Brown, 2007).

The assessment of the CEPA Writing section is scored by trained and accredited markers, who undergo retaining and re-accreditation every year via the on-line marking program that was developed in-house at NAPO. Each script is rated independently by two markers, and the scores are then analyzed using a Rasch analysis that account for raters' effects. Also, scripts flagged for disagreement between raters are graded by a third maker before being transformed it into scores. The CEPA Writing score is reported separately from the overall test scores, and it ranges from 1 to 6. A score between two points on the scale (e.g., 5.5, 4.5, 3.5, 2.5, and 1.5) can be reported (Brown, 2007).

The second semester CEPA-English exam is reported on a scale of 0-100, so that it can be combined with the other three components of the final grade (the Semester 1 exam and the Semesters 1 and 2 continuous assessments). To produce the second semester exam score, Writing scores are combined with the Multiple Choice Questions (MCQ) scores. Logit scores derived from IRT analysis were combined in the following proportions: MCQ 70%, Writing Task 1 20%, Writing Task 2 10%. These proportions were then converted to a 0-100 scale (Brown, 2007).

2.3.6 The CEPA-English Test Quality: Content Validity Evidence

Content related evidence refers to the extent to which the test provides an adequate and representative sample of the particular content domain that the test is designed to measure (AERA et, 1999). A group of English language specialists representing the three institutions along with the CEPA committee group evaluated the content validity of the test; they examined the extent to which each item of the CEPA-English exam represents the content and the level of English proficiency desired (Brown, 2007).

2.3.7 Institutional Analyses of the CEPA-English Test Data

Each institution conducted a study examining the success and failure rates of students entering with different levels of English proficiency. The result of the UAE University study, for example, found that between 15 and 20% of students did not pass or complete the preparatory English program. Results were compiled into a final report by NAPO, which stated that "the three tertiary institutions agree that a minimum score on CEPA of approximately 175-184, plus a writing score of at least 5.0 is an absolute minimum for students to be considered as direct entry students in undergraduate or Higher Diploma programs. The three institutions also agree that they would have difficulty preparing students with CEPA scores of less 140 for English medium further education" (Marsden, 2004, p. 28, as cited in Brown, 2008). As a result, the three institutions recommended using a minimum score of 150 on the CEPA-English test as a requirement for admission across the UAEU, ZU, and HCT. However, ZU found that the cutoff score of 150 as the minimum level of English proficiency was unlikely to have as much effect as in the UAEU (Brown, 2008).

2.4 AN OVERVIEW OF IRT

2.4.1 Basics Concept of IRT

IRT, also known as latent trait theory, is a powerful psychometric technique that is commonly used in education and psychological testing to analyze test data at the item level. IRT links observable examinee performance to items in a test to an unobservable trait(s) of interest via statistical models. More specifically, IRT consists of a set of mathematical models that use a latent trait (θ) and item parameters (difficulty, b_i, discrimination, a_i, and guessing, c_i) to predict the probability of a correct response to an item. The relationship between the examinee's item performance and the abilities underlying item performance is described by a nonlinear monotonic increasing function called an item characteristic curve (ICC). An ICC provides a graphical representation of the probability that examinees will answer an item correctly for given ability level. The shape of the ICC determines the mathematical function for the IRT model (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985; Hambleton, et al., 1991).

2.4.2 Advantages and Disadvantages of IRT

IRT was originally developed to overcome the problems associated with classical test theory (CTT). IRT offers substantial advantages over CTT for test construction. First, IRT is a test-free measurement in that the ability estimation is independent of the test items being administered; therefore examinees can be compared even though if they are administered different subset of items. Second, IRT is a sample-free calibration in that the item parameter estimates (a_i, b_i, and c_i parameters) are independent of the sample of examinees used to obtain the estimates; they do not depend upon particular characteristics of the examinees answering each item. These first two advantages are referred to as the invariance of ability and item parameters, which are the most important properties of IRT. Third, IRT focuses on individual items, rather than the entire test, as in CTT, and IRT has the capability of linking items and examinees on the same latent scale. Fourth, IRT provides a statistic that indicates the precision of each ability estimate. That is, different standard errors of measurement (SEM) can be estimated for each of the examinees' ability levels. Finally, because of the invariance property, IRT provides a useful framework for solving a variety of measurement problems that are difficult to solve in CTT. These include selecting items, building item banks, constructing new tests, equating scores from different test administrations, evaluating DIF, and using computer adaptive testing (CAT; Embretson & Reise, 2000; Hambleton, 1993; Hambleton & Swaminathan, 1985; Hambleton, et al., 1991). For example, selection of items in IRT is based on the amount of information each individual item contributes to the total score.

Despite the advantages of using IRT, it has three main limitations. First, IRT requires stronger assumptions about the data to which the model is applied than CTT. Additionally, IRT models are complex. The fit between the item response models and the test data needed to be examined because invariant of item and ability parameters will not holds if the IRT model does not fit the data (Hambleton, et al., 1991, p. 53). Finally, IRT requires large samples to achieve accurate and stable parameter estimates. For example, approximately 1000 subjects are often required as a minimum simple size to adequately estimate a three-parameter model (Kingston and Dorans, 1985). Hence, the successful application of IRT in analyzing and interpreting test results can be obtained, when assumptions are met, the item response model fits the data, and large sample sizes are used.

2.4.3 Dichotomous IRT Models

IRT models assume that an examinee's performance on a test can be predicted by one or more abilities; that the correct response to an item has a monotonically increasing relation with the abilities which is described by the ICC; and that a specific mathematical relationship exists between an examinee's performance, the examinee's ability, and test item parameters which is known as an item response function or IRF. IRFs represent the nonlinear regression of a response probability on a latent trait or ability (Hambleton, et al., 1991; Hambleton & Swaminathan, 1985).

The selection of the IRT model is based on the type of data, model fit and the assumptions of the models. A variety of IRT models have been developed for dichotomous items, which have only two possible response options (e.g., correct/incorrect). All dichotomous IRT models assume that at least one person parameter (ability), and one item characteristic (i.e., item difficulty) is related to a person's performance. Dichotomous IRT models are based on the number of parameters in the model. Three IRT models are commonly used for dichotomous items: one, two and three parameter logistic models. The 1-parameter logistic model (1PL) assumes that only one item parameter, item difficulty (b_i), interacts with an examinee's ability level to determine item performance; the 2-parameter logistic model (2PL) adds the discrimination parameter (a_i) to the 1PL model to describe the test item; and the 3-parameter logistic model (3PL) adds the psuedo-guessing parameter (c_i) to the 2PL model to describe the test item.

The 3PL model was developed in educational testing to extend the application of IRT to multiple choice test questions that may elicit guessing. It is likely that examinees who do not know the correct answer may guess the correct answer in multiple-choice items (Hambleton, et al., 1991). The 3PL model uses b_i, a_i, and c_i parameters to describe each item. The three parameter logistic (3PL) model is:

$$p_i(\theta) = c_i + \frac{(1-c_i)}{1+e^{-Dai\ (\theta-bi)}}, \qquad (2.1)$$

$pi(\theta)$ is the probability that an examinee with ability θ answers item i correctly, a_i is the discrimination parameter for item i, b_i is the item difficulty (location) parameter for item i; c_i is the psuedo-guessing parameter for item i, D is a scaling factor equal to 1.7 introduced to make the logistic function as close as possible to the normal ogive function, and e is a transcendental number that has the value 2.718 (Hambleton, et al., 1991).

The b_i parameter or location parameter sets the location of ICC relative to the θ scale on the horizontal axis. The value of b_i is indicated by the ability value at the point where the probability of a correct response is $(1 + c_i)/2$. Lower b values correspond to easier items whereas higher b values correspond to more difficult ones (Embretson & Reise, 2000).

The a_i parameter is the slope of ICC at the inflection point (the point where $\theta=b_i$). The a_i parameter is also called the discrimination parameter, which indicates how well an item distinguishes low ability examinees from high ability examinees. Items that are highly discriminating have steep slopes and can separate examinees into different ability levels more easily than items with less steep slopes. Thus, as the a_i parameter decreases, the curve gets flatter until there is virtually no change in probability across the ability continuum (Hambleton et al., 1991). The c_i parameter is the probability that an examinee with an extremely low ability level will get the item correct, and it is equal to the lower asymptote of the ICC. When c_i is equal to zero, the 2PL can be expressed as a special case of the 3PL (Yen & Fitzpatrick, 2006). Thus, the item parameters (a_i , b_i , and c_i) vary from item to item, and they determine the shape of the ICC

(Figure 3): a_i determines the steepness, b_i determines the placement on the horizontal axis, and c_i determines the lower asymptote.

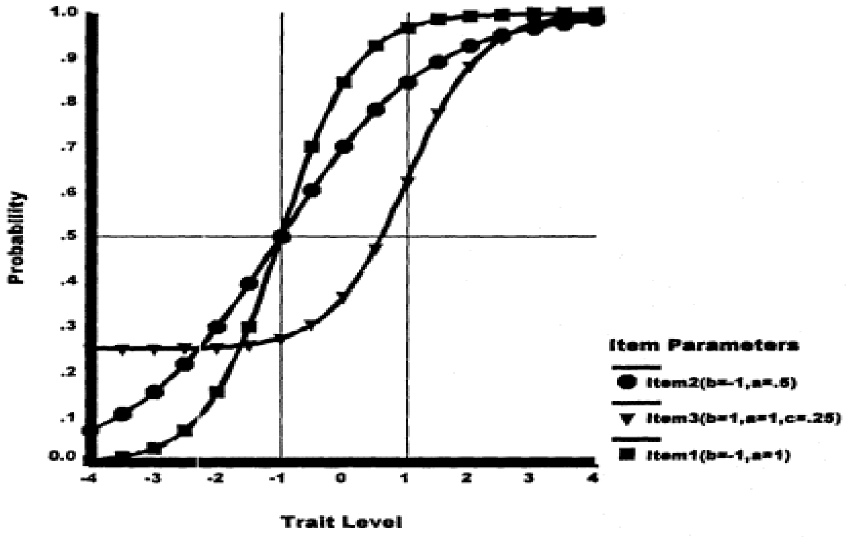


Figure 3. An example of ICCs for 3PL models

2.4.4 Examining the Assumptions Underlying IRT Models

In order for the IRT results to be valid, it is important to examine if the assumptions of the IRT model are met by the test items. There are four important assumptions underlying IRT models: 1) the number of dimensions that underlie examinee performance; 2) examinees' responses to the test items are independent; 3) the form of the IRT model is appropriate; and 4) the test is non-speeded (Hambleton, 1993; Hambleton & Swaminathan, 1985; Hambleton, et al., 1991).

2.4.4.1 Assumption of Unidimensionality

The first assumption of IRT is that an examinee’s performance can be predicted in relation to one or more underlying dimensions. Nearly all of the common IRT models assume unidimensionality, which requires that only a single underlying ability (θ) be measured by test items. According to this assumption, only one dominant factor is sufficient to account for examinee test performance (Embretson & Reise, 2000).

Assessing the dimensional structure of a test is important as it provides empirical evidence regarding the internal test structure. The assessment of dimensionality can support the number of scores to be reported (i.e., total score or subscores). For example, if there are two distinguishable dimensions (e.g., algebra and geometry), then it is appropriate to report two subscores. But when there is only one dominant dimension, a single total score is appropriate (Haladyna, 2004). In addition, the assessment of dimensionality is useful to ensure accurate evaluation of a test scoring method and related issues, such as equating and DIF (Nandakumar & Ackerman, 2005; Stout, 1987). Furthermore, unidimensionality is important for accurately interpreting test scores. If a score is composed of more than one dimension, it is difficult to determine what is contributing to the score. The validity of score interpretations are jeopardized if there is an irrelevant factor being measured in addition to the target factor. Finally, researchers indicate that it is essential to test the assumption of unidimensionality prior to examining model-data fit. If unidimensionality is violated, the results of other tests are difficult to interpret, and the estimation of item and ability parameters could be biased (Hambleton & Zaal, 1991; Hattie, 1984; Lord, 1980).

It is important to point out that the unidimensionality assumption cannot be strictly satisfied because several cognitive and non-cognitive factors affect test performance. These factors include level of motivation, test anxiety, speed of performance, test sophistication, and other cognitive skills (Hambleton, et al., 1991). Unidimensional models are often chosen despite this problem. This is because their parameter estimations are less complicated than multidimensional models, which take into account other factors, such as motivation and test anxiety (B´eguin & Glas, 2001).

There are two major methods for investigating the assumption of unidimensionality: parametric and nonparametric. The former includes linear and nonlinear factor analyses and the latter includes the DIMTEST procedure. These procedures are discussed in more details in the next section of chapter two.

2.4.4.2 Assumption of Local Item Independence

The second assumption of IRT is local item independence. IRT assumes that item responses are conditionally independent, or an examinee's responses to different items on a test are statistically independent, after controlling for the examinee's latent ability. Local item independence specifies that only the examinee's ability and the characteristics of test items influence test performance (Hambleton & Swaminathan, 1985). For this assumption to be true, an examinee's performance on one item *must not* affect his/her responses to any other item on the test. The assumption of local item independence does not imply that items are not correlated across all examinees, but only that there is no relationship among item scores at a fixed ability level (Hambleton & Swaminathan, 1985; Lord & Novick, 1968). This means that the ability specified in the model is the only factor influencing an examinee's performance on test items.

2.4.4.2.1 Assumption of Local Item Independence

The second assumption of IRT is local item independence. IRT assumes that item responses are conditionally independent, or an examinee's responses to different items on a test are statistically independent, after controlling for the examinee's latent ability. Local item independence specifies that only the examinee's ability and the characteristics of test items influence test performance (Hambleton & Swaminathan, 1985). For this assumption to be true, an examinee's performance on one item *must not* affect his/her responses to any other item on the test. The assumption of local item independence does not imply that items are not correlated across all examinees, but only that there is no relationship among item scores at a fixed ability level (Hambleton & Swaminathan, 1985; Lord & Novick, 1968). This means that the ability specified in the model is the only factor influencing an examinee's performance on test items.

2.4.4.2.2 Violation of the Local Item Independence Assumption

Local item dependence (LD), or violation of the local item independence assumption, occurs when examinee's response to an item depends not just on his/her ability level but also on his/her response to one or more other items in the test (Embretson & Reise, 2000). Therefore, the inclusion of items with LD may result in inaccurate estimation of item and person parameters (Tuerlinckx & De Boeck 2001; Chen & Thissen 1997; Ackerman 1987); overestimation of reliability and test information functions (Lee, 2004; Sireci, Thissen & Wainer 1991; Thissen, Steinberg & Mooney, 1989; Wainer & Lukhele, 1997); and introduction of additional dimensions (Wainer & Thissen, 1996), which violates unidimensionality.

There are a variety of possible causes of LD include unintended ability dimensions (e.g., verbal ability with math word problems) that are measured, external assistance or interference with some items, fatigue, speededness, practice, familiarity with item or response format (e.g., multiple-choice versus constructed-response), subpopulation membership (e.g., DIF among gender subpopulations), or scoring rubric or raters. Other sources include unmodeled item interactions: order of item presentation, items that are "chained" or organized into steps, items that share the same rubric, and reading comprehension items that share the same passage (i.e., testlet) (Yen, 1993).

Therefore, the problem of LD among passage-based test items must be addressed when using English language tests. One common approach to handle this problem with dichotomously scored data is to treat the items that share a common passage as a testlet and then fit the data to a polytomous IRT model (Thissen et al., 1989; Wainer, 1995; Wainer & Lewis, 1990; Wainer et al., 1991).

2.4.4.2.3 Modeling Testlet Dependencies

A testlet, which refers to an aggregation of items related to a single content area that is developed as a unit, comprises items that may or may not be locally dependent (Wainer & Kiely, 1987). For example, a reading passage on the CEPA-English test and its associated items (e.g., 3 or 4 items) could be construed as one testlet. Passage-based test items could consist of several such testlets (Thissen, et al., 1989). Although the use of testlets can help increase testing efficiency, it more likely violates the local item independence assumption of IRT (Wainer, Bradlow & Wang, 2007). This is because item responses within a testlet are not entirely independent, but are instead highly related through a common stimulus (Rosenbaum, 1988).

In using a polytomous IRT model to score testlets, the local item independence assumption of IRT holds across testlets, because the testlet is modeled as a unit (i.e., polytomous item). However, one shortcoming with this approach is that the information contained in the pattern of item responses is discarded since the testlet score is represented by the total number of correct items.

The graded response model (GRM; Samejima, 1969), a polytomous IRT model, can be used to score testlets using MULTILOG. The GRM is used when item responses can be characterized as ordered categorical responses (e.g., Likert-type rating scales). The GRM is an extension of the 2PL logistic model since it models K-1 separate 2PL models:

$$P_{ik}^{*}\ (\theta) = \frac{e^{a_i\,(\theta - B_{ik})}}{1+e^{a_i\,(\theta - B_{ik})}}, \qquad (2.2)$$

where a_i is the slope parameter that is introduced to reflect the relationship between the item and the trait being measured. B_{ik} is the K-1 between each the category threshold parameter that specifies the point on the θ axis or the level of θ at which an individual has a 0.5 probability of responding in category k or higher.

The GRM does not provide direct predictions of score responses, but the conditional probability of scoring at particular score levels is obtained by subtracting adjacent conditional cumulative probabilities using.

$$P_{ik}\ (\theta) = \ P_{ik}^{*}\ (\theta) - P_{ik+1}^{*}\ (\theta)\ , \qquad (2.3)$$

where $P_{\mathrm{ik}}(\theta)$ is the probability of responding in category k, $P^*_{ik}(\theta)$ is the cumulative probability of responding in category k or higher, and $P^*_{ik+1}(\theta)$ is the probability of responding in the next category or higher.

2.4.4.2.4 Local item Independence and Unidimensionality

Some researchers have argued that for unidimensional models, the assumption of local item independence is equivalent to assumption of the unidimensionality—items found to be locally dependent will appear as a separate dimension in a factor analysis (Hambleton & Swaminathan, 1985; Lord, 1980; Lord & Novick, 1968). Therefore, when the assumption of unidimensionality is true, the assumption of local item independence holds. Stout (1987, 1990) stated that a test is essentially unidimensional if covariances between items conditional on the ability are approximately zero. Thus, unidimensionality is obtained if responses to items are locally independent and a single latent trait accounts for the relationship between the items.

2.4.4.2.5 Examining the Assumption of Local Item Independence Using Statistical Indices

Several statistical indices can be used to examine local item independence assumption. These indices include Yen's (1984) Q_3, Pearson's chi-square test (χ^2), the likelihood ratio test (G^2), standardized Φ coefficient difference (Φ_{diff}), and standardized log-odds ratio difference (τ). These indices are based on a process that involves examining the residual covariation between pairs of items after fitting an IRT model to the data (Embretson & Reise, 2000).

2.4.4.2.5.1 Yen's Q_3 Statistic

Q_3 is the most commonly used index for detecting LD. Q_3 is the correlation of the residuals for a pair of items after partialling out the ability estimate (θˆ). To calculate Q_3, first, the expected examinee response to each test item is calculated by using item parameters and estimated ability levels. Next, for each examinee and each item, the difference between expected and observed item performance is calculated as:

$$d_{ik} = u_{ik} - \hat{p}_i(\hat{\theta}_k), \qquad (2.4)$$

u_{ik} is the score of the k^{th} examinee on the i^{th} item, and $\hat{P}_i\left(\hat{\theta}_k\right)$ is the probability that an examinee with θ level will answer the item correctly. Finally, the Q_3 statistic is calculated by correlating the residual scores among item pairs and can be expressed as:

$$Q_{3ij} = r_{didj}\,, \qquad (2.5)$$

r_{didj} is the correlation between examinees' deviation scores from the two items. If local item independence holds between any pair of items, then the expected value of Q_3 should be equal to -1 / (n -1), where n is the number of items on the test. As n increases, the value of Q_3 is expected to be around zero, thus, a large positive value of Q_3 indicates that item pairs that share some other factor may be a cause of concern (Embretson & Reise, 2000). However, the Q_3 method has two problems. One problem is that assumptions of linearity and bivariate normality among the residuals may not always be met. Another problem is that the empirical Type I error rates for the Q_3 statistic are higher than the nominal Type I error rates (Chen & Thissen, 1997).

2.4.4.2.5.2 Chen and Thissen's Chi-Square LD Statistic Indices

In addition to the Q_3 statistic, Chen and Thissen (1997) proposed four indices (i.e., χ^2, G^2, Φ_{diff}, and τ) for identifying LD. To compute these four LD indices, a 2 x 2 table for each pair of test items with binary responses and across all examinees is formed as follows:

Table 1. A 2 x 2 Table for Each Pair of Test Items with Binary Responses and across All Examinees

		Item Y	
		0	1
Item X	0	O_{11}	O_{12}
		E_{11}	E_{12}
	1	O_{21}	O_{22}
		E_{21}	E_{22}

In this 2 x 2 table, 1 and 0 represent the correct and incorrect responses, respectively; *O* represents the observed frequency, and *E* represents the expected frequencies under the IRT model. The expected frequency for a correct response for one item pair (item *i* and *j*) is the integral of the product of the ICCs or trace lines for both items and the standard normal curve (Φ (θ)):

$$\boldsymbol{E_{ij}} = \int \boldsymbol{p}\,(\boldsymbol{u} = \boldsymbol{i} \mid \boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})\ \boldsymbol{p}\,(\boldsymbol{u} = \boldsymbol{i} \mid \boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) \boldsymbol{\Phi}(\boldsymbol{\theta}) \boldsymbol{d\theta}\,, \qquad (2.6)$$

These expected frequencies are predicted from the IRT model using maximum marginal likelihood (MML) estimation. Then, observed and expected frequencies for a pair of items are used to compute the four LD indices. Pearson χ^2 is computed as follows:

$$X^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(O_{ij} - E_{ij}\ )^2}{E_{ij}}, \qquad (2.7)$$

Likelihood ratio test (G^2) is computed as follows:

$$G^2 = -2\ \sum_{i=1}^{2} \sum_{j=1}^{2} O_{ij}\ \ln\left(\frac{E_{ij}}{O_{ij}}\right), \qquad (2.8)$$

Standardized $\boldsymbol{\Phi}$ coefficient difference (Φ_{diff}) is computed as follows:

$$\Phi_{diff}\ = \frac{\Phi_{obs} - \Phi_{exp}}{\sqrt{var\ (\Phi_{obs}\ )}}, \qquad (2.9)$$

Standardized log-odds ratio difference (τ) is computed as follows:

$$\tau = \frac{\tau_{obs} - \tau_{exp}}{\sqrt{\sum_p \sum_q {}^{1}/_{O_{pq}}}}, \qquad (2.10)$$

These four LD indices compare the observed and expected frequencies for pairs of items by examining the covariation between items that is not accounted for by the IRT model. For example, when no LD exists between a pair of items, then the covariation of the observed and expected frequencies for the pair of items should be approximately the same, indicating perfect independence between the pair of items. These indices (i.e., χ^2, G^2, Φ_{diff}, and τ) have three important advantages over the Q_3 index. The first advantage is that they do not require using the θ estimates and correlation between residuals, as Q_3 index does, instead they used the χ^2 table. They also require much less computing time than Q_3. Finally, they can be computed from a selected subset of items, unlike Q_3 which requires the whole set of data to compute θ.

Pearson's χ^2 and G^2 indices tests for independence are distributed normally as χ^2 with degrees of freedom equal to one. However, the inclusion of the slope or discrimination parameters results in what “may be described as the loss of a fraction of the one degree of freedom for the test of independence” (Chen & Thissen, 1997, p. 269).

The Φ_{diff} and τ indices are expected to be distributed normally with a mean of zero and a standard deviation of 1. The Φ_{diff} and τ indices expected values are zero under the null hypothesis of LD (Chen & Thissen, 1997). The main advantage of Φ_{diff} and τ indices over χ^2 and G^2 statistics is that they have signs to indicate the direction of association. A positive value of Φ_{diff} and τ indicates greater dependence of the observed frequencies than the IRT model predicts, while a negative value indicates less dependence of the observed frequencies than the model predicts. Despite this advantage, the primary shortcoming with the Φ_{diff} and τ indices is that they are undefined when some of the cells have zero observed frequency. Specifically, the Φ_{diff} index is undefined when both cells of the same row or column have zero observed frequency, while the τ index is undefined when any of the cells have zero observed frequency. In contrast, G^2 is well defined with empty observed cells—where the contribution of a zero cell to G^2 is defined as zero (Chen & Thissen, 1997).

Chen and Thissen (1997), in their simulated data, have evaluated χ^2, G^2, and Q_3 indices for detecting LD. The authors investigated the distribution and power of these statistics under two conditions: the null condition of LD and the conditions in which LD is introduced. The results show that under the null condition of LD, both the χ^2 and G^2 indexes have distributions very similar to the χ^2 distribution with one degree of freedom, and that the G^2 is slightly more powerful than the χ^2. Under the LD conditions, both the G^2 and χ^2 indices were extremely sensitive in detecting LD and multidimensionality among items. Results further show that the Q_3 index tends to outperforms both the χ^2 and G^2 LD. Finally, Chen and Thissen (1997) point out that “Any meaningful interpretation of the LD indexes requires skill and experience in IRT analysis and close examination of the item content. Examination of the pattern of the LD indexes across item pairs is as important as the magnitude of any single LD index” (p. 288).

2.4.4.3 Assumption of the Appropriateness of the IRT Model

The third assumption of IRT is that the form and shape of the IRT model, which is defined by the item parameters, is preferred. IRT models assume that an examinee’s performance on a test can be predicted by one or more abilities. They also assume that there is monotonicity relationship between ability and performance. That is, as the examinee’s ability (θ) increases, the probability of correctly answering an item increases. The relation between examinees’ item performance and the ability is modeled by ICCs. An ICC reflects a nonlinear relation for the regression of item score on the ability measured by the test. The shape of the ICC determines the mathematical function for the IRT model. IRT models assume that there is a specific mathematical relationship between an examinee’s performance, the examinee’s abilities and test item parameters (Hambleton, et al., 1991; Hambleton & Swaminathan, 1985).

2.4.4.3.1 Examining the Assumption of Appropriateness of the IRT Model

Checking the form and shape of the IRT model involves comparing models with different numbers of item parameters. In other words, the model to be used (the 1PL, 2PL, or 3PL) is determined statistically by the minimal number of parameters that offer a maximal amount of information via likelihood ratio (Hambleton & Swaminathan, 1985; Hambleton, et al., 1991). For example, to assess if the addition of extra parameters add significantly more information, a researcher can compute the significance of the difference between -2log likelihood ratio for 2PL versus 3PL models. The difference between the statistics for 2PL versus 3PL models, distributed as chi-square, is used to evaluate the significance of specific additional parameters in improving model-data-fit. Statistically comparing two nested models yields a difference chi-square (G^2) with degrees of freedom equal to the number of additional parameters that are estimated. Then, the simplest model that offers the most information is chosen. A large and significant difference between the two 2PL and 3PL models helps the researcher select the preferred model, but it does not indicate if the selected model would fits the data. This can be determined by using a chi-square goodness-of-fit statistic to compare the correspondence between model predictions and observed data.

2.4.4.4 Assumption of Speededness

The last assumption of IRT is non-speededness. Speededness refers to testing situations where the time limits on a standardized test, such as the CEPA-English test, do not allow substantial numbers of examinees to answer all questions. As a result, examinees may either rush through the questions, skipping items they fail to reach or do not have the ability to answer, or randomly guess on items, usually at the end of the test. Hence, in a speeded test, it is assumed that examinees may omit items at the end of the test due to the time limit, not to their limited ability. If examinees fail to answer test items because of time limit rather than their limited ability, then two dominant factors influence their test performance: speed and ability. In this case, when speed does affect test performance, the unidimensionality assumption is essentially violated since the ability measured by a test is not the only factor impacting test performance. The local independence assumption of the IRT models is also being violated.

Furthermore, in the presence of test speededness, ability parameter estimates for end-of-test items are often under-estimated, and item difficulty parameters for end-of-test items are often over-estimated (Douglas et al., 1998; Oshima, 1994). For example, in his simulated data using the 3PL model, Oshima (1994) examined the effect of failing to answer not-reached items at the end of tests and the effect of randomly responding to late items; both effects represent the two possible types of behavior caused by time limit in speeded tests. Oshima found that when omitted responses were treated as incorrect, both *a* and *b* parameters were overestimated, while *c* parameter was underestimated for the items toward the end of a speeded test. However, the ability estimates were least affected by the speededness of the test in terms of correlation between true and estimated ability parameters.

Because of the systematic pattern of omitted responses among end-of-test items in speeded tests, these omitted responses cannot be classified as missing at random. The amount of a pattern of missing data may be indicators of whether the test is speededness. Thus, speededness introduces a severe threat to the validity of interpretations of the test scores if the test is not intended to measure the speed of examinee in responding to the items. Therefore, it is important to check whether the non-speededness assumption is met on test data.

2.4.4.4.1 Examining the Assumption of Speededness

To check for the non-speededness assumption, a researcher can compare the variance of the number of omitted items to the variance of the number of items incorrectly answered. If the ratio of the two variances is close to zero, then the assumption of non-speededness is met (Gulliksen, 1950). Another way to examine speededness is to compare the percentage of examinees completing the total test, the percentage of examinees completing 75% of the test, and the number of items completed by 80% of the examinees. If nearly all examinees complete nearly all of the items, speed is not an important factor in test performance. Another way to examine this assumption is by comparing item parameter estimates for groups of examinees tested under a specific time limit and without a time limit. If item parameter estimates for each group are similar to each other, the assumption is met (Hambleton, et al., 1991).

2.4.5 Assessing Model Data Fit at Test and Item Levels

2.4.5.1 Model-Data-Fit at Test Level

The evaluation of model-data-fit is an important step in choosing the appropriate IRT models since the advantages of IRT can be attained *only* when an IRT model fits the test data. Hambleton and Swaminathan (1985) suggested three ways to evaluate the model-data-fit: a) validity of the model assumptions for the test, b) accuracy of the model predictions using real and simulated data, and c) the degree to which the expected properties of the IRT model (i.e., invariance of item and ability parameters) are obtained.

In order to assess the IRT model-data fit, a researcher should examine several possible sources of misfit. Lack of fit may occur for different reasons. First, failure to meet underlying IRT model assumptions, such as dimensionality, local independence, and monotonicity may result in having some items that are more likely not to fit the model. Second, misfit may occur because of failure to achieve invariant item and ability parameter estimates. Third, failure to select an appropriate IRT model may result in item misfit. Finally, other misfit reasons include failure to obtain a large enough sample size, nonmontonicity of item-trait relations, or poor item construction (Embretson & Reise, 2000; Hambleton, 1993; McKinley & Mills, 1985). Thus, validating the use of an IRT model requires checking each of these possible sources of misfit.

2.4.5.1.1 Methods for Assessing Model-Data-Fit at Test Level

Assessing the IRT model-data fit involves evaluating the degree to which the model predicts the observed data. The fit of the IRT model can be examined at the test and item levels (Hambleton & Swaminathan, 1985; Hambleton, 1989; Hambleton, et al., 1991). For example, IRT model-data fit can be assessed at the test level by comparing observed and expected total score distributions, after fitting different IRT models to the data set, using the chi-square goodness-of-fit statistic and/or using a graphical representation (Hambleton & Swaminathan, 1985; Hambleton, et al., 1991).[9]

[9] Expected total score distributions are obtained using item parameter estimates and assumed ability distribution.

Another way to assess the overall fit of the model to the data is by comparing expected values for item response patterns. This includes examining all possible response patterns for *n* items where there are only 2^n response patterns. In this procedure, observed frequencies are counts of individuals with particular response patterns. The joint likelihood of response patterns for the total scores is calculated by multiplying the conditional probabilities associated with the score levels for items. Expected frequencies are based on the likelihood function and item parameter estimates. Then, observed and expected frequencies are compared using the likelihood G^2 test. The G^2 statistic is tested for significance by comparing it to the χ^2 distribution with degree of freedom (*df*) = $2^{\boldsymbol{n}}$- *kn* - 1, where *k* equal to number of item parameters in the model. The G^2 test is given by the formula:

$$G^2 = 2\sum_{r=1}^{2^n} f_r \log \frac{f_r}{NP\ (U_r)}, \qquad (2.11)$$

U_r is the r^{th} response pattern, f_r is the observed frequency, *N* is sample size. However, it is not possible to use G^2 test for assessing the fit of model at test level when *n* is large ($n \leq 10$).

2.4.5.2 Model-Data-Fit at Item Level

In addition to examining the overall fit of the model to the data, it is also possible to examine the fit of item. This can be done by comparing observed and predicted score distributions across a range of discrete ability levels for each item (Hambleton & Swaminathan, 1985; Hambleton, et al., 1991). The assessment of item fit is an important complement to the overall fit of the model. Even when the test, as a whole, fits the model, some of the items may misfit the model. Therefore, it is important to examine the fit for individual items. The examination of item fit could be used to measure test dimensionality, which affects the validity of the test results (Reise, 1990); to indicate errors occurring in the calibration of the items, which determines the validity of the items' parameters; and to detect a variety of measured disturbances such as guessing and DIF (Smith, 1991). Assessing item fit is useful in identifying item that need to be deleted or revised which will improve overall model-data fit for the test.

2.4.5.2.1 Methods for Assessing Model-Data-Fit at Item Level

2.4.5.2.1.1 Goodness-of-Fit Statistics

There are two general methods for assessing item fit, namely goodness-of-fit statistical tests and graphical representations (Embretson & Reise, 2000). Goodness-of-fit methods involve comparing the observed and expected score distributions, and this typically involves creating a two-way contingency table for each item, where the rows of the table correspond to ability subgroups (θ) and the columns correspond to score response categories. This can be done in five steps. First, the parameters of an IRT model for a set of items are estimated, and then examinee ability (θ) levels are estimated based on these item parameters. Second, examinees are then sorted by their θ estimate, and divided into a number of θ subgroups (e.g., 10) based on their ranking. Third, the observed score response distribution across score categories for a specific item is constructed by cross-classifying examinees to one cell of a two-way table using their θ estimate and their score responses. Fourth, the expected score response distribution across score categories for a specific item is obtained using the IRT model to predict the number of examinees who should fall into each of the score categories. This prediction is obtained using estimated item parameters and θ level representing the discrete ability subgroups (e.g., midpoint of subgroup). Finally, the observed and expected score response distributions are compared using a chi-square goodness-of-fit statistic or a graphical representation (Embretson & Reise, 2000). The observed and expected score response distributions are typically summarized in an item fit table.

2.4.5.2.1.1.1 Traditional IRT Goodness-of-Fit Statistics

Commonly used goodness-of-fit statistical methods for assessing item fit are Pearson χ^2, likelihood ratio (G^2), Yen's Q_1 (Yen, 1981), and Bock's χ^2 (Bock, 1972) goodness-of-fit statistics. The Pearson χ^2 fit statistic is used to test how much observed frequencies deviate from expected frequencies. The Pearson χ^2 fit statistic for dichotomous item responses is given by the formula:

$$X^2 = \sum_{j=1}^{j} N_j \frac{(O_j - E_j)^2}{E_j(1-E_j)}, \qquad (2.12)$$

where J is the number of ability (θ) subgroups; O_{jk} and E_{jk} are the observed and expected proportions for θ subgroup j; N_j is the number of examinees in subgroup j.

The likelihood ratio (G^2) statistic for dichotomous item responses is defined as:

$$G^2 = 2\sum_{j=1}^{J} N_j \left(O_{jk} \log\frac{O_j}{E_j} + (1 - O_{jk})\log\frac{1-O_j}{1-E_j}\right), \qquad (2.13)$$

where j is the category or cell, J is the number of categories, *ln* is the natural logarithm function, and O_j and E_j are the observed and expected proportion of responses in category j, respectively.

Bock's chi square, $\chi^2{}_B$, proposed to assess item-fit that utilized a Pearson χ^2 test statistic. $X^2{}_B$ for dichotomous item i is given by the formula:

$$X^2{}_B = \sum_{j=1}^{J} \frac{N_j \left(Q_{ij} - E_{ij}\right)^2}{E_{ij}\left(1-E_{ij}\right)}, \qquad (2.14)$$

where i is the item number, J is the number of ability (θ) subgroups, j is the counter for the θ subgroups, N_j is the number of examinees with estimates falling within θ subgroup j, O_{ij} and E_{ij} are the observed and expected proportion of correct responses on item i within θ subgroup j, and E_{ij} =ˆ P (ˆ$_{med-j}$), where, ˆ$_{med-j}$ is the median of the ˆ values for examinees in subgroup j. The significance of $X^2{}_B$ is tested by comparing it to the χ^2 distribution degrees of freedom (*df*) equal to number of θ subgroups J minus the number of estimated item parameters (m), or *df* are $J - m$.

Yen (1981) modified $X^2{}_B$ test statistic. Yen's Q_1 test for dichotomous item i is defined as follows:

$$Q_1 = \sum_{j=1}^{10} \frac{N_j \left(Q_{ij} - E_{ij}\right)^2}{E_{ij}\left(1-E_{ij}\right)}, \qquad (2.15)$$

where i is the item number, j is the interval created by grouping examinees on the basis of their ability (θ) estimates, N_j is the number of examinees with estimates falling within θ subgroup j, O_{ij} and E_{ij} are the observed and expected proportion of correct responses on item i within θ subgroup j, and E_{jk} are computed from the model as the mean predicted probability of a correct response in each interval. Similar to $X^2{}_B$ statistic, the significance of Q_1 is tested by comparing it to the χ^2 distribution *df* equal to number of θ subgroups *J* minus the number of estimated item parameters (*m*), or (*10 - m*).

Pearson χ^2, G^2, X^2_B and Q_1 test statistics are 'traditional' methods for assessing goodness-of-fit. They are based on the calculation of a χ^2 test statistic, which is used to compare observed and expected score distributions. For traditional goodness-of-fit statistics, an interval is created by grouping examinees on the basis of their θ estimates. X^2_B differs from Q_1 in that the number of intervals varies, while in Q_1 the number of intervals is equal to 10. In addition, the expected proportions in X^2_B are computed using the median (rather than using the mean) of the estimates within each interval. A major problem of these traditional goodness-of-fit statistics that use χ^2 statistics as measures of fit, is their sensitivity to sample size. If the sample size is large, then a χ^2 test is too sensitive and has high statistical power that will often discount models even if the model fit is acceptable (Hambleton, 1993; Hambleton & Rodgers, 1986). On the other hand, if the sample size is too small, then a χ^2 test lacks statistical power and will fail to detect item misfit when it exists.

2.4.5.2.1.1.2 Alternative Goodness-of-Fit Statistics

Orlando and Thissen (2000) introduced an alternative fit statistic (S-X^2 and S-G^2) for dichotomous items. In this method, the observed and expected proportions of individuals responding with a particular response (e.g., correct) for each total score group are compared based on the joint likelihood distributions using either a Pearson (S-X^2) or likelihood-ratio (S-G^2) chi-square goodness-of-fit statistic. S-X^2 or S-G^2 for dichotomous items is computed as following:

$$S - X_i^2 = \sum_{k=1}^{n-1} N_k \frac{(O_{ik} - E_{ik})^2}{E_{ik}(1 - E_{ik})}, \qquad (2.16)$$

and

$$S - G_i^2 = 2\sum_{k=1}^{n-1} N_k \left(O_{ik} \log\frac{O_{ik}}{E_{ik}} + (1 - O_{ik}) \log\frac{1-O_{ik}}{1-E_{ik}}\right), \qquad (2.17)$$

where O_{ik} and E_{ik} are observed and expected proportions, and *df* for the *S*-X^2 or *S*-G^2 are the number of score levels minus the number of estimated item parameters (*m*). If all score levels are used in an *n*-item test, then the *df* are (*n* – 1) – *m*.

The main advantage of Orlando and Thissen statistics (*S*-X^2 and *S*-G^2) over the traditional item fit statistics is that observed frequencies are obtained using observed data only instead of using θ estimates. That is, examinees are not grouped based on their θ estimates. Grouping examinees into equal-size intervals according to their θ estimates is highly sample dependent (Orlando & Thissen, 2000). In addition, Reise (1990) indicated that the number of ability intervals chosen is arbitrarily. Finally, how the intervals are created and the number of intervals can impact the value and the statistical significance of the fit statistic (Yen, 1981; Orlando & Thissen, 2000).

Orlando and Thissen (2000) conducted a simulation study evaluating the performance of *S*-X^2 and *S*-G^2 for tests of length 10, 40, and 80 items with fixed sample size of 1000. Item response data were generated under three IRT models (1PL, 2PL and 3PL). They examined Type I error rates and empirical power of their statistics. They found that the Type I error rates for the Yen Q_1 and likelihood ratio (G^2) statistics were unacceptably high in short tests. In longer tests, the Type I error rates for Q_1 were slightly above the nominal values. Meanwhile, the Type I error rates for $S\text{-}X^2$ remained close to the nominal rejection rate of 0.05 regardless of test length and across sample size. Type I error rates were somewhat higher for $S\text{-}G^2$ as sample size increased. Overall, the authors state that $S\text{-}X^2$ is a promising for detecting item misfit for dichotomous items, and $S\text{-}G^2$ is not very useful because of the inflated type I error rate.

2.4.5.2.1.2 Graphical Representations

In addition to using goodness-of-fit tests for assessing model fit, it is also possible to obtain a graphical display of the fit between observed and expected score distributions (Hambleton, 1993; Hambleton & Rodgers, 1986; Hambleton & Swaminathan, 1985). Graphical representations involve analyses of residuals and standardized residuals for each test item at various ability levels. The residual analyses involve obtaining the observed and expected score distributions across 10-15 ability (θ) subgroups for an item. Residuals, which are the differences between actual and expected item performance for a subgroup, are obtained. Further, the residual is standardized by dividing the raw residual by the standard error of the observed proportion of correct responses (Swaminathan & Rogers, 1991):

$$SR_j = \frac{(O_j - E_j)}{\sqrt{\frac{E_j(1-E_j)}{N_j}}}, \qquad (2.18)$$

Where O_j is the observed proportion of correct responses in ability interval j, E_j is the expected proportion of correct responses in the interval under the fitted model, and N_j is the number of examinees in the ability interval j.

Finally, the standardized residuals are then plotted against the θ scale. The standardized plots that show random scatter about zero indicate item fit, whereas the scatter-plots that show patterns indicate item misfit (Hambleton & Swaminathan, 1985; Hambleton, et al., 1991). When the selected IRT model fits a test's data, expected model features such as invariance of item and ability parameters are obtained

2.4.6 Examining the Invariance of Item and Ability Parameters

After determining the most appropriate IRT model for test data, it is important to assess item and ability parameters invariance. This invariance property is central to IRT applications. Invariance of item parameters means that item parameter estimates (a_i, b_i, and c_i parameters) do not depend on the ability distribution of the examinees. Item parameters will not change depending on which group was used to calibrate items. Invariance of ability parameters, on the other hand, means that examinee ability estimates does not depend on a particular test; it does not matter which set of items (hard or easy items) are administered to examinees to estimate their ability from the pool of items (Hambleton, et al., 1991; Lord, 1980). Although invariance is an all-or-nothing property in the population, it is not always observed due to the nature of drawing samples. Instead the degree to which invariance holds can be assessed by examining the correlation and the scatterplots of parameter estimates (Hambleton, Swaminathan & Rogers, 1991). If the correlations are reasonable and the plots are linear, the property of invariance is met. According to Wright (1968), the property of invariance exists to some degree when associated correlation coefficients are .80 and higher.

2.4.6.1 Examining the Invariance of Item Parameters

The degree to which the invariance of the item parameter holds is tested by first separating the data into two groups and then comparing the item parameter estimates for these two groups under the selected IRT model. The two groups are generally formed by splitting the data into high and low ability groups using theta (θ) values. MULTILOG is run to separately estimate the IRT item parameter for each subgroup using the selected IRT model. To determine the extent to which the invariance of item parameter estimates holds under the selected IRT model, item parameter estimates for the two subgroups are correlated. In addition to correlation analysis, the scatter plots of the estimated parameters are investigated in order to check the strength of the relationship. The estimated parameters are considered to be invariant if the correlations are relatively high and the plots are linear (Hambleton & Swaminathan, 1985; Hambleton, et al., 1991).

2.4.6.2 Examining the Invariance of Ability Parameters

To investigate the degree to which the invariance of the ability parameter holds, the total test items are divided into relatively hard versus easy items using the item difficulty estimated. Then, MULTILOG is run to separately estimate the ability parameter of the selected IRT model for the hard and easy item sets. The correlations between the ability parameter estimates obtained from the easy and hard items are computed. In addition, the pairs of the ability parameter estimates obtained from the easy and hard items on each form are plotted. The invariance of the ability parameters of the selected IRT model will hold true if the correlations are high and the scatter plots are linear (Hambleton & Swaminathan, 1985; Hambleton, et al., 1991).

2.4.7 Ability and Item Parameter Estimation

2.4.7.1 Ability Estimation

There are two main methods for estimating ability parameters: maximum likelihood estimation (MLE) and Bayesian-based methods (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985; Hambleton, et al., 1991; Yen & Fitzpatrick, 2006).

2.4.7.1.1 Maximum Likelihood Estimation

Maximum likelihood estimation (MLE) of θ estimates values of examinees' ability that generate the greatest probability or likelihood given the item response pattern (Hambleton & Swaminathan, 1985). The likelihood function for a given examinee of ability (θ) is the likelihood of a particular item response vector U = (u_1, u_2…, u_n), where u_i is "1" if the answer to item j is correct, or "0" otherwise. The likelihood function is expressed by the following formula (Embretson & Reise, 2000; Hambleton, et al., 1991; Hambleton & Swaminathan, 1985):

$$L(u_1\ , u_2, \ldots , u_i\ |\theta) = \prod_{i=1}^{n} P_i^{u_i}\ Q_i^{1-u_i}, \qquad (2.19)$$

The logarithm of the likelihood function is typically used to identify the ML estimate of θ because it simplifies the computation (Hambleton & Swaminathan, 1985):

$$\log L(\boldsymbol{u}|\theta) = \sum_{i=1}^{n}[u_i\ \log P_i\ + (1 - u_i)\log Q_i], \qquad (2.20)$$

Usually, the likelihood function is plotted across values of θ, and from the graph, the MLE of θ can be determined. However, the maximum likelihood (ML) estimate of θ can be obtained mathematically by setting the 1st derivative of the likelihood function or the log- likelihood function with respect to the parameter being estimated equal to zero (Hambleton & Swaminathan, 1985). To determine the point where the function reaches a maximum, the 2nd derivative or the slope of 1st derivative is set equal to zero. Finally, the ML estimate of θ is solved using the Newton-Raphson iterative procedure (Hambleton, et al., 1991).

MLE is commonly used because it has four desirable properties: First, it is consistent. That is, it is unbiased asymptotically—as the number of items/examinees increase and the IRT model holds, the estimates converge onto true values. Second, it is an efficient estimator. This means that it has the smallest variance. Third, it is asymptotically normally distributed. Finally, for the 1PL model, the MLE is a sufficient statistic. However, the estimation of parameters is positive or negative infinity for items or examinees with perfect or zero scores. In this case, Bayesian estimators may be used as an alternative method to MLE (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985; Hambleton, et al., 1991).

2.4.7.1.2 Bayesian Estimation

Similar to MLE, Bayesian estimation uses examinees response patterns to estimate ability. However, unlike MLE, Bayesian estimation makes an assumption about the distribution of ability in the group, where information about prior distribution is combined additionally with information from the examinee's item responses. In other words, the likelihood associated with the response patterns is combined with the information about the prior distribution of θ creating an adjusted distribution called the *posterior distribution* (Embretson & Reise, 2000; Hambleton

& Swaminathan, 1985). This relationship along with the information from the prior ability distribution are used to estimate the ability parameters. Without using a prior distribution of θ, the Bayesian estimators are similar to MLE in which Bayesian estimators have the same uniform value for all θ (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985). The Bayesian procedure is based on *Bayes theorem,* and it is used to express the $P\ (\theta \mid \boldsymbol{u})$ as:

$$P(\theta|\boldsymbol{u}) = P(\boldsymbol{u}|\theta)\frac{\mathrm{g}(\theta)}{[\sum^{\mathrm{J}} P(\boldsymbol{u}|\theta_j)*\mathrm{g}(\theta)]}, \text{where } \sum P(u|\theta_j) * \mathrm{g}(\theta) = P\ (u), \qquad (2.21)$$

$P(\theta \mid \boldsymbol{u})$ is the posterior distribution of θ, $P(\boldsymbol{u} \mid \theta)$ is the likelihood function for the response pattern **u**, g(θ) is the prior distribution of θ (typically, N(0,1)), and $\sum$ P(**u** | θ_j)*$g(\theta)$ = $P(\boldsymbol{u})$ is the unconditional or marginal probability distribution of response pattern **u** across θ for an examinee of unknown θ, randomly sampled from $g(\theta)$. When the mean of the posterior distribution is taken as the ability estimate, it is called the *Bayes mean* or *expected a posteriori* (*EAP*). When the ability estimate is the mode of the posterior distribution, it is called *Bayes mode estimator* or *maximum a posteriori* (*MAP*) (Embretson & Reise, 2000).

The EAP and MAP estimators are similar in that prior information is used in both; however, they differ since the EAP divides the discrete prior distribution of θ into many quadrature points, rather than using it as a continuous distribution. EAP is mathematically easier to implement than MAP because it does not require the user to know the first and the second derivatives of the likelihood function (Embretson & Reise, 2000). The EAP and MAP estimators typically have a smaller mean squared error (MSE) than the MLE when a prior distribution is known (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985). The MAP estimator has a somewhat larger MSE than the EAP estimator, unless the mode and the mean of the posterior distribution are the same. Furthermore, the EAP and MAP estimators exist for any response patterns—including all incorrect and correct—since the θ estimates are restricted to a reasonable range, due to the use of priors. However, the EAP and MAP estimators are generally biased towards the population mean, especially when the number of items is small (e.g., less than 20) (Embretson & Reise, 2000; Lord, 1986).

2.4.7.2 Item Parameter Estimation

Three types of ML procedures are commonly used to estimate parameters in IRT, namely Conditional Maximum Likehood (CML), Joint Maximum Likehood (JML), and Marginal Maximum Likelihood (MML). These three procedures handle the unknown ability levels differently (Embretson & Reise, 2000).

2.4.7.2.1 Conditional Maximum Likehood (CML)

CML estimation handles unknown ability by expressing the probability of the response patterns without including an ability parameter. In CML, the likelihood function for the sample is factored into the ability parameter and the item difficulty parameter, which is conditional on the sufficient statistic (total score). That is, the number of correct responses to an item is a sufficient statistic for the item difficulty parameter and the number of items correct is a sufficient statistic for the ability parameter (Embretson & Reise, 2000).

The CML procedure provides consistent and efficient parameter estimates. However, CML has some limitations: 1) it is only relevant to the 1PL logistic model; 2) no parameter estimates can be obtained for zero or perfect scores.; 3) examinees that have the same number of items correct but different response patterns will be given the same ability estimate; and 4) it has problems estimating parameters for long tests, complicated patterns of missing data, or polytomous data with many response categories (Embretson & Reise, 2000).

2.4.7.2.2 Joint Maximum Likehood (JML)

In the JML estimation, the ML for item and ability parameters is determined by finding values that jointly maximize the log likelihood across all examinees and items (Embretson & Reise, 2000). For example, the log likelihood function for the 3PL logistic model under the JML procedure is:

$$\ln L(\boldsymbol{u}|\boldsymbol{\theta},\boldsymbol{\omega}) = \sum^{N}\sum^{I} \ln P_{ni} + \ (1-u_{ni})\ln Q_{ui}\,, \qquad (2.22)$$

N is examinees; I is items; and $\boldsymbol{\theta}$, $\boldsymbol{\omega}$ are vectors of examinee and item parameters, respectively. The Newton-Raphson two-stage iterative method is usually applied to estimate item and ability parameters. Iterations are started by estimating the ability parameters with the initial values estimated for item parameters, which are treated as known or fixed; then using the ability estimates from stage one, item parameters are estimated, given that abilities are known or fixed. This two-stage procedure is repeated until both the estimates of the ability and item parameters converges—that is, when the difference between estimates from two successive stages is very small (Embretson & Reise, 2000).

The JML estimation procedure is applicable to many IRT models, and it is computationally efficient. Despite these advantages, it has two shortcomings. First, the estimates of parameters tend to be infinity or negative infinity for items or examinees with perfect or zero scores. Furthermore, the item parameter's estimates in JML are inconsistent and biased; they are estimated jointly with the ability parameter (Embretson & Reise, 2000).

2.4.7.2.3 Marginal Maximum Likelihood (MML)

MML estimation does not estimate the examinee's ability scores when estimating the item parameters, but models the response-pattern probabilities for an examinee of unknown θ drawn at random from a population in which θ is distributed with a know distribution of ability level (Embretson & Reise, 2000):

$$P(\boldsymbol{U}) = \int_{-\infty}^{\infty} L\ (\boldsymbol{U}|\theta)\,\mathrm{g}(\theta) d\theta \rightarrow \int_{-\infty}^{\infty} \prod^{I} P_i^{Ui} Q_i^{1-u_i}\ \mathrm{g}(\theta) d\theta, \tag{2.23}$$

*P (**U**)* is the unconditional or marginal likelihood of ***U*** across θ. Based on the marginal distribution, the item parameters are estimated, and then item parameters that maximize the log of the marginal likelihood distribution are chosen (Embretson & Reise, 2000).

To estimate the item parameters, MML applies the EM algorithm. The EM algorithm has two stages, namely the expectation and the maximization stage. In the expectation stage, expected values of the frequencies at quadrature points and expected frequencies of examinees passing the items are computed. These expected values are then submitted to the estimation equations to maximize likelihood estimation in the maximization stage. The E and M stages are repeated until the estimates converge. Then, when estimates converge the Newton-Gauss procedure is used to solve the maximum likelihood equation. In MML, ability parameters cannot be estimated directly. Nevertheless, by assuming item parameter estimates are known, ability parameters may be estimated using procedures such as MLE, MAP or EAP (Embretson & Reise, 2000).

MML has some advantages over other ML procedures. It is applicable to all types of IRT models and efficient for both short and long tests, whereas JML produces biased estimates for short tests. In addition, it provides estimates for perfect scores and thus no loss of information occurs by deleting examinees with perfect scores. Finally, the standard error estimates for items in MML are good approximations of expected sampling variance of the estimates. Despite these advantages, MML has some limitations. First, MML estimation is computationally involved and sophisticated. Second, an ability distribution must be assumed, and it is assumed to be normal if the prior ability distribution is not known. Therefore, a large number of examinees are needed to meet this assumption. Third, the discrimination parameter estimates near zero can result in a very large absolute value of difficulty (Embretson & Reise, 2000).

In conclusion, MML has the most desirable feature among the maximum likelihood methods (i.e., JML and CML); it applies to many IRT models, provides estimates that are consistent and unbiased, and provides estimates for perfect or zero scores. In contrast, CML applies only to the 1PL logistic model, and JML fails to provide estimates that are consistent and unbiased.

2.4.8 Item and Test Information Function for the 3PL Model

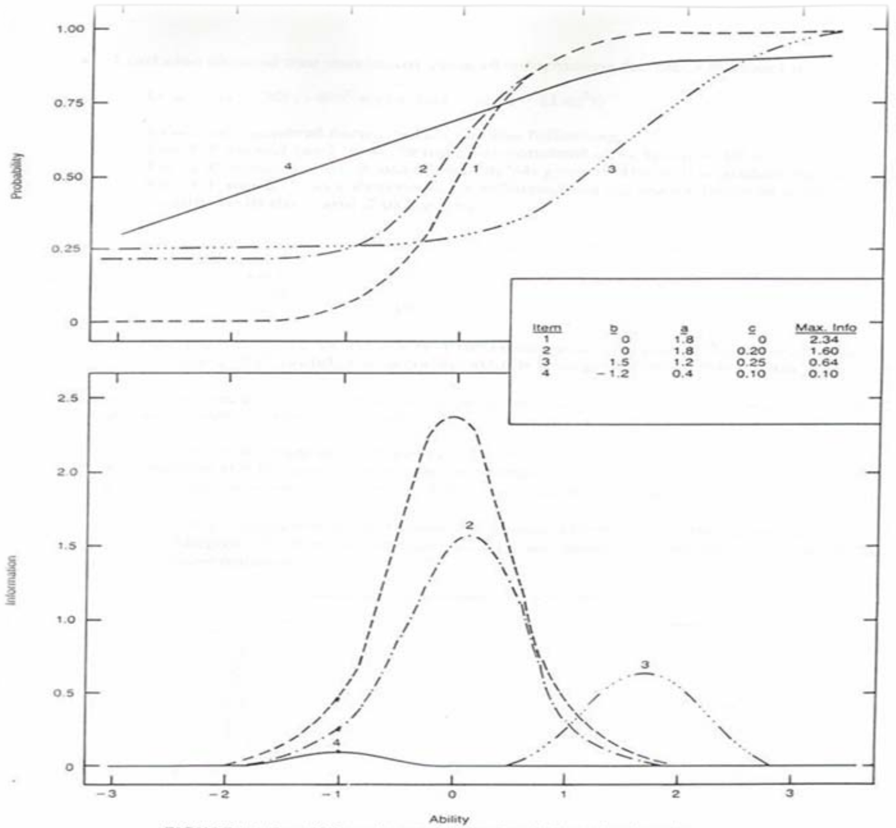
2.4.8.1 Item Information Function

The item information function indicates the amount of information or precision of measurement an item or test provides conditional on the ability level and is useful for describing, comparing and selecting items (Embretson & Reise, 2000; Hambleton, et al., 1991). For the 3PL IRT model, the IIF is calculated as (Lord, 1980):

$$I\ (\theta, u_i) = \frac{P_i^2}{P_i\ Q_i} = \frac{D^2 a_i^2 (1-c_i)}{\left[c_i + e^{D a_i\ (\theta - b)}\right]\left[1 + e^{-D a_i\ (\theta - b)}\right]^2}, \qquad (2.24)$$

P_i (θ) is the probability that a randomly selected examinee with ability θ will answer item *i* correctly; Q_i (θ) is the probability that a randomly selected examinee with ability θ will answer item *i* incorrectly; and P'_i (θ) is the first derivative of P_i (θ) with respect to θ.

The IIF shows the amount of information produced by each individual item on a test as a function of ability. The amount of information an item supplies is maximized when the difficulty parameter (b_i) is close to θ; the discrimination parameter (a_i) is large; and the pseudo guessing parameter (c_i) approaches zero (Hambleton, Swaminathan, & Rogers, 1991). In the 3PL model, item information is constrained by the *c* parameter—that is, the larger the *c* value, the smaller the amount of information supplied, assuming parameters *b* and *a* do not change. In this case, an item provides maximum information at an ability level slightly higher than its *b* parameter (Embretson & Reise, 2000). Generally, the amount of information provided by each item depends on the *a* parameter. This means that an item with high discrimination has a "peaked" information curve and provides more information within a narrower range of ability, whereas an item with low discrimination has a "flatter" information curve and provides less information within a wider range of ability (Embretson & Reise, 2000). Illustration of the IIF's for four items is given in Figure 4.



Item	b	a	c	Max. Info
1	0	1.8	0	2.34
2	0	1.8	0.20	1.60
3	1.5	1.2	0.25	0.64
4	−1.2	0.4	0.10	0.10

FIGURE 4.4 Four ICCs and corresponding item-information function.

Figure 4. Four ICCs and corresponding item-information function

Figure 4 shows the four ICCs, while the bottom figure shows the corresponding IIF's. The highest point on the IIF's curves represents the ability level at which the item provides the most information. From Figure 4, it is clear that the information function curve for items 1 and 2 are much higher, indicating that they give more information at their maximum points. However, item 3 shows more information at the "high end" of the ability continuum. Finally, item 4, which has low discrimination power, provides less information over a wide range of ability and has a considerable flatter information curve. Therefore, different items provide different amounts of information at different levels of θ.

2.4.8.2 Test Information Function

A test information function (TIF), or the contribution of the test toward ability estimation, is simply obtained by adding the IIFs for all items on a test. The TIF is defined as (Embretson & Reise, 2000; Hambleton, et al., 1991):

$$I\ (\theta) = \sum_{i=1}^{I} I(\theta), \qquad (2.25)$$

In this formula, *Ii* (*θ*) is the item information and *n* is the number of test items. The more information each item contributes, the higher the test information function (Embretson & Reise, 2000; Hambleton, et al., 1991).

Item and test information functions are useful for test development and for describing and selecting test items (Hambleton, 1993). In general, three methodological steps are required for constructing a useful test. The first step is determining the regions of the ability scale for which fine discrimination among scale points is desirable. The second step is determining the point on the ability scale at which an item provides its maximum information. The third, and most important step, is selecting items that provide more information over a range of ability values. The final step is selecting items that match the purposes of the test (Crocker & Algina, 1986). In order to do this, assessing the shape of the information function over a range of abilities is vital (Hambleton, et al., 1991; Lord, 1977). For example, if the test developer is concerned with measuring an examinee's abilities across a wide range of ability levels, as in a norm-referenced test, the test developer should select various items which provide more information over a range of ability values (e.g., the range $-3 \leq \theta \leq 3$). On the other hand, if the test developer is concerned with a cutoff score or pass/fail decision, the test developer should select items that provide most information at the cutoff score. In other words, if the purpose of the test is to select examinees with high ability levels, then test items that provide most of their information in the high ability range should be selected. In this case, the shape of the TIF “peaks” at or near the cutoff score. The more information a test provides at a particular ability level, the closer the ability estimates the true ability level, and hence the more precise the test will be with less associated error of estimation (Baker, 1992).

2.4.8.3 Standard Error

The amount of information a test provides at a θ level is inversely related to the standard error of the estimate (Embretson & Reise, 2000; Hambleton, et al., 1991; Hambleton, 1993).

$$SE(\theta^{\wedge}) = \frac{1}{\sqrt{TI(\theta)}}, \qquad (2.26)$$

In this formula, *SE* ($\hat{\theta}$) is the standard deviation of the distribution associated with estimates of ability for examinees given θ. *SE* ($\hat{\theta}$) serves as a good measure of the precision of measurement (Hambleton, 1993). That is, the higher the item discrimination and the smaller the variance at each ability level, the greater the information provided by the test at a given θ, and the smaller the *SE* of measurement. As a result, the ability estimation would be more precise (Hambleton, et al., 1991; Hambleton, 1993). Generally, the size of *SE* depend a number of factors: the number of test items, the quality of test items, and the match between item difficulty and examinee ability. That is, smaller *SE*’s are associated with longer tests, and with highly discriminating items. In addition, smaller *SE*’s are associated with tests composed of items with *b* values close to the examinee’s true ability (Hambleton, et al., 1991; Hambleton, 1993).

2.5 AN OVERVIEW OF METHODS FOR ASSESSING DIMENSIONALITY USING EFA AND DIMTEST

2.5.1 The EFA Procedure for Dichotomous Data

2.5.1.1 Basics Concept of Factor Analysis

In the social sciences, *latent variables* or *factors* cannot be directly observed; information about them can only be obtained indirectly by ascertaining their effects on observed variables. Factor analysis is a statistical technique used to explain correlations among variables (or item responses) in terms of a smaller number of dimensions or factors that determine the relation among the variables (Hair et al., 1992).

Factor analysis is mainly used to detect the underlying structure of a set of items and to summarize them into a smaller set of factors (or dimensions) with minimal information loss. Additionally, factor analysis is used to examine the internal structure of a test—that is, examining the pattern of correlations among items in order to identify domains that are being measured, relationships between the domains, and potential sources of construct-irrelevant variance. Such an examination also provides support for test score interpretations (Stone & Yeh, 2006).

There are two types of factor analyses: exploratory and confirmatory. The primary objective of exploratory factor analysis (EFA) is not to verify a factor structure, but rather to explore the underlying factor structure of a set of observed variables that could account for the relationship between the latent factors and observed variables with no prior hypothesis regarding this relationship (Stevens, 1996). In contrast, confirmatory factor analysis (CFA), is typically used to verify the factor structure of a set of observed variables based on a prior theoretical model, to explain the relationship between the factors, and to determine which variables correlate to which factors and how factors correlate to each other (Stevens, 1996). In CFA, the researcher hypothesizes relationships among a set of variables and these relationships derived from the theory that the researcher seeks to verify. For example, a researcher may determine whether the correlations between a set of variables in a personality test (such as extroversion and introversion) can be accounted for by a two-factor structure. In this case, the researcher hypothesizes which items are indicators of Factor 1 (extroversion) and which items are indicators of Factor 2 (introversion).

2.5.1.2 Linear Exploratory Factor Analysis Procedure

Linear exploratory factor analysis assumes that items are linearly correlated to one another and that items are linearly associated to factors. It also assumes that observed variables (i.e., item scores) and latent variables are continuous (Gorsuch, 1983; Harman, 1976). Yet, data in the social and behavioral sciences is often categorical (dichotomous or polytomous), which may not meet the assumptions of linear EFA.

In general researchers, agree that linear EFA may poorly assess dimensionality for dichotomous data (Hulin, et al., 1983; McDonald, 1981). Performing LFA on dichotomously scored items using Pearson correlations may cause several problems. First, spurious factors may arise due to both differences in distributions for items and to low reliability of item-level data (Ackerman, et al., 2003; Green, 1983). It is also possible to underestimate factor loadings and overestimate the number of dimensions when using LFA. Several researchers have suggested using tetrachoric correlations rather than Pearson correlations with factor analysis for dichotomous data to estimate its latent structures (Bock & Lieberman, 1970; Crocker & Algina, 1986; Kim & Mueller, 1978). However, estimates of the matrix of tetrachoric correlations is often not “positive definite” (Bock, et al., 1988; Knol & Berger, 1991; Lord & Novick, 1968). Furthermore, tetrachorics are inappropriate when the distribution of the latent ability variables is not normal or when guessing exists in the item response functions (Lord, 1980).

Two general methods are proposed as a remedy for the problems that may arise from using LFA with dichotomous data. The first method is derived in the framework of traditional LFA and involves using tetrachoric correlation and the Generalized (or Weighted) Least-Squares (GLS/WLS) estimation method (Christoffersson, 1975; Muthén, 1978). The second method is derived in the framework of the Multidimensional Item Response theory (MIRT) and involves using the ML (Bock & Aitkin, 1981) and the Unweighted Least-Squares (ULS) methods (McDonald, 1967, 1994).

2.5.1.3 Generalized Least-Squares Estimation

Christofferson (1975) and Muthén (1978) used a GLS estimator to conduct factor analysis of dichotomous variables using the matrix of tetrachoric correlations. This approach results in models that express the probability of correctly responding to dichotomously scored items as nonlinear functions of some ability. That is, response variables (X_i) are accounted for by the latent continuous unobserved variables (Y_i) and threshold variables (τ_i), where X_i takes a value of:

$$X_i = \begin{cases} 1, & if\ \mathrm{Y_i} > \tau_i \\ 0, & Otherwise. \end{cases} \qquad (2.27)$$

The factor analysis model for dichotomous variables can be described as:

$$\boldsymbol{Y} = \boldsymbol{\Lambda\theta} + \mathbf{E}\,, \qquad (2.28)$$

Here, ***Y*** is the latent continuous variables ($Y_1, \dots, Y_n$); **Λ** is a matrix of factor loadings of items ($\lambda_{i1}, \dots, \lambda_{im}$); **θ** is ability values of examinees ($\theta_1, \dots, \theta_n$); and ε is residuals matrix. According to De Champlain (1999), the factor analysis model described in the above equation is the same as the common factor model, with the exception that ***Y*** is unobserved. Assuming that ***θ*** ~ MVN (0, ***I***) and ***E*** ~ MVN (0, $\boldsymbol{\Psi}^2$) are multivariate normal, $\boldsymbol{\Psi}^2$ is a diagonal matrix of residual covariance, and Cov (***θ***, ***E***) = **0**.The variance-covariance matrix of the latent variables, denoted as **Σ**, is given by:

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda\Phi\Lambda}' + \boldsymbol{\Psi}.^2 \qquad (2.29)$$

The factor analysis model for dichotomous variables assumes further that *Y* ~ MVN (0, $\boldsymbol{\Lambda\Phi}\ \boldsymbol{\Lambda}' + \boldsymbol{\Psi}^2$) (De Champlain, 1999). The factor analysis model for dichotomous variables can be estimated using tetrachoric correlations and the limited-information GLS estimator (Muthén, 1984).

Christofferson’s (1975) GLS method estimates the parameters of the factor model by minimizing the “fit function”, whereas Muthén’s (1978) GLS procedure estimates the parameters of the factor model by minimizing the "weighted least-squares fit function".

Christoffersson's and Muthén's GLS procedures have several advantages. The first important advantage is that the procedures provide consistent parameter estimates. Secondly, these procedures provide statistical tests of model-fit. Asymptotically, *F* functions are minimized in the GLS procedure and follow a chi-square (χ^2) distribution, with a degree of freedom (*df*) = *k* (*k*-1)/2 – *t*, where *k* is equal to the number of items and *t* the number of parameters estimated in the model. Finally, these procedures also provide standard errors of estimation (De Champlain, 1999; Mislevy, 1986).

Despite these advantages, Christoffersson's and Muthén's GLS procedures have some limitations. First, the information of these procedures is limited because it is based on lower-order joint proportions of examinees who correctly answer one to four items taken at the same time which are one-way, two-way, three-way, and four-way margins (Bock et al., 1988). In other words, the GLS estimator ignores higher level interactions in the data; thus, GLS estimator does not use all of the available information (De Champlain, 1999). Despite this limitation, De Champlain (1999) notes that McDonald (1994) and Muthén (1978) have suggested that one should not lose too much information in the absence of higher-order marginals in most practical situations. Second, the GLS estimator displays an asymptotic covariance matrix of the estimated tetrachoric correlation coefficients. Third, the computation of the GLS estimator requires the production of a weight matrix that becomes heavy as the number of items on the test increases (Bock et al., 1988). Finally, due to the large size of the weight matrix, GLS limits the number of items to 20-25 (Bock et al., 1988; Muthén, 1978).

2.5.1.4 Programs for Performing the EFA of Dichotomous Data

Muthén’s GLS procedure is implemented in the Mplus software program (Muthén & Muthén, 2001). Mplus can be used to analyze tetrachoric correlation matrices using the weighted least squares with robust standard errors and mean-adjusted (WLSM) estimators for dichotomous data. Mplus uses a probit regression of items on factors, modeling a nonlinear relationship between the factors and items. However, Mplus does not model guessing. “There is no attempt in Mplus to ‘smooth’ in the computation of the tetrachoric correlations” (Tate, 2003, p.164). As a result, Mplus may overestimate the number of factors or dimensionality (Hulin, et al., 1983; Tate, 2003).

In addition to Mplus, various parametric factor analysis procedures have been implemented to explore the dimensionality of dichotomous data in software programs, such as Mplus (Muthén & Muthén, 2001), NOHARM (Fraser & McDonald, 1988), and TESTFACT (Wilson, et al., 2003). These programs are based on nonlinear models, but they differ in the sample statistics that are analyzed, estimation methods and fit statistics used, and how guessing is accommodated (Stone & Yeh, 2006).

NOHARM and TESTFACT are based on compensatory multidimensional item response theory (MIRT) models, and they estimate the item parameters for multidimensional normal ogive models for dichotomous items. The compensatory models in MIRT assume that low ability in one dimension can be compensated for with greater ability on another dimension (Embretson & Reise, 2000). The multidimensional compensatory three-parameter logistic model (MC3PL) is an extension of the 3PL model for dichotomous items and defines the probability of a correct response for the i^{th} item as follows (Embretson & Reise, 2000; Reckase, 1997):

$$P\left(X_{ij} = 1 | a_i, d_i, \theta_j \right) = c_i + (1 - c_i) \frac{\exp D\left(a_i^{'} \theta_j + d_i\right)}{1 + \exp D\left(a_i^{'} \theta_j + d_i\right)}, \qquad (2.30)$$

X_{ij} is the response of person *j* to item *i* (0 or 1); θ_j is a vector of latent abilities; a_i' is a vector of discrimination examinees for item *i* in dimension *k*; d_i is the easiness intercept for item *i* that is related to item difficulty; and c_i is the lower asymptote or the guessing parameter of item *i*. When d_i is added, easier items have (manifest) higher values for *d*. Therefore, the MC3PL model is referred to as an additive model (Reckase, 1997). When c_i is set to zero, the MC3PL becomes a multidimensional compensatory two-parameter logistic (MC2PL) model. The compensatory MIRT and NLFA models are mathematically equivalent (Knol & Berger, 19991; McDonald & Mok, 1995; Takane & De Leeuw, 1987) and NOHARM and TESTFACT can be used to perform NLFA of dichotomous data.

NOHARM analyzes the corrected sample proportion for item pairs instead of a tetrachoric correlation matrix, and it minimizes the unweighted least squares (ULS) discrepancy between the observed and the expected proportions (Knol & Berger, 1991). Unlike Mplus and NOHARM, which analyze summary information available from the item covariance or correlation matrix, TESTFACT analyzes full-information item factor analysis of the tetrachoric correlations matrix (Bock, et al., 1988; Wilson, et al., 2003). TESTFACT uses full-information marginal maximum likelihood (MML) estimation combined with an expectation-maximum (EM) algorithm to estimate the item parameters (Bock, et al., 1988; Gibbons & Hedeker, 1992; Muraki & Engelhard, 1985).

All three programs provide orthogonal (Varimax) and oblique (Promax) rotations of the initial solution. Orthogonal rotations produce factors that are uncorrelated while oblique methods allow the factors to correlate. In social sciences, some correlation is expected among factors; therefore, oblique rotation methods are appropriate. However, if the factors are truly uncorrelated, orthogonal rotation and oblique rotation produce nearly identical results. If the relationship is unknown, orthogonal rotations may be beneficial to use. All three programs also provide a Root Mean Square Residual (RMSR), which calculates the average standard deviation of the difference between observed and expected correlation matrices given by the EFA model. However, only Mplus provides Root Mean Square Error of Approximation (RMSEA), which corrects the χ^2 statistic for model complexity and is not affected by sample size. The RMSR statistic in TESTFACT is consistent with Mplus; it is based on differences between the observed and expected item correlations implied by the factor model. Mplus also provides a χ^2 fit-statistic that evaluates whether the EFA model (expected data) fits the observed data.

2.5.1.5 Evaluation of Mplus, NOHARM, and TESTFACT Methods

Given that MIRT models are equivalent to NLFA models, most simulated studies that compared Mplus, NOHARM, and TESTFACT estimation methods focused on estimating factor solutions of NFLA or MIRT models. For example, Knol and Berger's (1991) simulation study compared TESTFACT, NOHARM, and other estimation methods (e.g., ULS, maximum likelihood [ML], GLS). They found that NOHARM and TESTFACT performed similarly in terms of recovering factor analytic parameters. In other words, the results showed that full-information models (i.e., TESTFACT) were equivalent to the limited-information models (e.g., NOHARM or GLS estimator).

A study by Gosz and Walker (2002) compared the probability of a correct response based on the true item parameters and item parameters estimated in NOHARM and TESTFACT. Item responses were generated from the two parameter logistic (2PL) model, and exploratory analyses were used to recover the item parameters. The researchers found that NOHARM provided solutions that predicted item performance better than TESTFACT. In other words, the probabilities of a correct response based on the NOHARM item parameter estimates were closer to the true probabilities than those based on the TESTFACT estimates.

Tate (2003) conducted a study comparing a number of empirical methods for assessing test structure with dichotomous items using exploratory and confirmatory procedures. Tate used real test data obtained from a 62-item grade eight reading ability test, which consisted of 8 reading comprehension passages. In the EFA approach, Tate evaluated Mplus, NOHARM, and TESTFACT methods; all methods yielded similar factor structures for the real data set. Also, all methods performed reasonably well, except when guessing was modeled in item responses for the simulated data sets. Mplus performed less accurately than NOHARM and TESTFACT when the data assumed guessing. In general, the results of Tate's study were consistent with the results of Knol and Berger. Whereas NOHARM and TESTFACT performed well using the data with or without guessing, Mplus *only* performed well using data without guessing.

Zhang and Stone (2004) demonstrated similar findings to Knol and Berger's study. They found that NOHARM and TESTFACT performed similarly in terms of recovering multidimensional item parameters in an exploratory mode with items measuring two uncorrelated factors.

Stone and Yeh (2006) compared Mplus, NOHARM, and TESTFACT methods to assess the dimensionality and internal structure of the Multistate Bar Examination (MBE), a multiple-choice test with 200 four-option items, under two conditions: Condition 1, when guessing was *not* modeled (c = 0); and Condition 2, when guessing *was* modeled (c > 0). Under Condition 2, only a comparison between NOHARM and TESTFACT was provided, as Mplus cannot accommodate guessing in its model. The results for Condition 1 demonstrated that Mplus, NOHARM and TESTFACT provided a similar number of factors. Two factor solutions were uncovered by the three methods, although more than half of the items did not load substantially on any factor. The results for Condition 2 illustrated that greater eigenvalues were obtained, indicating that modeling guessing increases the proportion of explained variance for the factors. This finding was due to the correction for guessing that adjusts for the tetrachoric correlation matrix. Findings from Stone and Yeh's study were consistent with previous studies (Knol & Berger, 1991; Tate, 2003), but conflicted with Gosz and Walker's study.

In summary, several simulation studies have compared Mplus, NOHARM, and TESTFACT methods for the specific recovery of MIRT and factor model parameters (Gosz & Walker, 2002; Knol & Berger, 1991; Stone & Yeh, 2006; Tate, 2003). Only Gosz and Walker's study indicated that NOHARM is more efficient than TESTFACT. However, the results of other simulation studies (Knol & Berger, 1991; Stone & Yeh, 2006; Tate, 2003) have indicated that both NOHARM and TESTFACT methods yield very similar factor structures. The results also have indicated that—without modeling guessing—Mplus, NOHARM, and TESTFACT methods provide similar factor solutions. Yet, Mplus cannot accommodate guessing in its model (Stone & Yeh, 2006).

2.5.1.6 Identification of Number of Dimensions using EFA Method

After choosing the software programs to detect dimensionality for dichotomous data, the researcher must then decide how many factors to retain for interpretation. The number of factors determines the overall dimensionality of a test. This determination is important for evaluating a test's internal structure, which can be derived via examining the pattern of factor loading.

Reckase (1994) argued that the purpose of an assessment may affect exploration of the internal structures of a test. If the purpose is to examine the structure of the assessment, overestimating the dimensionality may be more desirable than underestimating the dimensionality. However, if ability estimation is the focus, "then the dimensionality of interest is the minimum dimensionality that provides the greatest information provided by the item responses" (p. 90). If dimensionality is overestimated, then more parameters are estimated, resulting in increased estimation error.

Both over-extraction and under-extraction of retained factors for rotation can result in inaccurate conclusions (Comrey & Lee, 1992; Ford et al., 1986; Gorsuch, 1983; Harman, 1976). Generally speaking, more factors are preferable– that is, a larger number of factors provides more information about the relationship between variables. In addition, empirical research suggests that over-factoring introduces less estimated errors than under-factoring (Fava & Velicer, 1992; Wood et al., 1996).

Different methods have been proposed to identify the number of factors/dimensions in EFA: Kaiser's criterion, Cattell's scree plot, parallel analysis, the amount of variance explained by a factor, the number of substantial factor loadings, and summary statistics for residuals such as the RMSR, and RMSEA. Yet, these are arbitrary criteria; they do not have strict cutoff values and often lead to different solutions (Ford et al., 1986; Humphreys & Montanelli, 1974).

The Kaiser criterion is one of the most frequently used methods for determining the number of factors to retain (Kaiser, 1960). According to this criterion, only the factors with eigenvalues greater than one are retained for interpretation. Despite the simplicity of this criterion, it often retains too many factors. Many researchers agree that it is one of the least accurate approaches for determining the number of factors (Velicer & Jackson, 1990; Tucker et al., 1969).

The second commonly used method to identify the number of factors to retain using EFA is the graphical analysis of the eigenvalues or “Cattell’s scree plot” (Cattel, 1966). This method involves examining a plot of the eigenvalues associated with each of the extracted factors and, subsequently, looking for the natural “breaking point” or “substantial drop” where the scree of the eigenvalue flattens out. The number of eigenvalues above the “elbow” indicates the number of factors to be retained.

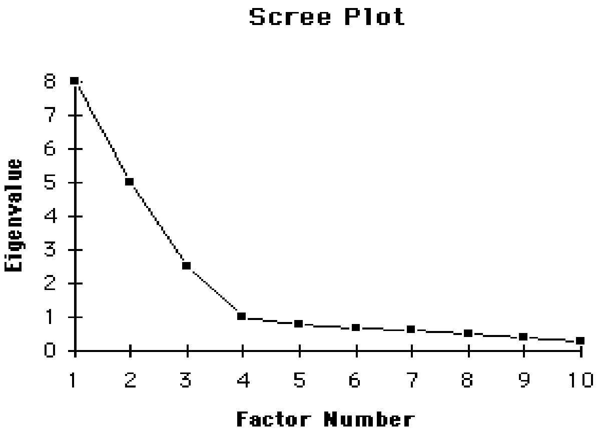


Figure 5. An Example of a Scree Plot

The scree plot in Figure 5 yields three factors which could be retained for rotation. Although the scree plot is relatively straightforward and more accurate than the Kaiser criterion, it tends to over-extract factors and to demonstrate subjectivity—that is, there is no clear way to identify the elbow or the substantial drop in the magnitude of eigenvalues. Consequently, the criterion performs well only during the presence of a strong common factor.

Another method that uses scree plot is the Parallel Analysis (PA), proposed by Horn (1965). In the PA approach, eigenvalues are computed from random numbers with the same amount of cases and items. Then, eigenvalues from real data and random data are compared in order to decide the number of factors to extract, which are equal to the number of eigenvalues in the real data that is greater than the number of eigenvalues from random data. Horn (1965) suggested using the mean of eigenvalues of random data as a baseline for the comparison between real and random data. Recently, though, several researchers have suggested using the desired percentile of the distribution of random eigenvalues data (e.g., 95th-percentile) as the basis for comparison. The PA method sometimes is arbitrary in that a factor meeting the criterion is retained, while a factor falling below is ignored. Before 2000, the PA procedure was unavailable in major statistical programs, such as SPSS and SAS. However, in 2000, O'Connor developed the SPSS and SAS programs for PA (O'Connor, 2000). Nonetheless, several researchers suggest that the PA method may successfully recover dimensionality. Today, it is often recommended as the most efficient method to assess the true number of factors (Velicer et al., 2000; Lance et al., 2006).

The fourth method for determining the number of factors is the amount of variance explained by each factor, indicated by its eigenvalue. The proportion of variance will be calculated by taking eigenvalue and dividing it by the sum of all eigenvalues, and then multiplied by 100; generally, the larger the eigenvalue, the more variation that is explained by the factor. Factors are retained to the model until a certain amount of total variance explained is achieved (e.g., 70%), or until the proportion of variance less than a certain proportion (e.g., less than 10%) is explained by each factor. The amount of variance serves as an index for the substantive importance of factors. Although this method is easy to interpret, it also uses a subjective criterion for determining the number of factors to retain for interpretation.

Yet another method for determining the number of factors is the examination of factor loadings for rotated solution in order to find the simplest and most easily interpreted factor structure. The factor loadings determine the items that are used as indicators for interpreting each factor. Typically, an absolute value of factor loading that is greater than .3 can be used to identify factor loadings or pattern coefficients that are considered “substantial” or “salient” (Gorsuch, 1983). Factors with only a few variables that have salient loadings are considered “trivial” factors.

The final index for determining dimensionality examines the difference between item relationships implied by the factor model and observed relationships using RMSR and RMSEA residuals statistics. The RMSR index summarizes the differences between the observed and expected item correlations implied by the factor model. This index calculates the average standard deviation of the difference between observed and expected correlation matrices given by the EFA model. Larger values of RMSR indicate less fit between the observed data and data expected under the model. The RMSR value of .05 or less is used as a criterion for acceptable factor solutions for Mplus and TESTFACT (Muthen & Muthen, 2001). However, the RMSR statistic in NOHARM is based on differences between observed and expected proportions. In analyzing a correlation matrix, the RMSR statistic should equal a standardized RMSR statistic, and values close to 4 times the reciprocal of the square root of the sample size indicate an acceptable factor solutions solution (Stone & Yeh, 2006). Therefore, in this study, values less than .028 will be used as indicate acceptable factor solutions. RMSR is not provided in TESTFACT, but it will be calculated by taking the residual matrix provided by TESTFACT.

While the RMSR is available for Mplus and TESTFACT, the RMSEA index is only available for Mplus. The RMSEA corrects the χ^2 statistic for model complexity and is not affected by sample size. Larger values of RMSEA also indicate less fit of the factor model to the observed data. An RMSEA value less than .05 is used to indicate an acceptable factor solution related to Mplus (Browne & Cudeck, 1993).

It is generally recommended to use multiple methods for determining the appropriate number of factors to retain. Therefore, five major criteria were used to determine number of factors/dimensions in the CEPA-English test: 1) scree plots, 2) proportion of variance accounted for by the largest eigenvalues, 3) the number of substantial loadings for factors, 4) the RMSR statistics, and 5) the *p*-value of DIMTEST.

2.5.2 The DIMTEST Procedure

The most common nonparametric technique for investigating the assumption of unidimensionality is DIMTEST, which assesses the presence of more than one dominant dimension in test items. DIMTEST uses two subtests to test the hypothesis that the test is unidimensional: Assessment Subtest (AT) and Partitioning Subtest (PT), taken by the same examinees. DIMTEST tests the null hypothesis that AT is dimensionally similar to PT, or H_0: d = 1, against the alternative hypothesis that AT is dimensionally homogeneous and distinct from PT, or H_1: d >1, where d is the number of dimensions. DIMTEST method is based on Stout’s (1987) concept of “essential unidimensionality,” which holds when only one dominant dimension influences the examinees’ performance on a set of test items. This method is based on the idea that when AT and PT are measuring the same dimension, then the covariances among the AT items, conditional onθ, are equal to zero, or $E\ [Cov\ (U_j,\ U_k)|\ \theta PT] = 0$, indicating that a test is unidimensional. If AT is measuring a different dimension than PT, then the conditional covariances among the AT items will be greater than zero (positive), or $E\ [Cov\ (U_j,\ U_k\ \theta PT] > 0\ j \neq$ k, indicating that a test is multidimensional (Zhang & Stout, 1999).

The DIMTEST procedure consists of two stages. First, the N test items must be split into two subtests: AT and PT. The AT subtest is of length M (4 ≤ M < half the test length), and the PT subtest is of length N – M items or K items. The AT subtest consists of items that are relatively homogeneous in terms of dimensionality, and the PT subtest consists of the remaining items of the test that are as heterogeneous as possible. PT is chosen to assign examinees to different K subgroups according to their PT scores. Selecting AT and PT subtests items can be done using expert judgment or using LFA of the tetrachoric correlation matrix, where items loading on the same dimension are selected into the AT subtest (Hattie, et al., 1996; Nandakumar & Stout, 1993). Once AT and PT have been chosen, the second stage is to calculate the DIMTEST statistic, T, which has an asymptotic normal distribution as the number of examinees and items approach infinity (Stout, 1987, 1990).

T is calculated as follows (Stout, 1987):

$$T = \frac{(T_l - T_B)}{\sqrt{2}}, \qquad (2.31)$$

T_L is sensitive to departures from unidimensionality and is based on the sum of the estimated conditional covariances between the AT items for examinees that have obtained the same score, *k*, on the PT items. While T_B is a correction for statistical bias in T_L due to the finite length of the test or set of AT items that are overly homogeneous with respect to item difficulty. T_L is computed as follows (Nandakumar & Stout, 1993; Stout, 1987):

$$T_L \ = \frac{1}{\sqrt{K[\Sigma_K(S_K^2 - \sigma_K^2)]}} \ / \ S_K \ , \ \ w \qquad (2.32)$$

her

$$X_K \ = S_K^2 - \sigma_K^2, \qquad (2.33)$$

X_k represents the difference between the "observed" variance estimate (S_k^2) and the "unidimensional" variance estimates (σ_k^2), when conditioned on K homogeneous score groups, as determined by the PT. If unidimensionality holds, the differences between these variances should be small suggesting that they are estimation of the same variance.

It is important to note that the DIMTEST *T* statistic only assesses whether the test is essentially unidimensional, but does not give an indication of the amount of dimensionality that exists on the test. The DIMTEST statistical procedure is performed through the DIMTEST computer program.

2.5.3 Evaluation of LFA, NLFA, and DIMTEST Procedures

Many simulation studies have evaluated LFA, NLFA, and DIMTEST for assessing the dimensionality of dichotomous test data. For example, Hambleton and Rovinelli (1986), in their simulation study, compared LFA and NLFA methods to determine the dimensionality of a set of test items. Hambleton and Rovinelli's results demonstrated that NLFA was the most promising method in assessing dimensionality. Results further demonstrated that LFA overestimated the number of underlying dimensions in the data.

Nandakumar (1994) compared the performance of LFA, NLFA, and DIMTEST for assessing unidimensionality, using both simulated and real data sets. Nandakumar found that all three methods correctly identified the unidimensionality, but they showed different abilities in detecting a lack of unidimensionality. The NLFA method exhibited especially good power in detecting multidimensionality when the correlation between abilities was high (p = .7) for the simulated two-dimensional data, whereas DIMTEST exhibited highest power in detecting multidimensionality overall.

Findings from Gessaroli and De Champlain (1996) were consistent with Nandakumar's study. In their simulation study, they compared the performance of DIMTEST and NLFA. They found that NLFA (with an approximate X^2) had comparable power to DIMTEST (with Stout's T statistic) as well as superior control of the Type I error rate for two-parameter logistic (2PL) models. They also found that while DIMTEST may not be able to detect multidimensionality with small sample sizes and short tests lengths, NLFA showed greater power in doing so. In other words, NLFA had a much lower Type I error rate than DIMTEST with a small sample sizes (n =500) or small (15) or moderate (30) number of items, and had equal power to DIMTEST for the largest sample size of 1,000 and the largest number of items of 45.

In a more recent study, Pyo (2000) replicated Nandakumar (1994)'s study comparing the performance of NLFA and DIMTEST methods for assessing dimensionality using both simulated and real data sets. Results showed that both methods were able to confirm unidimensionality for simulated unidimensional data even though they showed different power in detecting multidimensionality for two-dimensional test data. This finding agreed with the findings of the Nandakumar (1994) study. Pyo's results further showed that NLFA has the highest power in detecting multidimensionality, whereas DIMTEST's varied power, depending on different conditions (e.g., sample size, number of items, and correlation between abilities). For example, NLFA was more powerful than DIMTEST with a sample size of 500 and a test length of 20 items, and performed as well as DIMTEST for a test with 50 items. The finding that DIMTEST was extremely powerful in detecting multidimensionality conflicted with Nandakumar's study, but supported Gessaroli and De Champlain's study.

In summary, a number of simulation studies that compared DIMTEST with NLFA procedures found that DIMTEST was the most promising method in identifying essential unidimensionality (Hattie, et al., 1996; Nandakumar, 1994; Nandakumar & Stout, 1993). On the other hand, NLFA was more powerful than DIMTEST in detecting multidimensionality, especially if the number of items was small (Gessaroli & De Champlain, 1996; Pyo, 2000).

2.6 AN OVERVIEW OF DETECTING DIF USING MANTEL-HAENSZEL PROCEDURE

The issue of fairness has become increasingly important in the field of language testing, especially with the increased use of standardized English language tests for making *selection* decisions that impose serious consequences on a student's academic future. In fact, it is crucial that test items are fair to all examinees and not exhibit DIF against any particular group with certain characteristics (e.g., males vs. females; Caucasians vs. African-Americans).

2.6.1 Basics Concept of DIF

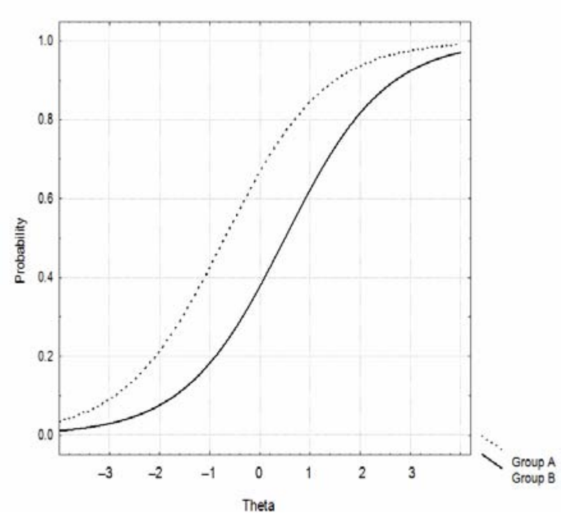
DIF is the statistical term commonly used to describe items that *function differently* in two or more groups of comparable ability (Camilli & Shepard, 1994). For example, if certain items on the CEPA-English test function differently for particular subgroups of equal level of ability, the examinees from one subgroup may receive lower scores failing to reflect their true abilities on the measured domain. Thus, such items may measure a different construct from what they intend to measure; they may reflect a bias that is not related to the domain. Such biased items become a source of error in measurement, consequently distorting inferences from examinees' scores (Camili & Shepard, 1994).

DIF is usually employed to determine whether there is a possible bias at the item level. It is considered in the process of developing new measures, adapting existing measures, or validating test score inferences. The assumption behind DIF is that when examinees have the same ability level—as indicated by a test score—the probability of answering an item correctly should be the same for every examinee. DIF can be an indicator of irrelevant source of variance such as, content of items that results in systematically lower or higher scores for members of particular groups (Messick, 1989).

It is important to point out that the presence of DIF does not necessarily guarantee that the item is biased (Angoff, 1993). In other words, DIF is necessary, but not sufficient for item bias (Clauser & Mazor, 1998). Therefore, when an item is flagged as potentially biased, further investigation by test developers and content specialists is necessary to try inform why items may be flagged as DIF. A logical analysis using expert judgment on test content can be conducted to evaluate plausible reasons for DIF including questioning why certain items are relatively easier or more difficult than others. Based on this combination of analyses, such items may be identified as biased and may be eliminated from the test (Clauser, 1998). Yet, items that exhibit DIF may have implications for curriculum and instructional changes (Harris & Carlton, 1989).

2.6.2 Types of DIF

There are two types of DIF: uniform and nonuniform. The former occurs when the probability of answering an item correctly is consistently greater for one group than another across the ability scales. That is, there is no interaction effect between the ability level and group membership (Swarninathan & Rogers, 1990). For example, an item with uniform DIF may favor males, regardless of their ability. Conversely, nonuniform DIF occurs when the probability of answering an item correctly between the comparison groups is not the same across all ability levels, indicating an interaction effect between the matching criterion and the group membership variables (Swarninathan & Rogers, 1990). For example, an item with nonuniform DIF may favor females with low ability while favoring males with higher ability. Illustration of uniform and nonuniform is given in Figures 6 and 7 respectively.



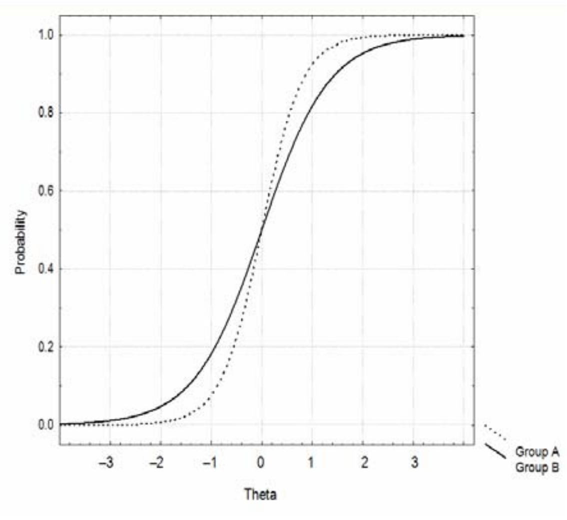


Figure 6. Example of uniform DIF

Figure 7. Example of non-uniform DIF

2.6.3 Focal and Reference Groups

DIF analysis is an initial step in determining whether a possible bias at the item level is present. The first step in DIF analysis involves identifying the focal and reference groups whose performances on an item are to be compared. The focal group is the group of primary interest, which may be at a disadvantage in answering the test items correctly. The reference group is the group whose performance is used as the standard against which the focal group is compared (Holland & Thayer, 1988). Typically, these subgroups are identified based on demographic variables, such as gender, culture, or ethnicity.

2.6.4 Matching Variables

The second step of DIF analysis involves matching members of the reference and focal groups based on their ability on the domain being measured. This is done to ensure that groups are comparable *prior* to the comparison of their test performance (Holland & Wainer, 1993). There are two types of matching variables: internal and external (Clauser & Mazor, 1998). An examinee’s observed total score on the test is commonly used as an internal matching variable for establishing comparability between the groups. Conversely, an examinee’s performance on other tests measuring similar construct as the test of interest can be used as the external matching variable. However, external criteria are rarely available (Clauser & Mazor, 1998). The matching criterion must be a valid and reliable measurement; it should be free of DIF (Angoff, 1993).

The observed total score is often used as a matching criterion because test scores are available, reliable, valid, and administered under the same conditions for all examinees (Dorans & Holland, 1993). Despite these advantages, the item (s) flagged as showing DIF might contaminate the total test score. To remedy such contamination, a purification process can be used where flagged items displaying DIF are excluded from a second analysis. Consequently, this leads to a substantial increase in the power of detecting DIF (Zenisky et al., 2003).

2.6.4.1 Thin and Thick Matching

Along with determining the types of matching variables (i.e., internal or external variables), it is also important to determine how many score groups should be used for the matching criterion; this can be accomplished by using thin and thick matching. Thick matching occurs when scores are categorized into wide intervals (e.g., 1-10, 11-20), whereas thin matching occurs when scores are grouped into narrower intervals (e.g., 1-3, 4-6). In other words, thin matching results in more categories of ability levels with fewer examinees in each category; thick matching results in fewer categories with more examinees (Clauser & Mazor, 1998). Holland and Thayer (1988) suggested that k + 1 as the optimal score group should be used for thin matching, where k is the number of test items.

A simulation study by Raju et al., (1989) investigated the use of thick matching on the MH procedure using a vocabulary test with 40 items. They suggested that four or more score groups are needed to achieve stable estimates of the common odds ratio from the MH procedure. Also, Donoghue and Allen (1993) conducted a simulation study to determine the advantages and/or disadvantages of thin versus thick matching on the MH procedure. The results showed that thin matching inflated Type I error rates when the tests were short (10 items or less).The authors determined that thin matching is superior for long tests (at least 40 items) and is similar to thick matching for moderate length tests (20-40 items). They concluded that for short tests (10 items or less), thick matching based on creating groups of equal numbers or equal percentages improved DIF detection more than thin matching.

2.6.5 Procedures for Detecting DIF

There are two major frameworks for DIF detection methods: IRT and non- IRT. Several non-IRT-based procedures have been proposed for detecting DIF in dichotomous items. The most commonly used non-IRT DIF methods is the Mantel-Haenszel (MH).

2.6.5.1 Mantel-Haenszel Procedure

The MH procedure, developed by Mantel and Haenszel in 1959 and proposed by Holland and Thayer in 1988, is a nonparametric method, which based on an analysis of contingency tables, and is commonly used in educational testing to detect DIF in dichotomous items (Clauser & Mazor, 1998; Holland & Thayer, 1988). To perform the MH procedure, answers to each test item are coded as correct (1) or incorrect (0). An observed total number of correct responses for each individual is subsequently calculated and, then the total test score is used to match an individual's abilities with both the focal (*f*) and the reference groups (*r*). A 2 x 2 *K* contingency table is then constructed for each item at each observed score level. Within each observed total score, the individual is cross-classified as belonging to the reference or focal groups and as giving a correct or incorrect response to each item as follows (Camilli & Shepard, 1994):

Table 2. A 2 × 2 Contingency Table for a Particular Item at jth Score Level

Group	Score on Studied Item		
	1	*0*	*Total*
Reference (R)	A_j	B_j	n_{Rj}
Focal (F)	D_j	n_{Fj}	n_{Fj}
Total	*m1j*	m_{0j}	T_j

A_j and B_j represent the frequency of correct and incorrect responses, respectively, for individuals in the reference group at score level *j*; C_j and D_j represent the frequency of correct and incorrect responses, respectively, for individuals in the focal group at score level *j*. $n_{R\mathrm{j}}$ and n_{Fj} are the total number of individuals in the reference group and focal group, respectively, at score level *j*; $m_{1\mathrm{j}}$ and $m_{0\mathrm{j}}$ are the total number of correct and incorrect responses, respectively, at score level *j*; and T_j is the total number of individuals at the j^{th} score level.

After constructing the *K* contingency table, the MH chi-square ($MH\ \chi^2$) statistic test with one degree of freedom is applied to test the null hypothesis that DIF is not present in the item.

That is, the odds of getting an item correct are the same for both focal and reference groups across ability levels. The $MH\,\chi^2$ statistic is calculated using the following formula (Camilli & Shephard, 1994):

$$MHX^2 = \frac{\left\{\left|\sum_{j=1}^{s}[A_j - E(A_j)]\right| - \frac{1}{2}\right\}^2}{\sum_{j=1}^{s} VAR\ (A_j)}, \qquad (2.34)$$

and the variance is computed using the following equation:

$$VAR\ (A_j) = \frac{n_{Rj}\ n_{Fj}\ m_{1j}\ m_{0j}}{T_j^2 (T_j - 1)} \quad \text{and} \quad E\ (A_j) = \frac{n_{Rj}\, m_{1j}}{T_j}. \qquad (2.35)$$

After testing the null hypothesis, the MH-Alpha or MH common odds ratio ($\alpha\hat{}_{MH}$) is calculated to estimate the DIF effect size using the following formula (Camilli & Shephard, 1994):

$$\alpha\hat{}_{MH} = \frac{\sum_{j=1}^{s} \frac{A_j D_j}{T_j}}{\sum_{j=1}^{s} \frac{B_j C_j}{T_j}}, \qquad (2.36)$$

The $\alpha\hat{}_{MH}$ is the ratio of the odds in which a member of the *reference group* will answer the item correctly, compared to the odds that a member of the *focal group* will answer the same item correctly. The value of $\alpha\hat{}_{MH}$ ranges from zero to ∞. A value of α equal to one indicates that there is no DIF on the studied item between the focal and reference groups; a value of α greater than one indicates that there is DIF against the focal group; and a value of $\alpha\hat{}_{MH}$ less than one indicates that there is DIF against the reference group (Holland & Thayer, 1988).

Because the common odd ratio, $\hat{\alpha}_{MH}$ is difficult to interpret, the $\hat{\alpha}_{MH}$ was transformed to the "MH delta" scale ($\hat{\alpha}_{MH}$) to make it symmetrical about zero using the following formula (Clauser & Mazor, 1998):

$$\hat{\Delta}_{MH} = -2.35\, ln\left(\hat{\alpha}_{MH}\right), \qquad (2.37)$$

The $\hat{\alpha}_{MH}$ is the "MH delta scale" which is used as an index of difficulty by the ETS. A value of $\hat{\alpha}_{MH}$ equal to zero indicates an absence of DIF. A positive value indicates DIF against the reference group; and a negative value indicates DIF against the focal group. An estimated standard error for $\hat{\Delta}_{MH}$ was obtained by (Dorans & Holland, 1993):

$$SE\left(\hat{\Delta}_{MH}\right) = 2.35\left(SE\left(ln\left(\hat{\Delta}_{MH}\right)\right)\right), \qquad (2.38)$$

and a *z* test statistic for the $\hat{\Delta}_{MH}$ was obtained by:

$$Z_{\hat{\Delta}_{MH}} = \frac{\hat{\Delta}_{MH}}{SE(\hat{\Delta}_{MH})}. \qquad (2.39)$$

The *z* test approximately follows a normal distribution with a mean of zero and a standard deviation of 1 (Donoghue & Allen, 1993).

ETS uses both the statistical significance test, χ^2, and the magnitude of DIF,| $\hat{\Delta}_{MH}$|, when classifying DIF items into three categories—A, B, and C—to interpret the results of DIF analysis (Camilli & Shepard, 1994; Zieky, 1993; Zwick & Ercikan, 1989) as following:

- Items with negligible DIF, or Category A, have a $\hat{\Delta}_{MH}$ value not significantly different from zero ($p \geq .05$), or $|\hat{\Delta}_{MH}| < 1$.
- Items with moderate DIF, or Category B, have a $\hat{\Delta}_{MH}$ value significantly different from zero ($p < .05$) and $1 \leq |\hat{\Delta}_{MH}| < 1.5$.
- Items with large DIF, or Category C, have a $\hat{\Delta}_{MH}$ value significantly greater than 1.0 ($p < .05$) and $|\hat{\Delta}_{MH}| \geq 1.5$.

According to ETS criteria, items with large DIF need to be removed from the test or revised. Items with moderate DIF are considered suspicious and need further examination and testing (Zieky, 1993).

There are three important advantages for using the MH procedure. First, it is computationally simple and relatively easy to implement. Second, it does not require a large sample size (at least 200 per group) (Hills, 1989). Third, it provides a χ^2 test of significance as well as an estimate of the effect size (Clauser & Mazor, 1998; Mazor et al., 1994). Despite these advantages, the MH procedure has two major weaknesses. First, the MH procedure is not effective in detecting nonuniform DIF (Hambleton & Rogers, 1989; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990). Second, when a test contains a small number of items (fewer than 20), observed scores may not represent true scores accurately, resulting in poor estimation of statistics (Zwick, 1990; Uttaro, 1992).

2.6.5.2 Purification

A study from 1988 conducted by Holland and Thayer proposed a two-stage MH procedure for purifying the matching criterion (i.e., total test score) which may be contaminated by DIF items. In the first stage, all items that exhibited DIF are removed from the total test score and then the total score is recalculated. This *new purified* total score is subsequently used as the matching criterion for the second stage of analysis. Again, all items on the test are individually tested for DIF via the inclusion of the purified items, plus one item to be studied in the separate DIF analyses. Examinees from the reference and focal groups are matched by the purified criterion and *one* studied item. Holland and Thayer (1988) recommend that when a purification process is used, the studied item should be included in the matching criterion, even if it was flagged as displaying DIF at the first stage and excluded from the matching criterion while studying other items. Several simulation studies demonstrated that failure to adhere to this recommendation may result in inflating the Type I error rates (Lewise, 1993; Zwick, 1990).

Clauser et al., (1993) compared the one-stage versus two-stage MH methods proposed by Holland and Thayer (1988). The results demonstrated that the two-stage method is superior to the one-stage method, with respect to the Type I error rate and power. This advantage depends on the number of items that are detected at the first stage of DIF. The greater the number of DIF items that are identified at the first stage, the greater the advantage of performing the two-stage MH approach.

More recently, Zenisky et al., (2003) evaluated the two-stage MH approach of detecting DIF using actual data. Results support the advantage of using the two-stage approach in detecting DIF, as demonstrated by Clauser et al., (1993) along with Holland and Thayer (1988), especially when the items that are flagged show relatively large DIF at the first stage (at least 30% of the items displayed DIF).

2.6.6 Factors Influencing the Power of Detecting DIF using MH Method

There are many factors that might influence the power of detecting DIF, such as sample size, ability differences between groups, percentage of DIF items, DIF magnitude, test length, item characteristics, and the type of DIF (uniform and nonuniform). Sample size in the focal and reference groups is the first important factor affecting the power of detecting DIF for all DIF procedures. Previous DIF simulation studies indicate that the power of detecting DIF increases as the sample size increases (Swaminathan & Rogers, 1990; Mazor et al., 1992; Rogers & Swaminathan, 1993; Tian et al., 1994; Stout et al., 1997; Gierl et al., 2004). So when the sample size is large, a small difference between the focal and reference groups results in a significant statistical test. In general, sample sizes with 200-250 subjects per group have been shown to ensure the satisfactory power of the MH procedure in detecting DIF, while controlling the inflation of Type I error (e.g., Mazor et al., 1992; Clauser & Mazor, 1998).

Another factor that influences the power of DIF detection is the ability distribution differences between the reference and focal groups. Simulation researches have shown that the power to detect DIF increases as the ability distribution for the comparison groups become more comparable (e.g., Rogers, 1989; Mazor et al., 1992; Shealy & Stout, 1993; Zwick et al., 1993; Jodoin & Gierl, 2001). For example, Mazor et al., (1992) studied the effects of equal and unequal ability distributions on the MH procedure using simulated data, and they recommended that large sample sizes should be utilized when comparing groups of differing abilities.

The third influencing factor is the amount of DIF contained in an item, or the DIF effect size. As the DIF effect size increases, the detection power of DIF is also likely to increase. A DIF effect size of .25 is common in actual testing situations (Zwick et al., 1993; Spray & Miller, 1994; Chang et al., 1996).

The percentage of items containing DIF is another important factor because it can reduce the validity of the matching variable. That is, as the percentage of DIF items increases, the contamination of the matching variable increases. Consequently, the ability estimates are less reliable and the matching variable is less accurate. Therefore, the power of the DIF procedures is likely to decrease, thus increasing the need for purification of the matching variable. For instance, Miller and Oshima (1992) found that the two-stage purification procedure did not have a substantial impact on MH DIF detection when the proportion of DIF items was small (i.e., 5 or 10%); but it did when the proportion of DIF items was large (i.e., 20 or 40%). Several researchers have reported that as percentage of DIF items increases to 10% or 15%, the MH method starts to lose control over Type I error (Miller & Oshima, 1992; Rogers & Swaminathan, 1993; Narayana & Swaminathan, 1994; Uttaro & Millsap, 1994; Fidalgo et al., 2000). However, more recent results have shown that higher percentages of DIF items (e.g., 5% and 10%) do not necessarily lead to inflated Type I error for the MH DIF detection method (Wang & Yeh, 2003; Wang, 2004; Wang & Su, 2004). The proportion of DIF items of 5% and 10% is common when DIF is related to gender and race, whereas 20% may occur when DIF is related to differential instructional opportunities to learn the curriculum (Miller & Oshima, 1992). In practice, it is common place for 10 to 15% of the items in a standardized achievement test to exhibit DIF (Clauser, 1993).

The fifth influencing factor is related to item characteristics, such as difficulty and discrimination. Previous studies found that DIF procedures are strongly influenced by the overall item difficulty (e.g., Donoghue et al., 1993; Donoghue & Allen, 1993; Linn, 1993). What can be detected as DIF may be “differences in item discrimination or a statistical artifact of a ceiling or floor effect in the item difficulty” (Linn, 1993. p. 361). The solution for this problem, as Linn suggested, is to depend on DIF procedures that do not rely on the difference in item discrimination levels between compared groups. Donoghue et al (1993) also found that DIF is affected by both the item discrimination and the difference in item difficulty between compared groups that have “the same non-zero guessing parameter” (p. 165), and is also affected by the inclusion of the studied item. Mazor et al (1994) demonstrated that items therefore actually displaying DIF, and are most likely to be detected as DIF are those with the lowest discrimination values, the highest difficulty indexes, and when there is a small difference between ability distributions.

Another factor related to DIF detection is test length—generally, the longer the test, the more reliable the total score, which results in increasing the power of detecting DIF. Several DIF simulation studies manipulated the test length ranging from 20 to 80 items (e.g., Rogers & Swaminathan, 1993; Clauser et al., 1993; Narayanan & Swaminathan, 1994; Jodoin & Gierl, 2001).

The final factor affecting the power of detecting DIF is the type of DIF (uniform DIF and nonuniform DIF). Several DIF simulation studies that compared the MH procedure with the logistic regression (LR) procedure reveal that the MH procedure is more accurate in detecting uniform DIF, whereas the LR procedure is more accurate in detecting nonuniform DIF (e.g., Rogers & Swaminathan, 1993; Lopez-Pina, 2001; Swaminathan & Rogers, 1990).

2.6.7 DIF Approach in Language Testing

To determine whether group differences in test performance reflect *true* differences or are due to the existence of items bias, DIF studies with respect to gender, ethnicity, and language have been conducted. Only a few studies have examined how different language groups perform differently on language proficiency testing (Chen & Henning, 1985; Sasaki, 1991; Ryan & Bachman, 1992).

Chen and Henning's (1985) study may be one of the first DIF studies in language testing. They employed an adapted Angoff delta-plot method to detect DIF items on the UCLA English as a Second Language Placement Examination (ESLPE) across Chinese (*n* = 77) and Spanish (*n* = 34) language groups. The ESLPE consisted of five subtests: listening, reading, grammar, vocabulary, and writing error correction. They plotted item difficulty estimates calibrated by the Rasch model for each item across the two groups, and they considered items beyond the 95% confidence interval around the regression line as DIF items. The results indicated that four DIF items were detected from the vocabulary test and five from the grammar test favoring the speakers group.

Sasaki (1991) also detected DIF on the UCLA ESLPE across Chinese (*n* = 262) and Spanish (*n* = 81) language groups. To investigate DIF, they used the same modified delta-plot method employed in Chen and Henning's study in addition to Scheuneman's chi-square method. Results of the study identified 22 of ESLPE items as showing DIF (4 listening, 1 reading, 4 grammars, 7 vocabularies, and 6 writing error corrections). The substantive analyses of the DIF results showed that vocabulary items with English–Spanish cognates favored the Spanish group, and items with idiomatic expressions favored the Chinese group.

The MH technique has been used as the main DIF detection approach in two large-scale language testing studies (Ryan & Bachman, 1992; Elder, 1996). For example, Ryan and Bachman (1992) detected DIF in the Test of English as a Foreign Language (TOEFL) and the First Certificate in English (FCE), across Indo- European (*n* = 792) and Non-Indo-European (*n* = 632) language groups using the MH procedure. The results showed that the TOEFL and FCE did not demonstrate gender DIF. However, 45% and 63% of items displayed significant DIF on the TOEFL and FCE, respectively when language groups were used as the grouping criterion (Indo-European vs. non-Indo-European). Elder (1996) also used MH approach to examined DIF on the reading and listening tests of the Australian Language Certificate for examinees with different backgrounds (compared Chinese, Italian, and Greek languages vs. heritage speakers). A large number of DIF items were detected, particularly in the Chinese exam.

2.7 AN OVERVIEW OF EQUATING METHODS: EQUIPERCENTILE AND IRT EQUATINGS

2.7.1 Basics Concept of Equating

In large-scale testing, multiple forms are often used to ensure test security, particularly for admission tests where items cannot generally be used more than once. Although multiple test forms are carefully constructed to be as similar as possible in content and statistical specifications, the forms differ somewhat in difficulty so that scores from the forms are not interchangeable without some type of equating. Equating procedures ensure that scores are equivalent across test forms so that examinees neither have advantages nor disadvantages from taking a relatively easy or difficult version of a test. Because test forms are equated, it does not matter which test form the examinee completes. Equating adjusts the differences in difficulty among forms containing the same content and statistical specifications, so that the forms can be used interchangeably. In other words, equating is used to achieve comparability by placing scores from multiple test forms on a common scale (Kolen & Brennan, 2004).

2.7.2 Equating Properties

Test equating must meet the following five conditions before being successfully employed (Kolen & Brennan, 2004):

1. Equal constructs: the equated test forms must measure the same construct.
2. Equal reliability: the equated test forms must have the same reliability.
3. Equity: examinees of identical performance levels on the underlining ability must obtain the same scores, no matter what test forms they are taking.
4. Symmetry: the equating transformation should be symmetric, meaning that the equating transformation from Forms Y to X must be the inverse of the equating transformation from Forms X to Y.
5. Population invariance: the equating transformation must be the same, regardless of the group of examinees used to perform the equating; the equating transformation should be invariant across groups of examinees from which it is derived. No matter which group of examinees is used for equating, the equating result should not change due to the characteristics of the particular groups, except for the underlying construct the test is measuring.

2.7.3 Data Collection Designs for Equating

A variety of designs are used in data collection for equating, and the choice of a design involves considering both practical and statistical issues. The four commonly used data collection designs are (see Table 3): (a) single-group design, in which two or more test forms are administered to the same group of examinees; (b) single-group design with counterbalancing, in which the test forms to be equated are assigned to the participants, but the order of administration is counterbalanced and randomly assigned; (c) random-groups design, or equivalent groups design, in which the two randomly selected groups of equivalent ability take different forms of the test; and (d) anchor-item design, in which a group of examinees from different populations is administered different test forms with common sets of items, V (Kolen & Brennan, 2004). In the current study, random-group and anchor-item designs were used to equate the CEPA- English test forms.

Table 3. Illustration of the Four Data Collection Designs for Test Equating

		Design		
		Single–group design		
		Test form		
Sample	X			Y
P1	√			√
		Single-group design withcounterbalancing		
		Test form		
	X	*Y*	*X*	Y
Sample	1st	2st	1st	2st
P1	√			√
P2		√	√	
		Random-groups design		
		Test form		
Sample	X			Y
P1	√			
P2				√
		Anchor-item design		
		Test form		
Sample	X	Y		V
P1	√			√
P2		√		√

2.7.3.1 Random-Groups Design

In the random-groups design, the test forms are randomly assigned to examinees through a spiraling process. For example, the booklets are handed out so that the first examinee receives Form X, the second examinee Form Y, the third examinee Form X, and so on. This spiraling process produces randomly equivalent groups taking Form X and Form Y. In addition, this design assures that examinees will be divided equally between the two forms. Thus, when spiraling is utilized for random assignment, any differences in the average performance between groups on the two forms are attributed to the differences in difficulty between the forms. Using the random-groups design, has many advantages: the elimination of fatigue and practice effect, the minimization of testing time since each examinee takes only one form; and the ability to equate more than one new form in the spiraling process. However, because this design requires that all test forms should be available and administered at the same time, test security might be a concern. In addition, because different examinees take only one form, larger samples are generally required for this design.

2.7.3.2 Anchor-Item Design

Although all four data collection designs have been used in IRT vertical equating, the anchor-item design—also called the common-item nonequivalent groups design—is the most widely used. Anchor-item design typically is used when it is impossible to administer more than one form of a test due to test security. In this design, different test forms can be used, regardless of whether the group is equivalent or not (Kolen & Brennan, 1995, 2004).

Within the anchor-item design, each test form contains a set of anchor or common items that may be internal or external to the test forms. *Internal* anchor-items refer to items that are part of the test itself and that contribute to the examinees' total scores; *external* anchor-items refer to items that do not contribute to the examinees' total scores. External anchor-items are frequently administered as a separate and timed block of items (Kolen & Brennan, 1995).

The use of anchor-items requires specific statistical assumptions (von Davier et al., 2004). First, anchor-item design assumes that there are two populations of examinees *P* and *Q*, in which Form *X* is administered to population *P*, Form Y is administered to population *Q*, and the anchor set of items V is administered to both populations. It also assumes that the two samples are independently and randomly drawn from *P* and *Q*, respectively. Finally, the content of the anchor-items should adequately represent the entire test in both content and statistical characteristics.

2.7.3.2.1 Selecting Anchor-Items

Because anchor-items are used to equate test forms, they must be carefully selected. Even when an IRT equating study is well-designed and satisfies the equating requirements, acceptable equating results can be inaccurate because anchor-items differ from one form to another. IRT equating using an anchor-item design can be successful if it satisfies five important requirements. First, the anchor-items must be carefully selected. According to Kolen and Brennan (2004) anchor-items must be representative of the entire test in terms of the content and difficulty level of the items. Second, anchor-items must maintain the same context (e.g., wording must be identical on the test forms) and positioned across the multiple forms. Third, the number of anchor-items should be long enough to adequately represent the entire test. A rule of thumb for the minimum length of the anchor-items is 20-25% of the total test length (Kolen & Brennan, 2004). Fourth, anchor-items must reflect the full range of examinees' abilities by including a mixture of easy and difficult items; so that, the equating is accurate at the low and high end of the score range. Finally, the quality of IRT equating using an anchor-item design depends on the similarity of the groups taking the new and old forms of the test. Therefore, the groups must be similar to each other in their ability distribution at least for the anchor items (Kolen & Brennan, 2004).

2.7.4 Equating Methods

There are several equating methods that can be used to yield comparable test scores. These methods can be categorized into vertical equating and horizontal equating. Vertical equating involves equating tests that have the same construct but are administered to groups of students with different abilities (nonequivalent groups), such as students in different grade levels (e.g., Grades 3 and 5), whereas horizontal equating involves equating tests within the same content from a different administration of the test for a single grade (Kolen & Brennan, 2004). Equating methods also can generally be classified as traditional equating or item response theory equating. Traditional equating methods are based on CTT, and are rooted in the observed-score equating. CTT has three equating methods: mean, linear, and equipercentile, whereas IRT has two equating methods: IRT true-score equating, and IRT observed-score equating (Kolen & Brennan, 2004).

2.7.4.1 CTT Equating Methods

2.7.4.1.1 Mean Equating

In mean equating, the means on the two forms are set equal to one another for a particular group of examinees; that is, the Form X scores are converted so that their mean will equal the mean score on Form Y. This type of equating assumes the differences in difficulty between the forms are constant along the score scale. For instance, under mean equating, if Form X is 3 points easier than Form Y for high-scoring examinees, it is also 3 points easier than Form Y for low-scoring examinees (Kolen & Brennan, 2004).

2.7.4.1.2 Linear Equating

Linear equating, the second type of CTT equating, is a special case of equipercentile equating. In linear equating, the mean and standard deviation on the two forms are set equal. Specifically, the raw total scores that are an equal (signed) distance from their means in standard deviation units are set to be equal. Unlike mean equating, therefore, linear equating allows both forms to differ in difficulty along the score scale. For example, this type of equating allows Form X to be relatively more difficult than Form Y for low-scoring examinees than for high-scoring examinees (Kolen & Brennan, 2004).

2.7.4.1.3.1 Equipercentile Equating

In equipercentile equating, scores on multiple test forms are considered to be equivalent if they have the same percentile rank for a given group of examinees (Kolen & Brennan, 1995). As a result, the score distributions of the new Form X equated to the scale of the old Form Y will be equal to the score distributions of the old form in the population of examinees (Kolen &

Brennan, 1995). Like linear equating, equipercentile equating allows for differences in difficulty between the two forms to vary along the score scale; for instance, Form X could be more difficult than Form Y at low and high scores, but less difficult at the middle scores (Kolen & Brennan, 2004).

The equipercentile equating function, $e_Y(x)$, is developed when the distribution of the Form X scores converted to the Form Y scale is the same as the distribution of Form Y (Kolen & Brennan, 2004):

$$G^* \,[e_Y(x)] = G[y], \qquad\qquad (2.40)$$

$G^*[e_Y(x)]$ the cumulative distribution function of $e_Y(x)$ in the population, and $G[y]$ is the cumulative distribution function of Y in the same population (who took Form X). When X and Y are continuous random variable, the equipercentile equating function is (Kolen & Brennan, 2004):

$$e_Y(x) = \; G^{-1}[F(x)], \qquad\qquad (2.41)$$

$e_Y(x)$ is the cumulative distribution function used to convert scores on Form X to the scale of Form Y.G^{-1} $G^{-1}$is the inverse of the cumulative distribution function G, which is the cumulative distribution function of Y in the population. $F(x)$ is the cumulative distribution function of X in the population. Likewise, by the symmetry property (Kolen & Brennan, 2004):

$$e_x(y) = F^{-1} \,[G(y)], \qquad\qquad (2.42)$$

Where $e_x\,(y)$ is the cumulative distribution function used to convert scores on Form Y to the scale of Form X. F^{-1} $F^{-1}$is the inverse of the cumulative distribution function F. For a population of examinees, the equipercentile equivalent for a given Form X score can be constructed by finding the proportion of examinees in the population earning a score at or below that Form X score. Then, the Form Y score that has the same percentile rank will be equivalent. The equipercentile equating function assumes that test scores (X and Y) are continuous random variables (Kolen & Brennan, 2004).

When X and Y are discrete random variables, which is typically the case, usually no score on Form Y has exactly the same percentile ranks as a score x on Form X (Kolen & Brennan, 2004). To convert discrete test scores (i.e., number-correct test scores) into continuous ones, percentiles and percentile rank are traditionally used. This convention can be done by using percentiles and percentile rank which uniformly spread the density at each discrete test score point over the range of the score +/-.05 (Kolen, 2006). For example, examinees with an integer score of 27 are assumed to be uniformly distributed in the range of 26.5-27.5. Consequently, this linear interpolation makes the discrete score distributions piecewise linear and, therefore, continuous (Lee & von Davier, 2008). As a result, the inverse of the percentile rank (P) is calculated as follows (Kolen & Brennan, 2004):

$$P^{-1}[P^*] = \frac{P^* \,/\, 100\, - F(x_U^* - 1)}{F\,(x_U^*) - F(x_U^* - 1)} + \ (x_U^* - .5)\ , \qquad (2.43)$$

Equipercentile equating can be conducted in two steps (Kolen, 1984). The first step involves tabulating or plotting the relative cumulative frequency (i.e., proportion of examinees who score at or below each raw score point) distributions for the two forms to be equated. The second step involves finding the score on the old form that has the same relative cumulative frequencies on the new form and declaring that to be equated scores. To discern this, discrete score distributions need to be made continuous. Percentiles and percentile ranks are usually used to continuize a discrete score distribution (Kolen & Brennan, 2004).

2.7.4.1.3.2 Properties of Equipercentile Equating

Equipercentile equating has two desirable properties. First, equipercentile equating will always result in equated scores within the range of possible scores as defined by percentiles and percentile ranks (- .5 $\leq x \leq K_{x+}$ .5). In addition, equated scores have the same mean, standard deviation, and distributional shapes (skewness, kurtosis, etc.). If test scores are continuous, then these distributions will be the same; if test scores are discrete, then the equated Form X score distribution will differ slightly from the Form Y distribution (Kolen & Brennan, 2004). Equipercentile equating, typically requires larger sample sizes than does linear or mean equating. It is also substantially more computationally complex than the linear or mean methods, especially for the common item nonequivalent groups design (Kolen & Brennan, 2004).

2.7.4.1.3.3 Smoothing in Equipercentile Equating

The use of sample percentiles and percentile ranks to estimate equipercentile relationships is not sufficiently precise. Therefore, smoothing methods are used to improve equipercentile equating by reducing the equated random error. In general, two types of smoothing are used: presmoothing and postsmoothing, while in presmoothing, the test score distributions for the two test forms to be equated are smoothed before the equipercentile equating is conducted. In postsmoothing, the equipercentile equivalent, $e\hat{}_{Y(x),}$ is smoothed directly. However, smoothed equipercentile equating can introduce systematic errors (i.e., bias) and may result in less error total (random error plus systematic error) than unsmoothed method, although the expectation is that any increase in the systematic errors due to smoothing would be offset by a decrease in the random error (Kolen & Brennan, 1995). Kolen (1984) conducted that a postsmoothing based on cubic splines is preferred to unsmoothed equipercentile equating.

2.7.4.1.3.4 Postsmoothing Using Cubic Splines

A smoothing cubic spline function is fitted to the equipercentile equating function is computed from unsmoothed raw score distributions relating scores on the new form to scores on the old form. Postsmoothing using cubic spline fits a continuous, cubic, function between adjacent integer scores over the range of scores. For integer scores, x_i, the spline function is (Kolen & Brennan, 2004):

$$\hat{d}_Y(x) = v_{0i} + v_{1i}\,(x - x_i) + v_{2i}(x - x_i)^2 + v_{3i}(x - x_i)^3, x_i \leq x < x_i + 1. \tag{2.44}$$

The weights (v_{0i}, v_{1i}, v_{2i}, v_{3i}) change across score points so a different cubic equation is defined between each integer score. At each score point, x_i, cubic spline is continuous (continuous second derivatives). Over score points for which the spline is fit, the spline function is minimized to have a minimum curvature and to satisfy the following constraint (Kolen & Brennan, 2004):

$$\frac{\sum_{i=low}^{high}\left[\frac{\hat{d}_Y(x_i) - \hat{e}_Y(x_i)}{\widehat{se}[\hat{e}_Y(x_i)]}\right]^2}{x_{high} - x_{low}} \leq S, x_{low} \leq x \leq x_{high} \leq K_x, \tag{2.45}$$

where x_{low} is the lowest integer score in the range, x_{high} is the highest integer score in the range, and $\hat{se}[\hat{e}_{Y(xi)}]$ is the estimated standard error of equipercentile equating that is used to standardize the differences between unsmoothed and smoothed relationships.

The cubic spline method focuses on choosing among degrees of smoothing of the equipercentile relationships. The degree of smoothing is controlled by the value of the parameter *S*, where $S \geq 0$. When $S = 0$, the fitted spline equals the unsmoothed equivalents at each integer score point. When *S* is very large, then the spline function is a straight line. In practice, values of *S* between 0 and 1 are used; this usually produces adequate results (Kolen & Brennan, 2004). The spline function is fit over a restricted range of scores x_{low} to x_{high}. As a result, the spline function is not influenced by score points where there are few examinees and large or poorly estimated standard errors (Kolen & Brennan, 2004). Kolen (1984) suggested excluding score points with percentile ranks lower than .5 and above 99.5. Furthermore, a linear interpolation procedure can be used to obtain equivalent scores outside the range of the spline function (see Kolen & Brennan, 2004, p.87).

An important requirement of the equating method is that an equating relationship should satisfy symmetry (Lord, 1998). However, the spline function is not symmetric because the spline function, $\hat{d}_Y(x)$ that is used to convert Form X scores to the Form Y scores is different from the spline function used to convert Form Y to Form X. To satisfy the symmetry property, it is necessary to define the $\hat{d}_Y(x)$ as the $\hat{d}_x^{-1}(x)$; this inverse can be used to convert the Form X scores to the Form Y scale. Then, a symmetric equating function, $\hat{d}_{\mathrm{Y}}^{*}(x)$, can be defined as the average of the two splines (Kolen & Brennan, 2004):

$$\hat{d}_Y^*\ (x) = \frac{\hat{d}_Y\ (x) + \hat{d}_x^{-1}(x)}{2},\ -.5\ \leq x \leq K_x +\ .5., \qquad (2.46)$$

2.7.4.2.1 IRT Equating Method

Equating with IRT offers more benefits than equating with CTT. One important advantage of IRT is that items are on the same scale—that is, the ability estimated from different subsets of items is on the same scale. Not only IRT equating is useful for the nonequivalent group design, but it is more appropriate when the relationship is nonlinear, the sample size is large, and accuracy is needed along the score scale. However, IRT equating demands strong statistical assumptions, such as unidimensionality and local independence, for unidimensional IRT models. Lastly, the selected IRT model must fit the observed data. If the assumptions hold true and the IRT model is appropriate, then the IRT equating methods satisfy the equating properties or requirements (Kolen & Brennan, 2004).

Essentially, the IRT equating method involves four steps (Hambleton & Swaminathan, 1985): 1) choosing an appropriate equating design based on the nature of the test, the group of examinees, and the statistical assumptions required to achieve the desired degree of equating precision; 2) determining an appropriate IRT model; 3) placing parameter estimates on a common scale; and 4) equating the test scores—that is, making a decision on the scale for reporting test scores, either by reporting the ability score, estimated true score, or observed score.

2.7.4.2.2 IRT Scaling: Placing Parameter Estimates on a Common Scale

When the examinees in each groups of the anchor-item design differ in their ability levels, item parameter estimates are not on the same scale. However, the common items are used to place or link IRT parameter estimates from multiple test forms on the same scale using separate, concurrent, or fixed calibration (Kolen & Brennan, 2004).

2.7.4.2.2.1 Separate Calibration

In separate calibration, the item parameters for common-items for each test form are separately estimated in a single run with MULTILOG group analysis. Placing item and ability parameters on the same scale causes indeterminacy between the scales for these parameters (Hambleton & Swaminathan, 1985). Thus, each test form calibration has a unique scale, but item parameter estimates from different calibration-runs are on ability scales that are linearly related to one another (Kolen, 2007; Kolen & Brennan, 1995). A linear transformation of IRT vertical scales is needed in order to compare results from different calibrations and to produce score scales with a meaningful interpretation. More specifically, “linking” constants (*A*, slope, and *B*, intercept) are estimated and used to transform the scale from one calibration to that of another (Kolen & Brennan, 2004).

In estimating the 3PL IRT model, it is assumed that the θ on scale *I* is linearly transformed into scale *J*:

$$\theta_{ji} \ = A\theta_{Ii} + B, \qquad (2.47)$$

A is the slope, and *B* is the intercept; they are constants in a linear transformation equation. θ_{ji} and θ_{Ii} are θ values for an individual *i* on scale *J* and scale *I*. Furthermore, the item parameters on scale *I* are transformed:

$$a_{ji} = \frac{a_{Ii}}{A}, b_{ji} = Ab_{Ii} + B, and\ c_{ji} = c_{Ii}, \qquad (2.48)$$

In this formula, the *c* or guessing parameter is independent of the scale transformation. Substituting a_{ji} for a_{Ii}, b_{ji} for b_{Ii}, and θ_{ji} for θ_{Ii} produce exactly the same probability of getting item *i* correct.

Various methods have been proposed to find these linking transformations with the dichotomous IRT models. These separate calibrations are the moment's methods (i.e., the mean/mean and the mean/sigma methods) and the characteristic curve methods (i.e., Haebara and Stocking-Lord methods).

In the mean/mean method, the mean of the common items' discrimination (*a*) and difficulty (*b*) parameter estimates are used to find the transformation constants (*A* and *B*) as follows:

$$A = \frac{\mu\ (a_I)}{\mu\ (a_J)} \qquad B = \mu(b_J) - A\,\mu(b_I), \qquad (2.49)$$

where $\mu\ (a_I)$ and $\mu(a_J)$ are the means of *a* parameters for the common items on Scale *J* and on Scale *I*, respectively; *A* is the slope, and *B* is the intercept of the linear transformation line (Loyd & Hoover, 1980; as cited in Kolen & Brennan, 2004).

In the mean/sigma method, on the other hand, the means and the standard deviations of the common items *b* parameter estimates are used to calculate the transformation constants as follows:

$$A = \frac{\sigma(b_J)}{\sigma(b_I)} \qquad B = \mu(b_J) - A\,\mu(b_I) \ , \tag{2.50}$$

where σ (b_J) and σ (b_I) are standard deviations of the *b* parameters for the common items on Scale *J* and on Scale *I*, respectively; μ (b_J) and μ (b_I) are the means of the *b* parameters for the common items on Scale *J* and on Scale *I*, respectively; and *A* is the slope and *B* is the intercept of the linear transformation line (Marco, 1977; as cited in Kolen & Brennan, 2004).

For both the mean/mean and mean/sigma methods, the value of *A* and *B* is used in the following way:

$$\theta_{Ji} = A\theta_{Ii} + B;\ a_{ji} = \frac{a_{Ii}}{A}; b_{Ji} = AB_{Ii} + B, c_{Ji} = c_{Ii}, \tag{2.51}$$

Because both the mean/mean and mean/sigma methods *do not* simultaneously consider all of the item parameter estimates (*a, b,* and *c*), almost identical ICCs are produced by different combinations of *a, b,* and *c* parameter estimates over the range of abilities at which most examinees score. For example, two items with different *b* parameter estimates can have very similar ICCs. In this case, the mean/sigma methods may be affected too much by the difference between the *b* parameters (Kolen & Brennan, 2004).

As a response to this problem, Haebara (1980) and Stocking and Lord (1983) developed the characteristic curve (ICC) methods that simultaneously consider all of the item parameters. Haebara's approach (1980) involves identifying the linear transformation that minimizes the sum of squared difference between the common items' characteristic curves summed across every examinee. This function can be expressed as:

$$Hdiff(\theta_i) = \sum_{j:V} \left[p_{ij}\left(\theta_{Ji}; \hat{a}_{Jj}; \hat{b}_{Jj}; \hat{c}_{Jj}\right) - p_{ij}\left(\theta_{Ji}; \frac{\hat{a}_{Ij}}{A}; A\hat{b}_{Ij}; \hat{c}_{Ij}\right)\right]^2, \qquad (2.52)$$

By squaring and summing the difference between each ICC on the two scales for common items (*j: V*), *Hdiff* is cumulated over examinees to find the transformation constants *A* and *B* that minimize the following criterion:

$$Hcrit = \sum_i Hdiff(\theta_i), \qquad (2.53)$$

In contrast, Stocking and Lord (1983) use the sum of the ICCs over the common items to find the transformation constants *A* and *B* that would minimize the following function:

$$SLdiff(\theta_i) = \left[\sum_{j:V} p_{ij}\left(\theta_{Ji}; \hat{a}_{Jj}; \hat{b}_{Jj}; \hat{c}_{Jj}\right) - \sum_{j:V} p_{ij}\left(\theta_{Ji}; \frac{\hat{a}_{Ij}}{A}; A\hat{b}_{Ij}; \hat{c}_{Ij}\right)\right]^2, \qquad (2.54)$$

In the SL_{diff} (θ_i) formula, each ICC is summed over the common items to compute a TCC. By squaring the difference between each TCC for the common items, the SL_{diff} is cumulated over examinees to find the transformation constants *A* and *B* that minimize the following criterion:

$$SLcrit = \sum_i SLdiff\ (\theta_i). \qquad (2.55)$$

2.7.4.2.2.2 Concurrent Calibration

Unlike the separate calibration method, the concurrent calibration approach simultaneously estimates item and ability parameters in a single computer run. In this method, the items that are not completed by a particular group are treated as *not reached* or *missing data* (Hambleton et al., 1991; Lord, 1980). Then, parameters from all the different test forms are simultaneously estimated and placed on a common scale via a single computer run (Kolen & Brennan, 2004).

2.7.4.2.2.3 Fixed Calibration

The fixed calibration procedure combines the features of both the separate and concurrent calibration methods. In the fixed calibration, the *a, b,* and *c* parameter estimates of common items whose parameters are known (either from a previous year's calibration or a separate calibration) are fixed at their previously estimated values. Then, item parameters for the remaining non-common items are estimated with the common items. Thus, the item and ability estimates for these non-common items are on the same scale as the common items (Kolen & Brennan, 2004).

2.7.4.2.2.4 Comparison of Separate and Concurrent Calibration Methods

Through simulation studies, a few researchers have compared the item parameter estimates using separate and concurrent calibration methods (Wingersky et al., 1987; Petersen et al., 1983; Kim & Cohen, 1998; Hanson & Béguin 2002). Simulation studies on separate calibration for dichotomous IRT models generally found that the Stocking-Lord method produces more stable and accurate equating results than the mean/mean and mean/sigma methods (e.g., Baker & Al-Karni, 1991; Hung et al., 1991; Way & Tang, 1991; Kim & Cohen, 1992).

After comparing the concurrent and separate estimation procedures, Petersen et al. (1983) and Wingersky et al. (1987) found that, in general, concurrent calibration produces somewhat more accurate equating results than those of separate calibration methods. Both these studies used the LOGIST computer program (Wingersky et al., 1982), which uses joint maximum likelihood to estimate the item parameters.

Kim and Cohen (1998), in their simulated unidimensional data, examined separate (Stocking-Lord method) and concurrent calibration methods. They used BILOG (Mislevy &Bock, 1990) for separate item parameter estimations and MULTILOG (Thissen, 1991) for concurrent estimations based on a marginal maximum a posteriori estimation and MMLE. Kim and Cohen found that separate and concurrent calibration provided similar results, but separate estimation with the Stocking-Lord method provided more accurate results when the number of common items was small (5 common items). They also found that concurrent calibration procedures produced more accurate equating results when the data fit the IRT model than separate calibration with the Stocking-Lord method, but concurrent calibrations may be less robust to the violation of IRT assumptions. As a result, Kim and Cohen argue that separate calibration using the Stocking-Lord method may be the best alternative IRT transformation method.

Unlike Kim and Cohen (1998), Hanson and Béguin (2002), in their simulated unidimensional data, used BILOG-MG (Zimowski et al., 1996) and MULTILOG for both concurrent and separate estimation in comparing the performance of separate (Mean/Mean, Mean/Sigma, Haebara, and Stocking-Lord) versus concurrent item parameter estimation in an anchor-item equating design. They found, that in general, BILOG-MG and MULTILOG performed similarly. Within separate calibration, the Stocking-Lord and Haebara methods produced lower error than the mean/mean and mean/sigma method, and concurrent calibration resulted in lower error than separate estimation, except for the MULTILOG nonequivalent group condition.

2.7.4.2.3 Equating Test Scores in IRT Scaling

2.7.4.2.3.1 IRT Ability-Score Equating

Once item parameter estimates for test forms are placed on the same scale, the ability estimate obtained for an examinee will be the same, regardless of which form of the test the examinee actually takes. Therefore, if a test is scored using the estimated IRT abilities, the equating is completed, meaning no additional steps are needed to develop a relationship between scores on Form X and Form Y. Using a linear transformation, the ability scores can be converted so that they are positive integers, which are easier to interpret for reporting purposes (Kolen & Brennan, 2004). However, there are several problems in using the estimated 3PL IRT abilities. First, examinees with the same number of correct scores (0-36) often obtain different estimated abilities because of pattern scoring. Another problem is that estimated abilities at the low and high ends of the scale tend to have more measurement errors associated with them than those in the middle (Kolen & Brennan, 2004).

2.7.4.2.3.2 IRT True-Score Equating

Because of practical problems associated with the IRT ability-score equating, test forms are typically scored using number-correct scores. However, when number-correct scores are used, the IRT equating processes require an additional step that entails equating true scores or observed scores (Kolen & Brennan, 2004).

In true-score equating, once the item parameter estimates from Forms X and Y are on the same scale, then the IRT true-score equating can be used to equate true scores on Form X to true scores on Form Y. That is, true scores on one form associated with a given θ are equivalent to the true scores on another form associated with that θ. The number-correct true score corresponding to θ_x and θ_y defines as (Kolen & Brennan, 2004):

$$\tau_{x\,(\theta_{xi})} = \sum_{j=1}^{N_x} p_{ij\,(\theta_{xi};a_j,b_j,c_j)} \text{ and } \tau_{y\,(\theta_{yi})} = \sum_{j=1}^{N_y} p_{ij\,(\theta_{yi};a_j,b_j,c_j)}, \qquad (2.56)$$

where N_X refers to the number of items on Form X, and N_y refers to the number of items on Form Y. However, there is a practical problem in using IRT true-score equating with the 3PL IRT model: an extremely low true-score cannot be estimated, because as θ approaches negative infinity, the probability of answering an item correctly approaches the c not 0. Therefore, the true scores, τ_x and τ_y, are associated with a value of ability only for the following ranges (Kolen & Brennan, 2004):

$$\sum_{j=1}^{N_x} c_j < \tau_x\,(\theta_{xi}) < N_x \;\; and \;\; \sum_{j=1}^{N_y} c_j < \tau_y\left(\theta_{yi}\right) < N_y\,, \qquad (2.57)$$

where $c\hat{}_j$ is the estimate of c_j, and k is a raw score in which k = 0, 1, 2,… , N.

To find $\tau_y\left(\theta_{yi}\right)$ equivalents to the IRT true-score equating involves three steps. First, a true score, τ_x, on Form X within the range $\sum_{j=1}^{N_x} c_j < \tau_x(\theta_{xi}) < N_x$ is specified. Then, the θ_i that corresponds to that true score (τ_x) is found. Finally, the true score on Form Y, τ_y , that corresponds to that θ_i is determined. For a given true score on Form X, the Newton Raphson iterative process can be used to determine the examinee θ_i corresponding to the true score (Kolen & Brennan, 1995).

2.7.4.2.3.3 IRT Observed-Score Equating

The IRT observed-score method, on the other hand, is conducted by estimating the frequency distributions of observed correct-scores on each form, and then by using equipercentile equating to approximately equate these estimated observed scores. More specifically, for Form X, the compound binomial is used to get the distribution of observed number-correct scores for each examinee at a given ability. These observed score distributions are then cumulated over a population of examinees to get a number-correct observed score distribution for Form X. The same procedures are used to get a number-correct observed score distribution for Form Y. Finally, by using equipercentile equating, the number-correct observed score distribution is equated. For IRT observed-score equating, the distribution of ability in the population of examinees must be specified (Kolen & Brennan, 1995).

2.7.4.2.3.4 Comparison of IRT True-Score and Observed-Score

In general, both true and observed scores equatings have advantages and disadvantages. IRT true-score equating is easier to compute and does not depend on the ability distribution of examinees. However, this method equates true-scores, which are not available in practice. In addition, with the 3PL IRT model, equivalents cannot be calculated at raw scores at the lower end of the score scale. Conversely, observed-score equating has the advantage that it depends only on the availability of the examinees’ observed scores. Also, its increased computational complexities are feasible if the posterior theta distribution from the IRT calibration program is used. Generally, the largest differences between true score and observed score equating methods are at the low and high end of the scale because these are the points where the true-score equating method does not produce equivalents (Kolen & Brennan, 2004).

2.7.5 Equating Accuracy: Standard Error of Equating

2.7.5.1 Random Errors

Whenever equating is performed to estimate equating relationships, two major sources of error typically occur: random and systematic errors. A major goal in designing and conducting equating is to minimize such equating error. Random equating errors occur when samples of examinees are used to estimate population parameters (e.g., means, standard deviations, percentile ranks, or item difficulties) rather than using the whole population. Random errors are especially a major concern when the sample size is small. Therefore, random errors can be reduced by increasing the sample size and by choosing an appropriate data collection design (Kolen & Brennan, 2004).

2.7.5.2 Systematic Errors

On the other hand, systematic errors in estimating equating relationships can occur in four ways. First, systematic errors occur when assumptions, or conditions, of a particular data collection design or an equating technique are violated. For instance, when the equating relationship and relationship between scores on different forms is not linear, then using linear equating may lead to systematic errors. Second, systematic errors can occur when the equating method introduces bias in estimating the equating relationship. For example, in the equipercentile equating method, even though smoothing techniques are used to produce smoother functions that contain less random errors than unsmoothed functions, these smoothing techniques can introduce systematic errors. Third, systematic errors can occur with an improper implementation of the data collection design. For example, in the random groups design, suppose that the test center assigned Form Y to examinees near the front of the room and Form X to those near the back of the room. This way of distribution of the test forms will defeat the spiraling process and may lead to systematic errors. Furthermore, in the anchor-item design, placing the common-items on the beginning of the test in Form X and near the end of the test in Form Y may lead these items to behave very differently on the two test forms. Lastly, systematic errors can occur when the group(s) of examinees used in the equating study is not representative of the population of examinees who took both forms. Although the use of large sample sizes reduces the magnitude of the random error, it does not reduce the magnitude of the systematic error (Kolen & Brennan, 2004).

Therefore, to control for systematic error, test forms must be built on the same test content specifications and on the same statistical specification, data collection designs must be appropriately chosen, an equating design must be properly implemented, and statistical techniques must be appropriately selected. Hence, random and systematic errors should be considered when designing and conducting equating (Kolen & Brennan, 2004).

2.7.5.3 Standard Error of Equating

Random equating errors are quantified through the standard error of equating. The standard error of equating, which indicates the amount of random error in equating, is defined as the standard deviation of equated scores when the equating procedure is replicated a large number of times using samples drawn from a population or a population of examinees for each replication (Kolen & Brennan, 1995).

For a given sample estimated, equating errors at a particular score level on form X, x_i, are defined as equal to the difference between the sample Form Y equivalent and expected equivalent in the population. Random equating error is defined as (Kolen & Brennan, 2004):

$$e_y^{\wedge}(x_i) – E\,[e_y^{\wedge}(x_i)], \qquad (2.58)$$

where $e^{\wedge}_{Y(xi)}$ is an estimate of the Form Y equivalent of a Form X score in the sample. $E\,[e^{\wedge}_{Y(xi)}]$ is the expected equivalent over random samples from the population(s). When the equating is done repeatedly, the equating error variance at score point x_i can be calculated as (Kolen & Brennan, 2004):

$$var\,\left[e_y^{\wedge}(x_i)\right] = E\left\{e_y^{\wedge}(x_i) - E\left[e_y^{\wedge}(x_i)\right]\right\}^2, \qquad (2.59)$$

Standard error of equating defines as (Kolen & Brennan, 2004):

$$se\,[\hat{e}_Y(x_i)] = \sqrt{\left[var\,[\hat{e}_Y\,(x)]\right]} = \sqrt{E\{\hat{e}_Y(x_i) - E[\hat{e}_Y(x_i)]\}^2.} \qquad (2.60)$$

Equating bias defines as the difference between the true and sample equating and is calculated using the following formula (Kolen & Brennan, 2004):

$$Bias\ \left(\hat{e}_y(x_i)\right) = E\ [\hat{e}_Y\ (x_i)] - e_Y(x_i), \qquad (2.61)$$

Then, the mean square errors (MSE), which are the sum of the squared variance of equating of $e\hat{}_{Y(xi)}$ and bias, are calculated as (Kolen & Brennan, 2004):

$$MSE\ [\hat{e}_Y\ (x_i)] = var[\hat{e}_Y\ (x_i)] + \ \{bias\ [\hat{e}_Y\ (x_i)]\ \}^2. \qquad (2.62)$$

2.8 SUMMARY

The CEPA-English exam, the first national English language test in the UAE, is used for three purposes. First, it accounts for 25% of the students' overall GSC English grade. Second, it is also used to determine admission in that students must achieve a minimum score of 150 on the test to be accepted into the undergraduate programs of the three institutions. Third, it is used to place students into the appropriate levels of English proficiency in the remedial program prior to the start of their academic courses. The CEPA-English test score also decides students' eligibility to take more challenging tests (the TOEFL or the IELTS) that the three institutions require to exempt students from the remedial program and to directly enroll them in the undergraduate courses. It is crucial therefore to ensure the technical quality of the CEPA-English test for these purposes.

IRT was used to provide evidence for the CEPA-English test's technical quality in relation to Forms A and B, including the psychometric properties of Forms A and B of the CEPA-English test, and the amount of information the test provided at the cutoff score of 150, which is used as a basis for admission to higher education. Before using IRT, it was necessary to verify the validity of the model. First, four important assumptions of IRT needed to be satisfied: 1) one or more trait dimensions underlie each examinee's performance, 2) all examinees' responses to the items on a test are independent, 3) the form of the IRT model is appropriate, and 4) the test is non-speededness. Second, the fit of the model at the test and item levels must be evaluated using goodness-of-fit statistical tests, as well as graphical representations, in order to determine whether the advantages of IRT can be attained. Third, the degree to which invariance of item parameters are obtained must be checked by using plots of item and ability parameter estimates, as well as the correlation of the estimated parameters. Fourth, to ensure the validity of all test applications using IRT, it was essential that both ability and item parameters be accurately estimated using the methods that fit the data. Finally, test items that provide the most information across and around the cutoff score should be selected using the item and test information functions method.

The successful application of IRT models requires determining the dimensionality of test data. This is typically achieved by the EFA method, a statistical technique that is commonly used to determine the overall dimensionality of the test and to evaluate a test's internal structure. However, performing linear EFA to assess dimensionality for dichotomous data such as the CEPA-English test may cause problems, such as the presence of the spurious factors, the underestimation of the factor loadings and overestimation of the number of dimensions, and the chance of success through guessing (Ackerman, et al., 2003; Carroll, 1945; Green, 1983). To overcome these limitations, several researchers have suggested using others factor analysis procedures for dichotomous data that are implemented in software programs, namely Mplus (Muthen &Muthen, 2001), NOHARM (Fraser &McDonald, 1988), and TESTFACT (Wilson et al., 2003).

After choosing the software programs to detect dimensionality for dichotomous data, the researcher must then decide how many factors to retain for interpretation. The number of factors determines the overall dimensionality of a test. Different methods have been proposed to identify the number of factors to retain using EFA. These methods include Kaiser's criterion, Cattell's scree plot, parallel analysis, the amount of variance explained by a factor, the number of substantial factor loadings, and the summary statistics for residuals such as the RMSR and RMSEA. These indices provide subjective decisions on the number of factors to retain.

Therefore, it is generally recommended to use multiple method*s* for determining the appropriate number of factors to retain.

Another common technique for assessing dimensionality of dichotomous test data is DIMTEST. DIMTEST assesses whether the test is *essentially* unidimensional through its *T* statistic test. The DIMTEST procedure consists of two stages. First, the N test items must be split into two subtests: Assessment Subtest (AT) and Partitioning Subtest (PT). Based on cluster analysis, AT items are selected to reflect one trait. The PT part, comprised of the remaining items of the test, is used to form K subgroups based on the scores of the PT items. Then, the DIMTEST *T* statistic is performed using the DIMTEST computer program. If the *p*-value of the *T* test is significant, then the unidimensional assumption of the test will not be rejected.

It is crucial that items in a test are fair to all examinees and are not biased against any particular group. Because the CEPA-English test imposes serious consequences on 12th grade students, it is important to examine whether the items on this test exhibit DIF. DIF analysis is an initial step in determining whether a possible bias at the item level is present. DIF can be an important indicator of irrelevant variance that can influence test scores. There are two types of DIF: uniform and nonuniform.

The MH method is the most commonly used non IRT procedures for detecting DIF in dichotomous variables. This method does not require a large sample size (e.g., 200 for each group with MH), and it is relatively easy to perform with computer software. In addition, MH procedure provides effect size measures to interpret the magnitude of DIF and to determine whether DIF items are negligible (Rogers & Swaminathan, 1993). The MH procedure is a more powerful test for detecting uniform DIF items (Rogers & Swaminathan, 1993; Lopez-Pina, 2001; Swaminathan & Rogers, 1990). Previous research, however, has determined that some factors might influence the power of detecting DIF: sample size, ability differences between groups, percentage of DIF items, DIF magnitude, test length, item characteristics, and the type of DIF (uniform and nonuniform). Hence, the researcher needs to take these factors into consideration when running the MH procedure.

The CEPA-English test is administered repeatedly each year; this increases threats to the test's security. To ensure test security, NAPO uses multiple forms for the CEPA-English test. These forms are constructed to be similar to each other in content and statistical characteristics. Although multiple test forms are carefully constructed, the forms differ somewhat in difficulty; therefore, scores from forms are not interchangeable without some type of equating.

The first step in the equating process involves selecting an appropriate equating design. Four data collection designs are commonly used in equating: (a) single-group design; (b) single-group design with counterbalancing; (c) random-group design; and (d) anchor-item design. In the current study, both random-group and anchor-item designs were used to equate the CEPA-English test forms. After choosing the appropriate design, the second step is to select the statistical equating methods. Various equating procedures, including procedures based on CTT and IRT, have been utilized to maintain comparable test scores. CTT has three equating methods: mean, linear, and equipercentile. IRT has two equating methods: IRT true-score equating and IRT observed-score equating. The final step in the equating process involves evaluating the results of equating using the standard error of equating (*SEE*) (Kolen & Brennan, 2004). *SEE* is used to evaluate the accuracy of equating. *SEE* is the most important evaluation criteria for controlling equating errors, which often occur in estimating relationship when equating is conducted (Kolen and Brennan, 2004).

NAPO used the IRT equating method with the anchor-items design to adjust differences in difficulty among the CEPA-English test forms so that forms can be used interchangeably. NAPO only used five anchor-items in equating. However, equating with anchor-items design is successful *only* if the anchor-items are properly chosen. This means that anchor-items are representative of the overall test, maintain the same location across test forms, are at least as long as 20%-25% of the total test length, and contain enough easy and difficult items (Kolen and Brennan, 1995, 2004). However, the present study will not replicate the equating method that was used by NAPO because it is only used five common items in Forms A and B and a different set of five common items in Forms C and D. In this study, equipercentile equating method under the random-groups design was used to equate Forms A and B of the CEPA-English test.

It is essential to examine whether the CEPA-English test provides the most of information at cutoff score of 150, which is the mean of the NAPO test distribution. The test information function (TIF) is a measure of the precision of the test scores. The item information function shows the contribution of each item to the total score. Item information functions can be summed to provide the test information function. The more information each item contributes, the higher the test information function. The more information a test provides, the more precise the test and the lower the *SE* will be, as the amount of information a test provides at a θ level is inversely related to the *SE* (θ) (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985; Hambleton, et al., 1991). Generally, higher TIF's and, smaller *SE*'s are associated with longer tests with highly discriminating items and with tests composed of items with *b* values close to the examinee's true ability (Hambleton, et al., 1991; Hambleton, 1993).

3.0 METHODOLOGY

Chapter 3 provides a description of the methodology used in the study. The overall design of the study is first described, followed by the data source, the instrument, the sample, the data collection procedure, and the statistical analyses conducted to answer the research questions.

3.1 OVERALL RESEARCH DESIGN

The CEPA-English test score is used as an achievement measure that accounts for 25% of the students' overall GSC English grade. The test score is also used for admission purposes, in that each student must achieve a minimum score of 150 on the test to be admitted into the UAE's public higher education institutions. This study is intended to ensure the technical quality of the test for the aforementioned purposes. To accomplish these goals, the study first examined the psychometric properties of Forms A and B of the CEPA-English test using IRT. Second, using the MH DIF detection method, the study investigated whether any items on Forms A and B exhibit DIF. Third, the study examined the extent to which the CEPA-English test scores are equivalent across Forms A and B by using the equipercentile equating method under the random-groups design. This also involved evaluating the quality of equating Forms A and B through examining the error associated with this design. Finally, using IRT, the study examined how much information is provided at the cutoff score of 150 for Forms A and B.

3.2 DATA SOURCE

The data for the current study were obtained from NAPO, taken from the 2007 administration of the CEPA-English test, which had four forms (A, B, C, and D). Forms A and B were the focus of this study. Data consisted of the 12th grade students' responses for 120 multiple-choice items; 90 items from the Grammar and Vocabulary Section and 30 items from the Reading Section. The data also included the students' demographic information (i.e., gender, study type, and school type), which is presented in Tables 4 and 5 (see pages 147-148).

3.3 INSTRUMENT

The CEPA-English test, a standardized paper-and-pencil exam, is designed to measure the English proficiency of 12th grade students. For the 2007 administration, the test has four forms (A, B, C, and D), each consisting of three sections: Grammar and Vocabulary, Reading, and Writing (see Appendix A). The first two sections together have a total of 120 items, all of which were multiple-choice questions with four options. Of these 120 items, 115 items are unique to each section and form, while one set of five common items are in Forms A and B and another set of five common items are in Forms C and D. These two sets of five items, which only represent grammar and vocabulary domains, were used for the purpose of equating the four forms.

The Grammar and Vocabulary Section, which lasts 45 minutes, consists of 90 items: 10 are parts of speech items, 40 are grammar items, and 40 are vocabulary items. The parts of speech items measure the students' knowledge of English word forms. The grammar items measure the students' ability to recognize common English grammatical patterns, such as agreement, verb forms, and word order. The vocabulary items measure knowledge of common English vocabulary words. The vocabulary words consist of the 400-499 most frequently used words, or the first 1000 or second 1000 most frequently used words from the General Service List (GSL), or from one of the first five Academic Word Lists (AWL).[10]

The Reading Section, which lasts 45 minutes, consists of three descriptive or narrative prose texts and one non-prose text, with a total of 30 multiple-choice questions. This section measures the students' ability to understand academic reading material. Each of the three prose texts focuses on a different subject domain: a simple descriptive passage on an everyday topic, a passage on social science or humanities, and one on science or technology.

The Writing Section consists of two tasks: Task 1 is writing an essay, while Task 2 is writing a formal letter. Each task lasts 30 minutes. This section requires students to provide their point of view related to an assigned issue, using varied and accurate grammatical and lexical resources, to employ reasoning and evidence to support their ideas. The students' writing is

[10] The GSL is a list of approximately 2,000 words created by West in 1953. The words were selected to represent the most frequent words of English and were taken from a corpus (a large and structured set of texts) of written English. The AWL is a list of words that are commonly used in English-language academic texts, not including words that are among the most frequent 2000 words of English. It contains 570 word families divided into 10 sublists.

evaluated based on content, fluency and coherence, grammar, vocabulary, spelling, and punctuation.

3.4 ADMINISTRATION OF THE CEPA-ENGLISH TEST

The 2007 CEPA-English exam was administered to all 12th grade students on May 19, 2007. Examinees either took the test in the morning or afternoon; all examinees, however, had only *a single* opportunity to take the test. NAPO and the higher education institutions collaborated with the Ministry of Education to administer the test. All examinees received a letter from NAPO through the school informing them of the date, time, and location of the test for which they were scheduled.

On the test day, examinees brought the letter of invitation from NAPO and a valid photo. Those who did not have these credentials were not allowed to complete the exam. The officials at the test center distributed all test materials, including pencils, erasers, and an answer sheet, needed for test completion. They then randomly distributed the four forms (A, B, C, and D) of the CEPA-English test booklets to the examinees, with Forms A, B, and C administered in the morning and Forms C and D administered in the afternoon. NAPO instructed all supervisors at each test center to distribute the test forms in such a way as to avoid collusion /cheating. This typically meant that test forms were distributed in alternate rows (front to back of room), so that students sitting beside each other would have different forms.

In addition, the site supervisors instructed the examinees to indicate their answers to the questions by marking the letter A, B, C, or D on a scannable answer sheet. Although the supervisors informed the examinees that they only had two-and-a-half hours to finish the test, they did not give any instructions regarding guessing.

Not only did the morning and afternoon sessions differ in terms of the CEPA-English form administrated, but they also had differed by the types of student in each session. Private and public school students were in both sessions, home-schooled students were mostly but not entirely in the afternoon session, and the majority of the part-time students were primarily in the afternoon session. Part-time students only received Form C or D, yet those who had received Form C may or may not have been assigned to the afternoon session.

3.5 SAMPLE

The subjects in this study were all 12th grade students who followed the Ministry of Education English curriculum and who took only *one* of the four forms of the CEPA-English test at any of the 21 test centers in the UAE. The entire sample in this study consisted of 18, 765 students for Forma A and B (*N* = 9**,** 496 for Form A; *N* = 9, 269 for Form B) of the 2007 CEPA-English test. The students' demographic information, namely students' gender, study type, and school type for Forms A and B, is illustrated in Tables 4 and 5, respectively.

Table 4. Demographic Information of the Sample of Form A

Gender	Study Type	School Type			Total
		Public	Private	home-schooled	
	Arts	2435	134	329	2898
Male	Science	1740	340	36	2116
	Total	4175	474	365	5014
	Arts	2226	103	411	2740
Female	Science	1397	287	58	1742
	Total	3623	390	469	4482

As Table 4 shows, 9, 496 students were included in Form A, with 5, 014 (52.8%) of whom were females and 4, 482 (47.2%) of whom were males. In terms of study type, 5, 638 (59.4%) students were in the Arts stream and 3, 858 (40.6%) students were in the Science stream. The students represented the following school types: 7, 798 (82.1%) in public schools; 864 (9.1%) in private schools; 834 (8.8%) in home-schooling.

Table 5. Demographic Information of the Sample of Form B

Gender	Study Type	School Type			Total
		Public	Private	home-schooled	
	Arts	2455	163	320	2938
Male	Science	1573	371	31	1975
	Total	4028	534	351	4913
	Arts	2053	75	578	2706
Female	Science	1517	126	7	1650
	Total	3570	201	585	4356

As Table 5 shows, the study included 9, 269 students for Form B, with 4, 913 (53.0%) of whom were females and 4, 356 (47.0%) of whom were males. In terms of study type, 5, 644 (60.9%) students were in the Arts stream and 3, 625 (39.1%) students were in the Science stream. The students represented the following school types: 7, 598 (82.0%) in public schools; 936 (10.1%) in home-schooled; and 735 (7.9%) in private schools.

According to Tables 4 and 5, the distributions of students for the different demographic characteristics were similar for Forms A and B. It was important to note that although home-schooled students technically belong to public schools, they were a different population from both private school and public students who were regular school attendees. They differed from regular students in their ages, school attendance, and personal responsibilities/ commitments (e.g., work, family care, household, etc). Thus, even though all public school students studied the same English curriculum, home-schooled students have shorter hourly class periods than others.

3.6 STATISTICAL ANALYSES

Data from Forms A and B of the CEPA-English test were analyzed using several statistical software programs (SAS 9.1, SPSS 17.0, MULTILOG 7.0, NOHARM, DIMTEST 2.0, EZLID SAS macro, IRTFIT SAS macro, ResidPlots-2, RAGE-RGEQUATE, and EZDIF).

3.6.1 Omitted and Not-Reached Responses

Using the SAS program, examinee responses to each item in Forms A and B of the CEPA-English test were scored as either correct or incorrect. Items can be omitted or not-reached. Omitted responses are assumed to be ones the examinees had intentionally skipped because they did not know the correct answers for them; not-reached responses are assumed to be ones that examinees did not answer due to lack of time. In general, omitted and not-reached items may operationally be distinguished by the fact that omitted items occur earlier in the test, whereas not-reached items fall in a sequence at the end of the test. Thus, this study treated those items at the end of the test that were not responded to by examinees as not-reached items.

In this study, correct responses were scored as “1” and incorrect responses as “0.” For all data analyses, except for analyses of examining speededness and the presence of guessing behavior, omitted and not-reached responses (missing data) in Forms A and B of the CEPA-English test were treated as incorrect responses, which was consistent with the operational scoring of the test by NAPO. Prior to data analyses, missing data were examined on Forms A and B. The frequency tables for not-reached and missing responses for Forms A and B are provided in Tables B1-B4 in Appendix B. As can be seen from Tables B1-B4, 99% of the examinees reached the end of the test in each form. The percentage of the examinees who completed all items was approximately 83% (7905) in Form A and about 82% (7595) in Form B.

3.6.2 Examining the Psychometric Properties of the CEPA-English Test using Classical and IRT Analyses

3.6.2.1 Classical Analyses

The SPSS program was used to obtain a summary of the classical analyses at item and test levels for Forms A and B. This included the number of examinees, number of items, item difficulty (*p*-values), point-biserials (r_{pbis}) and KR-20 correlation coefficients, mean, standard deviation, skewness, and kurtosis for the total scores were computed and evaluated. Additionally, frequency distributions of the total score curves for Forms A and B were computed and evaluated.

3.6.2.2 IRT Analyses

3.6.2.2.1 Fixing the *c* Parameter Values under the Unidimensional 3PL IRT Model

It was expected that the unidimensional 3PL IRT model would provide a better fit for each form of the CEPA-English test sine each is multiple choice, in which examinees can obtain correct answers simply by guessing. For a multiple-choice test with four alternatives, the probability of getting the item correct from random guessing is .25.

To ensure stable estimates of the *b* and *a* parameters, the *c* parameter values for some items were fixed at either 0 or .25 for Forms A and B of the CEPA-English items. There were two conditions under which the *c* parameter values were fixed at 0. The first was when the value of the *c* parameter did not meet Lord’s (1980) criterion (estimate *c* when $b – 2/a > –3.5$). The second was when a standard error (*SE*) for *c* could not be estimated. There were two conditions under which the *c* parameter values were fixed at .25. The first condition was when the value of the *c* parameter value was .4 or above. This is because, in standard practice, *c* values greater than .4 may be both unrealistic for multiple-choice tests with four-options and may lead to higher measured error. The second condition was when items did not fit the 3Pl model.

3.6.2.2.2 Examining the Assumptions Underlying Dichotomous IRT Models

The benefits of using IRT model relies on the following three factors: 1) the extent to which the selected IRT model is appropriate for the test data; 2) the extent to which the underlying assumptions and the item and ability propitiates of the selected IRT model are met in the test data; 3) and how well the selected model fits the data**.** The degree to which the assumption of the unidimensional dichotomous IRT model holds true for Forms A and B of the CEPA-English test data was investigated through analysis of the appropriateness of the IRT model, the unidimensionality, local item independence, the speededness of the test, and the presence of guessing behavior. In addition, assuming that the selected IRT model is true, randomly generated data were used to examine the unidimensionality and local independence assumption for each test form. The results obtained from the CEPA-English data were compared against the results obtained from the simulated data. It was expected that Forms A and B test data would meet the assumptions of IRT and that the results using real data would be very similar to those using simulated data.

3.6.2.2.2.1 Model Testing: Choosing the Preferred IRT Model

As discussed earlier it was expected that the unidimensional 3PL IRT model would provide a better fit for Forms A and B of the CEPA-English test data than the unidimensional 1PL or 2PL IRT models because each form is multiple choice test in which examinees can obtain correct answers simply by guessing. Therefore, in this case, it was more accurate to use the 3PL model, which takes guessing into account. Additionally, in the study, the 2007 sample sizes for each form (*N* = 9, 496 for Form A; *N* = 9, 269 for Form B) was more than adequate for the 3PL MML estimation, which usually requires a minimum of 1000 subjects (Kingston and Dorans, 1985).

In deciding which model to use, the three hierarchical or nested IRT models (1PL, 2PL, and 3PL) were compared to evaluate the extent to which additional estimated parameters in one model significantly increase the model-data-fit. Item parameters for the 1PL, 2PL, and 3PL models were estimated separately for Forms A and B of the CEPA-English test using MULTILOG (Thissen, 1988), which employs the MML method (Bock & Aitkin, 1981). The statistical comparison of competing nested models (e.g., the 2PL vs. the 3PL) estimated by MULTILOG was obtained through the statistic (-2log likelihood statistic) reported for each IRT model of each test form. The difference between the statistics for nested models, distributed as chi-square, was used to evaluate the significance of specific additional parameters in improving model-data-fit. Statistically comparing two nested models yields a difference chi-square (G^2) with degrees of freedom equal to the number of additional parameters that are estimated. If the additional parameters included in the less parsimonious model afford a significant improvement in the fit of the model to Forms A and B test data against the alternative nested model, the observed difference will be significant.

3.6.2.2.2.2.1 Evaluation of Unidimensionality and the Internal Structure

To evaluate unidimensionality and the internal structure of Forms A and B of the CEPA-English test, a nonlinear exploratory factor analysis (NLEFA) model was performed using the NOHARM software program, which allows for the modeling of guessing with a MIRT model. As discussed earlier in section 2.5.1.5, both NOHARM and TESTFACT produce similar results in assessing dimensionality; therefore, TESTFACT was not included in this study. Essential unidimensionality was also assessed using DIMTEST as an alternative nonparametric statistical procedure.

To assess the unidimensionality (i.e., the presence of a general English ability) of Forms A and B of the CEPA-English test, four criteria were specified with respect to the NOHARM analyses and the eigenvalue analyses: 1) proportion of variance accounted for by the largest eigenvalues, 2) scree plots, 3) number of substantial loadings for factors, and 4) RMSR statistics. Due to the large sample size of each form, a chi-square fit-statistic was not used because a large sample size would increases the power of the chi-square fit-statistic.

To evaluate the dimensionality of Forms A and B, the proportion of variance explained by the largest eigenvalues along with the point where a break occurred in the plots of the eigenvalues were examined. In addition, looking for high factor loadings (i.e., loading greater than .3) revealed which items were loaded on which factors. Items with substantial loading were used to interpret a factor. Because the factors were expected to be correlated on each form of the CEPA-English test, a Promax rotation was used to extract factors. In this study, trivial factors—those with 5 or fewer items with substantial factor loading—were eliminated. The value of RMSR statistics was also examined. Since the RMSR statistic in NOHARM is based on the differences between observed and expected proportions, the values of RMSR 4 times the reciprocal of the square root of the sample size indicated an acceptable factor solution (McDonald, 1991). Therefore, in this study, for NOHARM, factors were added to the model until the percentage of RMSR reduction was less than 0.041 for Forms A and B (i.e., 4*(1/√9, 496 for Form A and 4*(1/√9, 269 for Form B).

Finally, to examine whether Forms A and B of the CEPA-English test were essentially unidimensional, the DIMTEST analyses were performed. First, the total number of items on Forms A and B was spilt into 2 parts: Assessment Subtest (AT) and Partitioning Subtest (PT). The AT part was of length M (4 ≤ M < 60 half the test length), and the PT part was of length N – M items. Based on cluster analysis, AT items were selected to reflect one trait. The PT part, comprised of the remaining items of the test, was used to form K subgroups based on the scores of the PT items (groups that were used to condition on ability). Then, the DIMTEST *T* statistic was formed by examining differences between the observed variance of proportion correct scores and the theoretical expectations within subgroups. If unidimensionality holds, the differences between these variances should be small, suggesting that they are estimation of the same variance.

The *T* test statistic was computed using the DIMTEST, version 2.0, computer program, which consisted of two programs: ATFIND and DIMTEST. The ATFIND was used to identify a set of items for the AT set based on cluster analysis. The items for the AT set were listed on the ATLIST.IN file, which is used to tell the DIMTEST which items are in AT. The DIMTEST was used to compute the *T* test statistic and *p*-value. If the *p*-value is significant, the unidimensional assumption of each form of the CEPA-English test would not be rejected.

The ATFIND manual recommends that the sample of examinees used to identify the AT set should differ from the dataset used with the DIMTEST. Therefore, each dataset of each form was randomly split into two data files; odd records, used for running the ATFIND, were put in one file, and even records, were placed in the DIMTEST.

3.6.2.2.2.2.2 Evaluation of Dimensionality Using Simulated Data

To further compare the consistency between the SPSS, NOHARM, and DIMTEST methods in evaluating the underlying factor structure of Forms A and B of the CEPA-English test, a simulation data were used to provide comparisons based on observed data and item responses simulated using the IRT.

There were three estimation methods (SPSS, NOHARM, and DIMTEST) used to evaluate unidimensionality of Forms A and B. To determine the number of factors, the following were examined: 1) proportion of variance accounted for by the largest eigenvalues, 2) scree plots, 3) number of substantial loadings for factors, 4) RMSR statistics, and 5) the *p*-value of DIMTEST.

3.6.2.2.2.2.2.1 Outline of the Simulation Data

The overall steps for the simulation data are shown in Figure 8 and are described below:

- Step 1: For each test form of the simulated data, test length, sample size, and item parameters were fixed. Specifically, the test length was fixed at 116 items for Form A and at 119 items for Form B. The number of responses was fixed at 2 (1 as correct and 0 as incorrect). The sample size was fixed at 9, 496 for Form A and at 9, 269 for Form B. Finally, the item parameters were fixed at the values that were obtained from each form of the CEPA-English test.
- Step 2: Using MULTILOG, item parameters for Forms A and B of the CEPA-English test were estimated under the unidimensional 3PL model using the MML

estimation method. Then, these estimated parameters were saved in file ".PAR".

- Step 3: Using SAS, item responses were generated for a simulated individual by computing the probability of a correct response for each item on each form using the item parameter estimates from Step 1 and randomly sampling ability parameter (*θ*) from a normal (0,1) distribution. Finally, these generated item responses were saved in the file "random.dat."
- Step 4: Using MULTILOG, the simulated item responses were calibrated in order to estimate the item parameters that were based on simulated responses.
- Step 5: Using the random data from Step 2, the dimensionality of Forms A and B of the CEPA-English test data were examined through SPSS, NOHARM, and DIMTEST methods. In SPSS, the proportion of variance explained by the largest eigenvalues and scree plots were examined. In NOHARM, the RMSR statistics and the number of substantial loadings for factors generated with Promax rotated factor were examined. In DIMTEST, the *p*-value of the *T* test statistic was compared with the .05 α level.

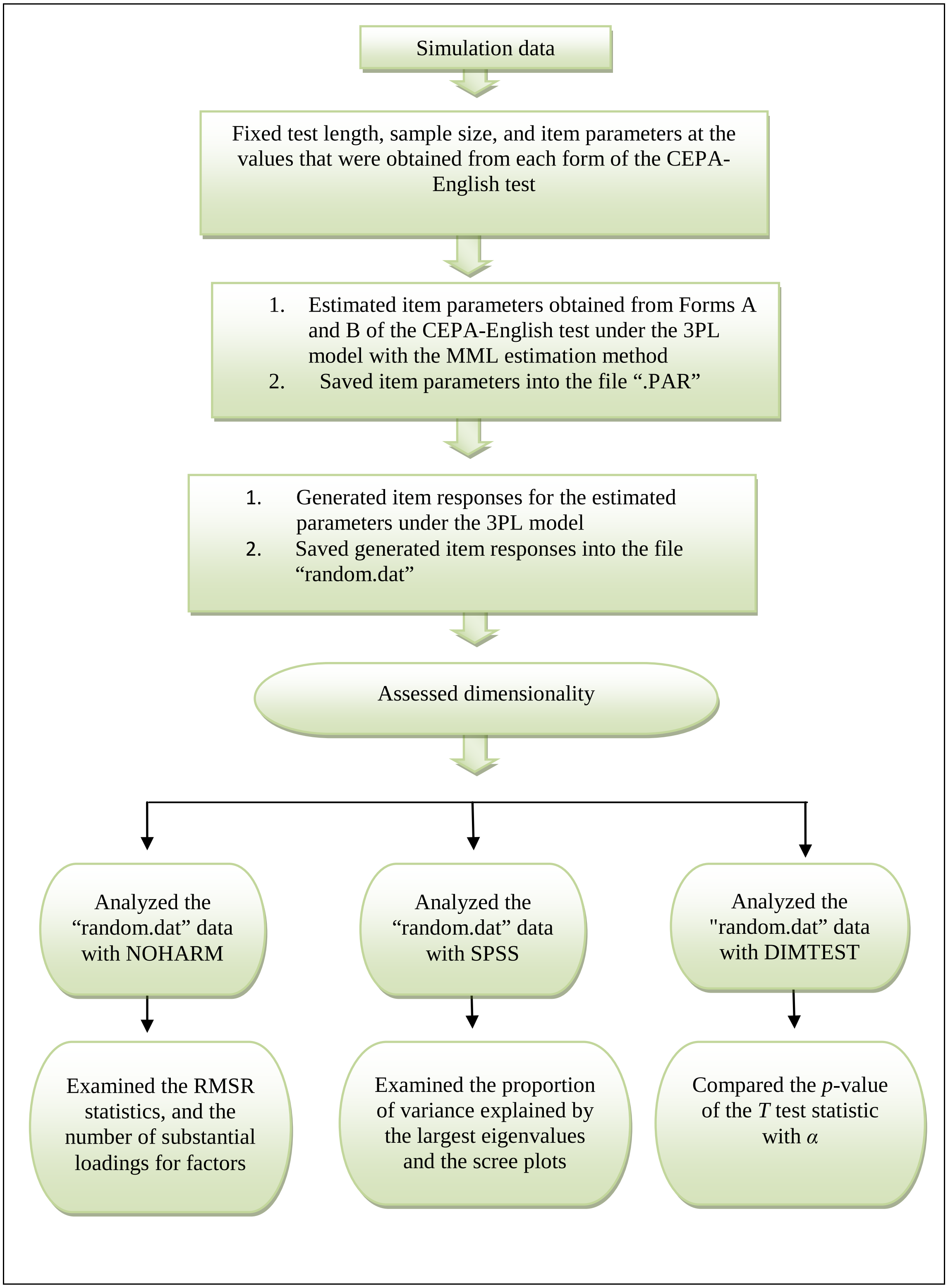

Simulation data
Fixed test length, sample size, and item parameters at the values that were obtained from each form of the CEPA-English test
1. Estimated item parameters obtained from Forms A and B of the CEPA-English test under the 3PL model with the MML estimation method
2. Saved item parameters into the file ".PAR"
1. Generated item responses for the estimated parameters under the 3PL model
2. Saved generated item responses into the file "random.dat"
Assessed dimensionality
Analyzed the "random.dat" data with NOHARM
Analyzed the "random.dat" data with SPSS
Analyzed the "random.dat" data with DIMTEST
Examined the RMSR statistics, and the number of substantial loadings for factors
Examined the proportion of variance explained by the largest eigenvalues and the scree plots
Compared the p-value of the T test statistic with α

Figure 8. Overall Steps in Conducting the Simulation data

3.6.2.2.2.2.2.3 Validation of the Data Generation

It is important to ensure that the SAS program used in this study properly generated ability parameters and item response data. To validate the generation of ability parameters, both the histogram with the normal curve of the ability parameters and the mean and standard deviations were examined. It was expected that the generated ability parameters would be normally distributed with a mean of 0 and standard deviation of 1. To validate the generation of item responses from the unidimensional 3PL IRT data, the proportions of examinees correctly responding for each item was computed. Then, these observed proportions were compared to the model-based expected proportions. The difference between the two proportions was expected to be small if the data generated were valid.

3.6.2.2.2.3.1 Assessing Local Item Independence

The assumption of local item independence holds true on each form of the CEPA-English test when the probability of correctly answering an item is not affected by the probability of correctly answering another item. Yen's chi-square test (Q_3) for dichotomous items was used to detect local item dependence (LD) for Forms A and B. Q_3 is the most commonly used index for detecting LD. Chen and Thissen (1997), in their simulated data, have evaluated χ^2, G^2, and Q_3 indices for detecting LD. They found that the Q_3 index tends to outperform both the χ^2 and G^2 LD. The Q_3 statistic is obtained by correlating the differences between students' observed and expected responses for pairs of items after taking into account overall test performance.

Local independence will be enhanced between any pair of items on Forms A and B if the expected value of Q_3 is equal to $-1 / (n -1)$, where n equals the number of test items. The local independence will also hold between any pair of items if the mean of Q_3 test statistics is close to zero, indicating that there is no correlation between item pairs after accounting for the examinee's ability (Embretson & Reise, 2000). As a rule of thumb, all pairs of items with Q_3 values equal to or greater than .20 were flagged as locally dependent. The sign of the Q_3 was used to infer whether the dependency between two items was negative or positive—a positive value indicating item pairs that share greater dependence, and a negative value indicating items that share less dependency.

The EZLID SAS macro was used to compute the Q_3 index of local dependence for all pairs of items for Forms A and B of the CEPA-English test. Because this macro requires a file of scored responses with ability estimates and item parameter estimates, it was used in conjunction with the MULTILOG computer program (Thissen, 1991). MULTILOG was used to compute MML item parameter estimates for Forms A and B under the unidimensional 3PL IRT model. The MULTILOG program was also used to compute examinee score estimates using MLE based on these item parameter estimates.

As discussed in Chapter 2, there are a variety of possible causes of LD. A common one is having multiple items relate to a shared stimulus, such as a reading passage (i.e., testlet) (Yen, 1993). In this study, the local item independence for items that have a common reading passage on each from of the CEPA-English test was examined. If the items were not locally independent, then they were treated as testlets. The graded response model (GRM; Samejima, 1969), a polytomous IRT model, was then used to score the four testlets on each form using MULTILOG.[11]

3.6.2.2.2.3.2 Assessing the Local Item Independence Using Simulated Data

In order to confirm whether all pairs of items of Forms A and B of the CEPA-English test are locally independent, data were simulated under the assumption that the IRT model was true.

3.6.2.2.2.3.2.2 Outline of the Simulation Data

The overall steps of the simulation data are shown in Figure 9 and are described below:

- Step 1: For each test form of the simulated data, test length, sample size, and item parameters were fixed. Specifically, the test length was fixed at 116 items for Form A and at 119 items for Form B. The number of responses was fixed at 2 (1 as correct and 0 as incorrect). The sample size was fixed at 9, 496 for Form A and at 9, 269 for Form B. Finally, the item parameters were fixed at the values that were obtained from each form of the CEPA-English test.
- Step 2: The simulated real item responses “random.dat” using MULTILOG were calibrated. Next, these estimated item parameters were saved in the “random_3PL.PAR” file.
- Step 3: The data generated were scored using MLE estimates, including the id file with item responses. Then, these score were saved in the“random_3PL.SCO” file.

[11] The Reading Section of each from of the CEPA-English test consisted of four passages with a total of 30 multiple-choice questions. There were 6-7 items per passage in form A, 5-8 items in form B, 5-8 items in form C, and 6-8 items in form D.

- Step 4: The results obtained from the randomly generated data for Forms A and B were compared against the results obtained from the real data of Forms A and B.
- Step 5: The “random.PAR” and “random.SCO” files were included in the EZLID SAS macro.
- Step 6: The local item independence of Forms A and B were assessed by examining then mean of the Q3 test statistic using the EZLID SAS macro.

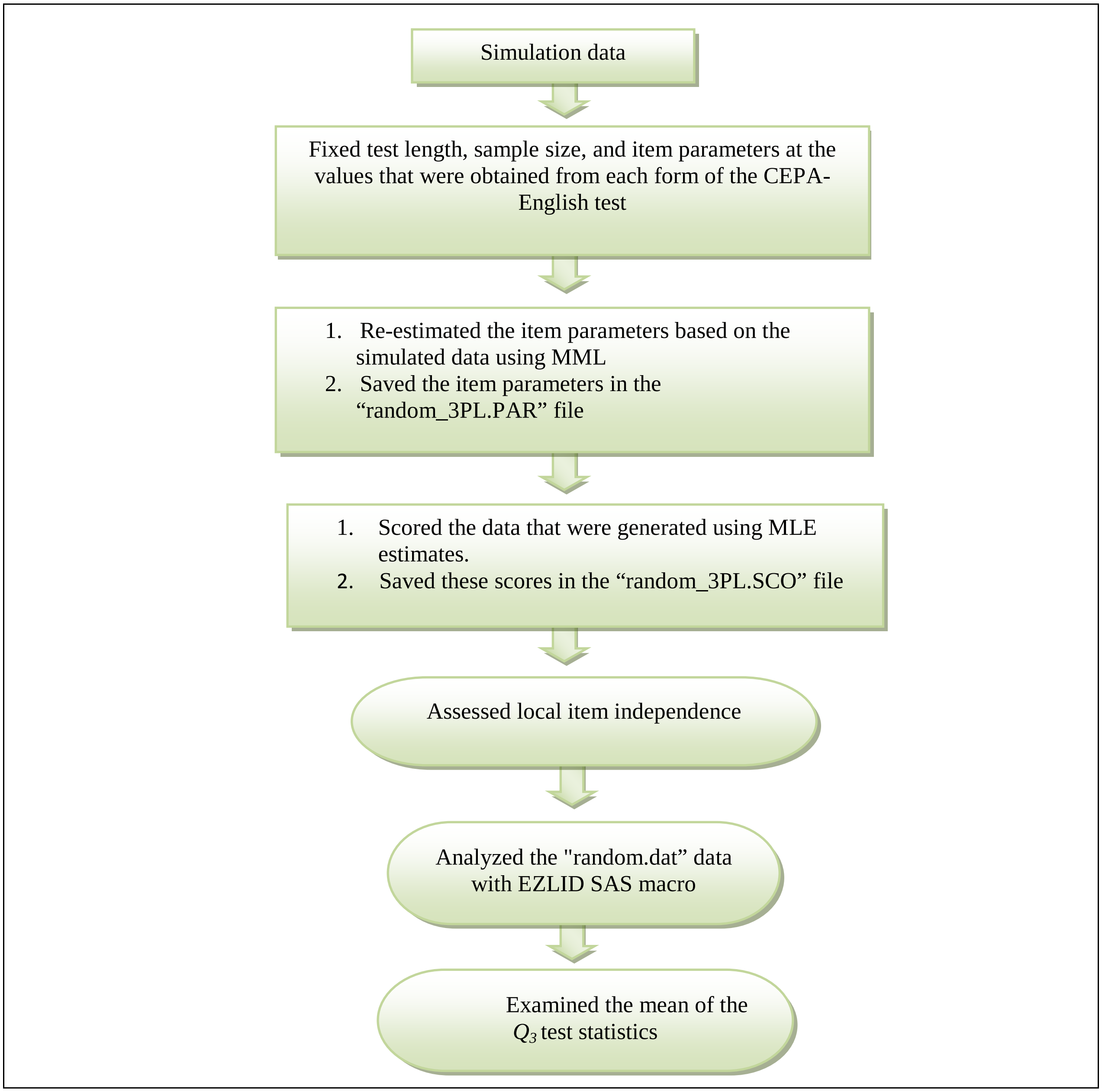

Simulation data
Fixed test length, sample size, and item parameters at the values that were obtained from each form of the CEPA-English test
1. Re-estimated the item parameters based on the simulated data using MML
2. Saved the item parameters in the "random_3PL.PAR" file
1. Scored the data that were generated using MLE estimates.
2. Saved these scores in the "random_3PL.SCO" file
Assessed local item independence
Analyzed the "random.dat" data with EZLID SAS macro
Examined the mean of the Q_3 test statistics

Figure 9. Overall Steps in Simulating Data

3.6.2.2.2.4 Examination of Speededness

Each form of the 2007 CEPA-English test was administered in two-and-a half hours. Because the test was administered in a specific time limit, it is important to examine the degree to which each form is non-speeded. The speededness of Forms A and B of the CEPA-English test was examined by comparing the percentage of examinees who completed all items, and the percentage of examinees who completed at least 75% of the items. If the majority of examinees completed nearly all of the items, then non-speedednees can be assumed on each form. In addition, the ratio of the variance of omitted items to the variance of incorrectly answered items was calculated as an index of test speededness. If the ratio of the two variances is close to zero, then the assumption of non-speededness is met on each form (Gulliksen, 1950). In order to examine speededness in this study, missing responses at the end of the test on Forms A and B were treated as not-reached items.

3.6.2.2.2.5 Assessing the Presence of Guessing Behavior

Guessing behavior is often observed on a multiple-choice test .The degree to which a multiple-choice test is speeded may motivate examinees to use guessing strategies when responding to test questions. That is, before the administration time is up, examinees may randomly guess on items at the end of the test rather than leave items unanswered. It is important to account for guessing since the probability that examinees will get the correct answer on multiple-choice items by chance through random guessing introduces measurement error and attenuates relationships between items (Carroll, 1945). Guessing behavior can also be a source of construct-irrelevant variance since it raises the possibility of correct responses based on abilities outside the area of assessment (Rogers, 1999). According to Stone and Yeh (2006), the influence of guessing on multiple-choice items is dependent on an unknown mechanism or process.

To model guessing behavior among low-ability examinees, either random guessing or partial-knowledge guessing can be assumed (Waller, 1989). A random guessing model also is assumed when examinees lack the needed knowledge to answer the item correctly. The proportion of correct responses is approximately what would be expected under the random guessing model (1/m, *m* is the number of options). For a multiple-choice test with four alternatives, the probability of success from random guessing is .25. In contrast, examinees with partial knowledge may not respond at random or may use this knowledge to eliminate some of the options and then randomly guess. In this case, the probability of a chance correct response will be greater than 1/*m*.

Random guessing behavior depends partly on the administration instructions and whether a correction for guessing is used to discourage random guessing (Linn & Gronlund, 2000). Because the 3PL model includes a parameter that can be used to model guessing behavior among low-ability students, it can remove the effect of random guessing and can make an adjustment for partial-knowledge guessing (Waller, 1989).

As earlier stated, examinees for the 2007 CEPA-English administration were not given any instructions regarding guessing. The extent to which examinees are guessing answers on Forms A and B of the CEPA-English test was evaluated by plotting the proportion of correct responses by total scores. If guessing behavior exists, the plots of proportion of correct responses should increase from 0 to 1 as the total score increases. If the plots for an item indicate a constant proportion for a correct response that is greater than 0 for low total scores, then guessing behavior can be assumed (Stone & Yeh, 2006).

In addition, the distribution of *c* parameters at the beginning (the first 10 items) of Forms

A and B of the CEPA-English test was compared to those at the end of the test (the last 10 items). If the two distributions are the same, then the guessing behavior is not present.

3.6.2.2.3 Assessing the Preferred IRT Model Fit at Item Level

Once the preferred IRT model is obtained, it is important to examine the extent to which each test item fits the preferred IRT model. As previously mentioned, item misfit exhibits if there is a difference between the observed and predicted score distributions across a range of discrete ability levels for each item. It was expected that the 3PL model would fit item on Forms A and B of the CEPA-English test data.

In the present study, the $S\text{-}X^2$ goodness-of-fit statistic was used to assess the fit of the unidimensional 3PL IRT model to each item on Forms A and B of the CEPA-English test data. The Orlando and Thissen chi-square statistics ($S\text{-}X^2$ and $S\text{-}G^2$) have several advantages over the traditional item fit-statistics (i.e., Pearson χ^2, G^2, X^2_B and Q_1 test statistics). Orlando and Thissen (2000), in their simulation study, found that $S\text{-}X^2$ is promising for detecting item misfit for dichotomous items, and $S\text{-}G^2$ is not very useful because of the inflated type I error rate.

The IRTFIT SAS macro was used to compute the $S\text{-}X^2$ fit- statistics. Before using the IRTFIT macro to examine the fit of the 3PL IRT model to each item on Forms A and B, the item responses of each form were first calibrated using the 3PL model in the MULTILOG program. Then, the item parameter estimates were used with the IRTFIT macro.

In addition to using the $S\text{-}X^2$ goodness-of-fit statistic, the fit of each item on Forms A and B to the preferred IRT model was also assessed using a graphical display of observed versus predicted score distributions. Plotting the observed versus predicted score distributions allows for a visual representation of the fit between the two distributions. The analyses of residual involved first dividing the ability scale into 15 intervals between $-3 \leq \theta \leq +3$, then, computing the proportion of examinees in each interval that provides a correct response. Next, the predicted score distributions across 15 ability subgroups for an item were obtained. After that, the residual, or the difference between the expected proportion of correct responses and the observed proportion of correct responses of 15 ability subgroups for each item were obtained. The residual was standardized by dividing the raw residual by the standard error of the observed proportion of correct responses. Finally, the residuals or standardized residuals were then plotted against the θ scale. These plots provide evidence of item fit or misfit. The residuals plots that show a random scatter about zero indicate item fit (Hambleton & Swaminathan, 1985; Hambleton, et al., 1991; Swaminathan & Rogers, 1991). The ResidPlots-2 program, developed by Liang, Han, and Hambleton (2008), was used to perform graphical residual analyses.

The ResidPlots-2 program provides residual and standardized residual plots. Four MULTILOG files are needed to run the ResidPlots-2 program: the data file, the .PAR file, the .SCO file, and the syntax file, which is used for the second run—that is, the MLG syntax used to score examinees using estimated item parameters. Users are required to put these files into one folder and to enter the name of the MLG syntax file. Finally, ResidPlots-2 requires that the name and/or path of the MLG syntax file not include any spaces.

3.6.2.2.4 Examining the Invariance of Item Parameters

It is imperative to assess the invariance of item parameters after determining the most appropriate IRT model for the test data. It was expected that the property of invariance of item parameters would hold for Forms A and B of the CEPA-English test data. The degree to which the invariance of the item parameter of the unidimensional 3PL IRT model holds true for each from was investigated using the unrestricted 3PL model within a multiple group analysis in MULTILOG. On each form, the groups were defined according to the ability level: Group 1 represented low ability, while Group 2 represented high ability. For each group, two hierarchical models were estimated: Model 1 was an unrestricted model in which the corresponding item parameters (slopes, *a*'s, thresholds, *b*'s, and guessing parameters, *c*'s) for all items on each Forms A and B were free to vary across Groups 1 and 2, but Model 2 was a restricted model in which the corresponding parameters for all 1items on each form were set to be equal across the two groups. In both models, the mean of Group 1 was estimated and the standard deviation was fixed at 1. The mean for the ability distribution of Group 2 was fixed at 0 and the standard deviation was fixed at 1. The unrestricted model was compared to the restricted model by subtracting the –2 times the log of the likelihood and comparing this value to the chi-squared distribution with degrees of freedom equal to the difference in the parameters estimated. If the difference chi-square between Models 1 and 2 is not significantly different, then the item parameters of the 3PL IRT model should be invariant across groups for Forms A and B of the CEPA-English test, regardless of possible differences in the ability distribution means of the groups.

In addition to the examination of the -2log likelihood statistics between the restricted and unrestricted models, the pairs of the item parameter estimates obtained from the unrestricted models for each subgroup were correlated and plotted. The invariance of the item parameters of the 3PL IRT model will hold for Forms A and B if the correlations are high (i.e., .80 and higher) and the scatter plots are linear (Hambleton & Swaminathan, 1985; Hambleton, et al., 1991).

3.6.3 Detecting DIF using the Mantel-Haenszel Procedure

Because the CEPA-English test impose serious consequences on 12th grade students, it is necessary to examine whether any item on each form of the CEPA-English test exhibits DIF. DIF is an important indicator of irrelevant source of variance, such as content of items that results in systematically lower or higher scores for members of particular groups (Messick, 1989). In this study, the Mantel-Haenszel (MH) procedure was used to flag DIF in all items on Forms A and B the CEPA-English test because of its simplicity and ease of interpretation (Rogers & Swaminathan, 1993). Specifically, A two-stage MH process suggested by Holland and Thayer (1988) was used to examine whether the items on Forms A and B of the CEPA-English test exhibit DIF between males and females, between study type (i.e., Arts and Science), and between school types (i.e., public, private, and home-schooling). It was expected that Forms A and B of the CEPA-English test would be free from the DIF items.

The program EZDIF version 1.0 developed by Waller (1998) was used to uniform perform DIF analyses on Forms A and B. For each item, the EZDIF program was used to obtain the MH chi-square ($MH\ \chi^2$) statistic, the MH common odds ratio ($\hat{\alpha}_{MH}$), the *p-value* of $MH\ \chi^2$, the MH delta scale ($\hat{\Delta}_{MH}$), 5) the standard error of $\hat{\Delta}_{MH}$, and 6) ETS classification of effect size, indicating negligible (*A*), moderate (*B*), or large (*C*) DIF.

The first step in perform DIF analyses using EZDIF involves identifying the focal and reference groups whose performances on an item are to be compared. For gender, male was the focal group, for study type, Arts stream was the focal. When comparing private and public schools, public was the focal group, and when comparing private and home-schooled, home-schooled was the focal group. When comparing public and home-schooled, home-schooled was the focal group. By contrast, the reference group was coded as “0” (i.e., female, Sciences stream, and private schools). The total test score, which was comprised of the number of items that were answered correctly, was used to as matching variable for this analysis. Also, 10 (0-10) scoring scale were used on Forms A and B of the CEPA-English test.

Sample size in the focal and reference groups is a critical factor that affects the power of detecting DIF (rejecting a true null hypothesis); with a large sample size, even a small DIF will be statistically significant. The sample sized required for the MH method is 200-500, with a minimum of 200 subjects in each group. Taking into account the effect of sample size, both random and equal sample sizes were used for the MH DIF analysis on each from of the CEPA-English test. The number of students in the original data was 9, 496 for Form A; 9,269 for Form B. In the DIF analyses for Forms A and B, a random sample of 2000 subjects with equal numbers of male and female subjects as well as equal number of students majoring in Arts or Sciences was used. Another random sample includes 1400 subjects with equal numbers of students in public schools, private schools, and home-schooled was also used.

Using EZDIF, in the first stage of DIF analyses, the total score was used as the matching variable. In the second stage of DIF analyses, item(s) identified as type *C* DIF items were removed from the matching variable in order to purify the matching criterion which may be contaminated by DIF items.

Two criteria were used to flag items as differentially functioning. The first criterion was a statistically significant *p* value of the $MH\ \chi^2$ test statistic. Because a statistically significant test does not necessarily indicate that the magnitude of DIF is statistically significant, a second method were used to flag DIF items; specifically, following guidance proposed by ETS, CEPA-English items were classified into the three levels (A, B, and C) of the DIF as follows:

- Items with negligible DIF, or Category A, have a $\hat{\Delta}_{MH}$ value not significantly different from zero ($p \geq .05$), or $|\hat{\Delta}_{MH}| < 1$.
- Items with moderate DIF, or Category B, have a $\hat{\Delta}_{MH}$ value significantly different from zero ($p < .05$) and $1 \leq |\hat{\Delta}_{MH}| < 1.5$.
- Items with large DIF, or Category C, have a $\Delta\hat{}_{MH}$ value significantly greater than 1.0 ($p < .05$) and $|\Delta\hat{}_{MH}| \geq 1.5$.

In addition to using the magnitude of the $\Delta\hat{}_{MH}$, the sign of $\Delta\hat{}_{MH}$ was also used to examine the direction of DIF (Zieky, 1993). An item with a positive value of the $\Delta\hat{}_{MH}$ means that the item was more difficult for the reference group. Conversely, an item with negative value means that the item was more difficult for the focal group.

3.6.4 Equating Using Equipercentile Procedure

The 2007 CEPA-English test had four forms (A, B, C, and D) that were randomly distributed to the examinees. NAPO used 10 internal anchor or common items for the purpose of equating the four forms: five common-items were used for equating Forms A and B (administered in the morning) and a different set of five common-items were used for equating Forms C and D (administered in the afternoon). Thus, A and B are not linked with C and D. NAPO used different five common-items in the morning and afternoon to prevent the morning anchors from becoming compromised.

These common-items, which only represented grammar and vocabulary domains, were placed in different positions across the test forms. Furthermore, NAPO used item parameter estimates on the five common-items from the IRT 3PL model calibration to place parameter estimates from Forms A and B on the same scale and another five common-items to equate Forms C and D. The IRT Item parameters for the 10 anchor-items across the CEPA-English test forms were selected from previous administrations of the CEPA-English test.

NAPO used the XCALIBRE computer program (Assessment Systems, 1997) with a marginal maximum-likelihood estimation (MMLE) to estimate the IRT parameter estimates for common-items under the 3PL model. Item parameters for common-items were separately generated for each test form. Specifically, to equate Forms A and B, NAPO created an anchor file with item parameters for the common-items. Then, NAPO anchored the item parameters of the common-items when it ran the IRT analysis for each form. NAPO fixed the *a*, *b,* and *c* item parameter estimates for the five common-items taken from the previous year's test at their previously estimated values (see Table 6) on both forms. Then, item parameters for the remaining 115 non-common items were estimated with the common-items. The fixed calibration was separately implemented with XCALIBRE for each form. The same process was also used to equate Forms C and D.

Table 6. The IRT Item Parameters for the Five Common-Items across the CEPA-English Test Forms Taken from the Previous Year's Test

		IRT Item Parameters		
Test Form	Common-Items	*a*	*b*	*c*
Form A	2	0.78	1.68	0.20
	10	1.28	0.23	0.25
	27	2.05	0.95	0.17
	30	1.22	0.80	0.16
	48	1.79	0.68	0.24
Form B	37	0.78	1.68	0.20
	13	1.28	0.23	0.25
	35	2.05	0.95	0.17
	20	1.22	0.80	0.16
	44	1.79	0.68	0.24
Form C	12	1.15	0.13	0.25
	14	1.56	1.33	0.20
	21	1.59	0.95	0.21
	27	1.28	1.16	0.29
	33	0.94	0.65	0.26
Form D	5	1.15	0.13	0.25
	15	1.56	1.33	0.20
	38	1.59	0.95	0.21
	2	1.28	1.16	0.29
	32	0.94	0.65	0.26

The present study did not replicate the equating method that was used by NAPO because it only used five common items in Forms A and B and a different set of five common items in Forms C and D. Thus, Forms A and B were equated in this study since the scale is different across forms. So any interpretation regarding student performance scores across forms will be misleading. As shown in Table 7, the mean *p*-values (item difficulty) for the five common-items suggested that performance of the 12th grade students were similar for Forms A and B and less so for Forms C and D. The results of *p*-values also indicated that the performance of the students were higher on both Forms A and B than on Forms C and D. So it appeared that the students who took Forms A and B were different than those who took Forms C and D.

Table 7. Means and Standard Deviations of *p*-values for the Five Common-Items by Forms

Test Form		Common-Items				
		Item 2	Item 10	Item 27	Item 30	Item 48
Form A	*N*	9469	9469	9469	9469	9469
	M	.440	.705	.495	.577	.575
	SD	.497	.456	.500	.494	.494
		Item 37	Item 13	Item 35	Item 20	Item 44
Form B	*N*	9282	9282	9282	9282	9282
	M	.471	.703	.529	.552	.602
	SD	.499	.457	.499	.497	489
		Item 12	Item 14	Item 21	Item 27	Item 33
Form C	*N*	5783	5783	5783	5783	5783
	M	.552	.645	.399	.603	.479
	SD	.497	.478	.490	.489	.500
		Item 5	Item 15	Item 38	Item 2	Item 32
Form D	*N*	5899	5899	5899	5899	5899
	M	.348	.488	.505	.643	.505
	SD	.476	.500	.500	.497	.500

It is important to point that the same five common-items for Forms A and B and for Forms C and D were placed in different positions (see Table 6). The mean of *p*-values ranged from .440 to .705 for Form A; from .471 to .703 for Form B; from .399 to .645 for Form C; and from .348 to .643 for Form D. The standard deviation scores of the p-values ranged from .456 to .500 for Form A; from .457 to .499 for Form B; from .478 to .500 for Form C; and from .476 to .500 for Form D.

As mentioned in Chapter 2, IRT equating using the anchor-item design can be successful only if the common items meet the following criteria: a) careful selection, b) representative of the entire test in terms of content and difficulty level of items, c) similar position across the multiple forms, d) cover a range of difficulty levels, and e) long enough to adequately represent the entire test. A rule of thumb for the minimum length of the common-items is 20-25% of the total test length (Kolen & Brennan, 1995, 2004). However, NAPO only used five common-items from the grammar and vocabulary section to equate the CEPA-English test forms; neither of these five common-items provides adequate stability nor allows for appropriate coverage of test content. In addition, these items are not in similar positions across the multiple forms.

Because it was expected that the anchor-item design with five common-items would not be appropriate for equating the CEPA-English test scores across Form A and B, the present study did not replicate the equating method that was used by NAPO. Using only 5 common-items could lead to large random equating error. In this study, equipercentile equating method under the random-groups design was used to equate Forms A and B of the CEPA-English test.

In this study, groups of examinees for Forms A and B of the CEPA-English test were defined by school type (i.e., public, private, and home-schooling).[12] Because test forms were randomly distributed to examinees, it was reasonable to assume that examinees who either took test Form A or B from different booklets were equivalent. That is, the examinees who took test Form A were randomly equivalent to those who took Form B. Therefore, it was reasonable to use a random groups design. After choosing the appropriate design, the second step was to select the statistical equating methods. In this study, equipercentile equating with the cubic spline postsmoothing method under the random-groups design was used to equate Forms A and B of the CEPA-English test.

Using cubic spline postsmoothing, the RAGE-RGEQUATE computer program was used to implement the random-groups equipercentile equating. This program requires a control file name (i.e., the input data file), which was created using Notepad in Windows system. This control file was created as follows: the first entry column was the raw score (*x*); the second entry column was the frequency for the new form (Form B); the third entry column was the frequency for the old form (Form A); and the fourth entry column was the raw-to-scale score equivalent for the old form.

The use of cubic smoothing spline in a postsmoothing method does not provide a statistical test. Therefore, it was important to carefully inspect the graphs and moments to choose the degree of smoothing. Equipercentile equating with the cubic spline postsmoothing method was done with nine different values of *S*, ranging from .01 to 1. Also, standard error (*SE*) was used to evaluate the accuracy of equating with postsmoothing method. The ± SE band of difference of unsmoothed frequency method was also graphed. It was expected that as *S*

[12] Home-schooled students mostly but not entirely attended the afternoon session. Thus, the majority received either Form C or D.

increased, the smoothed relationships would differ more than the unsmoothed, and some relationships would be outside of the *SE* band.

3.6.5 Assessing Test Information Function using IRT

It is important to examine the test information function (TIF) as it is a measure of the precision of the test scores. The TIF is obtained by summing the items information function for all items on a test. The amount of information a test provides at a θ level is inversely related to precision with which ability is estimated at that point of θ level.

The CEPA-English test score is used for admission purposes, in that each student must achieve a minimum score of 150 on the test to be admitted into one of the three institutions' undergraduate programs (i.e., UAEU, ZU and HCT). The test score is used for placing students with a score below 150 into the appropriate levels of English proficiency in the remedial program. Because of the potentially life change consequences of this cutoff score, it is critical to assess the extent to which the test information function for Forms A and B of the CEPA-English test is maximized at the cutoff score of 150, which is the mean of the test in the NAPO study.

MULTILOG was used to examine the amount of information Forms A and B provided at nine ability (θ) intervals, ranging from -3.0 to 3.0 for the 3PL model. A graph of the test information function and standard error was also provided. For each form, it was expected that each form would provides more precise information at the cutoff score of 150. In other words, the amount of the test information functions would peak at an ability score of zero and that the standard error would be minimized.

4.0 RESULTS

The main purpose of the present study was to explore the psychometric quality of the CEPA-English test. This study addressed the following research questions:

1. Do Forms A and B of the CEPA-English test data meet the assumptions of IRT?
2. Does the preferred IRT model fit each item on Forms A and B of the CEPA-English test data?
3. Does the property of invariance of item parameters hold true for Forms A and B of the CEPA-English test data?
4. Are there any DIF items on Forms A and B of the CEPA-English test?
5. To what extent are the CEPA-English test scores equivalent across Forms A and B?
6. To what extent is the test information function for Forms A and B of the CEPA-English test maximized at the cutoff score of 150?

Chapter 4 presents the results of the statistical analyses with respect to the above research questions. The results were organized into seven sections. Section one provides a summary of the classical analyses at item and test levels for Forms A and B. Section two discusses the results from examining the assumptions underlying dichotomous IRT models. This examination included 1) choosing the appropriate IRT model, 2) evaluating the internal structure and unidimensionality, 3) examining local item independence, 4) examining speededness, and 5) assessing the presence of guessing behavior. Sections three and four contain the results related to assessing the invariance of item parameters and assessing model-data-fit at item level, respectively. Section five presents results regarding detecting DIF items via the Mantel-Haenszel procedure. Section six provides results of equating Forms A and B using equipercentile with the cubic spline postsmoothing method. The last section illustrates the amount of information provided by Forms A and B.

4.1 EXAMINING THE PSYCHOMETRIC PROPERTIES OF THE CEPA-ENGLISH TEST USING CLASSICAL AND IRT ANALYSES

4.1.1 Classical Analyses

The analyses in this study were carried out on data obtained from NAPO, taken from the 2007 administration of Forms A and B of the CEPA-English test. A Summary of classical analyses results for Forms A and B are presented in Table 8.

Table 8. Summary of Classical Item and Test Analyses for Forms A and B

Statistics	Form A	Form B
N of Examinees	9496	9296
n of Items	116	119
M of r_{pbis}	.420	.399
SD of r_{pbis}	.123	.119
M of *p*	.542	.500
SD of *P*	.173	.159
M of TS	62.82	59.58
Range of TS	16-116	12-119
SD of TS	23.706	23.604
Skewness of TS	.282	.517
Kurtosis of TS	-1.028	-.765
KR-20 Reliability Coefficient	.963	.960

Note. *p*=the proportion of correct responses for each item. TS= total score.

The sample size of this study consisted of 9,496 students for Form A and 9,296 students for Form B. To determine the overall psychometric features of the two forms, first, item point-biserials (r_{pbis}) and KR-20 correlation coefficients were computed for Forms A and B. Any item with a negative r_{pbis} was removed from all analyses since it indicted that an item was difficult for high ability students and easier for low ability students. As a result, four items (47, 86, 87, and 97) were deleted from Form A and one item (2) was deleted from Form B. Excluding these items resulted in having 116 items in Form A and 119 items in Form B. The r_{pbis} values for Form A ranged from .04 to .66, with a mean of .420 and a standard deviation of .123. The r_{pbis} values for Form B ranged from .08 to .63, with a mean of .399 and a standard deviation of .119. Thus, the average r_{pbis} values for both forms were above the criterion of $r_{pbis} \geq .3$ (Nitko, 2004), suggesting that, overall, the items are consistent with the entire test. The KR-20 reliability coefficients for Forms A and B were .963 and .960, respectively, indicating very high internal consistency reliability.

For both Forms A and B, there was a large variation in the *p*-values (item difficulty), ranging from .11 to .87 for Form A, and .16 to .81 for Form B. An inspection of the *p*-values suggested that the majority of the test items in both forms were relatively moderate in their difficulty levels; overall, the mean *p*-values was .542 for Form A and .500 for Form B, indicating that Form A was slightly easier than Form B.

The overall total scores on Form A ranged from 16 to 116, while the total scores on Form B ranged from 12 to 119. These results indicated that the sample in both forms was heterogeneous in terms of English proficiency, as indicated by the large range of CEPA-English scores, and thus they might be somewhat heterogeneous in their general English ability. The mean of the total score for Form A was 62.82 with a standard deviation of 23.706. The mean of the total score for Form B was 59.58 with a standard deviation of 23.604.

The histograms shown in Figures 10 and 11, as well as the skewness and kurtosis statistics, revealed that the frequency distributions for Forms A and B were slightly different in their skewness and kurtosis. As Figure 10 displays, the distribution of the total scores for Form A was slightly positively skewed, with a value of .282, and flatter than a normal distribution, with a kurtosis of -1.028. The distribution of the total scores for Form B appeared to be positively skewed, with a value of .517, and had a slightly flatter distribution with a kurtosis of -.765. Thus, distributions of both forms were slightly positively skewed, but Form B seemed to be somewhat more skewed than Form A (see Figure 11).

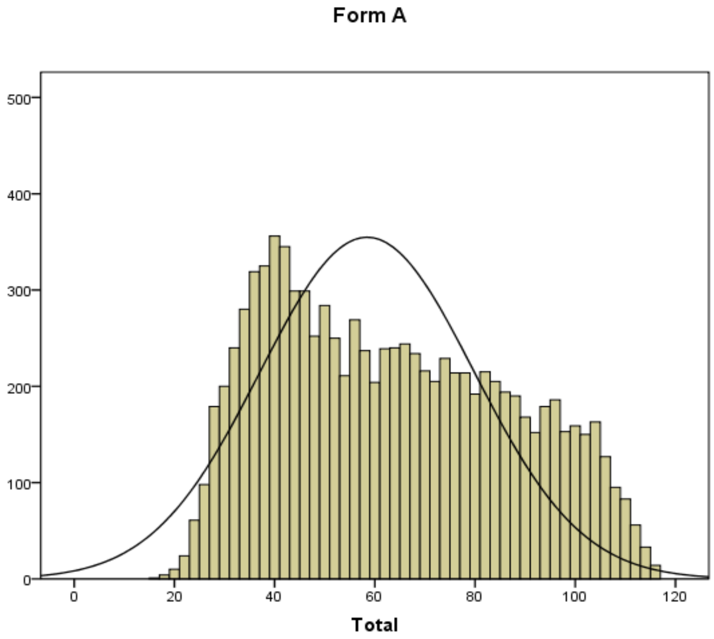


Figure 10. Frequency Distribution of Form A

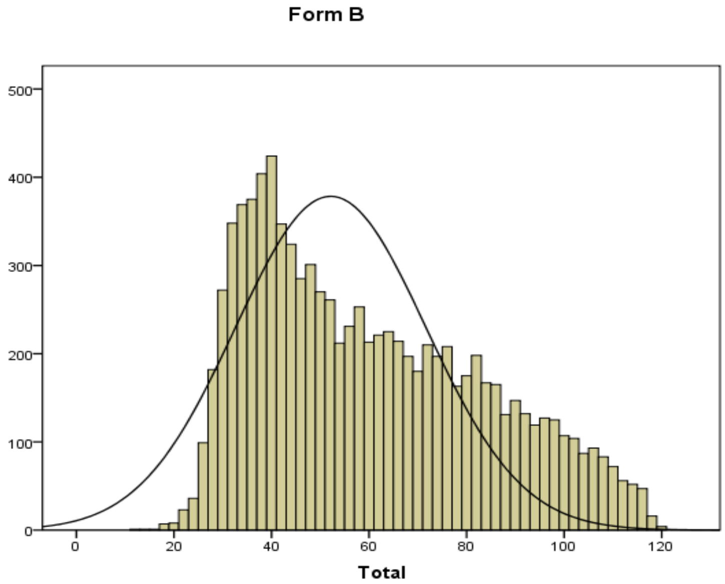


Figure 11. Frequency Distribution of Form B

4.1.2 IRT Analyses

4.1.2.1 Fixing the *c* Parameter Values under the Unidimensional 3PL IRT Model

Because each form of the CEPA-English test is multiple choice, in which examinees can obtain correct answers simply by guessing, it was expected that the unidimensional 3PL IRT model would provide a better fit for each form. The 3PL model estimates the item difficulty (*b*), discrimination (*a*), and guessing (*c*) parameters to describe each test item. To ensure stable estimates of the *b* and *a* parameters, the *c* parameter values for some items were fixed at either 0 or .25 for Forms A and B of the CEPA-English items. There were two conditions under which the *c* parameter values were fixed at 0. The first was when the value of the *c* parameter did not meet Lord's (1980) criterion (estimate *c* when $b – 2/a > –3.5$). The second was when a standard error (*SE*) for *c* could not be estimated. There were two conditions under which the *c* parameter values were fixed at .25. The first condition was when the value of the *c* parameter value was .4 or above. This is because, in standard practice, *c* values greater than .4 may be both unrealistic for multiple-choice tests with four-options and may lead to higher measured error. The second condition was when items did not fit the 3PL model.

In order to identify the *c* parameters that needed to be fixed, it was important to examine the fit for individual items prior to using the 3PL model to estimate the item parameters of Forms A and B of the CEPA-English test. Item misfit exhibits if there is a difference between the observed and predicted score distributions across a range of discrete ability levels for each item. A graphical display of observed (theoretical) versus (empirical) predicted score distributions was used to identify items that misfit via the ResidPlots-2 program, graphical residual analyses, developed by Liang, Han, and Hambleton (2008) (see Figures D1 and D2 in Appendix D).

Based on the theoretical and empirical ICC's for Form A and B, *c* value were fixed at .25 for Items 8, 28, 42, 51, and 94 in Form A and for items 5, 8, 21, 25, 27, 35, 59, 66, 69, 73, 79, 85 92,102, 106, and 116 in Form B. The theoretical and emprical ICC's differed greatly at the lower end of theta indicating item misfit. After re-computing the item misfit statistics, the results indicated that these items still did not fit the 3PL model, and therefore the *c* parameters were re-estimated if they were not initially equal to or above .4, or *SE*s could not be estimated. The item fit analyses will be discussed later in section 4.1.2.3. The values of the final estimation of the *c* parameter for those items that were fixed at either 0 or .25 for Forms A and Bare presented in Table 9. The values of the estimated item parameters (*a*, *b*, and *c*) for Forms A and B are listed in Tables C1 and C2 in Appendix C, respectively.

Table 9. Values of the Final Estimation of the *c* Parameters for Those Items that were Fixed

Test Form	Item	Condition	Before Fixing *c*	After Fixing *c*
Form A	1	Lord' criterion	.18	0
	12	Lord' criterion	.21	0
	20	*c* value	.40	.25
	25	*c* value	.47	.25
	26	SE	.00	0
	31	SE	.00	0
	36	SE	.00	0
	76	Lord' criterion	.05	0
	79	Lord' criterion	.11	0
	82	SE	.00	0
	84	SE	.00	0
	88	SE	.00	0
	93	*c* value	.40	.25
	101	*c* value	.46	.25
	102	SE	.00	0
Form B	2	Lord' criterion	.07	0
	10	*c* value	.45	.25
	14	SE	.00	0
	27	*c* value	.42	.25
	74	SE	.00	0
	82	SE	.00	0
	95	SE	.00	0
	107	Lord' criterion	.15	0
	111	Lord' criterion	.09	0

Note. Lord' criterion condition denotes that the value of the *c* parameter did not meet Lord's criterion. The SE condition denotes that a standard error for *c* could not be estimated. The *c* value condition denotes that the value of the *c* parameter was .4 or above.

4.1.2.2 Examining the Assumptions Underlying Dichotomous IRT Models

4.1.2.2.1 Model Testing: Choosing the Preferred IRT Model

Although it is reasonable to use the 3PL model with each form of the CEPA-English test, it is important to evaluate whether the 3PL is the most suitable model to describe the items in each form. To determine which IRT model—1PL, 2PL, or 3PL—best fits each form of the CEPA-English test data, the difference in the -2log likelihood statistic of the nested models were compared using MULTILOG. The results of comparing the three nested models are presented in Tables 10 and 11.

Table 10. Comparison of the 1PL versus the 2PL

Test Form	–2 log of 1PL model	–2 log of 2PL model
Form A	656055.2	701892.1
Form B	531204.7	585950.3

Note. - 2log = -2log likelihood statistic.

Table 11. Comparison of the 2PL versus the 3PL

Test Form	–2 log of 2PL model	–2 log of 3PL model
Form A	701892.1	724234.5
Form B	585950.3	630669.5

Note. 2log = -2log likelihood statistic.

As shown in Tables 10 and 11, the -2log likelihood statistics for Form A with 116 items was 656055.2for the 1PL and 701892.1for the 2PL models. The difference between the chi-square of the two models, or G^2, was -45836.9. The -2log likelihood statistics were 701892.1 for the 2PL and 724234.5 for the 3PL models. The G^2 observed value was -22342.4.

Similar findings also were observed on Form B with 119 items; that is, the -2log likelihood statistics were 531204.7 for the 1PL and 585950.3 for the 2PL models. The G^2 observed value was -54745.6. The -2log likelihood statistics were 585950.3 for the 2PL and 630669.5 for the 3PL models. The G^2 observed value was -44719.2.

Thus, for both Forms A and B the G^2 observed values were negative, which indicated a problem with estimating the -2log likelihood values. That is, the value of the -2log likelihood was smaller for a more-constrained model (1PL and 2PL) than for a less-constrained one (3PL). This problem is due to the large number of items in both forms (i.e., 116 items in Form A and 119 items in Form B). There appears to be a "bug" in the MULTILOG program when using item samples of 116 or more. Because of this problem with the -2log likelihood, additional evidence was obtained to justify the use of the 3PL model, which takes guessing into account. Therefore, the parameters were re-estimated using a reduced test with 100 randomly selected items obtained from each of Forms A and B of the CEPA-English test. Specifically, the -2log likelihood statistics of the 1PL, 2PL, and 3PL models were obtained for the reduced test with 100 random items for Forms A and B. In addition, descriptive statistics for the item parameters (*a*, *b*, and *c*) for the full and reduced tests were examined for Forms A and B. Furthermore, the correlations and the scatterplots of the item parameters for the reduced test and its corresponding items in the full test were also examined for each form. The results of comparing the three nested models using 100 random items for Forms A and B are presented in Tables 12 and 13.

Table 12. Comparison of the 1PL versus the 2PL

Test Form	–2 log of 1PL model	–2 log of 2PL model	G^2	*df*	*p*
Form A	814162.0	804886.5	9275.5	100	<.05
Form B	839599.9	828383.2	11216.7	100	<.05

Note. - 2log = -2log likelihood statistic.

Table 13. Comparison of the 2PL versus the 3PL

Test Form	–2 log of 2PL model	–2 log of 3PL model	G^2	*df*	*p*
Form A	804886.5	796496.0	8390.5	100	<.05
Form B	828383.2	817293.5	11089.7	100	<.05

Note. - 2log = -2log likelihood statistic.

According to Tables 12 and 13, the -2log likelihood statistics for Form A with only 100 random items was 814162.0 for the 1PL model and 804886.5 for the 2PL model. As expected, the G^2 observed value of 9275.5 was greater than its critical value of 124.3 for p=0.05 with 100 degrees of freedom, indicating that the additional parameters specified in the 2PL did afford a significant improvement in model-data fit. The -2log likelihood statistics were 804886.5 for the 2PL model and 796496.0 for the 3PL model. The G^2 observed value of 8390.5 was greater than its critical value of 124.3 for p=0.05 with 100 degrees of freedom, suggesting that the additional parameters specified in the 3PL did afford a significant improvement in model-data fit.

Similar findings also were observed on Form B with only 100 random items; that is, the -2log likelihood statistics were 839599.9 for the 1PL model and 828383.2 for the 2PL model. The G^2 observed value of 11216.7 was greater than its critical value of 124.3 for p=0.05 with 100 degrees of freedom, suggesting that the additional parameters specified in the 2PL did afford a significant improvement in model-data fit. The -2log likelihood statistics were 828383.2 for the 2PL model and 817293.5 for the 3PL model. The G^2 observed value of 11089.7 was greater than its critical value of 124.3 for p=0.05 with 100 degrees of freedom, which also suggested that the additional parameters specified in the 3PL did afford a significant improvement in model-data fit.

In addition to comparing the -2log likelihood statistics of the nested models, descriptive statistics for item parameters (*a*, *b*, and *c*) estimated from MULTILOG were examined and summarized in Table 14.

Table 14. Descriptive Statistics for Item Parameters for Forms A and B

Test Form	No. of Items	Statistics	*a*	*b*	*c*
Form A	*n* =116	*M*	1.289	.317	.206
		SD	.454	.968	.092
	n =100	*M*	1.312	.280	.204
		SD	.487	.954	.093
Form B	*n* =119	*M*	1.290	.530	.206
		SD	.480	.928	.085
	n =100	*M*	1.312	.504	.205
		SD	.496	.925	.087

Examination of the mean and standard deviation of item parameters (*a*, *b*, and *c*) estimated from the full and reduced tests for Forms A and B revealed that both tests had similar item parameter values (see Table 14). More significantly, in the full test both Forms A and B had slightly different *b* values. The *b* values typically range from -3.0 to 3.0; lower *b* values correspond to easier items whereas higher *b* values correspond to more difficult items (Baker, 2001). In Form A, the *b* values ranged from -1.85 to 3.70, with a mean of .317 and a standard deviation of .968. In Form B, the *b* values ranged from -1.70 to 2.55, with a mean of .530 and a standard deviation of .928. This finding indicated that most of the items on both forms were relatively moderate in their difficulty levels and that Form B (*M* of *b* =.530) was more difficult than Form A (*M* of *b* =.317).

Both Forms A and B had similar *a* values, which typically range from 0 to + 2.0; a high *a* value indicates that the item has a steep ICC and discriminates well (Baker, 2001). The *a* values for Form A ranged from .18 to 2.25, with a mean of 1.289 and a standard deviation of .454, and the *a* values for Form B ranged from .25 to 2.54, with a mean of 1.290 and a standard deviation of .480. This means that the items in both forms moderately discriminated between high-performing and low-performing students. The wide range of the *a* values in both forms suggested that the assumption of equal discrimination for Forms A and B of the CEPA-English test data was incorrect. Consequently, this excludes using the 1PL model, which assumes equal discrimination values for Forms A and B test data.

Both Forms A and B also had similar *c* values, which typically range from 0 to .30 or .40. In Form A, the *c* values for 116 items ranged from .00 to .38, with a mean of .206 and a standard deviation of .092. In Form B, the *c* values ranged from .00 to .38, with a mean of .206 and a standard deviation of .085. This means that for both forms, the probability for low ability examinees to get a correct response by guessing was moderate. The range of the *c* values suggested that it was likely that guessing occurred, as students tended to guess answers to items that they did not know. Hence, the 3PL model was the most appropriate model to use with Forms A and B test data.

Furthermore, the results of examining the correlations and scatterplots of the item parameters for both full and reduced tests for Forms A and B demonstrated that there was a high relationship between the estimated parameters. The scatterplots of all estimated parameters also showed linear relationships (see Figures 12-14).

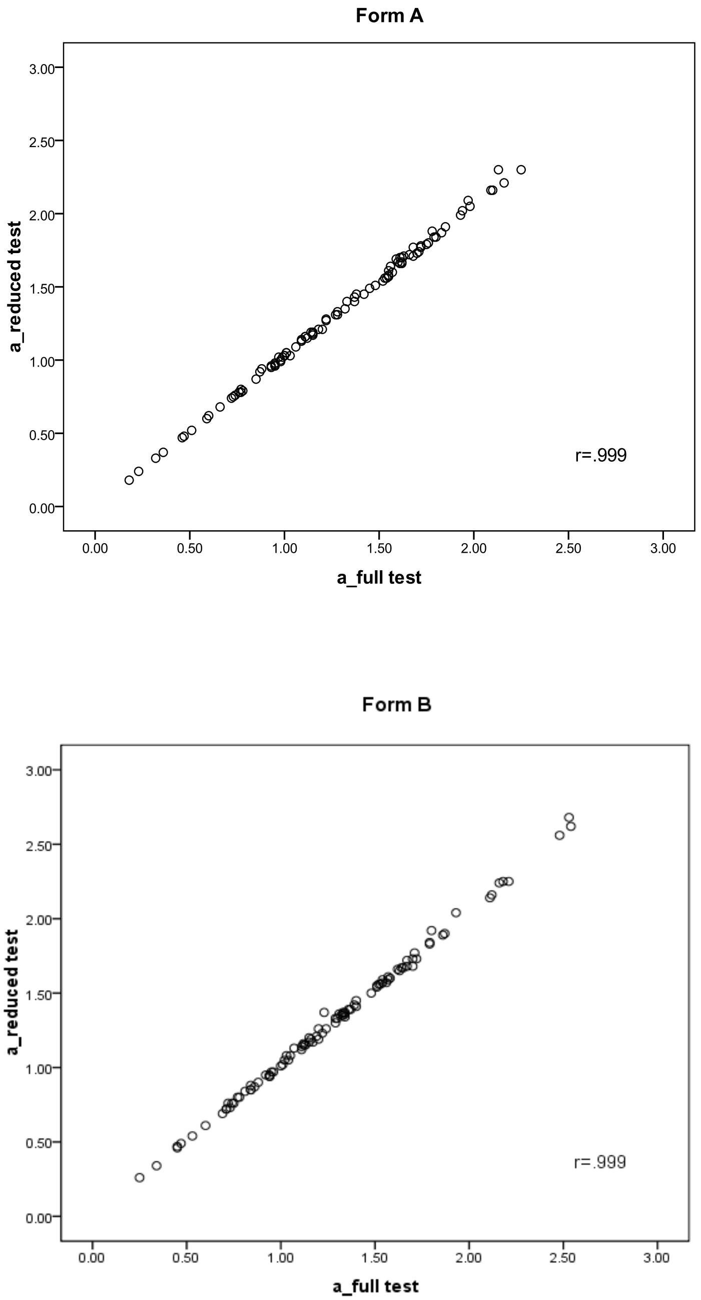
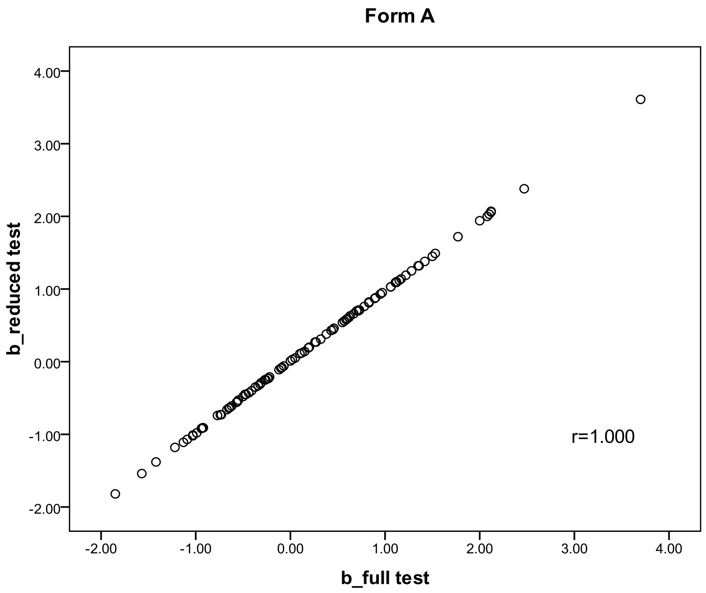


Figure 12. Scatterplots of *a*’s for full and reduced tests on Forms A and B



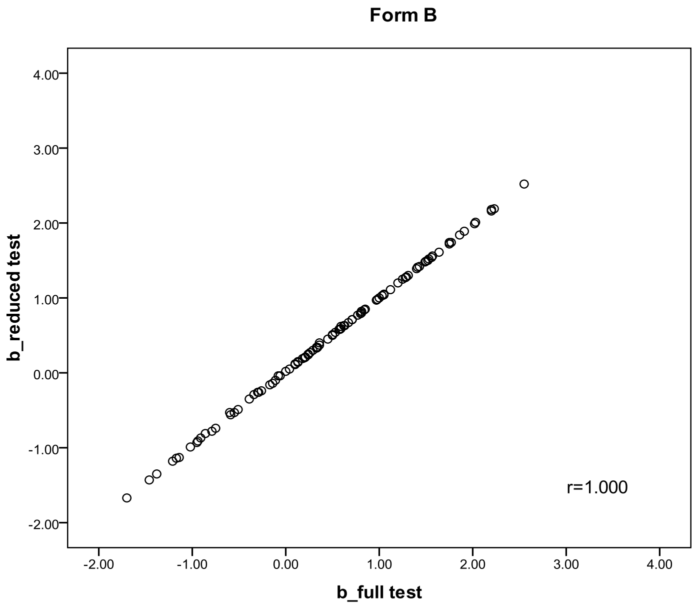


Figure 13. Scatterplots of *b*'s for full and reduced tests on Forms A and B

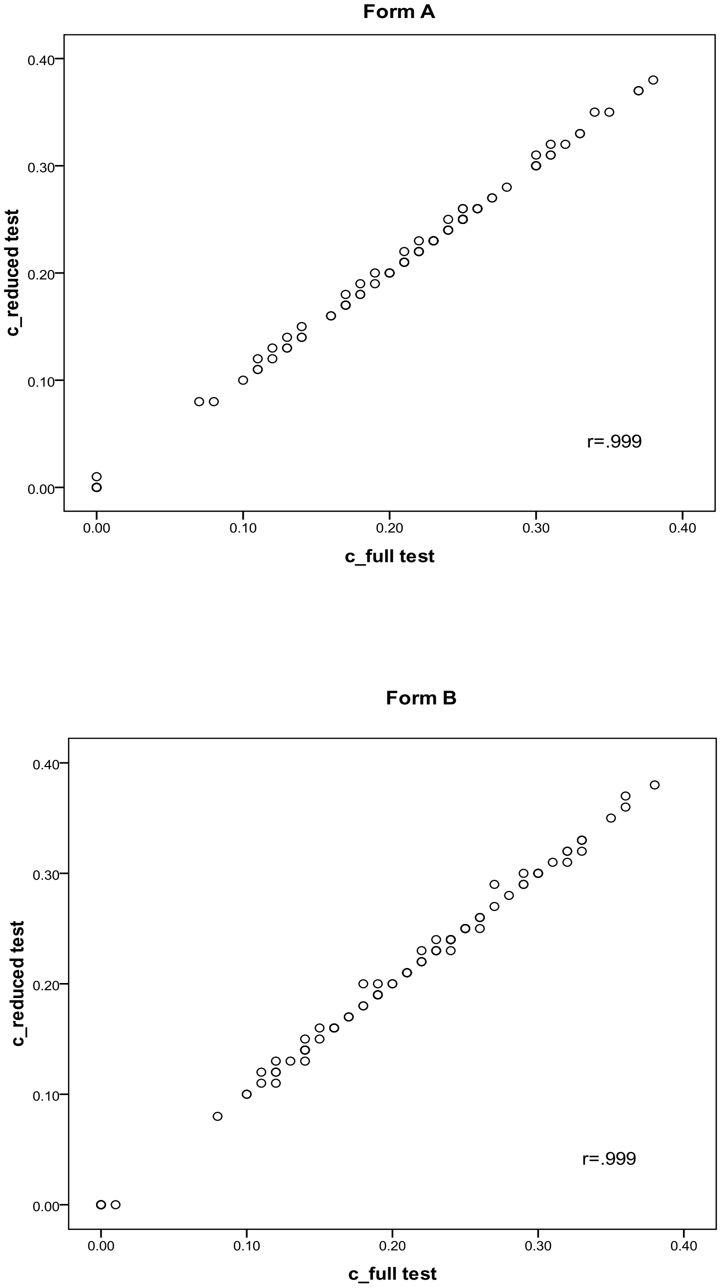


Figure 14. Scatterplots of *c*'s for full and reduced tests on Forms A and B

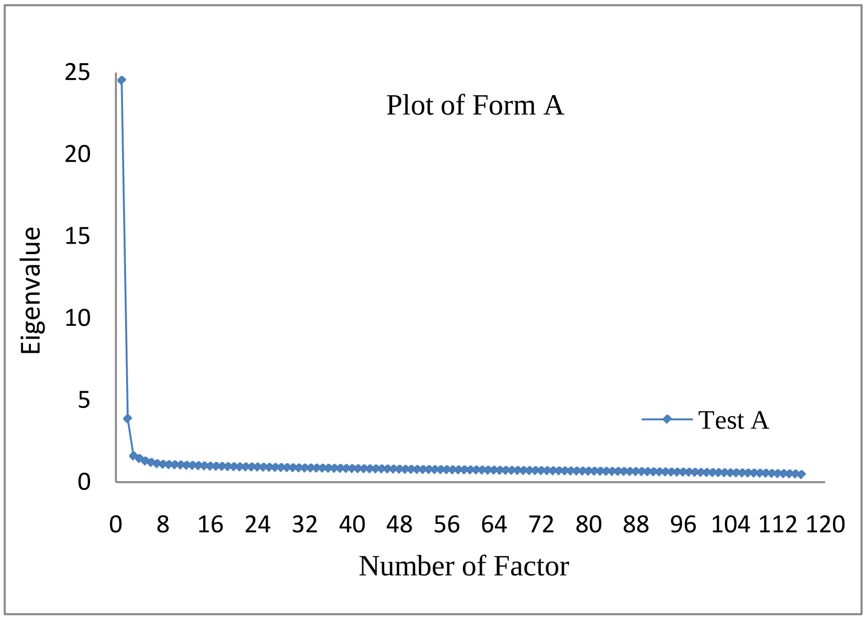
Thus, the 3PL model with reduced test had similar *a*, *b*, and *c* parameter estimates to those of the 3PL model with full test, providing evidence to support the use of the 3PL model. Thus, the 3PL model was used to examine the psychometric properties of the full test (with 116 items in Form A and 119 items in Form B).

4.1.2.2.2.1 Evaluation of the Unidimensionality and Internal Structure

After choosing the preferred IRT model, the second step was to evaluate unidimensionality and the internal structure of Forms A and B of the CEPA-English test data. First, in order to check the unidimensionality of each form, several indices were examined, including the proportion of variance explained by the largest eigenvalues and the eigenvalue plots.

For Form A, the first factor accounted for 21.138 % of the total variance, which is slightly greater than the 20% suggested by Reckase (1979) as adequate for IRT's assumption of unidimensionality. The second factor explained 3.337% of the total variance. The third and the fourth largest factors explained only 1.378 and 1.238 of the total variance, respectively. In addition, the examination of the magnitude of the difference between the eigenvalues of the first factor and the rest of the factors (with eigenvalues greater than 1) revealed that the eigenvalue of the first unrotated factor (24.520) was six times greater than the eigenvalue of the second factor (3.871) and 15 times greater than the eigenvalue of the third factor (1.599). The difference between the second and subsequent eigenvalues was small.

For Form B, the first factor accounted for about 20% of the total variance and had an eigenvalue of 23.058. The second factor explained 3.130% of the total variance and had an eigenvalue of 3.725. The third and the fourth largest factors explained only 1.645 and 1.468 of the total variance, respectively. In addition, the ratio of the first to the second eigenvalue was large (6:1) and the second eigenvalue was close to other eigenvalues. Thus, the eigenvalues obtained from both Forms A and B yielded similar results, supporting the assumption that the test is essentially unidimensional. Hence, there is strong evidence of one dominant dimension in the data set. After examining the eigenvalues, the scree plots for Forms A and B were examined, since the scree plot has traditionally been used to determine the presence of a dominant factor (see Figures 15).



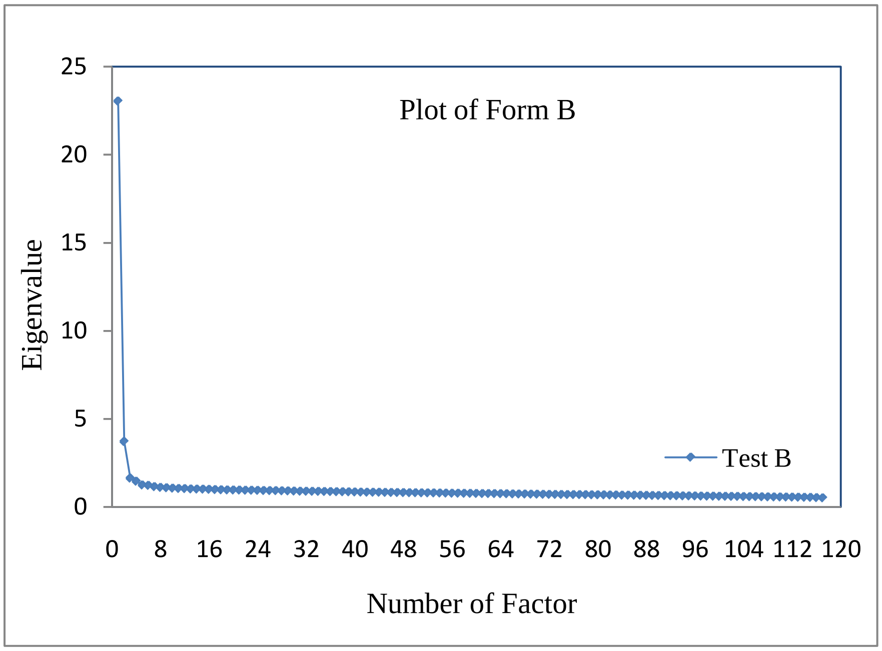


Figure 15. Eigenvalue plots for Forms A and B

The examination of the scree plots for both forms showed that natural breaks occurred after the second eigenvalue. However, the first eigenvalue was markedly higher than the second and the relative change for the second and consecutive eigenvalues was reasonably constant. Thus, the eigenvalue plots appeared to be corroborated with the proportion of variance explained by the largest eigenvalues, providing additional evidence to support one dominant factor underlying Forms A and B of the CEPA-English test—the test was essentially unidimensional.

To further assess the dimensionality and internal structure of Forms A and B of the CEPA-English test data, the root mean square residuals (RMSR) were examined using the NOHARM Program. A range of factor solutions (one-, two-, and three-models) was estimated, given that there was no a priori model with which to fit the data. To evaluate goodness of fit, Tanaka indexes were examined.

A RMSR value equal to or less than four times the reciprocal of the square root of the sample size indicates good model fit (Fraser, 1988). Given that the sample size was 9,496 for Form A and 9,269 for Form B, a RMSR value of 0.041 for both forms indicates an acceptable factor solution. The RMSR values across Forms A and B are given in Table 15.

Table 15. The RMSR Values across Forms A and B

Test Form	Number of Dimensions	RMSR	Tanaka's Index
Form A	1	.004	.983
	2	.003	.992
	3	.002	.993
Form B	1	.004	.987
	2	.003	.993
	3	.002	.994

According to Table 15, the Tanaka indexes for both Forms A and B were generally greater than the criteria value of .95, indicating a good model fit. The RMSR values of Forms A and B for a one-factor model were both .004, indicating that a one-factor model may provide an acceptable solution. A similar RMSR value was also found for both forms for two- and three factors, with small change in the RMSR values between the one- and two factor solutions, as well as between the two- and three factor solutions (see Table 15).

Because the RMSR analysis for both forms yielded similar results across the three solutions, the simplest one-factor solution may be preferred. This is consistent with the eigenvalue analysis because the first eigenvalue was markedly higher than the second and the relative change for the second and consecutive eigenvalues was reasonably constant.

In addition to the RMSR values, the resulting pattern of factor loadings produced by NOHARM was examined for Forms A and B. The Promax-rotated solution was used where an item was loaded on a factor when the absolute value of the loading was greater than .30. The number of substantial positive pattern coefficients and the factor correlation matrix for both forms are shown in Tables 16 and 17.

Table 16. Number of Substantial Pattern Coefficients across Forms A and B

Test Forms	One Factor	Two Factors		Three Factors		
	1	1	2	1	2	3
Form A	113	59	81	22	57	71
Form B	118	87	64	74	54	17

Table 17. Correlations between Factors across Forms A and B

Test Form	Dimension	No. of Substantial Positive Pattern Coefficients				
		Two Factors		Three Factors		
Form A	1	1	2	1	2	3
	2	.717		.642		
	3			.714	.667	
Form B	1					
	2	.740		.693		
	3			.699	.614	

The findings from Table 16 showed that both Forms A and B yielded a similar pattern in the number of substantial coefficients for the one- and two-factor solutions. For example, when one factor was extracted from Form A, all items loaded on the first factor (ranging from .173 to .972) except for three items: items 82 (.173), 29 (.194), and 51 (.241). When one factor was extracted from Form B, all items loaded on the first factor (ranging from. 323 to .980) except for Item 107 (.262).

Under a two-factor solution, 59 items loaded on the first factor in Form A (ranging from .314 to 1.403) and 87 items loaded on the first factor in Form B (ranging from .311 to 1.306), with 81 items loaded on the second factor in Form A (ranging from .316 to 1.146) and 64 items loaded on the second factor in Form B (ranging from .317 to 1.415). In both forms, few items either did not load on any factor or loaded on more than one factor. Items loaded on the second factor did not reflect a particular content area. In addition, with a two-factor solution for both forms, there was no difference in the pattern of the *c* parameter values for those items that loaded on one factor compared to those items that loaded on two factors. Furthermore, the correlation between the two factors was moderate in strength, at .717 for Form A and at .740 for Form B (see Tables 17).

On the other hand, both test forms demonstrated a significantly different pattern in the number of substantial coefficients for the three factor solutions (see Tables 16). In addition, there were few items that either did not load on any factor or that loaded on more than one factor. The correlations for the three factor solutions were generally moderate (see Tables 17).

The results of examining the pattern in coefficients in both test forms provided evidence for at least two dominant factors. However, there was also support for one dominant factor for both forms of the CEPA-test data since almost all the items loaded on the first factor.

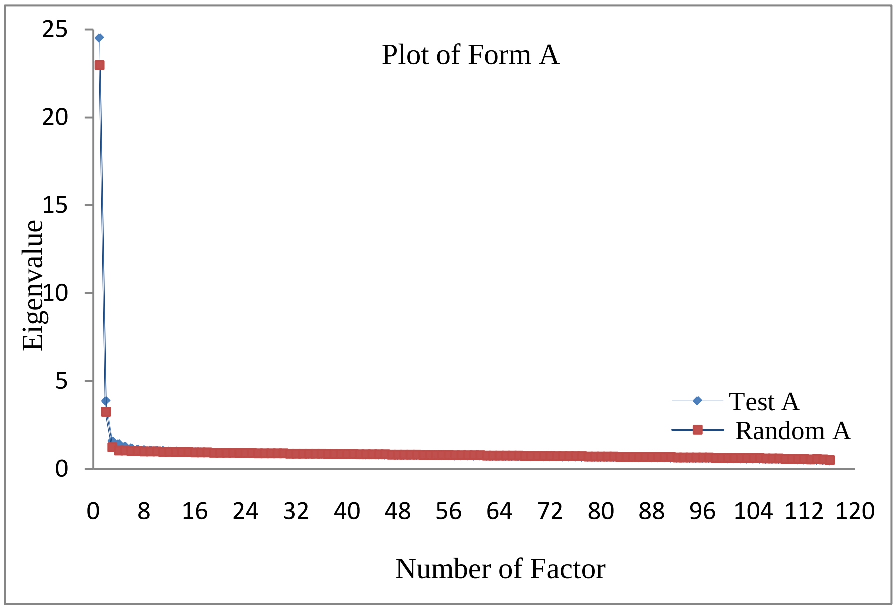
To further check whether Forms A and B of the CEPA-English test is essentially unidimensional, an exploratory DIMTEST was performed as an alternative nonparametric statistical procedure. The value of the DIMTEST *T* test statistic was 8.758 for Form A with a significant *p*-value of 0.000, and 7.616 for Form B with a significant *p*-value of 0.000. Thus, the null hypothesis of essential unidimensionality of both forms of the CEPA-English test data was rejected, indicating that the test has at least two dimensions.

In summary, the assessment of dimensionality employed in DIMTEST suggested that Forms A and B of the CEPA-test data were multidimensional. However, both nonlinear exploratory factor analysis using NOHARM and the eigenvalue analysis using SPSS showed evidence of essentially unidimensionality, that demonstrated the existence of a general dominant factor on both Forms A and B. Thus, it was reasonable to conclude that the unidimensionality assumption of the 3PL IRT model held for Forms A and B of the CEPA-English test data.

4.1.2.2.2.2 Evaluation of Dimensionality Using Simulated Data

In order to have a baseline for assessing the unidimensionality assumption for the real data of Forms A and B, data were simulated under the assumption that the IRT model was true (i.e., given unidimensionality). This was done by using the 3PL IRT item parameter estimates and by sampling randomly from an *N* (0, 1) ability distribution. For Forms A and B, the ability parameters were normally distributed, with a mean of 0 and standard deviation of 1, suggesting that SAS properly generated the ability parameters.

For each test form, the plots obtained from simulations were compared with those obtained from the real data. Figure 16 shows the eigenvalue plot of real versus simulated data for Forms A and B, respectively. For each form, the eigenvalue plots of simulated data appeared to corroborate the real data, providing additional evidence which supports one dominant factor underlying Forms A and B of the CEPA-English test data.



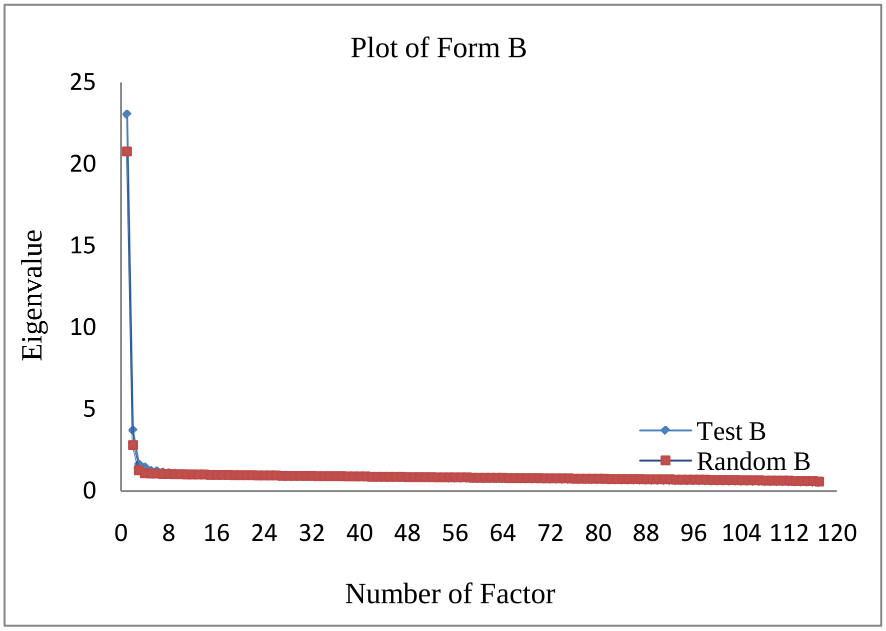


Figure 16. Eigenvalue plot of real versus simulated data for Forms A and B

Furthermore, for each form, the results of the RMSR analyses based on the real data were somewhat consistent with the results based on the simulated data. That is, the RMSR value was lower than the cutoff point, indicating that a one-factor model may provide an acceptable solution. When all three models were compared, a similar RMSR value was also found for two-, and three-solutions, with no change in the RMSR values, (see Table 18). Because the RMSR analyses yielded similar results across the three solutions, the simplest one-factor model may be preferred, and this was consistent with the results obtained from eigenvalue analysis.

Table 18. The RMSR Values across Forms A and B

	Number of Dimensions	RMSR	Tanaka's index
	1	.003	.991
Random A	2	.002	.996
	3	.002	.997
Random B	1	.002	.995
	2	.001	.997
	3	.001	.997

The results of simulated data using DIMTEST were different than those found using real data. For simulated data, a *T* value of .2297 with a non significant *p*-value of .4092 was obtained for Form A, and a *T* value of .1926 was generated for Form B with a non significant *p*-value of

.4237. According to this finding, Forms A and B of the CEPA-English test data were essentially unidimensional.

4.1.2.2.3.1 Assessing Local Item Independence

After meeting the assumption of unidimensionality, the local item independence of Forms A and B of the CEPA-English test was examined. This was done using the Yen's Q_3 statistic via the SAS program. The Q_3 values were calculated for all item pairs from each form. When local independence is true, the expected value is -1/ (*n*-1), where *n* is the total number of items. In this case, the expected value of the Q_3 statistic was -.01 for Forms A and B. As a rule of thumb, all pairs of items with Q_3 values equal to or greater than .20 were flagged as locally dependent. The mean of Q_3 was -.01 for both Forms A and B, which was very close to zero and equal to the expected value. In addition, the Q_3 values ranged from -.090 to .148 for Form A, and from -.095 to .148 for Form B. No pairs of items were found to be locally dependent since no pair of items was found to be equal to or greater than .20, suggesting that there was no correlation between residuals across examinees. Thus, based on these findings, the assumption of local independence did hold for Forms A and B.

4.1.2.2.3.2 Assessing the Local Item Independence Using Simulated Data

In order to have a baseline for assessing the assumption of local item independence for the real data of Forms A and B, the Q_3 values were calculated for each pair of items for Forms A and B using simulated data. The results of the Q_3 statistic based on the real data were consistent with the results based on real data. For simulated data, the mean of Q_3 was -.01 for Forms A and B.

Also, the Q_3 values ranged from -.061 to .031 for Form A and -0.073 to .036 for Form B, suggesting that the simulated data of Forms A and B was locally independent.

4.1.2.2.4 Examination of Speededness

The speededness of Forms A and B of the CEPA-English test was examined by comparing the percentage of examinees who completed all items, and the percentage of examinees who completed at least 75% of the items. In Form A, 99 % of examinees reached the end of the test, 83% of examinees completed all the items, and 0.01 % of examinees completed 75% of the test items (see Tables B1-B2 in Appendix B). In Form B, 99 % of examinees reached the end of the test, about 82% of the examinees completed all the items, and 0.0 % of examinees completed 75% of the test items (see Tables B3-B4 in Appendix B). Based on the results of Forms A and B, the CEPA-English test was not a speeded test, since all examinees completed more than 75% of the items, and 99% of them reached the end of the test.

In addition, the ratio of the variance of omitted items to the variance of incorrectly answered items was calculated as an index of test speededness. The variance of omitted items for Form A was 4.564 and the variance of incorrect answers was 561.985, which yielded a ratio of zero. The variance of omitted items for Form B was 6.294, and the variance of incorrect answers was 557.132, which yielded a ratio of .01. This ratio was very close to zero, indicating that both Forms A and B were not speeded tests.

4.1.2.2.5 Assessing the Presence of Guessing Behavior

Because the CEPA-English test is a multiple-choice test, the extent to which examinees were guessing on each item on Forms A and B was evaluated by plotting the proportion of correct responses by total scores. If guessing behavior exist, the plots of the proportion of correct responses should increase from 0 to 1 as the total score increases. If the plots for an item indicate a moderately constant proportion for a correct response that is greater than 0 for low total scores, then guessing behavior can be assumed. The proportion of correct responses is approximately what would be expected under the random guessing model (1/m, *m* is the number of options). The random guessing model for multiple-choice items with four options is .25. Figure 17 displays plots of proportion of correct responses by total scores for two items in Forms A and B, respectively.

As shown in Figure 17, Item 95 in Form A and Item 75 in Form B exhibited a fairly constant proportion for a correct response around .1 to .25 and .13 to .25, respectively, for examinees with total scores between 23 and 62 for Form A and between 21 and 50 for Form B. These constant proportions of correct responses provided a lower boundary for the probability of a correct response that was greater than 0. This provides evidence that low ability examinees might be using guessing strategies.

Conversely, Item 25 in Form A and Item 82 in Form B in Figure 17 demonstrated the problem that may arise from examining plots of proportion of correct responses for some items. Items 25 and 82 had no relative constant proportion correct for low total scores. Thus, it was assumed that there was guessing behavior operating, which might occur with very easy items (e.g., Item 25 in Form A with p = .87 and Item 82 in Form B with p = .81).

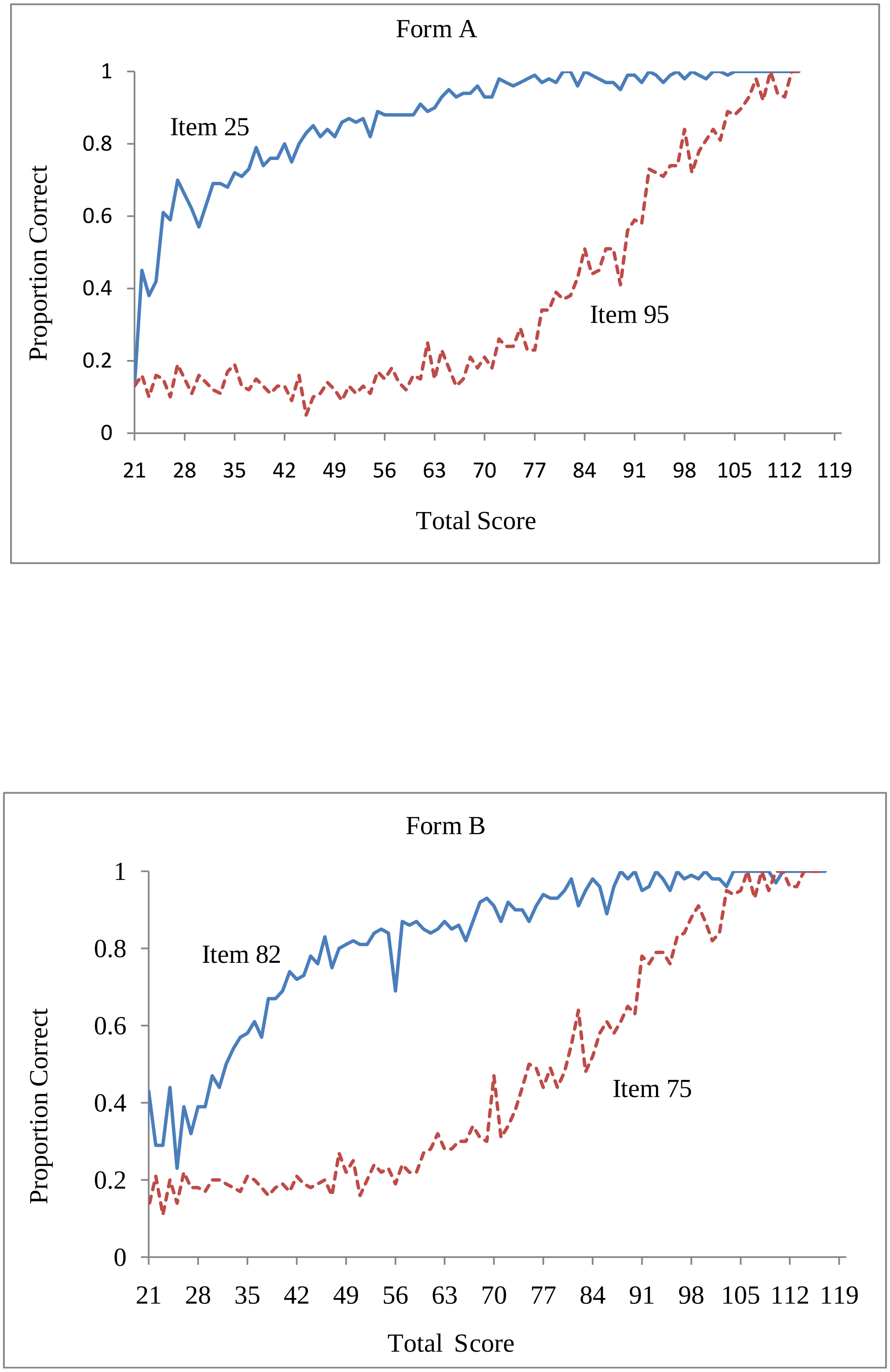


Figure 17. Total Score by Proportion of Correct Responses for Two Items from Forms A and B

To summarize the results for all items on Forms A and B, Table 19 provides the frequency of the average proportion of correct responses for low-ability examinees. To provide a useful range across all items, the analysis was restricted to examinees with total scores lower than 46 for Form A and lower than 48 for Form B. Because it was difficult to identify a constant proportion of correct responses for easier items, only items with overall item difficulty (*p*) equal to or less than .7 were included in Table 19.

Table 19. Average Proportion Correct (*p*) for Low-Ability Examinees on Items Where *p* ≤ .7

P	Form A	Form B
<.1	0	0
.1-.2	24	22
.2-.3	41	46
.3-.4	20	28
>.4	7	8
M	. 27	.28
SD	.09	.08
n of Items	92	104
N of Examinees	3040	3807

As shown in Table 19, the results were consistent with the random guessing model for multiple-choice items with four options. The overall pattern in the frequency distributions for Forms A and B was similar. The highest frequencies on both forms were in the range of .2 to .3. There were also several items in both forms with a value that was slightly below or above the average *c*-parameter (.25)—the value that would be expected under the random guessing model.

In addition to examining the plots of the proportion of correct responses by total scores, the distribution of the *c* parameters at the beginning (the first 10 items) of Forms A and B of the CEPA-English test was compared to those at the end (the last 10 items). If the two distributions are the same, then guessing behavior is not present. The frequency distribution of the *c* parameters at the beginning and end of each form are presented in Figures 18 and 19.

As indicated in Figure 18, it appeared that the two distributions generally differed in their skewness and kurtosis. For Form A, the frequency distribution of the *c* parameters at the beginning was almost normal with a value of -.022, but had a flatter distribution, with a kurtosis of -.918. The frequency distribution of the *c* parameters at the end of the test was positively skewed with a value of .593 and peaked with a kurtosis of .646. These findings suggested that guessing behavior did exist.

For Form B, the frequency distribution of the ***c*** parameters at the beginning of the test was slightly negatively skewed with a value of -.293 and had a flatter distribution with a kurtosis of -.679. The *c* parameters at the end had a normal distribution with a value of .005, but had a flatter distribution, with a kurtosis of -.940 (see Figure 19). Although the two distributions were generally similar in kurtosis, they differed in skewness, suggesting that guessing behavior did exist. Thus, the comparison of the distributions of the *c* parameters in both forms corroborated with the results from plotting the proportion of correct responses by total scores, which provided additional evidence that guessing occurred.

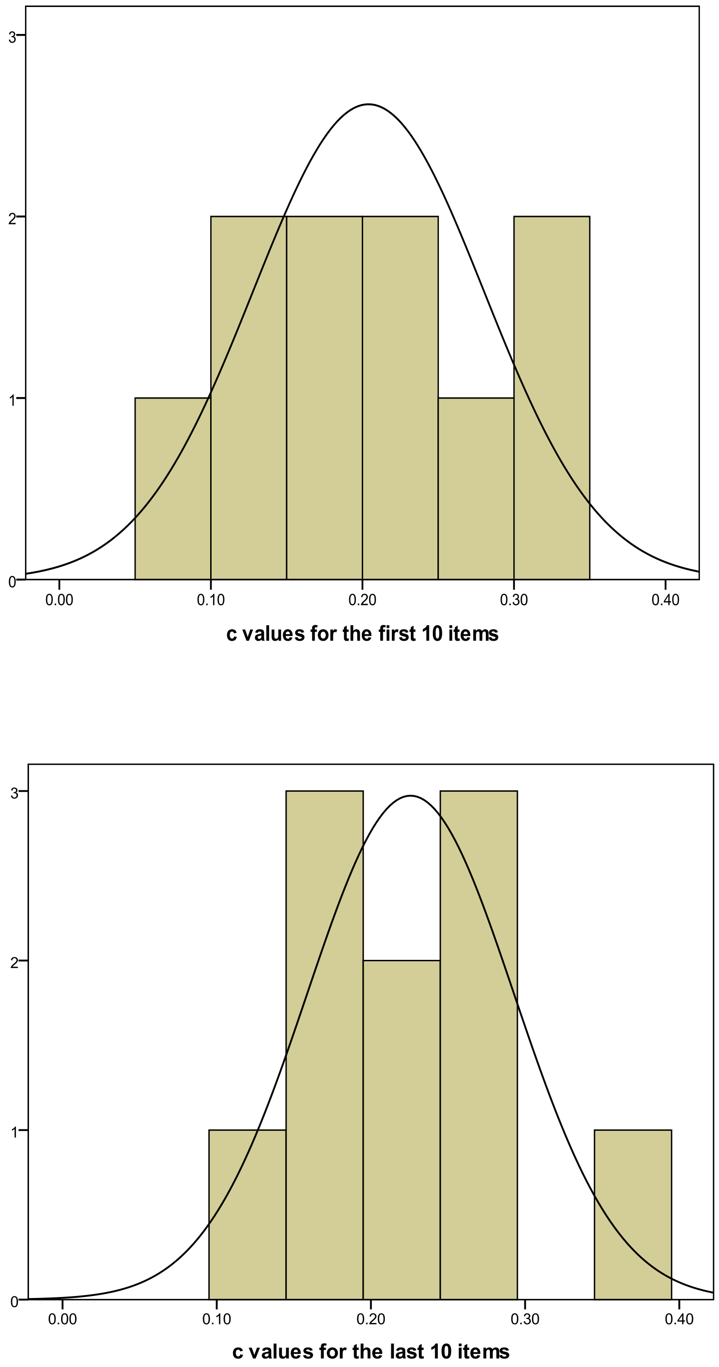


Figure 18. Frequency Distributions of the *c* Parameters at the Beginning and End of Form A

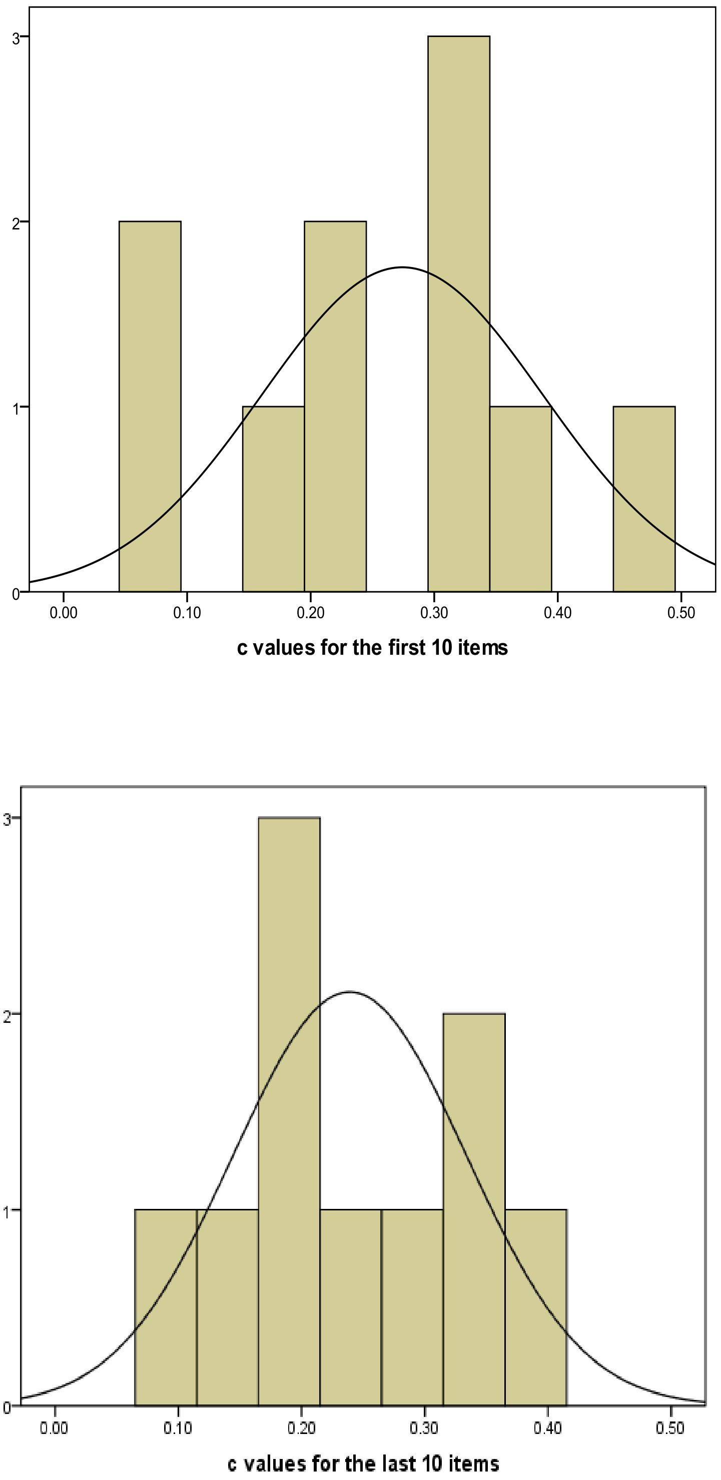


Figure 19. Frequency Distributions of the *c* Parameters at the Beginning and End of Form B

4.1.2.3 Assessing the Preferred IRT Model Fit at Item Level

To examine the fit of the 3PL IRT model to each item on each form of the CEPA-English test data, the *S-X*^{2} statistics were computed using the IRTFIT SAS macro. The test items that misfit the 3PL model on Forms A and B is provided in Tables 21 and 22. Additionally, Figures D1 and D2 in Appendix D show the plots of residuals for misfit items obtained with the 3PL model for Forms A and B, respectively.

The results of the *S-X*^{2} analyses showed that among the 116 items of Form A, 110 items misfit the 1PL model, 68 items misfit the 2PL model, and 23 items misfit the 3PL model (p <.05) (see Tables 20 and 21). Among the 119 items of Form B, 114 items misfit the 1PL model, 88 items misfit the 2PL model, and 25 items misfit the 3PL model (p <.05) (see Tables 20 and 22). According to Table 20, for both forms, the 3PL model had the smallest number of misfitting items, and consequently, the highest percentage of items that fit the model.

Table 20. Number of Misfitting Items Identified by the 1PL, 2PL, and 3PL Models for Forms A and B

Test Form	No. of Item	IRT Models		
		1PL	2PL	3PL
Form A	116	110	68	23
Form B	119	114	88	25

In addition to using the $S\text{-}X^2$ goodness-of-fit statistic, the fit of each CEPA-English item on Forms A and B to the 3PL model was also assessed using a graphical display of theoretical versus empirical score distributions via the ResidPlots-2 program. The residuals were computed in 15 equally spaced ability categories between -3 and +3. Figures D1 and D2 in Appendix D clearly showed that the theoretical and empirical ICC's differed at the lower end of ability, indicating item misfit. Hence, the plots of residuals for misfit items obtained with the 3PL model were consistent with the $S\text{-}X^2$ statistical results. Based on both the $S\text{-}X^2$ analyses and residuals analyses, the 3PL model was a better fit for each item on Forms A and B of the CEPA-English test data.

Table 21. Misfit Items Identified by the 3PL Model on Form A

Item No.	*df*	$S\text{-}X^2$	*P*
6	87	138.94	.0003
8	91	206.72	.0000
18	89	123.44	.0092
23	89	138.55	.0006
28	88	151.00	.0000
42	89	129.81	.0031
46	90	118.39	.0241
51	91	307.14	.0000
58	87	137.54	.0005
74	92	143.10	.0005
75	90	117.60	.0270
79	92	171.60	.0000
82	93	203.78	.0000
87	87	110.74	.0439
88	79	143.44	.0000
93	88	121.82	.0099
94	88	201.40	.0000
101	83	152.45	.0000
102	91	353.26	.0000
107	91	146.57	.0002
109	91	180.35	.0000
110	90	137.90	.0009
116	90	133.06	.0022

Table 22. Misfit Items Identified by the 3PL Model on Form B

Item No.	*df*	$S\text{-}X^2$	*P*
2	92	118.34	.0336
3	93	117.43	.0443
5	82	112.26	.0149
8	90	140.07	.0006
10	87	112.44	.0346
14	92	122.50	.0184
16	82	106.15	.0377
20	89	211.02	.0000
21	91	160.07	.0000
25	84	116.93	.0102
27	85	131.92	.0008
35	93	209.94	.0000
59	93	196.77	.0000
66	92	116.24	.0447
67	87	114.06	.0274
69	93	127.61	.0100
73	92	157.51	.0000
79	93	275.03	.0000
86	92	147.88	.0002
91	89	113.90	.0388
92	89	126.80	.0053
102	92	219.50	.0000
106	93	131.90	.0050
111	93	127.39	.0104
116	88	156.09	.0000

4.1.2.4 Examining the Invariance of Item Parameter Estimates

To examine the degree to which the invariance of the item parameters of the unidimensional 3PL IRT model held for Forms A and B of the CEPA-English test, examinees were subdivided into two groups: Group 1 represented low ability examinees where 75% had a theta level < 0 and 25% > 0, while Group 2 represented high ability examinees where 75% had a theta level > 0 and 25% < 0. After creating the two groups, the -2log likelihood statistics between the restricted and unrestricted models were examined. In the restricted model, the item parameter (*a*, *b*, and *c*) estimates on each form were set to be equal across the two groups. In the unrestricted model, the item parameters estimates on each form were free to vary across the two groups. The results of comparing the restricted and unrestricted models in Forms A and B are presented in Tables 23.

Table 23. Comparison of the Restricted versus the Unrestricted Models

Test Form	–2 log of restricted	*df*	–2 log of unrestricted	*df*	G^2	*df*	*P*
Form A	798343.2	300	797526.4	600	816.8	300	<.05
Form B	817241.0	300	817222.4	600	18.6	300	<.05

Note. - 2log = -2log likelihood statistic.

According to Table 23, the -2log likelihood value for the 100 random items of Form A for the restricted and unrestricted models was 798343.2 (with 300 parameters) and 797526.4 (with 600 parameters), respectively. The difference in chi-square, or G^2, between the two models was 816.8, which was greater than its critical value of 124.3 for p=.05 with 300 degrees of freedom. This indicated that there was a significant difference between the two models, which suggested that item parameter estimates were not invariant on Form A.

Different findings were observed on Form B using 100 randomly selected items. The -2log likelihood value for the items of Form B for the restricted and unrestricted models was 817241.0 (with 300 parameters) and 817222.4 (with 600 parameters), respectively. The G^2 observed value between the two models was 18.6, which was less than its critical value of 124.3 for p=.05 with 300 degrees of freedom. This indicated that there was no significant difference between the two models, suggesting that item parameter estimates were invariant on Form B.

To further explore invariance of item parameter estimates from the 3PL model, correlations and plots of parameter estimates on the unrestricted model for low and high ability groups were examined. Figures 20-22 display the scatterplots of *b*'s, *a*'s, and *c*'s for Form A.

According to Figure 20, the correlation of the *b* parameter estimates between Group 1 and Group 2 in both forms was strong (r =.973, $p < 0.01$ for Form A and r =.961, $p < 0.01$ for Form B). The scatterplots of all *b*'s also showed a linear relationship, suggesting that the *b* parameter estimates were invariant across the groups in both forms.

The correlation of the *a* parameters between the two groups was moderate in both forms (r = .845, $p < 0.01$ for Form A and r =.839, $p < 0.01$ for Form B). The relationship between the *a* parameter estimates for the two groups was quite linear in both forms, suggesting that the *a* parameter estimates were invariant across the groups in both forms (see Figure 21*).*

The correlation of the *c* parameters between the two groups was relatively moderate in both forms (r = .673, $p < 0.01$ for Form A and r =.704, $p < 0.01$ for Form B), but was less than .80, which was suggested by Wright (1968) as a minimal correlation to indicate that the property of invariance exists to some degree. Furthermore, the plots showed that the *c* parameter estimates were not the same for low and high ability groups, suggesting that the estimates were not invariant across the groups in both forms (see Figure 22*)*.

Thus, on the basis of the results of examining the correlations and plots of parameter estimates, it was reasonable to conclude that the assumption of item parameter invariance was not met for Form A primarily due to the estimation of the *c* parameters. The results of examining the -2log likelihood statistics further indicated that the assumption of item parameter invariance was not supported for Form A. While the results of examining the -2log likelihood statistic supported the assumption of item parameter invariance for Form B, some of the *c* parameter estimates were not very stable across the two groups.

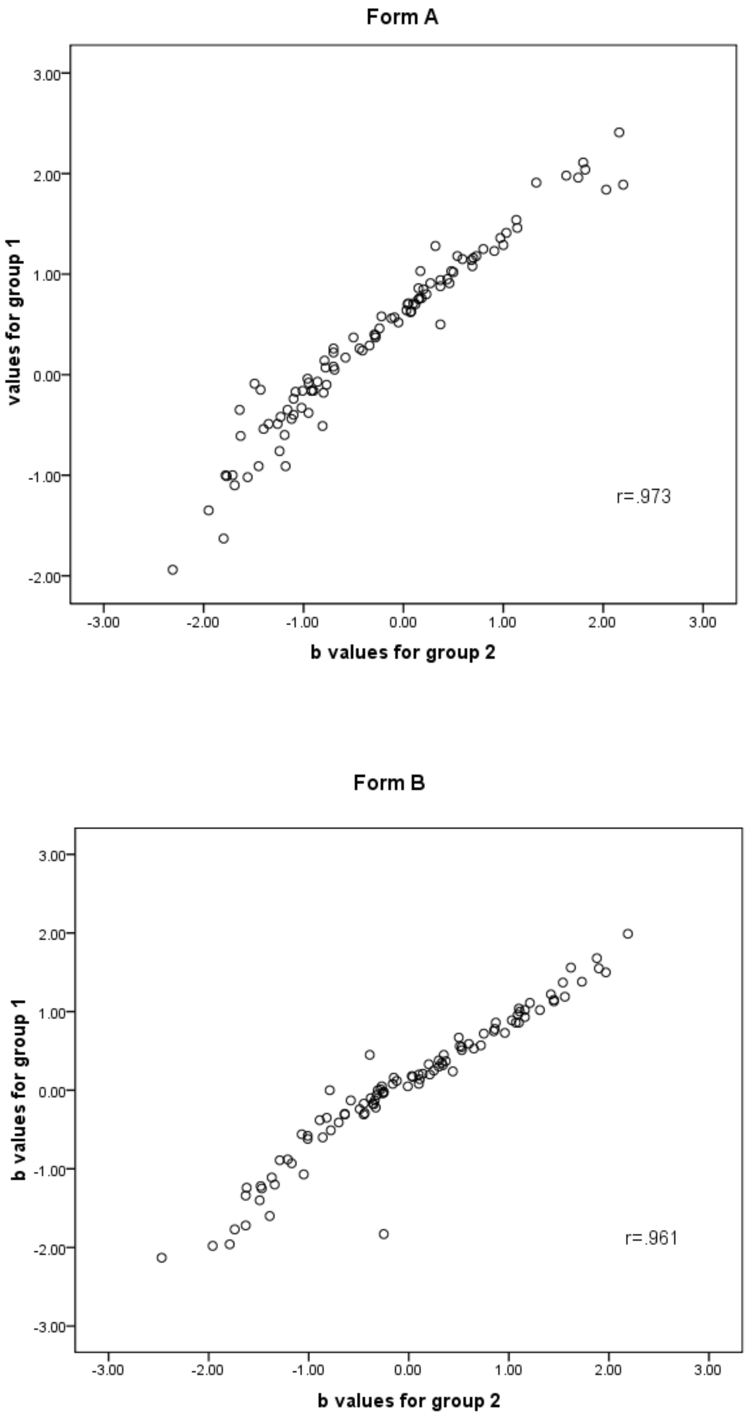
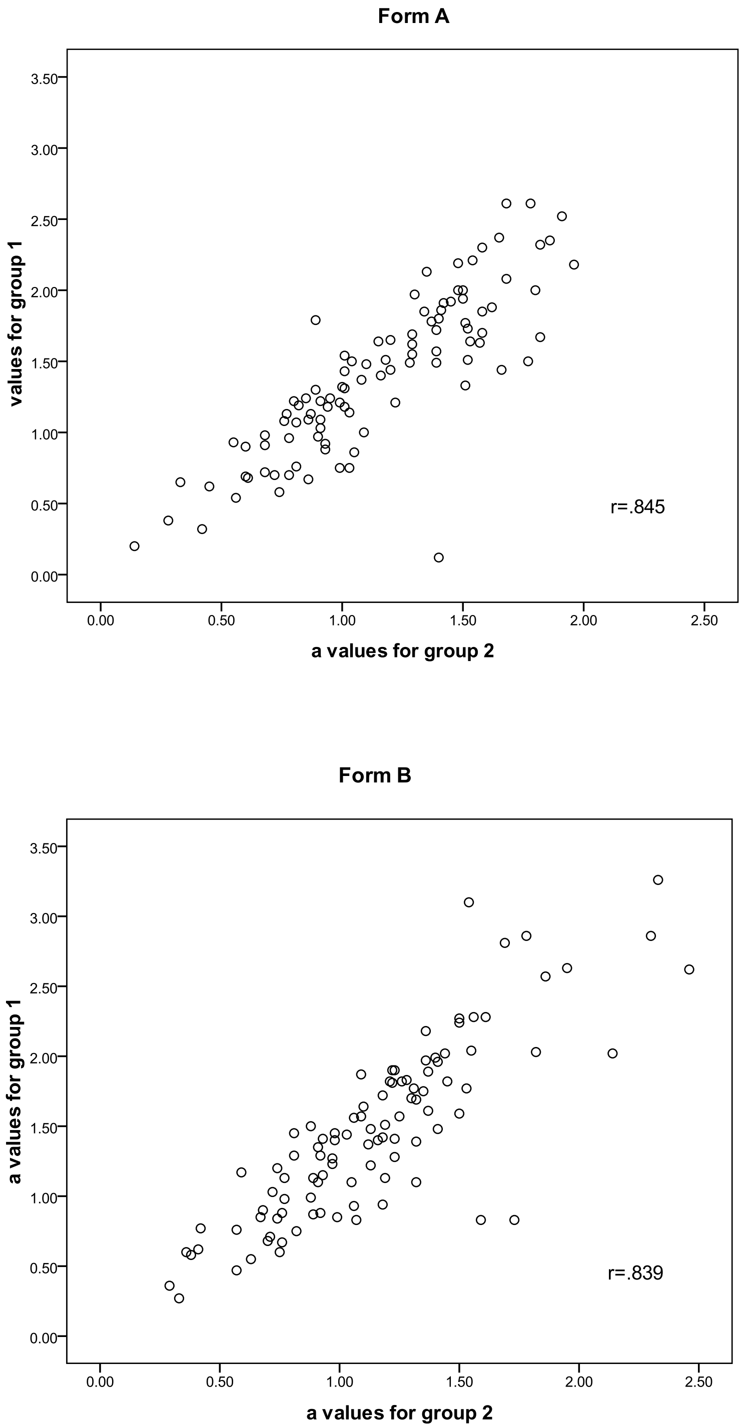


Figure 20. Scatterplots of *b*'s for Forms A and B


Form A

Form B

Figure 21. Scatterplots of *a*’s for Forms A and B

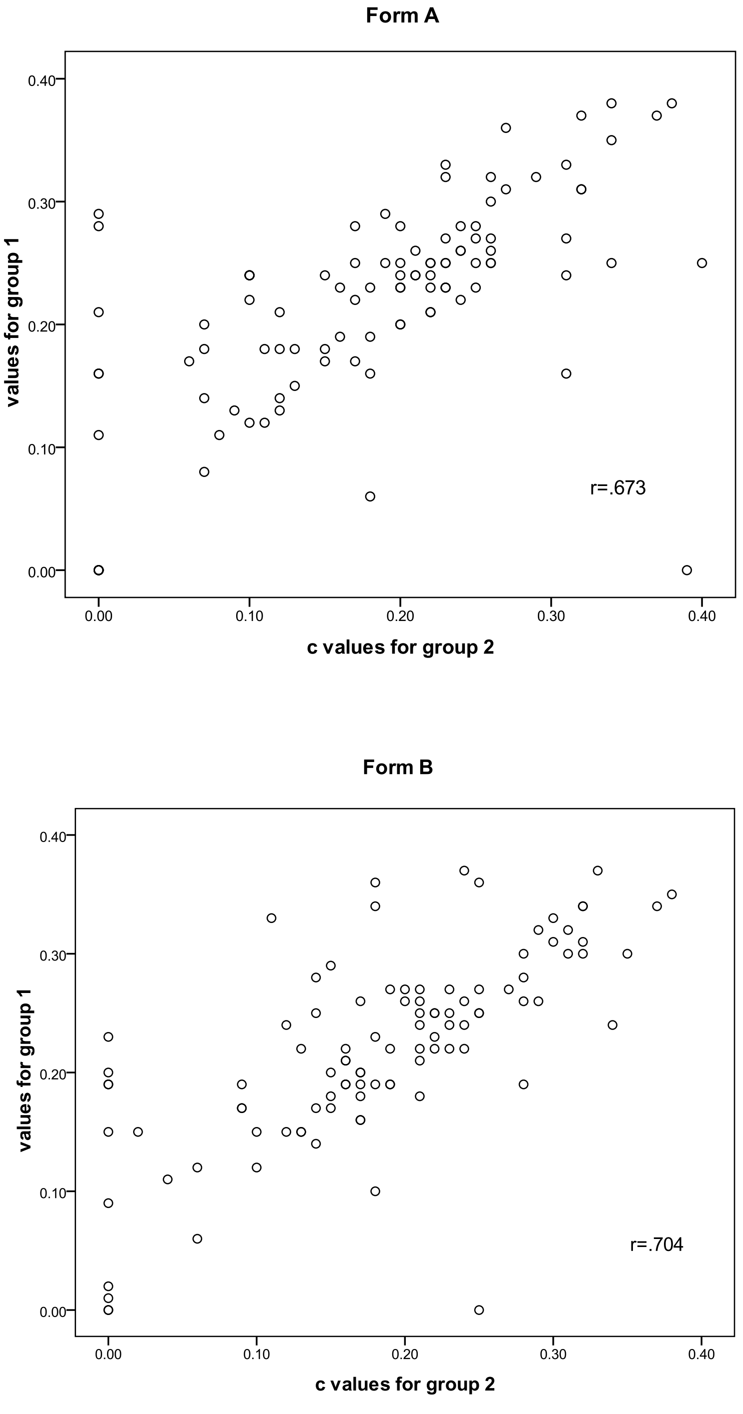


Figure 22. Scatterplots of *c*'s for Forms A and B

4.1.3 Detecting DIF using the Mantel-Haenszel Procedure

A two-stage MH process was performed, using the EZDIF program (Waller,1998), to examine whether each item on Forms A and B of the CEPA-English test exhibits uniform DIF between males and females, between study types (i.e., Arts and Sciences), and between school types (i.e., public, private, and home-schooled). For gender, male was the focal group, and for study type, Arts stream was the focal group. When comparing private and public schools, public was the focal group, and when comparing private and home-schooled, home-schooled was the focal group. When comparing public and home-schooled, home-schooled was the focal group. The focal group was coded as "1" and the reference group was coded as "0" (i.e., female, Sciences stream, and private schools). For each test form, the total scores were used as the matching variable, and scores were matched at 10 conditioning levels (0-10).

4.1.3.1 Descriptive Statistics for Forms A and B

In performing MH DIF analyses in Forms A and B of the CEPA-English test, a random sample of 2000 subjects with equal numbers of male and female subjects as well as equal number of students majoring in Arts or Sciences was used. Another random sample included 1400 subjects with equal numbers of students in public schools, private schools, and home-schooled. Descriptive statistics of the total scores for Forms A and B by gender, study type, and school type are provided in Tables 24 and 25.

The descriptive statistics indicated that, overall, female students performed better (M = 65.414 for Form A and M = 61.479 for Form B) in the CEPA-English test than males (M = 59.812 for Form A and M = 55.921 for Form B) and that students in the Sciences stream performed better (*M* =77.846 for Form A and *M* = 75.401 for Form B) than those in the Arts stream (*M* =52.659 for Form A and *M* = 50.464 for Form B). The descriptive statistics also indicated that students in private schools (*M* =79.713 for Form A and *M* = 77.720 for Form B) outperformed both those in the public schools (*M* =63.860 for Form A and *M* = 59.430 for Form B) and those who were home-schooled (*M* =42.573 for Form A and *M* = 42.205 for Form B) (see Tables 24 and 25).

Table 24. Descriptive Statistics of Form A by Gender, Study Type, and School Types

		N	*M*	*SD*	*Skewness*	*Kurtosis*
Gender	Females	1000	65.414	23.749	.218	-1.080
	Males	1000	59.812	23.947	.357	-1.006
Study Type	Arts	1000	52.659	18.719	.724	-.136
	Sciences	1000	77.846	21.258	-.454	-.564
School Types	Public	700	63.860	22.415	.205	-1.060
	Private	700	79.713	24.464	-.530	-.857
	Home-s	700	42.573	17.530	1.330	1.362

Note. Home-s = home-schooled.

Table 25. Descriptive Statistics of Form B by Gender, Study Type, and School Types

		N	*M*	*SD*	*Skewness*	*Kurtosis*
Gender	Females	1000	61.479	23.151	.461	-.779
	Males	1000	55.921	22.491	.640	-.469
Study Type	Arts	1000	50.464	18.873	-.137	-.755
	Sciences	1000	75.401	22.065	.967	.510
	Public	700	59.430	22.526	534	-.694
School Types	Private	700	77.720	25.979	-.287	-1.062
	Home-s	700	42.205	17.360	1.635	2.341

Note. Home-s = home-schooled.

4.1.3.2 DIF Analyses for Form A

4.1.3.2.1 DIF Analyses between Males and Females

In the first stage of MH DIF analyses between males and females, only Items 81 and 109 were flagged as *C* items, exhibiting a large DIF. To obtain a pure matching variable, these items were excluded from the total score in the second stage of DIF analyses. In the second stage, these items were still classified as type *C* DIF items.

Item 81 favored females (the reference group), while Item 109 favored males. The odds ratio, or α̂, values for these items were 2.19 and .47, respectively, where α larger than 1.00 favors the reference group and α less than 1.00 favors the focal group. Item 81 had a significant $MH\ \chi^2$ value of 29.35 (p < .001) and a $\hat{\Delta}_{MH}$ of -1.84 with a standard error of .339. Item 109 had a significant $MH\ \chi^2$ value of 28.43 (p < .001) and a $\hat{\Delta}_{MH}$ of 1.76 with a standard error of .330. Item 81 is a vocabulary item, measuring the knowledge of the most frequent words taken from the family of headwords in the first 2000 most frequent words, while Item 109 is a reading item (see Table 26).

Additionally, in the second stage, Items 8, 26, 31, 88, 113, 114, 115, and 116 were classified as *B* items, showing moderate DIF. Among the eight items showing a moderate DIF, five items (26, 31, 88, 113, and 115) favored females (the reference group). Three items (8, 114, and 116) favored males (the focal group). The α̂ for these three items were .64, .47, and .59, respectively. In addition, Item 8 had a significant $MH\ \chi^2$ value of 7.01(p =.008) and a $\hat{\Delta}_{MH}$ of 1.04 with a standard error of .383. Item 114 had a significant $MH\ \chi^2$ value of 14.82 (p < .001) and a $\hat{\Delta}_{MH}$ of 1.05 with a standard error of .270. Item 116 had a significant $MH\ \chi^2$ value of 22.09 (p < .001) and a $\hat{\Delta}_{MH}$ of 1.24 with a standard error of .262 (see Table 26).

Table 26. Final Result of the MH Analysis for Moderate and Large DIF Items on Form A between Males and Females

Item	Content	Label	Favored Group	$\hat{\alpha}_{MH}$	$MH\ \chi^2$	*P*	$\hat{\Delta}_{MH}$	$SE(\hat{\Delta}_{MH})$
8	Gram: Intensifiers	BF	Males	.64	7.01	.008	1.04	.383
26	Gram: Subordinating Conjunctions	BR	Females	1.83	15.13	.000	-1.42	.361
31	Gram: Coordinating Conjunctions	BR	Females	1.58	9.30	.002	-1.08	.348
81	Voc: First 2000 most frequent words	**CR**	Females	2.19	29.35	.000	-1.84	.339
88	Voc: First 1200 most frequent words	BR	Females	1.64	15.59	.000	-1.17	.293
109	Reading NA	**CF**	Males	.47	28.43	.000	1.76	.330
113	Reading NA	BR	Females	1.60	19.13	.000	-1.11	.250
114	Reading NA	BF	Males	.64	14.82	.000	1.05	.270
115	Reading NA	BR	Females	1.63	16.64	.000	-1.15	.279
116	Reading NA	BF	Male	.59	22.09	.000	1.24	.262

Note. The third column shows the degree of DIF detected. *B* = moderate DIF; *C* = large DIF; *CR/BR* = items favor a reference group; *CF/BF* = items favor a focal group; Gram = Grammar; Voc = Vocabulary; NA=No information available regarding what item measures.

4.1.3.2.2 DIF Analyses between Arts and Science Streams

The results of the first and the second stage of MH DIF analyses between Arts and Sciences streams showed that no item was flagged as a *C* item, exhibiting a large DIF. However, in the second stage, five items (1, 3, 38, 39, and 65) were classified as *B* items, showing a moderate DIF (see Table 27).

Among the five items showing moderate DIF, three items (1, 38, and 39) favored the Sciences stream (the reference group), while two items (3 and 65) favored the Arts stream (the focal group). The $\hat{\alpha}$ values for Items 3 and 65 were .63 and .65, respectively. In addition, Item 3 had a significant $MH\ \chi^2$ value of 14.10 (p < .001) and a $\hat{\Delta}_{MH}$ of 1.09 with a standard error of .290. Item 65 had a significant $MH\ \chi^2$ value of 0.65 (p = .006) and a $\hat{\Delta}_{MH}$ of 1.00 with a standard error of .351(see Table 27). Item 3 is a grammar item, measuring agreement, while Item 65 is a vocabulary item testing the knowledge of the most common English vocabulary words taken from the first 2000 most frequent words.

Table 27. Final Result of the MH Analysis for Moderate and Large DIF Items on Form A between Arts and Sciences Streams

Item	Content	Label	Favored Group	$\hat{\alpha}_{MH}$	$MH\chi^2$	*P*	$\hat{\Delta}_{MH}$	$SE(\hat{\Delta}_{MH})$
1	Gram: Comparatives and superlatives	BR	Sciences	1.59	14.71	.000	-1.09	.279
3	Gram: Agreement	BF	Arts	.63	14.10	.000	1.09	.290
38	Gram: Comparatives and superlatives	BR	Sciences	1.66	13.24	.000	-1.20	.324
39	Gram: Word order	BR	Sciences	1.74	5.52	.019	-1.31	.532
65	Voc: First 2000 most frequent words	BF	Arts	.65	7.66	.006	1.00	.351

Note. The third column shows the degree of DIF detected. *B* = moderate DIF; *BR* = items favor a reference group; *BF* = items favor a focal group; Gram = Grammar; Voc = Vocabulary.

4.1.3.2.3 DIF Analyses between Private Schools and Home-Schooled

In the first stage of MH DIF analyses between private schools and home-schooled, no item was flagged as a *C* item, exhibiting a large DIF. However, in the second stage, fourteen items (1, 8, 12, 18, 20, 26, 32, 47, 53, 59, 67, 69, 79, and 115) were classified as *B* items, showing a moderate DIF. Among the fourteen items showing moderate DIF, five items (1, 18, 26, 32, and 115) favored private schools (the reference group), while nine items (8, 12, 20, 47, 53, 59, 67, 69, and 79) favored home-schooled (the focal group) (see Table 28).

Items 1, 8, 12, 18, 20, 26, and 32 are grammar items measuring comparatives and superlatives, intensifiers, pronouns 2 (relative pronouns), quantifiers, modals, subordinating conjunctions, and coordinating conjunctions, respectively. Item 47 is a part-of-speech item, measuring the students' knowledge of English word forms taken from the first 2000 most frequent words. Items 53, 59, 67, 69, and 79 are vocabulary items, testing the knowledge of the most common English vocabulary words, while Item 115 is a reading item.

Table 28. Final Result of the MH Analysis for Moderate and Large DIF Items on Form A between Private Schools and Home-schooled

Item	Content	Label	Favored Group	$\hat{\alpha}_{MH}$	$MH\ \chi^2$	P	$\hat{\Delta}_{MH}$	$SE(\hat{\Delta}_{MH})$
1	Gram: Comparatives and superlatives	BR	Private	1.616	9.33	.002	-1.128	.358
8	Gram: Intensifiers	BF	Home-s	.609	4.33	.038	1.166	.547
12	Gram: Pronouns 2 (Relative pronouns)	BF	Home-s	1.664	10.87	.001	-1.197	.359
18	Gram: Quantifiers	BR	Private	1.591	4.68	.031	-1.091	.470
20	Gram: Modals	BF	Home-s	.616	7.05	.008	1.140	.418
26	Gram: Subordinating conjunctions	BR	Private	1.680	6.28	.012	-1.220	.477
32	Gram: Coordinating Conjunctions	BR	Private	2.017	20.23	.000	-1.648	.360
47	Pos: First 2000 most frequent words	BF	Home-s	.624	5.74	.017	1.107	.440
53	Voc: First 900 most frequent words	BF	Home-s	.598	8.21	.004	1.207	.414
59	Voc: First 1500 most frequent words	BF	Home-s	.594	9.73	.002	1.226	.387
67	Voc: First 1600 most frequent words	BF	Home-s	.595	8.55	.003	1.218	.409
69	Voc: Words taken from the second AWL	BF	Home-s	.599	6.52	.011	1.205	.455
79	Voc: First 1000 most frequent words	BF	Home-s	.621	9.27	.002	1.120	.362
115	Reading NA	BR	Private	1.721	9.69	.002	-1.275	.400

Note. The third column shows the degree of DIF detected. *B* = moderate DIF; *BR* = items favor a reference group; *BF* = items favor a focal group; Gram = Grammar; Pos = part-of-speech items; Voc = Vocabulary; NA=No information available regarding what item measures; Home-s = home-schooled.

4.1.3.2.3.4 DIF Analyses between Private and Public Schools

In the first stage of MH DIF analyses between private and public schools, only Items 47 and 58 were flagged as *C* items, exhibiting a large DIF. To obtain a pure matching variable, these items were excluded from the total score in the second stage of DIF analysis. In the second stage, these items were still classified as type *C* DIF items. Items 47 and 58 favored public, and the α̂ values for these items were .44 and .50, respectively. Item 47 had a significant $MH\ \chi^2$ value of 25.71 ($p < .001$) and a $\hat{\Delta}_{MH}$ of 1.93 with a standard error of .377. Item 58 had a significant $MH\ \chi^2$ value of 26.30 ($p < .001$) and a $\hat{\Delta}_{MH}$ of 1.65 with a standard error of .326 (see Table 29).

Item 47 is a part-of-speech item, measuring the students' knowledge of English word forms taken from the first 2000 most frequent words. Item 58 is a vocabulary item, measuring the knowledge of the most frequent words taken from the family of headwords in the first 1000 most frequent words.

Furthermore, in the second stage, six items (7, 19, 49, 59, 71, and 84) were classified as *B* items, showing a moderate DIF. Among the six showing moderate DIF, four items (19, 49, 59, and 84) favored public schools (the focal group), and two items (7 and 71) favored private schools (the reference group). The α for Items 7 and 71 were 1.73 and 2.00, respectively. In addition, Item 19 had a significant $MH\ \chi^2$ value of 9.94 (p=0.002) and a $\hat{\Delta}_{MH}$ of -1.28 with a standard error of .394. Item 71 had a significant $MH\ \chi^2$ value of 13.55 ($p < .001$) and a $\hat{\Delta}_{MH}$ of -1.63 with a standard error of .433 (see Table 29). Item 7 is a grammar item, testing pronouns 1, while Item 71 is a vocabulary item, measuring the knowledge of the most common English vocabulary words taken from the most 1400 frequent words.

Table 29. Final result of the MH Analysis for Moderate and Large DIF Items on Form A between Private and Public Schools

Item	Content	Label	Favored Group	$\hat{\alpha}_{MH}$	$MH\ \chi^2$	*P*	$\hat{\Delta}_{MH}$	$SE(\hat{\Delta}_{MH})$
7	Gram: Pronouns 1	BR	Private	1.73	9.94	.002	-1.28	.394
19	Gram: Verb forms 2 (infinitives/gerunds, etc)	BF	Public	.61	13.04	.000	1.17	.321
47	Pos: First 2000 most frequent words	**CF**	Public	.44	25.71	.000	1.93	.377
49	Pos: First 1000 most frequent words	BF	Public	.65	10.56	.001	1.01	.304
58	Voc: First 1000 most frequent words	**CF**	Public	.50	26.30	.000	1.65	.326
59	Voc: First 1500 most frequent words	BF	Public	.64	11.13	.001	1.04	.306
71	Voc: First 1400 most frequent words	BR	Private	2.00	13.55	.000	-1.63	.433
84	Voc: First 2000 most frequent words	BF	Public	.63	6.07	.014	1.08	.424

Note. The third column shows the degree of DIF detected. *B* = moderate DIF; *C* = large DIF; *BR* = items favor a reference group; *CF/BF* = items favor a focal group; Gram = Grammar; Pos = part-of-speech items; Voc = Vocabulary.

4.1.3.2.3.5 DIF Analyses between Public and Home-Schooled

The results of the first MH DIF analyses between public schools and home-schooled showed that three items (1, 71, and 88) were flagged as *C* item, exhibiting a large DIF. To obtain a pure matching variable, these items were excluded from the total score in the second stage of DIF analysis. In the second stage, these items were still classified as type *C* DIF items,

Items 1 and 88 favored public schools (the reference group), while Item71 favored home-schooled (the focal group). The αˆ values for Items 1, 71, and 88 were 2.24, .40, and 2.29, respectively. Item 1 had a significant *MH* χ^2 value of 39.22 (p < .001) and a $\Delta\hat{}_{MH}$ of -1.90 with a standard error of .302, while Item 71 had a significant *MH* χ^2 value of 28.69 (p < .001) and a $\Delta\hat{}_{MH}$ of 2.13 with a standard error of .403. Item 88 had a significant *MH* χ^2 value of 30.61 (p < .001) and a $\Delta\hat{}_{MH}$ of -1.95 with a standard error of .351 (see Table 30).

Item 1 is a grammar item, measuring comparatives and superlatives. Items 71 and 88 are vocabulary items, measuring the knowledge of the most common English vocabulary words taken from the most 1400 and 1200 frequent words, respectively.

Furthermore, in the second stage, seven items (8, 19, 20, 26, 32, 103, and 115) were classified as *B* items, showing a moderate DIF. Among the seven items showing moderate DIF, five items (19, 26, 32,103 and115) favored public schools. Two items (8 and 20) favored home-schooled. The αˆ for Items 8 and 20 were .48 and .62. In addition, Item 8 had a significant *MH* χ^2 value of 11.94 (p = .001) and a $\Delta\hat{}_{MH}$ of 1.75 with a standard error of .498. Item 20 had a significant *MH* χ^2 value of 10.36 (p = .001) and a $\Delta\hat{}_{MH}$ of 1.13 with a standard error of .346 (see Table 30). Items 8 and 20 are grammar items, measuring intensifiers and modals, respectively.

Table 30. Final Result of the MH Analysis for Moderate and Large DIF Items on Form A between Public Schools and Home-schooling

Item	Content	Label	Favored Group	$\hat{\alpha}_{MH}$	$MH\ \chi^2$	P	$\hat{\Delta}_{MH}$	$SE(\hat{\Delta}_{MH})$
1	Gram: Comparatives and superlatives	**CR**	Public	2.24	39.22	.000	-1.90	.302
8	Gram: Intensifiers	BF	Home-s	.48	11.94	.001	1.75	.498
19	Gram: Verb forms 2 (infinitives/gerunds, etc)	BR	Public	1.65	14.80	.000	-1.18	.303
20	Gram: Modals	BF	Home-s	.62	10.36	.001	1.13	.346
26	Gram: Subordinating conjunctions	BR	Public	1.67	9.96	.002	-1.20	.377
32	Gram: Coordinating Conjunctions	BR	Public	1.77	17.86	.000	-1.34	.311
71	Voc: 1400 frequent most words	**CF**	Home-s	.40	28.69	.000	2.13	.403
88	Voc: 1200 frequent most words	**CR**	Public	2.29	30.61	.000	-1.95	.351
103	Reading NA	BR	Public	1.75	16.98	.000	-1.31	.315
115	Reading NA	BR	Public	2.01	24.09	.000	-1.64	.329

Note. The third column shows the degree of DIF detected. *B* = moderate DIF; *C* = large DIF; *CR/BR* = items favor a reference group; *CF/BF* = items favor a focal group; Gram = Grammar; Voc = Vocabulary; NA=No information available regarding what item measures; Home-s = home-schooled.

4.1.3.3 DIF Analyses for Form B

4.1.3.3.1 DIF Analyses between Males and Females

For Form B, the first stage of MH DIF analyses between males and females reveled that only Item 32 was flagged as a *C* item, exhibiting a large DIF. To obtain a pure matching variable, this item was excluded from the total score in the second stage of DIF analysis. In the second stage, Item 32 was still classified as type *C* DIF item. Item 32 favored females, where α was 2.158. This item had a significant *MH* χ^2 value of 33.304 (p < .001) and a $\Delta\hat{}_{MH}$ of -1.808 with a standard error of .313. Item 32 is a grammar item, measuring verb forms 2 (infinitives/ gerunds, etc) (see Table 31).

Additionally, in the second stage, Items 21, 29, 40, 50, 58, 64, 68, 72, 82, and 107 were classified as *B* items, showing a moderate DIF. Among the ten items showing moderate DIF, eight items (29, 50, 58, 64, 68, 72, 82, and 107) favored females (the reference group). Two items (21 and 40) favored males (the focal group). The $\alpha\hat{}$ for Items 21 and 40 were .571 and .631, respectively. In addition, Item 21 had a significant *MH* χ^2 value of 21.496 (p < .001) and a $\Delta\hat{}_{MH}$ of 1.318 with a standard error of .283. Item 40 had a significant *MH* χ^2 value of 16.328 (p < .001) and a $\Delta\hat{}_{MH}$ of 1.082 with a standard error of .265 (see Table 31). Items 21 and 40 are grammar items, measuring verb forms 1 (tense/aspect/voice) and coordinating conjunctions, respectively.

Table 31. Final Result of the MH Analysis for Moderate and Large DIF Items on Form B between Males and Females

Item	Content	Label	Favored Group	$\hat{\alpha}_{MH}$	$MH\ \chi^2$	P	$\hat{\Delta}_{MH}$	$SE(\hat{\Delta}_{MH})$
21	Gram: Verb forms 1 (tense/aspect/voice)	BF	Males	.571	21.496	.000	1.318	.283
29	Gram: Subordinating conjunctions	BR	Females	1.546	12.409	.000	-1.023	.287
32	Gram: Verb forms 2 (infinitives/ gerunds, etc)	**CR**	Females	2.158	33.304	.000	-1.808	.313
40	Gram: Coordinating Conjunctions	BF	Males	.631	16.328	.000	1.082	.265
50	Pos: Second 2000 most frequent words	BR	Females	1.543	18.474	.000	-1.019	.236
58	Voc: 700 most frequent words	BR	Females	1.827	22.052	.000	-1.417	.300
64	Voc: Second 1000 most frequent words	BR	Females	1.570	11.203	.001	-1.060	.311
68	Voc: Second 1000 most frequent words	BR	Females	1.754	24.071	.000	-1.320	.270
72	Voc: Second 1000 most frequent words	BR	Females	1.648	13.378	.000	-1.174	.316
82	Reading NA	BR	Females	1.608	14.833	.000	-1.117	.286
107	Reading NA	BR	Females	1.552	21.158	.000	-1.033	.223

Note. The third column shows the degree of DIF detected. *B* = moderate DIF; *C* = large DIF; *CR/BR* = items favor a reference group; *BF* = items favor a focal group; Gram = Grammar; Pos = part-of-speech items; Voc = Vocabulary; NA=No information available regarding what item measures.

4.1.3.3.2 DIF Analyses between Arts and Science Streams

In the first stage of MH DIF analyses between Arts and Sciences streams, only one item (112) was flagged as a *C* item, exhibiting a large DIF. To obtain a pure matching variable, this item was excluded from the total score in the second stage of DIF analysis. In the second stage, Item 112 was still classified as type *C* DIF items (see Table 32).

Also, in the second stage, only on item (15) was classified as *B* items, showing a moderate DIF. Item 112 favored Sciences stream (the reference group), while Item 112 favored Arts stream (the focal group). The $\alpha\hat{}$ values for these items were 1.791 and 2.029, respectively. Items 15 had a significant $MH\,\chi^2$ value of 26.249 (p < .001) and a $\Delta\hat{}_{MH}$ of -1.369 with a standard error of .266. Items 112 had a significant $MH\,\chi^2$ value of 35.939 (p < .001) and a $\Delta\hat{}_{MH}$ of -1.663 with a standard error of .277 (see Table 32). Item 15 is a grammar item, which measures prepositions, while Item 112 is a reading item.

Table 32. Final Result of the MH Analysis for Moderate and Large DIF Items on Form B between Arts and Sciences Streams

Item	Content	Label	Favored Group	$\hat{\alpha}_{MH}$	$MH\ \chi^2$	P	$\hat{\Delta}_{MH}$	$SE(\hat{\Delta}_{MH})$
15	Gram: Prepositions	BF	Arts	1.791	26.249	.000	-1.369	.266
112	Reading NA	**CR**	Sciences	2.029	35.939	.000	-1.663	.277

Note. The third column shows the degree of DIF detected. *B* = moderate DIF; *C* = large DIF; *CR* = items favor a reference group; *BF* = items favor a focal group; Gram = Grammar; NA=No information available regarding what item measures.

4.1.3.3.3 DIF Analyses between Private and Home-Schooled

The results of the first and the second stage of MH DIF analyses between private schools and home-schooled showed that no item was flagged as *C* item, exhibiting a large DIF. However, in the second stage, twenty six items (5 ,10,14 ,16 ,21,26, 30, 32, 43, 47, 55, 59, 60, 62, 65, 66,77, 82, 83, 87, 97, 98, 99, 109, 113, and 117) were classified as *B* items, showing a moderate DIF. Among the twenty six items, fifteen items (5, 10, 14, 16, 26, 30, 32, 47, 62, 77, 82, 97, 98, 99, 117) favored private schools (the reference group), while eleven items (21, 43, 55, 59, 60, 65, 66, 83, 87, 109, and 113) favored home-schooled (the focal group) (see Table 33).

Items 5, 10, 14, 16, 21, 26, 30, and 32 are grammar items, measuring, intensifiers, pronouns 2 (relative pronouns), verb forms 2 (infinitives/gerunds, etc), conditionals, verb forms 1 (tense/aspect/voice), and verb forms 2 (infinitives/ gerunds, etc). Item 43 and 47 are part-of-speech items, measuring the students' knowledge of English word forms taken from the first 1000 and 2000 most frequent words. Items 55, 59, 60, 62, 65, 66, 77, 82, 83, and 87 are vocabulary items, testing the knowledge of the most common English vocabulary words, while Items 97, 98, 99, 109, 113, and 117 are reading items (see Table 33).

Table 33. Final Result of the MH Analysis for Moderate and Large DIF Items on Form A between Private Schools and Home-schooling

Item	Content	Label	Favored Group	$\hat{\alpha}_{MH}$	$MH\ \chi^2$	P	$\hat{\Delta}_{MH}$	$SE(\hat{\Delta}_{MH})$
5	Gram: Intensifiers	BR	Private	1.818	10.353	.001	-1.405	.429
10	Gram: Pronouns 2 (Relative pronouns)	BR	Private	1.564	5.977	.014	-1.051	.417
14	Gram: Verb forms 2 (infinitives/ gerunds, etc)	BR	Private	1.958	18.160	.000	-1.579	.370
16	Gram: Conditionals	BR	Private	1.649	9.840	.002	-1.176	.374
21	Gram: verb forms 1 (tense/aspect/voice)	BF	Home-s	.580	7.971	.005	1.280	.438
26	Gram: Verb forms 2 (infinitives/ gerunds, etc)	BR	Private	1.617	8.268	.004	-1.130	.386
30	Gram	BR	Private	1.539	6.572	.010	-1.013	.380
32	Gram: Verb forms 2 (infinitives/ gerunds, etc)	BR	Private	1.637	6.301	.012	-1.158	.447
43	First 1000 most frequent words	BF	Home-s	.562	9.881	.002	1.353	.423
47	First 2000 most frequent words	BR	Private	1.769	12.091	.001	-1.340	.374

Table 33 (continued)

Item	Content	Label	Favored Group	$\hat{\alpha}_{MH}$	$MH\ \chi^2$	P	$\hat{\Delta}_{MH}$	$SE(\hat{\Delta}_{MH})$
55	Voc: Words taken from the first AWL	BF	Home-s	.566	9.740	.002	1.338	.426
59	Voc NA	BF	Home-s	.623	7.447	.006	1.114	.410
60	Voc NA	BF	Home-s	.643	5.819	.016	1.039	.416
62	Voc: First 900 most frequent words	BR	Private	1.567	4.806	.028	-1.055	.456
65	Voc NA	BF	Home-s	.591	9.450	.002	1.237	.401
66	Voc NA	BF	Home-s	.517	15.109	.000	1.552	.403
77	Voc: First 1000 most frequent words	BR	Private	1.588	5.852	.016	-1.088	.423
82	Voc: Second 1000 most frequent words	BR	Private	1.651	6.471	.011	-1.179	.455
83	Voc: First 800 most frequent words	BF	Home-s	.548	7.594	.006	-1.179	.507
87	Voc: Words taken from the second AWL	BF	Home-s	.574	12.149	.000	1.414	.371
97	Reading NA	BR	Private	1.645	8.529	.003	1.303	.390
98	Reading NA	BR	Private	1.667	9.072	.003	-1.170	.390
99	Reading NA	BR	Private	2.082	20.710	.000	-1.200	.372
109	Reading NA	BF	Home-s	.530	13.053	.000	-1.723	.405
113	Reading NA	BF	Home-s	.489	17.480	.000	1.490	.404
117	Reading NA	BR	Private	1.880	16.339	.000	1.680	.366

Note. The third column shows the degree of DIF detected. *B* = moderate DIF; *BR* = items favor a reference group; *BF* = items favor a focal group; Gram = Grammar; Pos = part-of-speech items; Voc = Vocabulary; NA=No information available regarding what item measures; Home-s = home-schooled.

4.1.3.3.4 DIF Analyses between Private and Public Schools

In the first stage of MH DIF analyses between public and private schools, only Item 43 was flagged as a *C* item, exhibiting a large DIF. To obtain a pure matching variable, this item was excluded from the total score in the second stage of DIF analysis. In the second stage, Item 43 was still classified as type C DIF item. Item 43 was the only item favored public schools (the focal group) and had a α value of .423. This item had a significant $MH\ \chi^2$ value of 31.279 ($p <$ .001) and a $\Delta\hat{}_{MH}$ of 2.022 with a standard error of .368. Item 43 is a part-of-speech item, measuring the students' knowledge of English word forms taken from the first 1000 most frequent words (see Table 34).

Furthermore, in the second stage, seven items (5, 10, 35, 55, 82, 84, and 95) were classified as *B* items, showing a moderate DIF. Among the seven items showing moderate DIF, three items (5, 10, and 55) favored private schools (the reference group). Four items (35, 82, 84, and 95) favored public schools (the focal group). The values for these four items were .604, 1.831, .630, and .642, respectively. In addition, the values of $MH\ \chi^2$ for the four items were 13.525 ($p <$ .001), 9.813 ($p =$ .002), 8.831, ($p =$ .003), and 7.055 ($p =$ .008), respectively, with the $\Delta\hat{}_{MH}$ of 1.185, -1.421, 1.085, and 1.041, respectively. The standard errors of these items were .318, .443, .361, and .383, respectively (see Table 34). Of the four items that favored public schools, Items 35 is a grammar item, while Items 82 and 84 are vocabulary items, testing the knowledge of the most common English vocabulary words. Item 90 is a reading item.

Table 34. Final Result of the MH Analysis for Moderate and Large DIF Items on Form B between Private and Public Schools

Item	Content	Label	Favored Group	$\hat{\alpha}_{MH}$	$MH\ \chi^2$	P	$\hat{\Delta}_{MH}$	$SE(\hat{\Delta}_{MH})$
5	Gram: Articles and determiners	BR	Private	2.078	16.011	.000	-1.718	.425
10	Gram: Pronouns 2 (Relative pronouns)	BR	Private	1.607	8.288	.004	-1.115	.377
35	Gram NA	BF	Public	.604	13.525	.000	1.185	.318
43	Pos: First 1000 most frequent words	**CF**	Public	.423	31.279	.000	2.022	.368
55	Voc: Words taken from the second AWL	BR	Private	1.542	7.323	.007	-1.018	.364
82	Voc : Second 1000 most frequent words	BF	Public	1.831	9.813	.002	-1.421	.443
84	Voc NA	BF	Public	.630	8.831	.003	1.085	.361
95	Reading NA	BF	Public	.642	7.055	.008	1.041	.383

Note. The third column shows the degree of DIF detected. *B* = moderate DIF; *C* = large DIF; *BR* = items favor a reference group; *CF*/*BF* = items favor a focal group; Gram = Grammar; Pos = part-of-speech items; Voc = Vocabulary; NA=No information available regarding what item measures.

4.1.3.3.5 DIF Analyses between Public Schools and Home-Schooled

In the first stage of MH DIF analyses between public and home-schooled four items (55, 95, and 109) were flagged as a *C* item, exhibiting a large DIF. To obtain a pure matching variable, these items were excluded from the total score in the second stage of DIF analysis. In the second stage, Items 55, 95, and 109 still classified as type *C* DIF item (see Table 35).

Items 55 and 109 favored home-schooled (the focal group) while Item 95 favored private schools (the reference group). The $\hat{\alpha}$ values for Items 55, 95, and 109 were .472, 2.305, and .467, respectively. In addition, the values of $MH\ \chi^2$ for these items were 27.804 (p < .001), 35.414 (p < .001), and 27.724 (p < .001). These items had $\hat{\Delta}_{MH}$ of 1.765, -1.962, and 1.789, respectively. They had standard errors of .338, .327, and .340, respectively (see Table 35).

Also, in the second stage, twenty items (8, 14, 17, 21, 32, 35, 39, 46, 55, 62, 77, 85, 90, 93, 95, 99, 103, 105, 109, and 113) were classified as *B* items, showing a moderate DIF. Among the twenty items, twelve items (8, 14, 17, 32, 35, 39, 62, 77, 85, 90, 95, 99) favored public schools (the reference group), while eight items (21, 46, 55, 93, 103, 105, 109, and 113) favored home-schooled (the focal group) (see Table 35).

Items 8, 14, 17, 21, 32, 35, and 39 are grammar items, measuring, pronouns 1, verb forms 2 (infinitives/gerunds, etc), questions, verb forms 1 (tense/aspect/voice), verb forms 2 (infinitives/gerunds, etc), pronouns 2 (relative pronouns). Item 46 is a part-of- speech items, measuring the students' knowledge of English word forms taken from the second 2000 most frequent words. Items 55, 62, 77, and 85 are vocabulary items, testing the knowledge of the most common English vocabulary words, while Items 90, 93, 95, 99, 103, 105, 109, and 113 are reading items.

Table 35. Final Result of the MH Analysis for Moderate and Large DIF Items on Form B between Public Schools and Home-schooling

Item	Content	Label	Favored Group	$\hat{\alpha}_{MH}$	$MH\ \chi^2$	P	$\hat{\Delta}_{MH}$	$SE(\hat{\Delta}_{MH})$
8	Gram: Pronouns 1	BR	Public	1.789	20.234	.000	-1.366	.302
14	Gram: Verb forms 2 (infinitives/gerunds, etc)	BR	Public	1.586	12.452	.000	-1.083	.302
17	Gram: Questions	BR	Public	1.591	12.193	.000	-1.091	.311
21	Gram: Verb forms 1 (tense/aspect/voice)	BF	Home-s	.615	9.444	.002	1.141	.364
32	Gram: Verb forms 2 (infinitives/gerunds, etc)	BR	Public	1.828	17.056	.000	-1.418	.341
35	Gram NA	BR	Public	1.544	8.305	.004	-1.021	.343
39	Gram: Pronouns 2 (Relative pronouns)	BR	Public	1.581	12.460	.000	-1.076	.300
46	Pos: Second 2000 most frequent words	BF	Home-s	.622	11.189	.001	1.115	.327
55	Voc: First most frequent words	**CF**	Home-s	.472	27.804	.000	1.765	.338
62	Voc: 900 most frequent words	BR	Public	1.632	10.142	.001	-1.151	.352
77	Voc: 1000 most frequent words	BR	Public	1.773	13.934	.000	-1.345	.352
85	Voc NA	BR	Public	1.872	18.348	.000	-1.473	.339
90	Reading NA	BR	Public	1.998	24.219	.000	-1.627	.327
93	Reading NA	BF	Home-s	.584	8.721	.003	1.262	.421
95	Reading NA	**CR**	Public	2.305	35.414	.000	-1.962	.327
99	Reading NA	BR	Public	1.797	18.885	.000	-1.377	.312
103	Reading NA	BF	Home-s	.611	10.879	.001	1.156	.343
105	Reading NA	BF	Home-s	.640	9.544	.002	1.050	.331
109	Reading NA	**CF**	Home-s	.467	27.724	.000	1.789	.340
113	Reading NA	BF	Home-s	.555	16.807	.000	1.384	.336

Note. The third column shows the degree of DIF detected. *B* = moderate DIF; *C* = large DIF; *CR*/*BR* = items favor a reference group; *CF*/*BF* = items favor a focal group; Gram = Grammar; Pos = part-of-speech items; Voc = Vocabulary; NA=No information available regarding what item measures; Home-s = home-schooled.

4.1.3.4 Summary of DIF Analyses

Overall, in Forms A and B of the CEPA-English test, results of DIF analyses indicated that very few items (about 2% in Form A and 1% in Form B) were flagged as "*C*" between males and females. More items were flagged as "*B*" in Form A (7%) and in Form B (8%), but the direction of the DIF was not equally distributed between the reference (females) and the focal (males) groups. In both forms, more than half of the items that were flagged as "*B*" favored females, indicating that these items were easier for females (see Table 36). It is interesting that no items were flagged as "*C*" between Arts and Sciences streams in Form A, whereas only one grammar item was flagged in Form B. Very few items (about 4% in Form A and 1% in Form B) were flagged as "*B*." In Form A, more than half of these items favored students in the Sciences stream.

With respect to the DIF analyses between public and private schools, only one part-of-speech item was flagged as "*C*" in Form B. Few items (about 6% in Forms A and B) were flagged as "*B*". About half of these items favor public schools (the reference group) in both forms (see Table 36). With respect to the DIF analyses between private schools and home-schooled, in both forms, no item was flagged as "*C*." Both Forms A and B had a large number of items flagged as "*B*", while Form A had 12%, Form B actually had 22%. In both form, the direction of the DIF was generally equally distributed between the reference (private) and the focal (home-schooled) groups (see Table 36). With respect to the DIF analyses between public schools and home-schooled, in both forms, a large number of items (about 3% in Forms A and B) were flagged as "*C*." Both Forms A and B also had a large number of items flagged as "*B*," while Form A had 7%, Form B actually had 17%. In both forms, about more than half of items flagged as "*B*" favored public schools (the reference group) (see Table 36).

Table 36. Summary of Number of Large and Moderate DIF Items on Forms A and B by Gender, Study Type, and School Types

		Favored Group	Large DIF	Moderate DIF
	Gender	Females	1CR	6BR
		Males	1CF	3BF
Form A	Study Type	Arts		2BF
		Science		3BR
	School Type	Home-s		8BF
		Private		6BR
	School Type	Public	2CF	5BF
		Private		2BR
	School Type	Home-s	1CF	2BF
		Public	2CR	5BR
Form B	Gender	Females	1CR	8BR
		Males		2BF
	Study Type	Arts		1BF
		Science	1CR	
	School Type	Home-s		11BF
		Private		15BR
	School Type	Public	1CF	4BF
		Private		3BR
	School Type	Home-s	2CF	8BF
		Public	1CR	12BR

Note. *B* = moderate DIF; *C* = large DIF; *CR*/*BR* = items favor a reference group; *CF*/*BF* = items favor a focal group; Home-s = home-schooled.

4.1.4 Equating Using the Equipercentile Method

The RAGE-RGEQUATE program was used to implement equipercentile equating with the cubic spline postsmoothing method on Forms A and B of the CEPA-English test. The percentile ranks for both forms are plotted in Figure 23. In addition, the moments for raw score equivalents for equating Forms A and B are provided in Table 37.

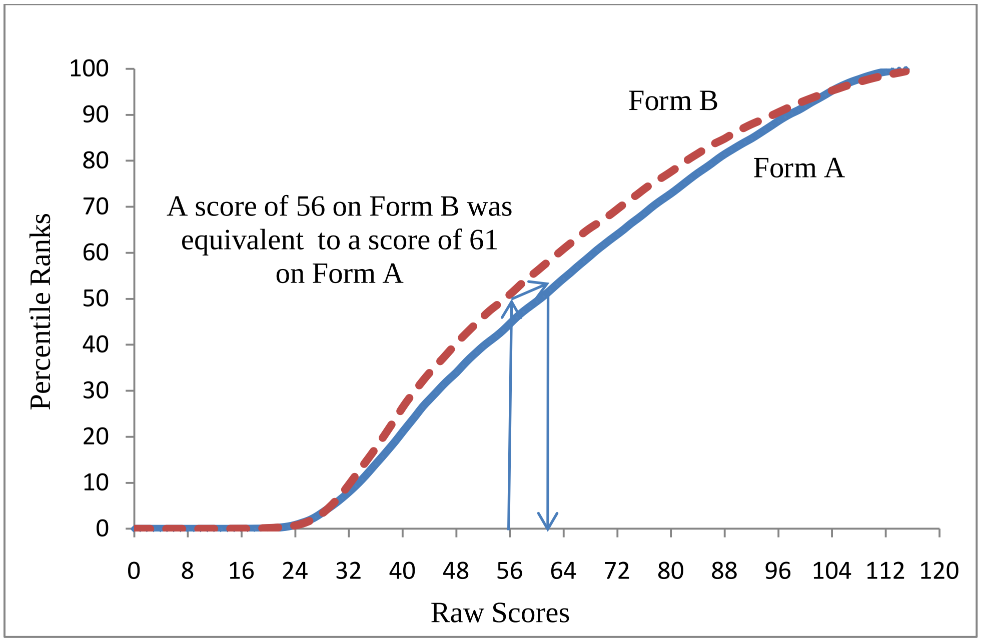


Figure 23. Percentile Ranks for Equating Forms A and B of the CEPA-English Test

Using the equipercentile equating method, scores on Form B were converted to the Form A scale. As shown in Figure 23, Form A appeared somewhat easier than Form B; for example, a raw score of 56 on Form B would correspond to a percentile rank of 49.69. To earn approximately the same percentile rank on Form A, a student would need a raw score of 61 on Form A (see Table E1 in Appendix E and Figure F1 in Appendix F for equivalent raw scores for equated Forms A and B using the postsmoothing with $S = 0.01$ level).

Furthermore, the mean of item difficulty (*p*-value) for Forms A and B suggested that both forms were relatively moderate in their difficulty levels, and that Form A (M of P =.542) was slightly easier than Form B (M of P =.500) (see Table 37). As indicated by the skewness values, both forms had a positively skewed distribution, but Form B seemed to be somewhat more skewed than Form A. Also, both forms had a slightly flatter distribution with a kurtosis of 1.9471 for Form A and 2.1791 for Form B (see Figures 10 and 11).

Table 37. Raw Score Moments for Equating Forms A and B

				Moments			
		N of Items	*M* of *p*	*M*	*SE*	Skewness	kurtosis
Test Form	Form A	116	.542	62.8241	23.7050	0.2820	1.9721
	Form B	119	.500	59.5815	23.6024	0.5173	2.2350
No Smooth	No Smooth			62.8245	23.7028	0.2823	1.9717
Cubic Spline Postsmoothing							
	S=0.01			**62.8234***	23.7215	**0.2821***	1.9784
	S=0.05			62.8208	23.7227	0.2816	1.9784
	S=0.10			62.8174	23.7220	0.2814	1.9782
	S=0.20			62.8056	23.7170	0.2817	**1.9770***
	S=0.30			62.7881	23.7116	0.2833	1.9777
	S=0.40			62.7660	**23.7043***	0.2862	1.9806
	S=0.50			62.7424	23.6965	0.2897	1.9847
	S=0.75			62.6893	23.6812	0.2987	1.9947
	S=1.00			62.6462	23.6715	0.3065	2.0030

Equipercentile equating with the cubic spline postsmoothing method was done with nine different values of *S*, ranging from .01 to 1. When *S* = .01, the mean raw scores for Form B were similar to those for Form A. Table 37 demonstrated that as *S* increased, the moments for smoothed equipercentile equating differed more than Form A moments. Table 37 also revealed that as *S* increased, the smoothed relationships differed more than the unsmoothed, and some relationships were outside of the standard error (*SE*) band. Graphs of the raw-to-raw score equivalents for cubic spline postsmoothing with *S*=.01 and at nine different values of *S* are provided in Figure F1 in Appendix F and Figure G1 in Appendix G, respectively. The graph presenting the smoothed curve with *S*=.01 yielded the best equating result and stayed within the standard error band. This graph also seemed to be smooth without deviating far from the unsmoothed values.

4.1.5 Assessing Test Information Function

MULTILOG was used to examine the amount of information Forms A and B of CEPA-English test provided under the 3PL model at nine ability (θ) intervals, ranging from -3.0 to 3.0. For Forms A and B, the test information function and the standard error of the ability estimates at each θ level are given in Figures 24 and 25. As illustrated in Figures 24 and Table 38, Form A provided most of its information between an ability level of 0 and 1.0, with information ranging from 47.371 to 49.102. The test information curve generally peaked at an ability level of .80 with the information equaling 50.840 and the smallest standard error being .140. At the extremes of the scales, the information was lower and the error of measurement higher. Most information in Form B was supplied between ability levels of .20 and 1.40, with information ranging from

46.857 to 50.042. The test information curve generally peaked at an ability level of .80, with the information equaling 54.436 and the smallest standard error being .136 (see Figures 25 and Table 39). At the extremes of the scales, the information was low and the error of measurement high. Thus, under the 3PL model, Forms A and B of the CEPA-English test provided the most information slightly above the cutoff score of 150, which is the mean of the NAPO test distribution. Overall, Form A had more information at the mean than Form B, but Form B had more information slightly above the mean than Form A.

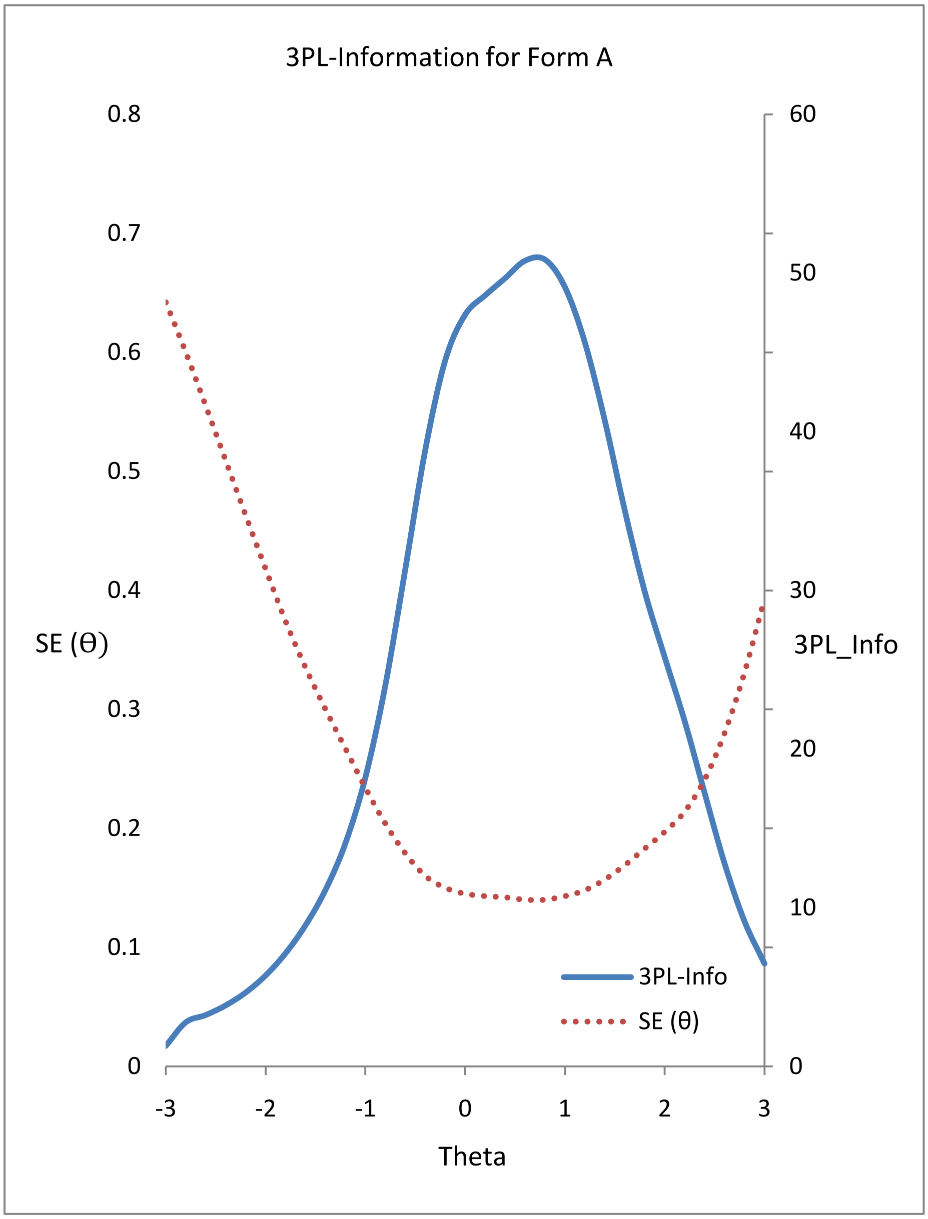


Figure 24. The Test Information Function and Standard Error for Form A

Table 38. The Test Information Function and Standard Error for Form A

Theta	3PL-Info	*SE* (θ)
-3.00	1.298	0.642
-2.80	2.782	0.600
-2.60	3.245	0.555
-2.40	3.852	0.509
-2.20	4.661	0.463
-2.00	5.736	0.418
-1.80	7.134	0.374
-1.60	8.901	0.335
-1.40	11.127	0.300
-1.20	14.068	0.267
-1.00	18.209	0.234
-0.80	24.015	0.204
-0.60	31.335	0.179
-0.40	38.839	0.160
-0.20	44.491	0.150
0.00	47.371	0.145
0.20	48.585	0.143
0.40	49.665	0.142
0.60	50.767	0.140
0.80	**50.840**	**0.140**
1.00	49.102	0.143
1.20	45.611	0.148
1.40	40.683	0.157
1.60	35.028	0.169
1.80	29.912	0.183
2.00	25.787	0.197
2.20	21.756	0.214
2.40	17.228	0.241
2.60	12.784	0.280
2.80	9.139	0.331
3.00	6.482	0.393

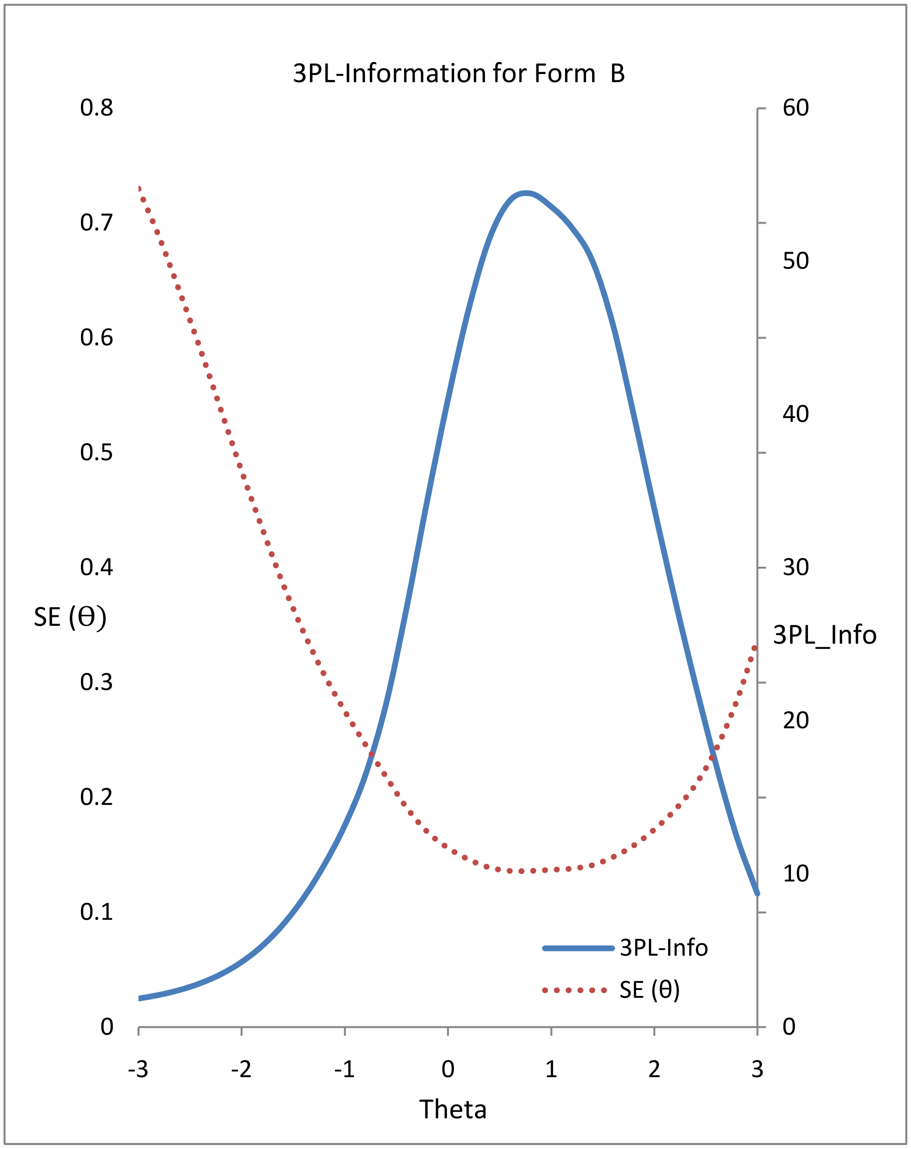


Figure 25. The Test Information Function and Standard Error for Form B

Table 39. The Test Information Function and Standard Error for Form B

Theta	3P-Info	SE (θ)
-3.00	1.875	0.730
-2.80	2.115	0.688
-2.60	2.439	0.640
-2.40	2.877	0.590
-2.20	3.469	0.537
-2.00	4.265	0.484
-1.80	5.321	0.434
-1.60	6.700	0.386
-1.40	8.451	0.344
-1.20	10.600	0.307
-1.00	13.184	0.275
-0.80	16.469	0.246
-0.60	21.099	0.218
-0.40	27.410	0.191
-0.20	34.423	0.170
0.00	40.991	0.156
0.20	46.857	0.146
0.40	51.440	0.139
0.60	53.971	0.136
0.80	**54.436**	**0.136**
1.00	53.582	0.137
1.20	52.215	0.138
1.40	50.042	0.141
1.60	45.861	0.148
1.80	40.015	0.158
2.00	33.878	0.172
2.20	27.958	0.189
2.40	22.373	0.211
2.60	17.130	0.242
2.80	12.431	0.284
3.00	8.728	0.338

5.0 DISCUSSION

This chapter summarizes the research findings and presents implications of the major findings. Following this summary, the chapter discusses the limitations of this study and recommendations for future studies.

5.1 SUMMARY AND DISCUSSION OF FINDINGS

The CEPA-English test, a standardized paper-and-pencil exam, is designed to measure the English proficiency of 12th grade students. It is the first national high-stakes test in the UAE and one of the most important exams students take in their academic career. This test is mandatory for all 12th grade students seeking undergraduate studies at the UAE's three major higher education institutions (UAEU, ZU, and HCT). The test score is used as a measure of achievement that accounts for 25% of the students' overall GSC English grades. The test score is also used for admission purposes, in that each student must achieve a minimum score of 150 on the test to be admitted into one of the three institutions' undergraduate programs. Finally, the test score is used for placing students with a score below 150 into the appropriate levels of English proficiency in the remedial program. The CEPA-English test consists of 120 multiple-choice items—90 items in the Grammar and Vocabulary Section and 30 items in the Reading Section.

The 2007 CEPA-English test had four forms (A, B, C, and D), Forms A and B were the focus of this study.

The main purpose of this study was to provide evidence for the quality of the CEPA-English test. To accomplish this goal, the study first examined the psychometric properties of Forms A and B of the CEPA-English test using IRT. Using the MH DIF detection method, the study then examined whether any items on Forms A and B exhibited DIF. Afterwards, using the equipercentile equating method under the random-group design, the study examined the extent to which the CEPA-English test scores are equivalent across Forms A and B. This also involved evaluating the accuracy of equating Forms A and B through examining the errors associated with this design. Lastly, using IRT, the study examined the amount of information gained at the cutoff score of 150 for Forms A and B.

5.1.1 Examining the Psychometric Properties of the CEPA-English Test using Classical and IRT Analyses

5.1.1.1 Classical Analyses

To determine the overall psychometric features of Forms A and B of the CEPA-English test, which helped in selecting the appropriate IRT model, the results obtained from classical and IRT analyses were examined. In this study, the point-biserials (r_{pbis}) values ranged from .04 to .66 for Form A and from .08 to .63 for Form B. Items with negative r_{pbis} values were excluded from the analyses, since a negative r_{pbis} indicates that an item is difficult for high ability students and easier for low ability students. As a result, four items (Items 47, 86, 87, and 97) and one item (Item 2) were deleted from Form B. Excluding these items resulted in 116 items in Form A and

119 items in Form B and produced an internally consistent test ($r_{R\text{-}20}$ = .963 for Form A and .960 for Form B).

Furthermore, results obtained from classical analyses showed that Form A was slightly easier than Form B since the mean difficulty level (*p*-value) was .542 for Form A and .500 for form B. English proficiency was heterogeneous based on the obtained CEPA-English overall total scores, which ranged from 16 to 116 on Form A and ranged from 12 to 119 on Form B. This large range of the CEPA-English scores suggested that the sample in both forms might be somewhat heterogeneous in their general English ability. The results also revealed that distributions of both forms were slightly positively skewed, but Form B seemed to be somewhat more skewed than Form A.

Based on the item parameters obtained from IRT, it was concluded that both forms were moderately difficult for the students (the *b* values ranged from -1.85 to 3.70 for Form A and from -1.70 to 2.55 for Form B), but that Form B (*M* of *b*=.530) was more difficult than Form A (*M* of *b* =.317). The *a* values ranged from .18 to 2.25 for Form A with a mean of 1.289, and from .25 to 2.54 with a mean of 1.290 for Form B, indicating that both test forms moderately discriminated between high-performing and low-performing students. Another important finding was that, as expected, some examinees with low ability tended to guess the correct answer on the most difficult items (the *c* values ranged from .00 to .38 with a mean of .206 for both Forms A and B). Hence, it was reasonable to conclude that the 3PL IRT model was the most suitable model to describe the items in each form. Additionally, the sample sizes (*N* = 9,496 in Form A and 9,296 in Form B) in the study were adequate for the 3PL MML estimation, which usually requires a minimum of 1000 subjects (Kingston and Dorans, 1985).

5.1.1.2 IRT Analyses

5.1.1.2.1 Examining the Assumptions underlying Dichotomous IRT Model

The 3PL IRT model was used to examine the psychometric properties of Forms A and B of the CEPA-English test. The successful use of the 3PL IRT model requires checking if the assumptions are confirmed by the CEPA-English test data. The assumptions of the 3PL model include: 1) determining whether the 3PL IRT model is the preferred model for each test form data; 2) assessing unidimensionality and the internal structure of each form; 3) investigating whether the items of each form are locally independent; 4) investigating whether each form is non-speeded; and 5) evaluating the extent to which examinees are guessing on items. The application of IRT also requires examining the extent to which the preferred 3PL IRT model fits each item on each form. Finally, the application also requires examining the degree to which the properties of the invariance of item and ability parameters hold true for each form of the CEPA-English test data.

5.1.1.2.1.1 Model Testing: Choosing the Preferred IRT Model

The choice of which IRT model—1PL, 2PL, or 3PL—best fits Forms A and B of the CEPA-English test data was determined statistically by assessing whether the additional parameters estimated under the 3PL model add more information than the 1PL or 2PL models (Embretson & Reise, 2000; Hambleton, 1993; Hambleton, et al., 1991; Hambleton & Swaminathan, 1985). If the addition of the extra parameters did not contribute to the fit of the model, then either the 1PL or 2PL models would be more appropriate.

When the parameters were estimated using all items (i.e., 116 items in Form A and 119 items in Form B), the value of the -2log likelihood was smaller for the 1PL and 2PL models than for the 3PL model. Because of this problem with the -2log likelihood, the parameters were re-estimated using a reduced test with 100 randomly sampled items in Forms A and B. The results with reduced test revealed that the 3PL model fit the data better than the 1PL or the 2PL models for both forms. In addition, findings from the descriptive statistics, correlations, and scatterplots of the item parameters for the full and reduced tests provided additional evidence supporting the use of the 3PL model. In summary, it was reasonable to conclude that the 3PL model was the most appropriate model to use with Forms A and B of the CEPA-English test data.

5.1.1.2.1.2 Evaluation of Unidimensionality and the Internal Structure

To evaluate the internal structure and unidimensionality of Forms A and B of the CEPA-English test data, four indexes were examined: the proportion of variance explained by the largest eigenvalues, the eigenvalue plots, residual statistics (RMSR), and the pattern of factor loadings. A dominant first factor underlying the test performance of the examinees is needed to satisfy the unidimensionality assumption. Reckase (1979) suggested that if the first factor explains 20% of the total test variance, it can be concluded that the test is unidimensional. In practice, however, it is difficult to strictly meet this assumption, as many factors affect test performance.

The results of the eigenvalues and the scree plots revealed one dominant dimension underlying Forms A and B of the CEPA-English test data, since the first factor in each form did in fact explain 20% of the total variance. In each form, the ratio between the first and second eigenvalues was six, which indicated that there was a significantly large difference between the first and second eigenvalues. Furthermore, the scree plot showed that there was a sharp drop from the first eigenvalue to the second and that the relative change for the second and consecutive eigenvalues was reasonably constant, providing additional evidence that the test was unidimensional (see Figure 15).

In both Forms A and B, the RMSR value for a one-factor model was .004, indicating that a one-factor model provided an acceptable solution. In addition, on each form, the inspection of the factor loadings for the one-factor model showed that almost all items loaded on the first factor. Further, the results of residual analyses and factor loadings suggested that two- and three-factor models were also acceptable solutions for each form. However, the simplest one-factor solution was preferred, and this was consistent with the eigenvalue analyses and scree plots. Furthermore, the results of the RMSR analyses and the scree plots of the simulated data corroborated with the real data. Therefore, it was reasonable to conclude that the unidimensionality assumption was confirmed in each form of the CEPA-English test data.

Iit is interesting that the results of the DIMTEST analyses using real data suggested that Forms A and B of the CEPA-English test had more than one dimension, whereas the simulated data suggested that the test data was essentially unidimensional. This inconsistency in findings can be interpreted to indicate that the CEPA-English items measure several different aspect of language ability (i.e., grammar, vocabulary and reading) in addition to the overall English ability level.

5.1.1.2.1.3 Assessing Local Item Independence

Local independence implies that only one latent trait influences the examinee’s response to the items. When unidimensionality is true, the assumption of local independence usually holds (Hambleton & Swaminathan, 1985; Lord, 1980; Lord & Novick, 1968). Yet, local independence can be obtained without satisfying the assumption of unidimensionality (Scherbaum, 2006).

When the local independence assumption is true, the expected value is -1/ (*n*-1) and the mean of Q_3 statistic equal to zero (Embretson & Reise, 2000).

The local item independence of Forms A and B of the CEPA-English test was examined using the Yen's Q_3 statistic. In this study, both the observed and simulated data revealed that the items on each form of the CEPA-English test data were locally independent.

5.1.1.2.1.4 Examination of Speededness

When examinees do not have enough time to complete a test, they may skip items they fail to reach or do not have the ability to answer. In a speeded test, it is assumed that *examinees may* omit items at the end of the test due to the time limit, not to their limited ability. If examinees fail to answer test items because of time limit rather than ability, then two dominant factors influence their test performance: speed and ability. In this case, when speed does affect test performance, the unidimensionality assumption is essentially violated, since the ability measured by a test is not the only factor impacting test performance. The local independence assumption of the IRT models is also being violated.

Each form of the 2007 CEPA-English test was administered in two-and-a half hours. Because the test was administered in a specific time limit, it is important to examine the degree to which each form is non-speeded. The results of the speededness analyses for Forms A and B of the CEPA-English test showed that the ratio of the variance of omitted items to the variance of the items answered incorrectly was 0.01, that all examinees completed more than 75% of the items, and that 99% of them reached the end of the test. These findings suggested that the test time was sufficient to allow for the majority of examinees to complete all the items. Therefore, it was reasonable to conclude that Forms A and B of the CEPA-English test data met the assumption of non-speededness.

5.1.1.2.1.5 Assessing the Presence of Guessing Behavior

The degree to which a multiple-choice test is speeded may motivate examinees to use guessing strategies when responding to test questions. The use of guessing strategies introduces measurement error and attenuates relationships between items and can be considered a source of construct-irrelevant variance. When examinees run out of time, they very likely tend to randomly guess answers to all of the items they previously skipped or did not reach. They also randomly guess on items when they lack the necessary knowledge to correctly answer an item. The proportion of correct responses was approximately what would be expected under the random guessing model (1/m, *m* is the number of options). For a multiple-choice test with four alternatives, the probability of getting the item correct from random guessing is .25. Guessing behavior of examinees depends partly on the administration instructions and whether a correction for guessing is used to discourage random guessing (Linn & Gronlund, 2000).

In each form of the 2007 CEPA-English test, no correction for guessing was used, and examinees were not given any instructions regarding guessing. Thus, the extent to which examinees are guessing answers on items on Forms A and B of the CEPA-English test was evaluated by examining plots of proportion of correct responses by total scores (Hambleton &Swaminathan, 1985). Based on this evaluation, several items in Forms A and B exhibited a fairly constant proportion for a correct response, providing evidence that low ability examinees might be using guessing strategies. However, it was difficult to identify a constant proportion of correct responses for easier items. As a result, the frequency of the average proportion of correct responses for low-ability examinees was examined *only* on items with $p \leq .7$. In both Forms A and B, the overall pattern of the frequency distributions was similar, and it was consistent with the random guessing model for multiple-choice items with four options (p~ .25). There were several items with a value that was slightly below or above the average *c*-parameter (.25)—the value that would be expected under the random guessing model. This might be due to the elimination of distracters, which may increase the expected proportion of correct responses. It might also be due to the use of well-designed distracters, which may lower the expected proportion of correct responses and may change the guessing strategy of examinees. Thus, the random guessing model may not be appropriate (Stone &Yeh, 2006).

Further, in both forms, the comparison of the distributions of the c parameters at the beginning and end of the test corroborated the results from plotting the proportion of correct responses by total scores. That is, the two distributions differed in skewness and kurtosis, providing additional evidence that the examinees are guessing answers in each form.

In summary, the evaluation of all the 3PL IRT model assumptions indicated that the assumptions were met. Guessing behavior did exist on Forms A and B of the CEPA-English test data. Consequently, it is important to account for guessing when exploring the psychometric properties of the CEPA-English test using IRT. This was achieved by using the 3PL model. Including this parameter not only can remove the effect of random guessing, but also can make an adjustment for partial-knowledge guessing (Waller, 1989). Test developers can attempt to eliminate random guessing behavior by instructing examinees not to guess.

5.1.1.2.2 Assessing the Preferred IRT Model Fit at Item Level

Once the preferred IRT model is obtained, it is important to examine the extent to which each test item fits the preferred IRT model. An item exhibits misfit if there is a difference between the observed and predicted score distributions across a range of discrete ability levels for each item. The fit of the 3PL IRT model to each item on Forms A and B of the CEPA-English test data was examined using the *S*-X^2 statistics. The *S*-X^2 analyses suggested that the 3PL model was a better fit for each form since the 3PL model has the smallest number (i.e., 23 items misfit the 3PL model in Form A 25 in Form B) of misfit items compared to the 1PL and 2PL models. The results of the plots of the item residuals for misfit items also corroborated with the *S*-X^2 statistical results, providing additional evidence for using the 3PL model.

5.1.1.2.3 Examining the Invariance of Item Parameters

After determining that the 3PL model is the most appropriate IRT model for Forms A and B of the CEPA-English test data, it was imperative to assess the invariance of item parameters on each form. Invariance of item parameters means that item parameter estimates (*a*, *b*, and *c* parameters) do not depend on the ability distribution of the examinees; rather item parameter estimates will not change depending on which group was used to calibrate items (Hambleton, et al., 1991; Lord, 1980). The degree to which invariance holds can be assessed by examining the correlation and the scatterplots of parameter estimates (Hambleton, Swaminathan & Rogers, 1991). If the correlations are reasonable and the plots are linear, the property of invariance is met. According to Wright (1968), the property of invariance exists to some degree when associated correlation coefficients are .80 and higher.

For each form, the degree to which the invariance of the item parameters of the 3PL model holds across low and high ability groups was examined by comparing the -2log likelihood statistics between the restricted and unrestricted models using 100 randomly selected items. The G^2 statistic between the restricted and unrestricted models was not significant for Form A but significant for Form B, suggesting that item parameter estimates were not invariant for Form A but were invariant for Form B.

The correlations and plots of parameter estimates for the unrestricted models were also examined. This examination revealed that the property of invariance of item parameter estimates from the 3PL was not strictly met for Form A and to some extent for Form B of the CEPA-English test data, mainly because some of the *c* parameters were not invariant across different ability groups (low and high) in both forms. However, the *b* and *a* parameters were invariant across groups in both forms, and the *b* parameters proved to be more invariant than the *a* parameters. This could be due to the fact that the *c* parameter is more difficult to estimate while the *b* parameter is easier to estimate.

5.1.2 Detecting DIF using the Mantel-Haenszel Procedure

The two-stage MH process was performed in order to examine whether each item on Forms A and B of the CEPA-English test exhibits uniform DIF between males and females, between study types (i.e., Arts and Sciences), and between school types (i.e., public, private, and home-schooled). There are some concerns that need to be addressed when using the MH procedure: sample size (e.g., Clauser & Mazor, 1998; Gierl et al., 2004), test length (e.g., Donoghue & Allen, 1993; Jodoin & Gierl, 2001), number of score groups, (Holland & Thayer, 1993), and inclusion of the item in the matching criterion (Holland & Thayer, 1993). Previous research has shown that the sample size, and test length, the use of 10 (0-10) scoring groups on each form, and the use of the total scores as matching criterion are adequate for these concerns. In this case, a random sample of 2000 subjects with equal numbers of male and female subjects as well as equal number of students majoring in Arts or Sciences was used. Another random sample includes 1400 subjects with equal numbers of students in public schools, private schools, and home-schooled was also used. Also, 116 items in Form A and 119 items in Form B were used in this study.

Overall, the results from the first and second stages were generally same. Several studies indicated that changes in the number of items being flagged as DIF as well as the degree of DIF before and after purification are influenced by the percentage of DIF detected in the first stage (Clauser & Hambleton, 1993; Clauser, et al., 1993; Zenisky et al., 2003). When tests contain a relatively large percentage (at least 30%) of DIF items in the first stage, the changes in the number of items being flagged as DIF between the two stages will be obvious. In this study, the number of items flagged as showing DIF at the first stage was relatively small (see Table 36 in page 253); consequently, the degree of DIF (classified as a *B* or *C* DIF item) and the number of DIF items did not change from the first to the second stages.

The results of the second stage of DIF analyses indicated the presence of a total of eleven items in form A and in Form B showing gender DIF, with a large DIF exhibited in two items in Form A (81 and 109) and one item in Form B (32). Nine items were flagged as "*B*" in Form A and ten items in Form B, and about half of these items favored females. It is interesting that no item was flagged as "*C*" between Arts and Sciences streams in Form A, only one grammar items

(112) was flagged in Form B. Five items in Form A and only one item in Form B were flagged as "*B*", and about half of these items favored students in the Sciences stream in Form A.

With respect to the analyses between public and private school students, in both forms, only one part-of-speech item was flagged as "*C*" in Form. Few items (about 6% in Forms and B) were flagged as "*B*". About half of these items favor public school students (the reference group) in both forms (see Table 36). With respect to the DIF analyses between private and home schooled students, in both forms, no item was flagged as "*C*." Both Forms A and B had a large number of items flagged as "*B*," while Form A had 12%, Form B actually had 22%. In both form, the direction of the DIF was generally equally distributed between the reference (private) and the focal (home schooled) groups (see Table 36). With respect to the DIF analyses between public and home schooled students, in both forms, a large number of items (about 3% in Forms A and B) were flagged as "*C*." Both Forms A and B had a large number of items flagged as "*B*," while Form A had 7%, Form B actually had 17%. In both forms, about more than half of items flagged as "*B*" favored public school students (the reference group) (see Table 36).

Thus, the results of DIF the analyses indicated that about the same number of DIF items were found in both forms by gender and study type. However, a significant proportion of items were flagged as "*B*" between private and home schooled students and also between public schools and home schooled, with more items in Form B than in Form A. However, the presence of items showing DIF does not guarantee that the item test is biased (Angoff, 1993). Further examination is needed to ensure whether the items measure other constructs that are not related to the ability being measured. The presence of DIF items by school type may indicates curriculum differences across private, public, and home schools.

5.1.3 Equating Using the Equipercentile Method

The CEPA-English test is administered repeatedly each year, which increases threats to test security. To ensure test security, NAPO uses multiple forms of the CEPA-English test that are constructed on the same specifications. Because Forms A and B were randomly distributed to examinees, it was reasonable to use a random groups design in this study, since the examinees in both forms had the same average ability levels.

To examine the extent to which the CEPA-English test scores are equivalent across Forms A and B, the equipercentile equating with the cubic spline postsmoothing method under the random-groups design was used via the RAGE-RGEQUATE program. This also involved examining the error associated with the equipercentile equating for Forms A and B.

The result of the plots of the percentile ranks for equating Forms A and B showed Form A was somewhat easier than Form B, given that the number of items on both form differed (see Figure 23). The results also demonstrated that the postsmoothed equipercentile equating with S = 0.01 provides the best equating results within the standard error band. That is, when using the postsmoothing with S = 0.01 level, the central moments for Forms A are almost the same as for Form B. Thus, using the postsmoothing with S = 0.01 level will help in obtaining a meaningful comparison of students' scores (see Table E1 and Figures F1 and G1 for equivalent raw scores for equated Forms A and B).

5.1.4 Assessing Test Information Function

The 3PL IRT model is used to generate an item information function that describes the amount of information produced by each individual item on a test as a function of ability. The item information function for each item at each θ level was added to produce the test information function. If the purpose of the test, as in CEPA-English, involves a cutoff score or pass/fail decision, then test items that provide most of their information at the cutoff score should be selected. The more information a test provides, the more precise the test, and hence the less error is associated with it (Baker, 1992).

Under the 3PL IRT model, the extent to which the test information function for equated Forms A and B is maximized at the cutoff score (150) was examined using MULTILOG. The amount of the test information functions provided by each form maximized or peaked at an ability level of .80 for Forms A and B, and the standard error was minimized at .140 and .136, respectively (see Figures 24 and 25). The most information was provided between ability of 0 to 1.0 for Form A and between ability of .20 to 1.40 for Form B.

Thus, the test information functions of Forms A and B of the CEPA-English test supplied most of their information at an ability level slightly above the cutoff score of 150, which is the mean of the test in NAPO study. This finding also suggested that it was important to add more easy items in order to gain more precise information at the cutoff score of 150.

5.2 FINAL CONCLUSION AND IMPLICATIONS OF THE MAJOR FINDINGS

The CEPA-English exam is used for achievement, selection, and placement purposes. Because this test imposes serious consequences on students, it is imperative to ensure that the test performs as intended and provides meaningful and interpretable results. Until now, there has been no study that extensively evaluates the technical quality of the CEPA-English exam. To this end, this study evaluated; 1) the psychometric quality of Forms A and B of the CEPA-English items and the test as a whole; 2) the extent to which DIF occurs on Forms A and B; 3) the comparability of Forms A and B of the test, and finally; 4) the amount of information provided at the cutoff score of 150 for Forms A and B, which is the mean of the test in the NAPO study. The cutoff score is used as a basis for admission to higher education.

Generally, the evidence collected in this study appeared to support that overall, the CEPA-English test demonstrated good psychometric properties. The psychometric properties of Forms A and B of the CEPA-English test were examined using classical and IRT analyses. The findings from classical analyses demonstrated the need to carefully examine the following items with negative r_{pbis}: Items 47, 86, 87, and 97 in Form A and Item 2 in Form B. These items were problematic and did not contribute to the overall effectiveness of the CEPA-English test. The problem might simply be that these items were miskeyed responses, or perhaps were poorly constructed. Thus, the implication of this finding is that the test developer may want to examine these items to see if they need to be revised in order to improve the design of the test.

High internal reliability ($r_{R\text{-}20}$ = .963 for Form A, and .962 for Form B) was obtained for each form, indicating that the items within each form were homogenous (Crocker & Algina, 1986). As expected, the high number of items (116 items in Form A and 119 items in Form B) produced a highly reliable test.

The findings from classical and IRT analyses indicated that the majority of the test items in both forms were relatively moderate in their difficulty levels and that Form B (the mean of *p*-values was .500 and the mean of *b* was .530) was more difficult than Form A (the mean of *p*-values was .542 and the mean of *b* was .317). The IRT analyses also indicated that both forms moderately discriminate between high-performing and low-performing students, where the mean of *a* parameter was 1.289 for Form A and 1.290 for Form B. The mean of *c* values was .206 for Forms A and B, suggesting that examinees with low ability might be able to choose the correct answer.

The results of model testing revealed that the unidimensional 3PL IRT model fits Forms A and B of the CEPA-English test data better than the 1PL or 2PL models. This is because the 3PL model includes the guessing parameter, which takes into account the performances of the low-ability examinees.

Although the assumptions of the 3PL model were met, the assumption of item parameter invariance was not met for Form A based on the results of examining the correlations and plots of parameter estimates. This lack of invariancy was primarily due to the estimation of the *c* parameters. The results of examining the -2log likelihood statistics further indicated that the assumption of item parameter invariance was not supported for Form A. While the results of examining the -2log likelihood statistic supported the assumption of item parameter invariance for Form B, the results of examining the correlations and plots of parameter estimates showed that some of the *c* parameter estimates were not very stable across the two groups.

After proving that the 3PL model fits Forms A and B of the CEPA-English test data at the test level, the degree to which the 3PL model fits each item on each form was examined. The results of fit analysis using both the $S\text{-}X^2$ statistics and residuals analyses demonstrated that the 3PL model improved the fit of each item on each form more than the 1PL or 2PL models. In addition, the results showed there were 23 items misfit the 3PL model in Form A and 25 in Form B. There were several reasons why these items might not fit the 3PL model. The first reason was that failure to meet the underlying IRT model assumptions, such as dimensionality, local independence, and monotonicity may result in having some items that are more likely to not fit the model. The second reason was that misfit may occur because of failure to achieve invariant item and ability parameter estimates. The third reason was that failure to select an appropriate IRT model may result in item misfit. Finally, other misfit reasons include failure to obtain a large enough sample size, nonmontonicity of item-trait relations, or poor item construction (Embretson & Reise, 2000; Hambleton, 1993; McKinley & Mills, 1985).

Since the evaluation of all the 3PL model assumptions indicated that the assumptions were met for Forms A and B of the CEPA- English test data, it might be the case that some items misfit the 3PL model because students with low ability guessed at random when they lacked the necessary knowledge to correctly answer an item. Another possible reason was that these items misfit the 3PL model because some of them showed medium or large DIF values. Also, these items misfit the 3PL model because the properties of invariance of item parameter estimates from the 3PL model was not strictly met for Form A and to some extent for Form B of the CEPA-English test data. A final possible reason was that these misfit items might be poorly constructed.

Overall, few DIF items were detected between males and females, between Arts and

Sciences students, and between private and public school students. However, a significant proportion of DIF items were flagged by school type. It is important to point out that the presence of DIF does *not* necessarily guarantee that the item is biased (Angoff, 1993). Yet, items with large DIF values can be an indicator that such items are measuring an additional unintended construct of the test—that is, they can indicate the existence of multidimensionality within the test (Camilli & Shepard, 1994). Therefore, further examination by the test developer and content specialists is needed to determine whether the items flagged as having a large DIF measure any ability irrelevant to the ability of interest—in this case, English. In addition, the choice of whether to revise or remove items with a medium or large DIF depends on whether the content experts consider such items essential for the purpose of the test.

An important point worth noting hear, is that DIF had a greater presence for school type in this study (between private and home schooled students and between public and home schooled students), which may indicate curriculum differences across schools. Given that the number of items differed by test sections, the majority of DIF items in Form A were from the grammar area. In Form B, the majority of DIF items between private and home schooled students were from the vocabulary area, whereas between public and home-schooled students the majority of DIF items were from the reading area.[13]

Given that the cutoff score was set at of 150, which is the mean of the test in the NAPO study, the CEPA-English test should be designed to provide maximum information at the cutoff score of 150. The test information functions illustrated that Forms A and B provided most of

[13] The CEPA-English test consists of 120 multiple-choice items: 10 are parts of speech items, 40 are grammar items, and 40 are vocabulary items.

their information at an ability level slightly above the cutoff score. The amount of information on both forms peaked at an ability level of .80. The results also showed that Form A had highest information at an ability level of 0 and 1.0, while Form B had highest information at an ability level of 1.20 and 1.40. At the extremes of both scales, the information was lower and the value of the standard error was higher. In general, the levels of information and the precision of items for both forms were higher in the middle region of the scales than at the end of scales. This suggested that the CEPA-English items discriminated more effectively between students at slightly above average levels of English skills. The fact that the CEPA-English test provides maximum information slightly above the cutoff score of 150 has an important implication for policy and decision makers in the UAE with respect to the effectiveness of using this cutoff score as a major indicator of candidates' performance or abilities. Because revising the cutoff score of 150 is not an option, to gain more precise information, it will be necessary to add easier items on the test.

5.3 LIMITATIONS OF THE STUDY

There are some limitations to the current research. The first limitation was in the generalizability of the results. Because the subjects for this study were 12th grade students from the UAE, the results of this study should not be generalized without considering the sample characteristics and performance of the 12th grade students. Further, all analyses in this study were based on the examinees' responses to multiple-choice items designed to assess the English proficiency level of 12th grade students. Therefore, the results may not be generalized to different types of items (e.g., polytomous) or to other content areas in achievement, selection, and placement tests.

Another limitation of the present study was that when comparing nested models MULTILOG cannot handle test with more than 100 items. The G^2 observed values were negative for Form A, with 116 items, and for Form B, with 119 items, indicating a problem with estimating the -2log likelihood values. It was important to point out that the problem with the -2log likelihood did not affect the estimation of the item parameters (*a*, *b*, and *c*) for Form A and B of the CEPA-English test. These estimations were accurate because results of examining the means and standard deviations of the item parameters (*a*, *b*, and *c*) for both the full test (with all items) and the reduced test (with 100 random items) for each form demonstrated that both tests had similar item parameters. Furthermore, the correlations and scatterplots between the estimated parameters were also very high and had linear relationships (see Table 14 and Figures 12-14).

The third limitation of the present study was that test items were not released and therefore, the content of the test could not be examined. Such an examination was important to more thoroughly evaluate the internal structure and dimensionality of the CEPA-English test data, to examine the possible sources of item misfit, and to investigate the items that were flagged as DIF.

The CEPA-English test is used for selection purposes, in that each student must achieve a minimum score of 150 on the test to be admitted into one of the three UAE's major institutions' undergraduate programs. It is therefore crucial to examine the predictive validity of the test. However, a further limitation of the present study was that the data needed to perform such an examination was not available (e.g., first-year college grade point averages or FGPA). Therefore, the predictive validity of the CEPA-English test could not be evaluated.

5.4 RECOMMENDATIONS FOR FUTURE RESEARCH

In spite of the outlined limitations of the current research, this study is significant because it is the first study to my knowledge that extensively evaluates the quality of the CEPA-English test in relation to Forms A and B. The evaluation of the psychometric properties of each item and the test as a whole is an essential part of any measurement process because it provides insight into the characteristics of the test and helps further development. Overall, the analyses reveal that the CEPA-English test demonstrates good psychometric properties, suggesting that the test is a suitable instrument to assess the English proficiency level of 12th grade students in the UAE. Yet, the CEPA-English test can be improved by eliminating items with negative discrimination and adding easier items in order to gain more precise information at the cutoff score of 150.

In addition, the test developer may want to evaluate items that misfit the 3PL model in Forms A and B of the CEPA-English test. They also may want to evaluate items with a medium and large DIF to determine whether to revise or eliminate them from the CEPA-English test. Thus, the findings in this study could help the test developer improve the design of the test. This study also contributes to DIF studies on language testing, which in turn will lead to the improvement of language testing design and construction.

Based on the findings and limitations of this study, the following recommendations are made for future research. First, the presence of items showing medium and large DIF on Forms A and B of CEPA-English test suggest the need to conduct a DIF study before the test is administered in order to flag test items that are statistically biased. Removing or retaining items with a medium and large DIF is another issue. Recommendations from the literature are that items with a medium DIF can be included in the test if there is no item with negligible DIF available to meet the content requirement of the test, while items with a large DIF can only be included in the test if the content experts consider this item essential for the purposes of the test (Camilli& Shepard, 1994; Zieky, 1993; Zwick & Ercikan, 1989).Thus, test developer may want to revise or eliminate items showing medium or large DIF in Form A and B of the CEPA-English test.

Second, the presence of a significant proportion of items with a medium or large DIF value by school type may indicate curriculum differences across private, public, and home schools. Consequently, content experts could attempt to examine why certain items are relatively easier or more difficult than others for the sub groups with the same ability. Content experts could evaluate 12th grade English curriculua and instruction for home-schooled students as compared to public and private school students.

Furthermore, in this study, the presence of the possible nonuniform DIF item was not examined by MH procedure. Past research has illustrated that MH procedure is a more powerful test for detecting uniform DIF items, but not nonuniform DIF (Hambleton & Rogers, 1989, Swaminathan & Rogers, 1990, Rogers & Swaminathan, 1993; Lopez-Pina, 2001). Mazor et al. (1994) modified the MH procedure using simulated data and reported that this modification improves the detection of non-uniform DIF over the standard MH procedure (Hidalgo & Lopez-Pina, 2004). Nevertheless, this modification has not been investigated extensively. Several DIF simulation studies reveal that the logistic regression (LR) procedure is a powerful method to detect nonuniform DIF (e.g., Swaminathan & Rogers, 1990; Rogers & Swaminathan, 1993; Lopez-Pina, 2001). Thus, further research could investigate the presence of nonuniform DIF in the CEPA-English test using the LR method.

Another important recommendation for further research is to examine the predictive validity of the CEPA-English test. The results of the CEPA-English test form a major part of admissions criteria for UAEU, ZU and HCT. Students must achieve a minimum score of 150 on the CEPA-English exam, along with a minimum average of 70% on the GSC exam or equivalent to be eligible for Bachelors' programs at UAEU, ZU and HCT, and Higher Diploma programs at the HCT. Because of the potentially life change consequences of this cut-off score, it is critical to assess the predictive validity of the CEPA-English test. Such evidence is needed to provide support for the use of the test in admissions decisions. It would also be important to examine the consistency of the predictive relationship across different groups of examinees by examining the presence of differential predictive validity evidence.

Further research could, thus, examine whether the CEPA-English test is an accurate predictor of students' academic success in college. Specifically, research could examine how accurately the total test scores of CEPA-English predict students' academic success measured by first-year college grade point averages or FGPA. Further research could also examine whether the prediction of students' FGPAs using the CEPA-English test scores are consistent across different groups of examinees: across gender, region (e.g., Abu Dhabi, Dubai), school type (e.g., private and public schools), and study streams (i.e., Arts and Science).

APPENDIX A

AN EXAMPLE OF THE CEPA-ENGLISH TEST

Grammar and Vocabulary (45 minutes)

Grammar

1 I went to summer school _______________ improve my English.

a) for

b) so

c) to

d) will

2 Wood_______________ when you place it in water.

a) floats

b) is floating

c) was floating

d) floated

3 Their ______________ often washes the car on Sundays.

a) brothers

b) brother

c) brother he

d) brother is

4 Did his mother give you _____________ tea?

a) a few

b) many

c) some

d) a lot

5 Her family visits her every weekend and takes her to _________ favorite restaurant.

a) them

b) their

c) theirs

d) they

6 You speak very good Arabic. How long you _____________ in Dubai?

a) do you live

b) are you living

c) live

d) have you lived

7 The speaker received a big round of applause after ____________ his speech.

a) will finish

b) finished

c) finishing

d) finishes

8 My car is not working. The engine needs to ________________.

a) be replaced

b) replace

c) replaced

d) been replaced

9 The students studied hard _________ they wanted to pass the exam.

a) because

b) but

c) so

d) although

10 If you ______________ harder, you would have been successful.

a) had worked

b) work

c) have worked

d) working

11 My father _______ in the military.

a) is a policeman

b) a policeman

c) policeman

d) is policeman

12 The United Arab Emirates ____ an independent country on 2 December 1971.

a) has become

b) became

c) will become

d) was becoming

13 The people of the UAE ____________ seen great changes since 1971.

a) had

b) was

c) are

d) have

14 I _______________ everywhere for Khalid. Have you seen him?

a) looking

b) have looking

c) have been looking

d) been looking

15 Would you like _________ milk in your coffee?

a) some

b) many

c) a few

d) a lot

Parts of speech

1 We all must _________ the laws of our country. a) some

a) obey

b) obediently

c) obedient

d) obedience

2 The _________ painting was much better than the copy.

a) originates

b) original

c) originally

d) originality

3 _________ football players can earn large amounts of money)

a) Profession

b) Professionally

c) Profess

d) Professional

Vocabulary

1 Faraj was _________ of the dark when he was young. He slept with the light on at night.

a) proud

b) afraid

c) silent

d) pleasant

2 If you are sick, you should sleep a lot and give your body a chance to

___________.

a) insult

b) sew

c) heal

d) obey

3 You are only allowed to carry two bags on to an airplane. That is the _________ for most airline companies.

a) labour

b) policy

c) response

d) assessment

4 There was a loud __________ from the crowd as the race finished.

a) greet

b) steer

c) urge

d) cheer

5 One way to ________________ your body is to exercise at the gym everyday.

a) strengthen

b) joke

c) shine

d) tremble

Reading (45 minutes)

Reading (Section of text)

Stamp collecting is a popular hobby for millions of people around the world. The British started the system of paying in advance for sending letters in 1840. The USA produced its first stamp in 1847. In 1864 the first stamp catalogue for people interested in stamp collecting, or philately, was printed in Paris. Stamps designed by different countries for use by the international postal service are interesting and informative. Rare stamps, or stamps with mistakes, can also be very valuable, so stamp collecting is not just for children.

Questions

1 The best title for this text is ____________________.

a) a children's hobby

b) a history of stamps

c) stamps of the world

d) a hobby for all

2 The system of pre-paid mail started in ____________________.

a) the USA

b) several countries

c) Britain

d) Paris

3 The first stamp was produced in ____________________.

a) 1840

b) 1847

c) 1864

d) 1804

4 Stamp collecting is also known as ____________________.

a) stamp cataloguing

b) philately

c) popular hobby

d) valuable

5 Stamp collecting is popular with ____________________.

a) a few adults

b) children

c) rich people

d) many people

APPENDIX B

FREQUENCY TABLES FOR NOT-REACHED AND MISSING ITEMS FOR FORMS A AND B

Table B1. Frequency Table for Not-Reached Items for Form A

No. of not-reached items	Frequency	Percent
0	9422	99.2
1	36	.4
2	1	.0
3	4	.0
4	2	.0
5	1	.0
6	2	.0
7	2	.0
8	3	.0
9	3	.0
10	2	.0
11	1	.0
15	3	.0
17	1	.0
18	1	.0
20	1	.0
21	2	.0
23	3	.0
28	1	.0
29	1	.0
50	1	.0
58	1	.0
59	1	.0
70	1	.0
TOTAL	9496	100.0

Table B2. Frequency Table for Missing Items for Form A

No. of missing items	Frequency	Percent
0	7905	83.2
1	992	10.4
2	247	2.6
3	115	1.2
4	57	.6
5	24	.3
6	24	.3
7	22	.2
8	14	.1
9	7	.1
10	5	.1
11	7	.1
12	8	.1
13	7	.1
14	1	.0
15	5	.1
16	6	.1
17	4	.0
18	4	.0
19	2	.0
20	4	.0
23	5	.1
24	5	.1

Table B2 (continued)

No. of missing items	Frequency	Percent
27	4	.0
28	2	.0
29	1	.0
30	1	.0
32	1	.0
34	1	.0
36	1	.0
37	1	.0
38	3	.0
44	1	.0
50	2	.0
51	1	.0
52	2	.0
58	1	.0
59	1	.0
67	1	.0
70	1	.0
75	1	.0
TOTAL	9496	100.0

Table B3. Frequency Table for Not-Reached Items for Form B

No. of not-reached items	Frequency	Percent
0	9201	99.3
1	17	.2
2	7	.1
3	4	.0
4	5	.1
5	5	.1
6	1	.0
8	4	.0
9	2	.0
10	3	.0
11	1	.0
12	1	.0
14	3	.0
15	1	.0
16	2	.0
20	3	.0
25	1	.0
26	2	.0
29	1	.0
30	2	.0
34	1	.0
40	1	.0
43	1	.0
TOTAL	9269	100.0

Table B4. Frequency Table for Missing Items for Form B

No. of missing items	Frequency	Percent
0	7595	81.9
1	1022	11.0
2	291	3.1
3	124	1.3
4	53	.6
5	36	.4
6	19	.2
7	14	.2
8	13	.1
9	4	.0
10	11	.1
11	6	.1
12	2	.0
13	10	.1
14	5	.1
15	8	.1
16	2	.0
17	3	.0
18	2	.0
19	2	.0
20	5	.1
21	3	.0
22	2	.0

Table B4 (continued)

No. of missing items	Frequency	Percent
23	3	.0
24	1	.0
26	3	.0
27	1	.0
28	1	.0
29	2	.0
30	5	.1
32	3	.0
35	2	.0
36	1	.0
40	1	.0
42	1	.0
43	1	.0
45	1	.0
46	1	.0
48	1	.0
49	2	.0
50	2	.0
52	1	.0
56	1	.0
61	1	.0
71	1	.0
73	1	.0
TOTAL	9269	100.0

APPENDIX C

ITEM PARAMETERS OF THE 3PL MODEL FOR FORMS A AND B

Table C1. Item Parameters of the 3PL Model for Form A

Item	*a*	*b*	*c*
1	0.46	-0.99	0.00
2	0.47	1.12	0.19
3	1.37	1.15	0.22
4	1.60	0.26	0.30
5	0.95	-0.64	0.21
6	0.76	-0.48	0.12
7	1.85	0.00	0.28
8	1.98	2.12	0.08
9	1.54	-1.09	0.13
10	1.28	-0.31	0.30
11	1.11	1.06	0.14
12	0.51	-0.92	0.00
13	1.22	0.59	0.18
14	1.62	-0.55	0.31
15	0.93	-0.47	0.23
16	1.73	-0.42	0.34
17	0.77	0.27	0.14
18	1.78	1.42	0.12
19	1.33	0.46	0.37
20	0.59	-1.03	0.25
21	0.95	-0.07	0.19
22	1.61	0.55	0.22
23	0.97	0.57	0.25

Table C1 (continued)

Item	*a*	*b*	*c*
24	1.68	0.78	0.32
25	0.78	-1.57	0.25
26	0.90	-1.72	0.00
27	1.62	0.43	0.21
28	1.09	0.38	0.21
29	0.23	3.70	0.25
30	1.09	0.02	0.18
31	0.77	-1.85	0.00
32	0.87	0.70	0.27
33	1.59	0.89	0.31
34	1.75	-0.44	0.24
35	1.38	0.63	0.25
36	0.74	-1.03	0.00
37	1.28	-0.02	0.26
38	1.15	-0.50	0.21
39	1.55	-1.42	0.07
40	0.88	1.50	0.20
41	1.97	0.67	0.23
42	1.94	1.17	0.27
43	1.32	0.97	0.24
44	1.22	0.62	0.23
45	0.72	-0.27	0.30
46	1.87	1.55	0.25
47	1.66	0.12	0.22
48	1.80	0.90	0.17

Table C1 (continued)

Item	*a*	*b*	*c*
49	1.15	2.00	0.26
50	1.06	0.19	0.25
51	2.13	2.08	0.21
52	1.67	0.78	0.23
53	1.12	-0.24	0.14
54	1.63	0.72	0.25
55	2.09	-0.34	0.33
56	1.11	1.12	0.11
57	1.57	-0.41	0.25
58	1.07	0.24	0.19
59	1.14	0.83	0.25
60	0.85	0.05	0.16
61	1.00	0.73	0.11
62	1.70	-0.65	0.20
63	2.16	-0.23	0.20
64	2.25	1.36	0.23
65	1.71	-0.26	0.21
66	1.55	0.83	0.20
67	1.76	-0.09	0.35
68	1.83	-0.62	0.30
69	1.68	-0.57	0.23
70	1.62	-0.12	0.22
71	2.10	-0.10	0.13
72	0.99	0.10	0.33
73	1.61	-0.37	0.26

Table C1 (continued)

Item	*a*	*b*	*c*
74	1.56	2.47	0.13
75	1.20	1.11	0.25
76	0.60	-0.93	0.00
77	1.01	1.06	0.17
78	1.79	0.20	0.26
79	0.32	0.15	0.00
80	1.93	-0.32	0.22
81	1.48	-0.94	0.26
82	0.18	2.10	0.00
83	1.34	0.81	0.32
84	0.93	-1.22	0.00
85	1.28	-0.22	0.24
86	1.72	0.60	0.31
87	1.29	0.47	0.24
88	0.98	-1.13	0.00
89	1.03	-0.73	0.30
90	1.00	2.12	0.37
91	1.51	0.95	0.24
92	1.37	0.27	0.17
93	0.66	-0.56	0.25
94	1.27	0.59	0.11
95	1.42	0.45	0.16
96	1.64	1.09	0.13
97	1.72	0.72	0.23
98	1.45	0.32	0.10

Table C1 (continued)

Item	*a*	*b*	*c*
99	1.15	1.22	0.23
100	1.33	-0.29	0.28
101	0.95	-0.67	0.25
102	0.36	-0.74	0.00
103	0.98	-0.27	0.34
104	0.89	0.79	0.33
105	1.52	0.57	0.26
106	1.01	0.95	0.18
107	0.73	0.83	0.17
108	1.55	1.28	0.38
109	2.05	2.16	0.12
110	1.45	1.08	0.23
111	1.53	0.87	0.26
112	1.09	1.77	0.24
113	0.92	-0.14	0.28
114	1.53	1.53	0.17
115	0.95	-0.77	0.24
116	1.18	1.35	0.17

Table C2. Item Parameters of the 3PL Model for Form B

Item	*a*	*b*	*c*
1	0.75	-0.55	0.08
2	0.45	-0.17	0.00
3	1.11	1.53	0.19
4	1.57	0.00	0.24
5	0.84	-0.75	0.30
6	1.39	0.10	0.23
7	1.54	-0.11	0.33
8	1.07	0.81	0.33
9	0.74	-0.39	0.36
10	0.53	-1.02	0.25
11	0.81	-0.34	0.18
12	1.31	-0.29	0.29
13	2.16	1.91	0.31
14	0.34	-1.70	0.00
15	1.02	0.62	0.19
16	1.33	-0.11	0.21
17	0.77	0.36	0.15
18	0.72	1.57	0.23
19	1.11	0.10	0.16
20	0.86	0.26	0.25
21	1.62	1.49	0.14
22	1.01	0.98	0.27
23	0.94	1.25	0.30

Table C2 (continued)

Item	*a*	*b*	*c*
24	1.29	0.97	0.36
25	2.18	0.59	0.30
26	1.32	-0.13	0.24
27	0.71	-0.59	0.25
28	1.17	0.80	0.32
29	0.94	-0.95	0.10
30	0.84	1.20	0.21
31	1.03	-0.60	0.27
32	1.00	-1.14	0.12
33	1.33	0.50	0.32
34	1.52	0.24	0.19
35	1.23	1.75	0.24
36	0.47	0.59	0.12
37	1.07	-0.06	0.19
38	1.40	0.20	0.33
39	1.20	0.53	0.29
40	1.20	1.50	0.16
41	1.37	0.67	0.14
42	1.79	1.20	0.20
43	1.57	0.04	0.23
44	1.58	0.45	0.29
45	2.53	1.52	0.26
46	1.70	0.56	0.27
47	0.92	-0.08	0.11
48	2.48	1.43	0.28

Table C2 (continued)

Item	*a*	*b*	*c*
49	1.34	0.77	0.23
50	1.40	1.03	0.26
51	0.84	0.21	0.14
52	1.06	0.93	0.21
53	1.64	0.71	0.22
54	0.60	2.55	0.25
55	2.21	0.33	0.24
56	0.45	2.23	0.17
57	1.65	0.81	0.19
58	1.23	-0.88	0.14
59	2.11	2.03	0.21
60	1.51	1.31	0.18
61	2.12	1.40	0.22
62	1.32	-0.91	0.14
63	1.16	0.36	0.15
64	2.54	-0.26	0.16
65	1.51	0.13	0.24
66	1.22	1.41	0.18
67	1.86	0.84	0.21
68	1.15	-0.51	0.12
69	1.72	1.75	0.26
70	1.63	0.21	0.24
71	1.29	0.34	0.21
72	1.67	-0.86	0.22
73	1.13	1.64	0.16

Table C2 (continued)

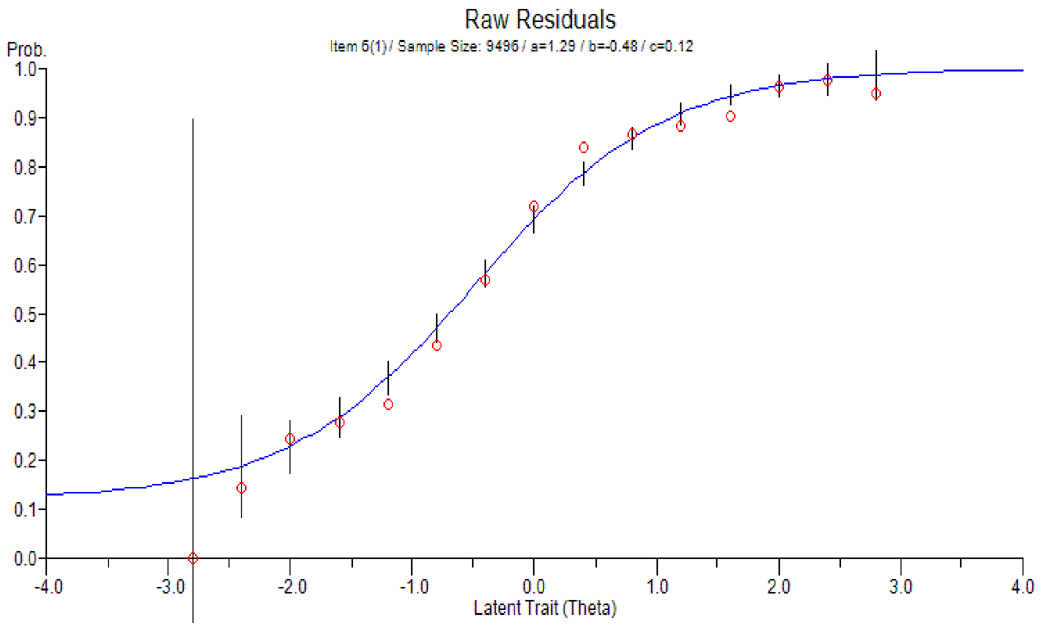
Item	*a*	*b*	*c*
74	0.71	-1.46	0.00
75	1.70	1.05	0.19
76	1.87	0.33	0.19
77	2.20	1.24	0.21
78	1.54	0.13	0.19
79	1.79	2.02	0.25
80	0.78	1.86	0.17
81	1.30	0.71	0.24
82	0.73	-1.38	0.00
83	2.26	2.49	0.15
84	1.56	-0.14	0.26
85	1.04	-1.17	0.12
86	1.34	1.29	0.30
87	1.49	0.76	0.36
88	1.80	2.20	0.17
89	1.71	0.29	0.20
90	0.94	-0.79	0.01
91	0.71	0.05	0.16
92	1.70	1.00	0.24
93	1.12	2.20	0.11
94	0.96	0.50	0.14
95	0.95	-0.94	0.00
96	0.76	1.02	0.26
97	1.67	-0.06	0.25
98	1.12	-0.30	0.23

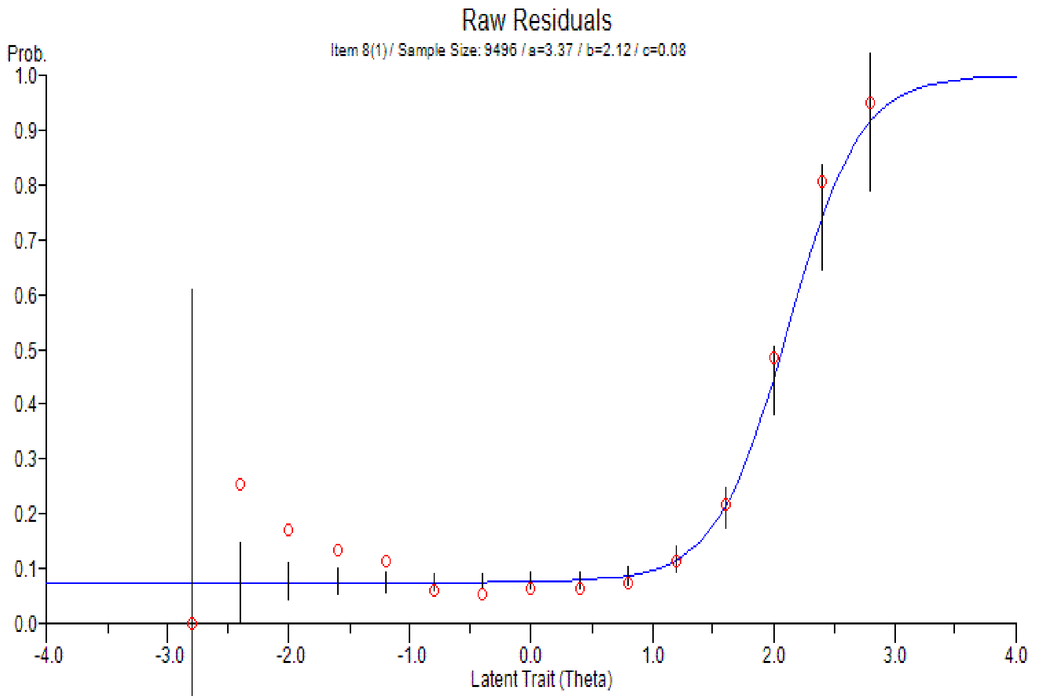
Table C2 (continued)

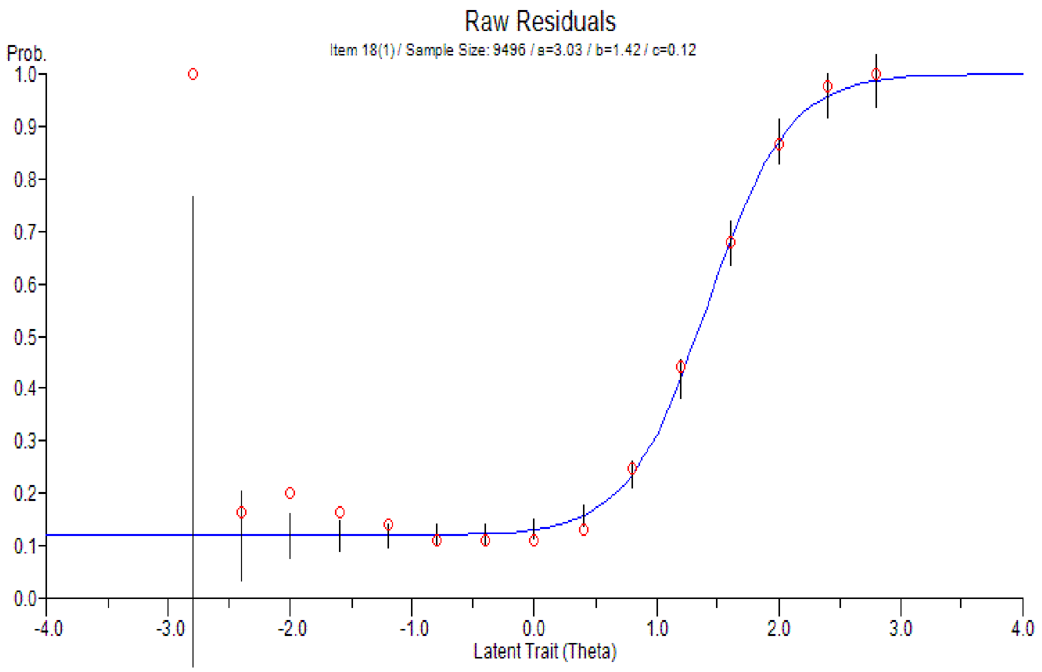
Item	*a*	*b*	*c*
99	1.15	1.56	0.33
100	1.34	0.00	0.30
101	1.24	0.58	0.16
102	1.05	1.28	0.14
103	0.69	1.05	0.10
104	1.33	1.29	0.23
105	1.36	0.85	0.21
106	1.93	1.77	0.22
107	0.25	-1.21	0.00
108	1.53	0.18	0.24
109	1.97	0.92	0.21
110	1.06	2.27	0.29
111	0.40	-0.30	0.00
112	1.34	0.33	0.20
113	1.43	1.16	0.21
114	1.48	0.24	0.35
115	0.88	0.81	0.13
116	1.19	0.57	0.38
117	1.79	0.63	0.19
118	1.33	1.12	0.32
119	0.93	0.75	0.23

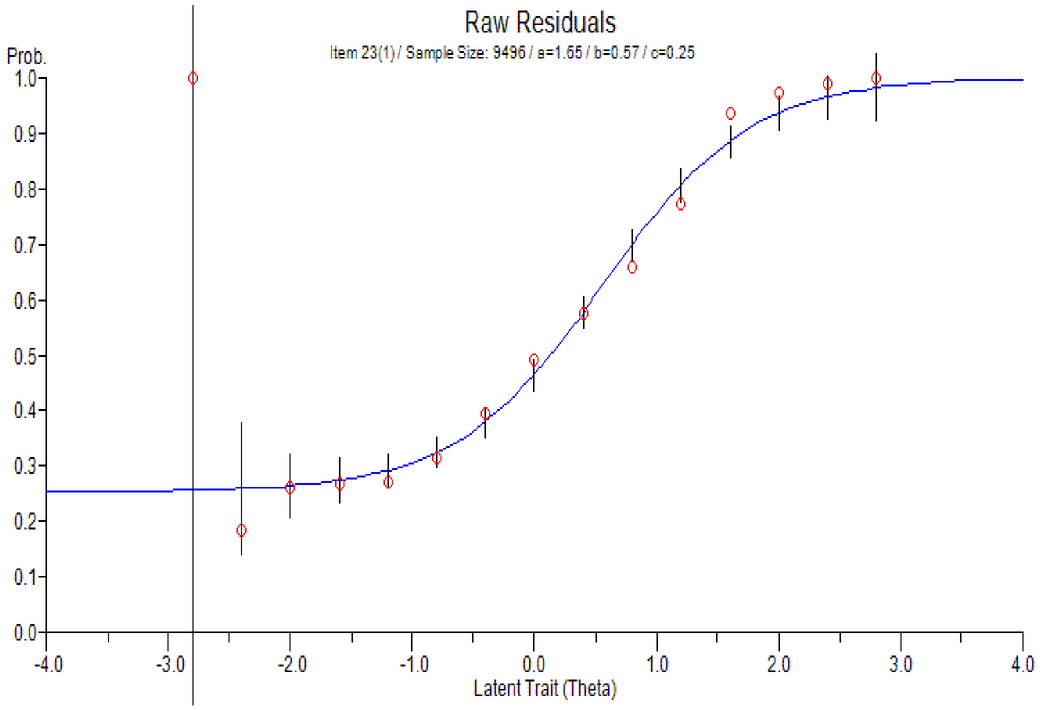
APPENDIX D

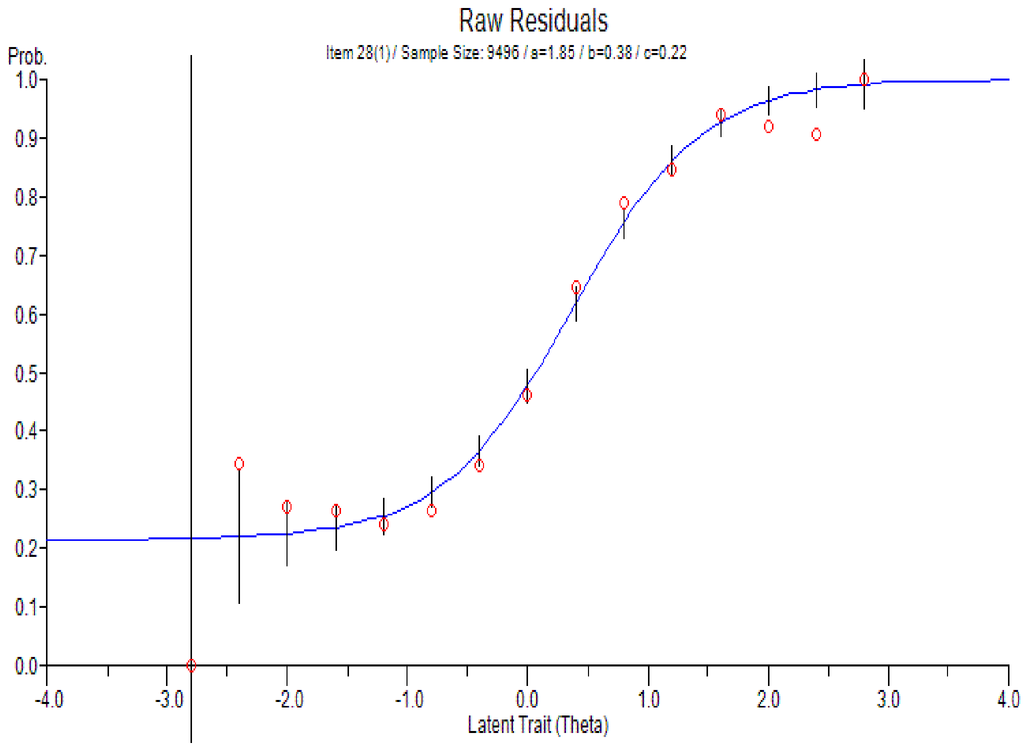
PLOTS OF THE ITEM RESIDUALS FOR THE MISFIT ITEMS FOR THE 3PL MODEL ON FORMS A AND B

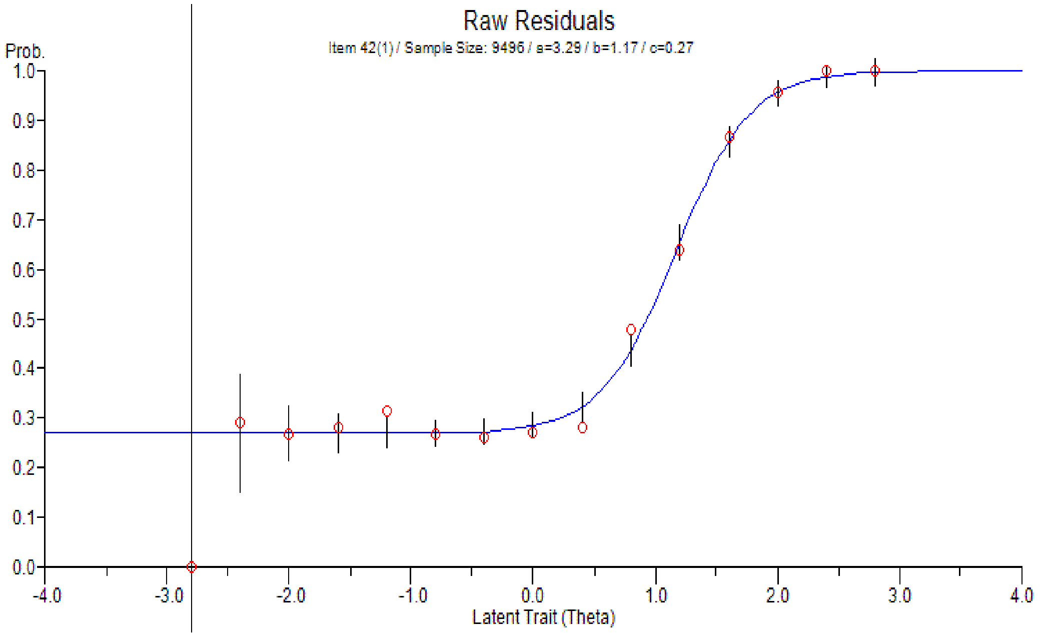

Raw Residuals
Item 6(1) / Sample Size: 9496 / a=1.29 / b=-0.48 / c=0.12
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

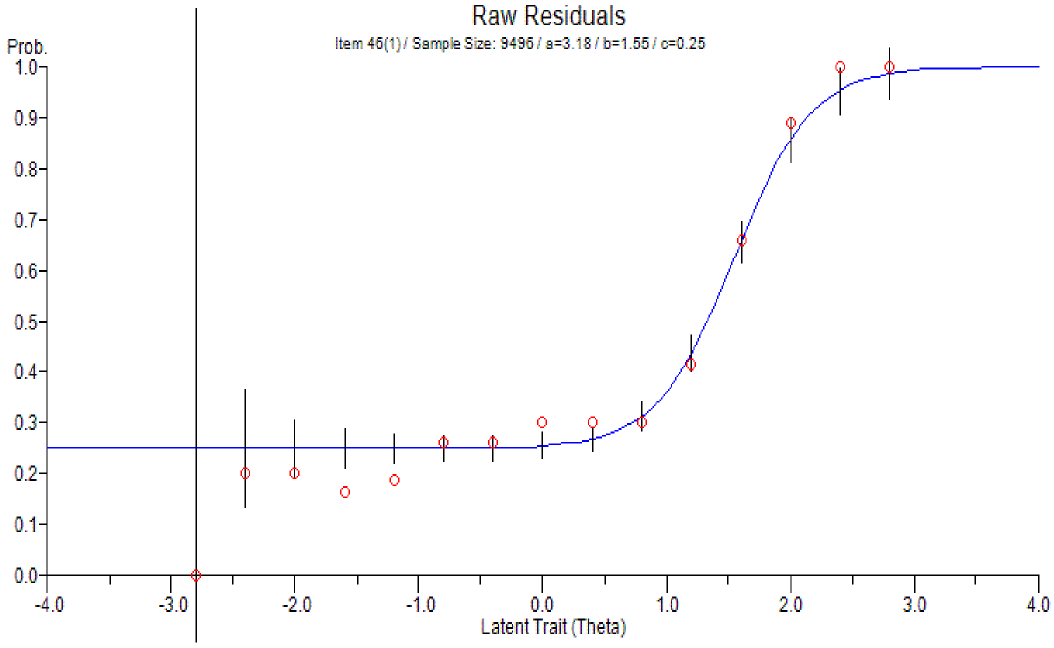

Raw Residuals
Item 8(1) / Sample Size: 9496 / a=3.37 / b=2.12 / c=0.08
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

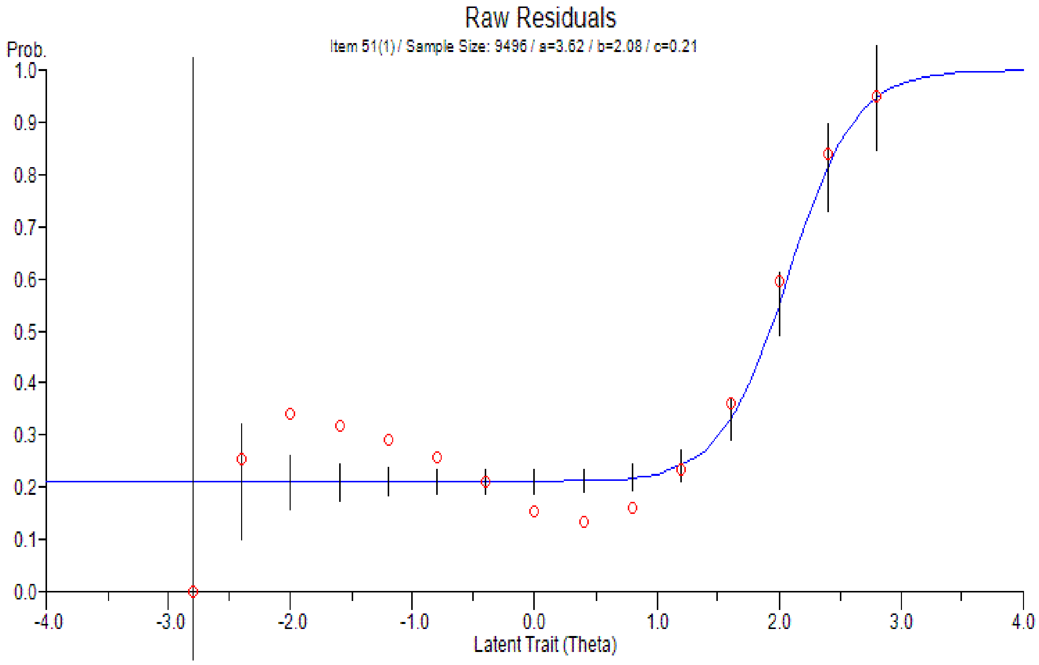

Raw Residuals
Item 18(1) / Sample Size: 9496 / a=3.03 / b=1.42 / c=0.12
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

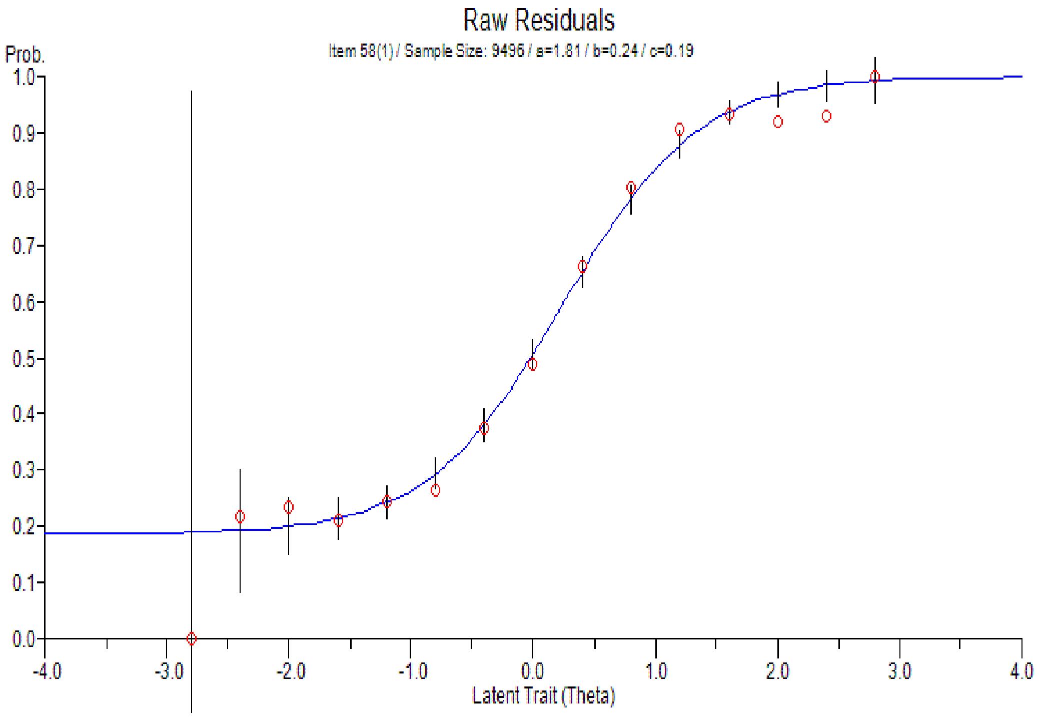

Raw Residuals
Item 23(1) / Sample Size: 9496 / a=1.65 / b=0.57 / c=0.25
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

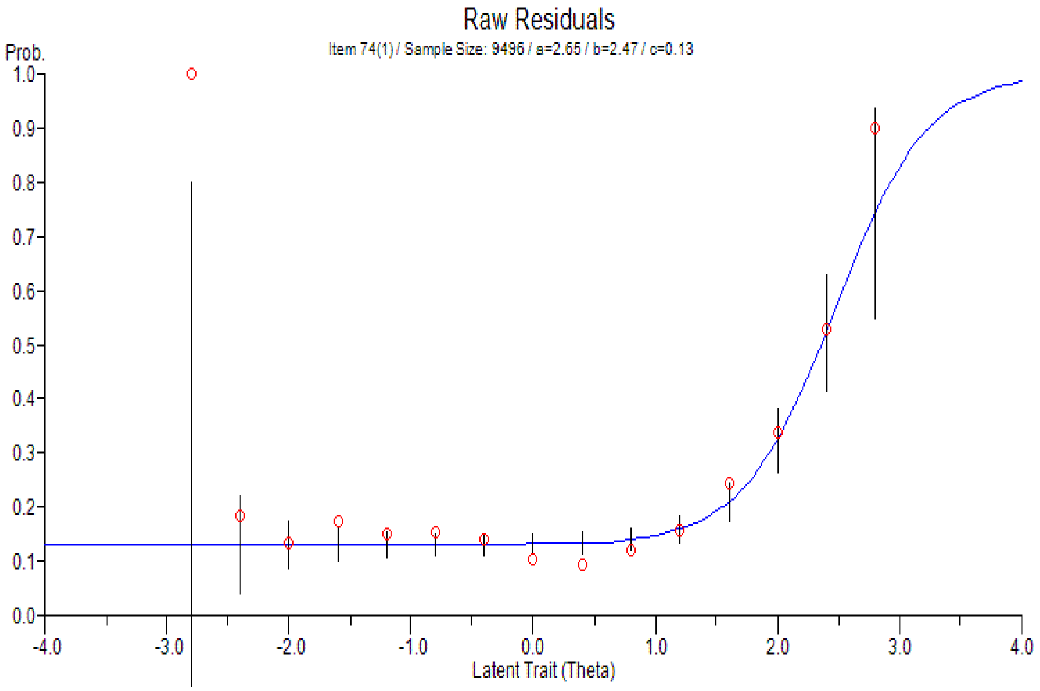

Raw Residuals
Item 28(1) / Sample Size: 9496 / a=1.85 / b=0.38 / c=0.22
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

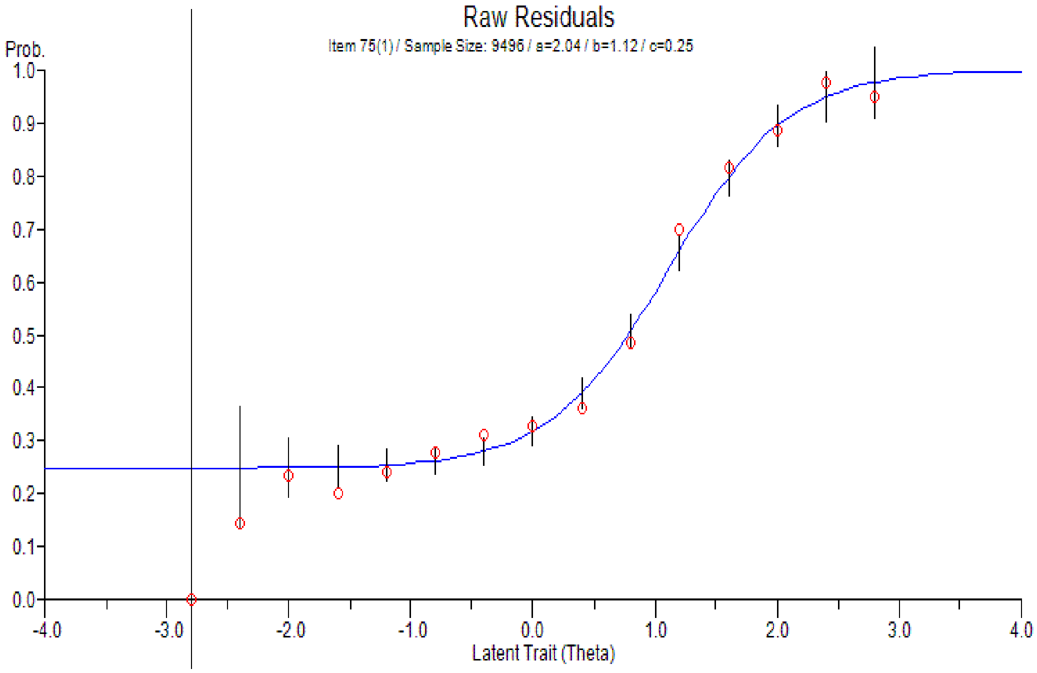

Raw Residuals
Item 42(1) / Sample Size: 9496 / a=3.29 / b=1.17 / c=0.27
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

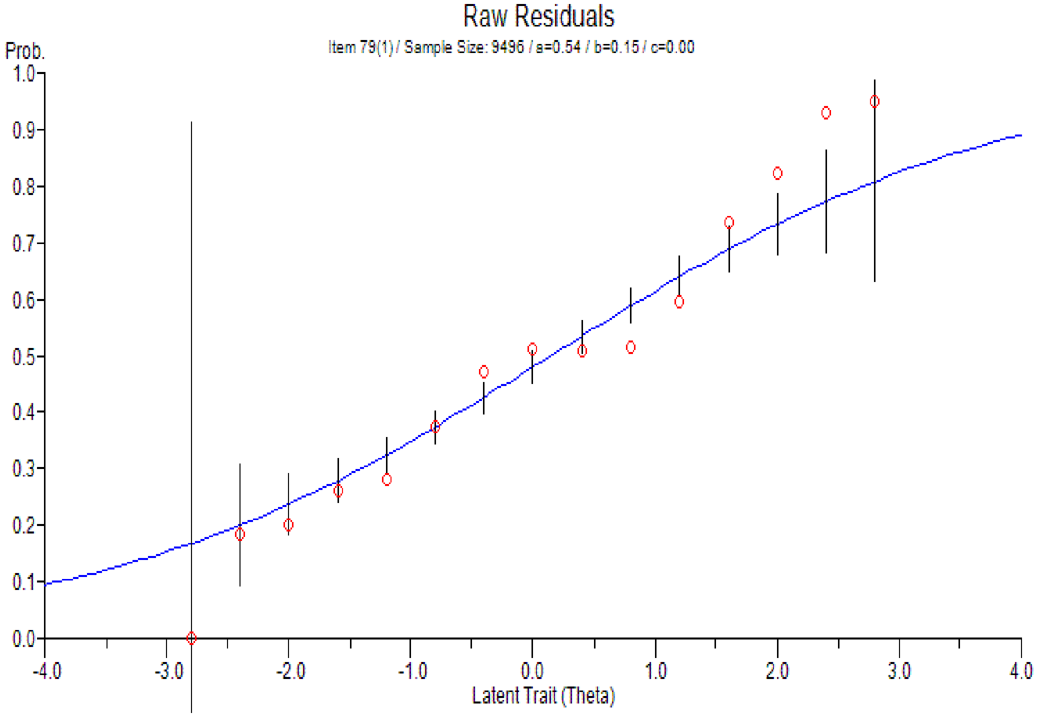

Raw Residuals
Item 46(1) / Sample Size: 9496 / a=3.18 / b=1.55 / c=0.25
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

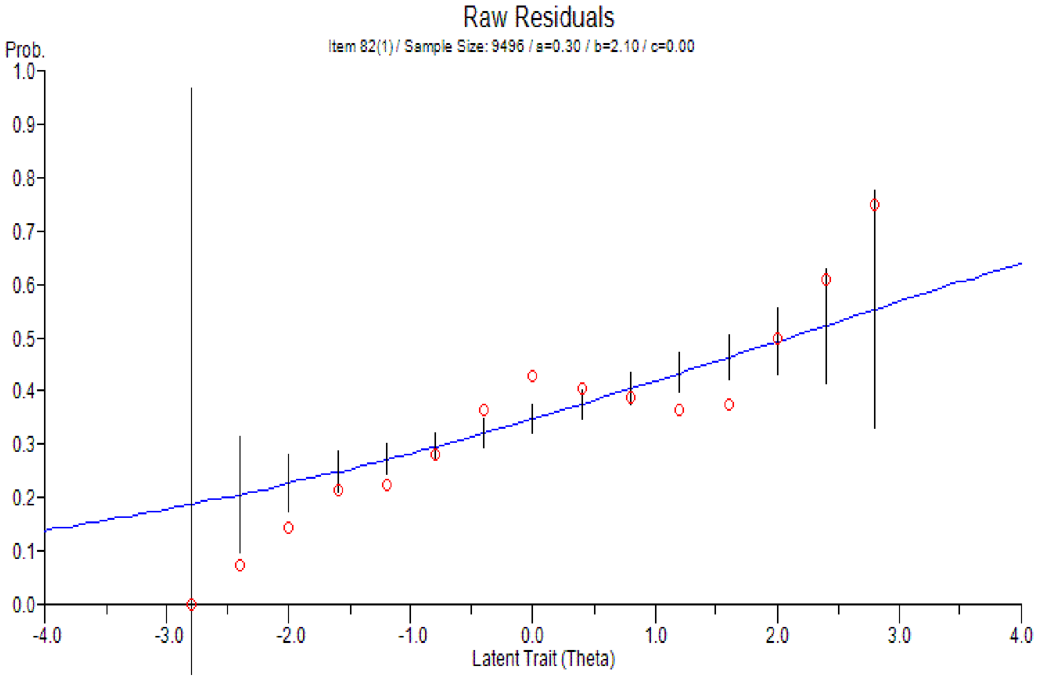

Raw Residuals
Item 51(1) / Sample Size: 9496 / a=3.62 / b=2.08 / c=0.21
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

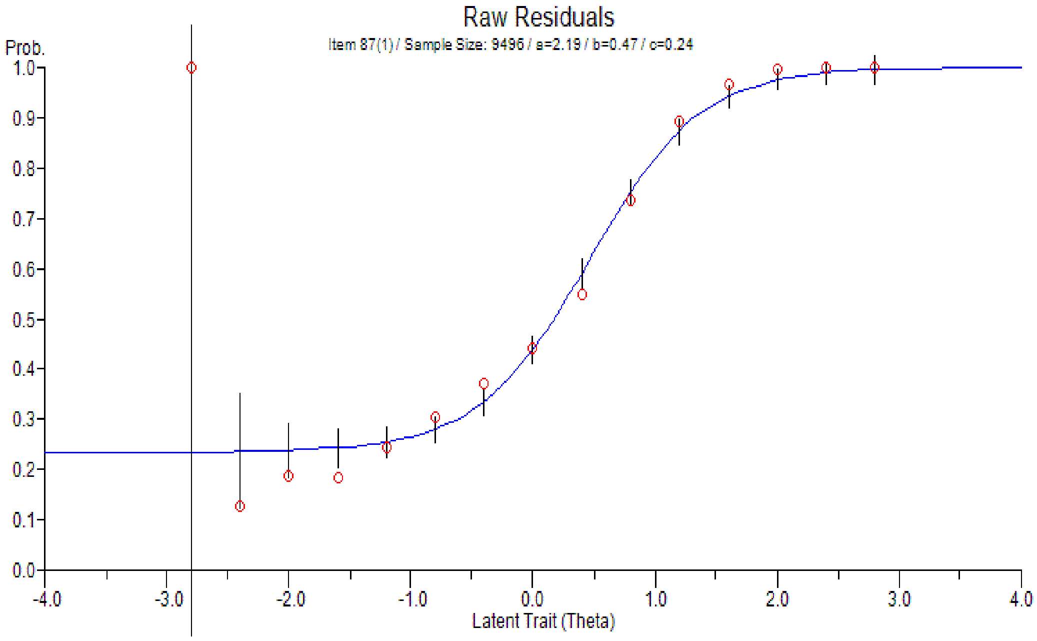

Raw Residuals
Item 58(1) / Sample Size: 9496 / a=1.81 / b=0.24 / c=0.19
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

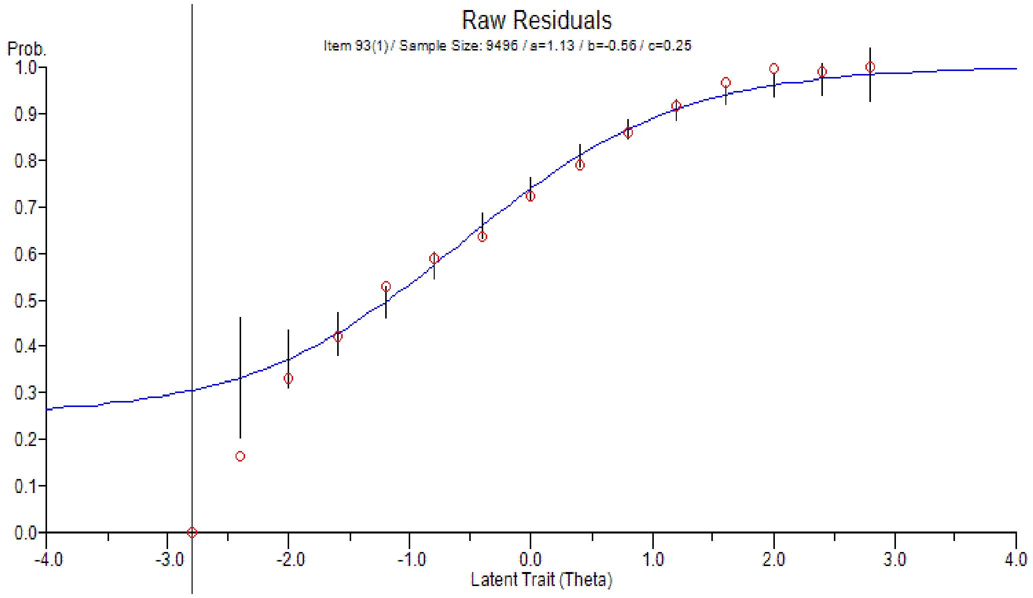

Raw Residuals
Item 74(1) / Sample Size: 9496 / a=2.65 / b=2.47 / c=0.13
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

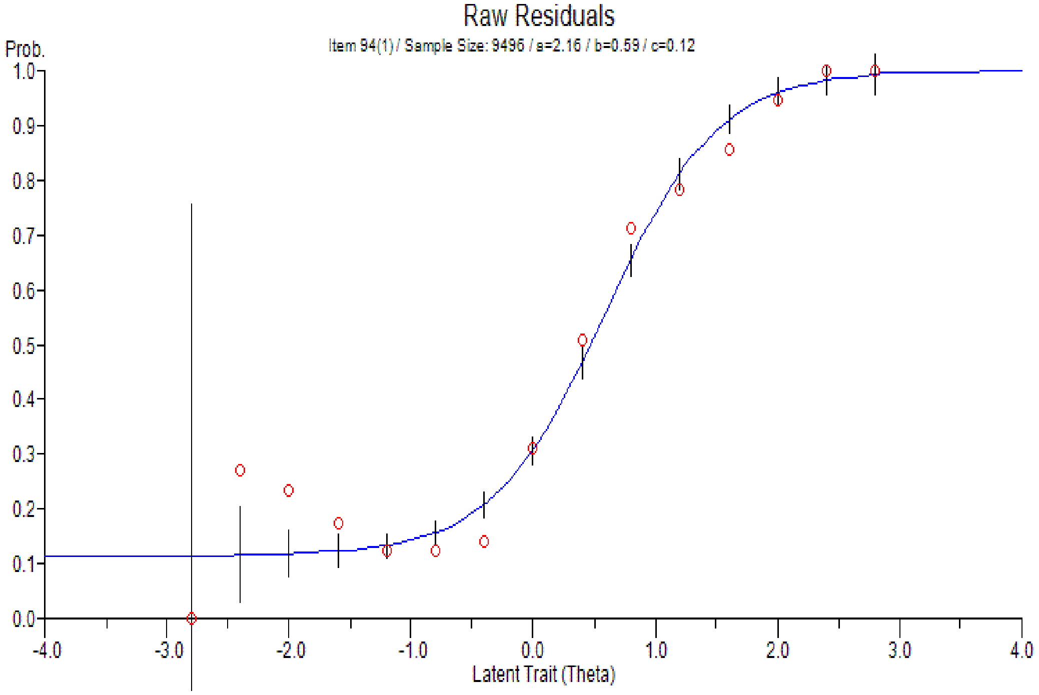

Raw Residuals
Item 75(1) / Sample Size: 9496 / a=2.04 / b=1.12 / c=0.25
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

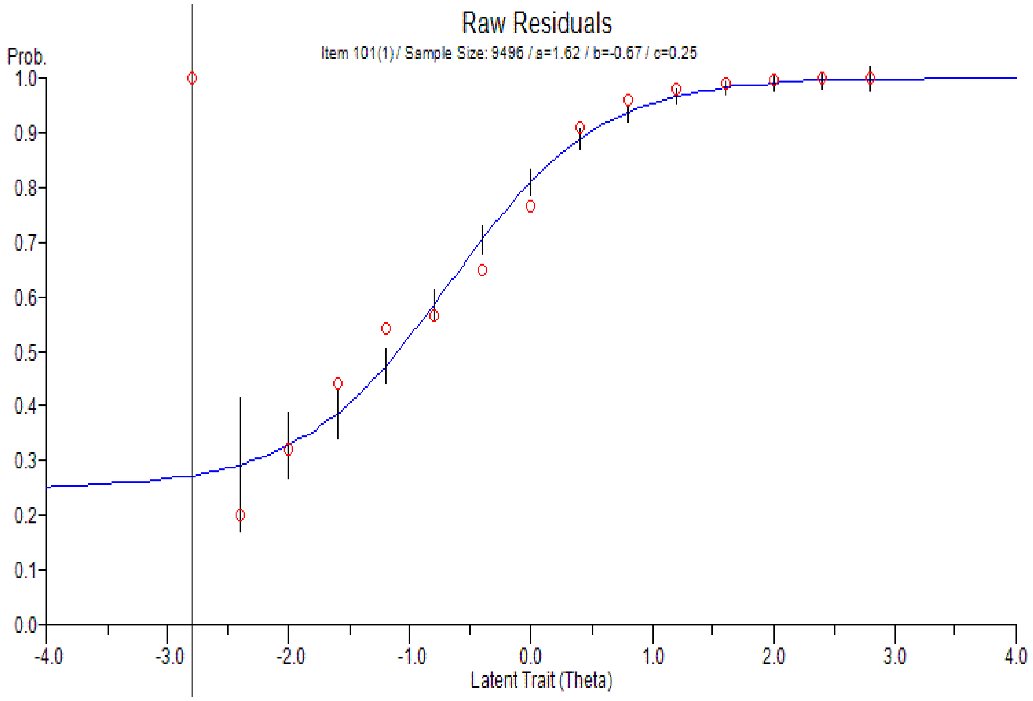

Raw Residuals
Item 79(1) / Sample Size: 9496 / a=0.54 / b=0.15 / c=0.00
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

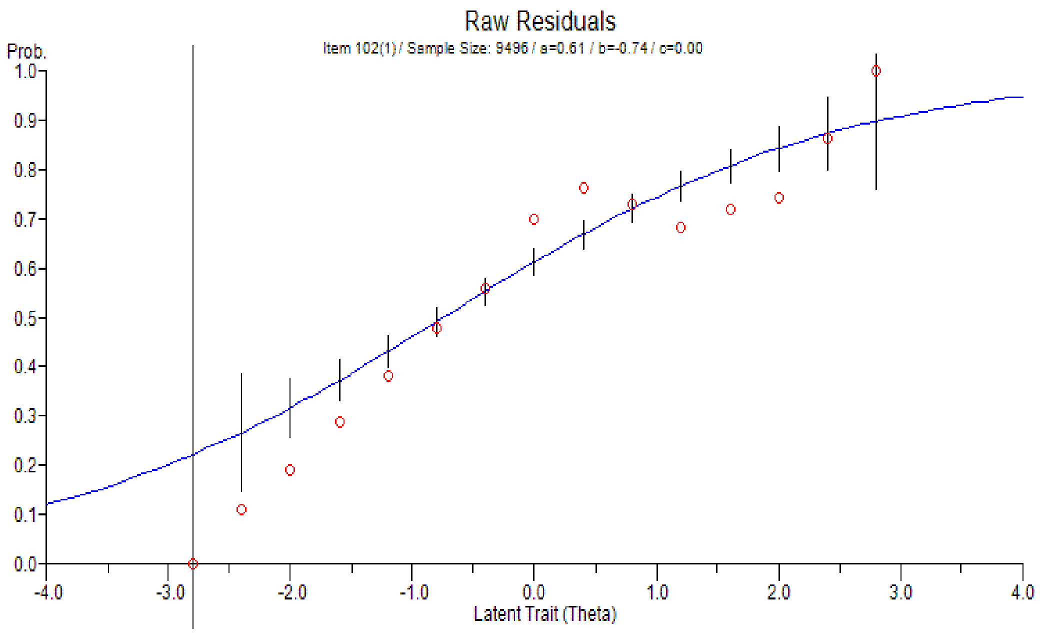

Raw Residuals
Item 82(1) / Sample Size: 9496 / a=0.30 / b=2.10 / c=0.00
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

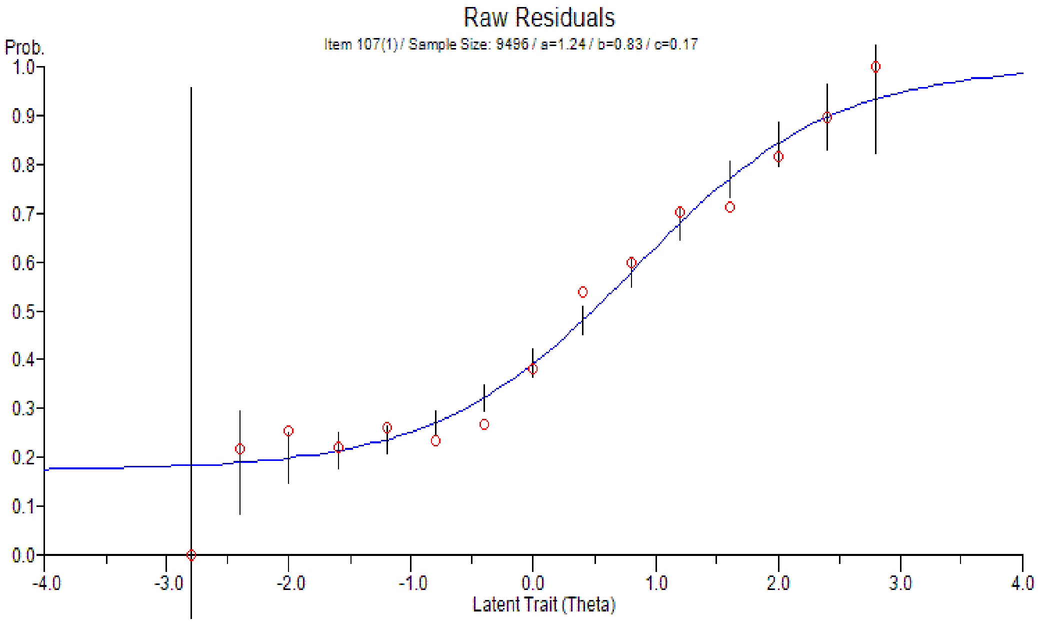

Raw Residuals
Item 87(1) / Sample Size: 9496 / a=2.19 / b=0.47 / c=0.24
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

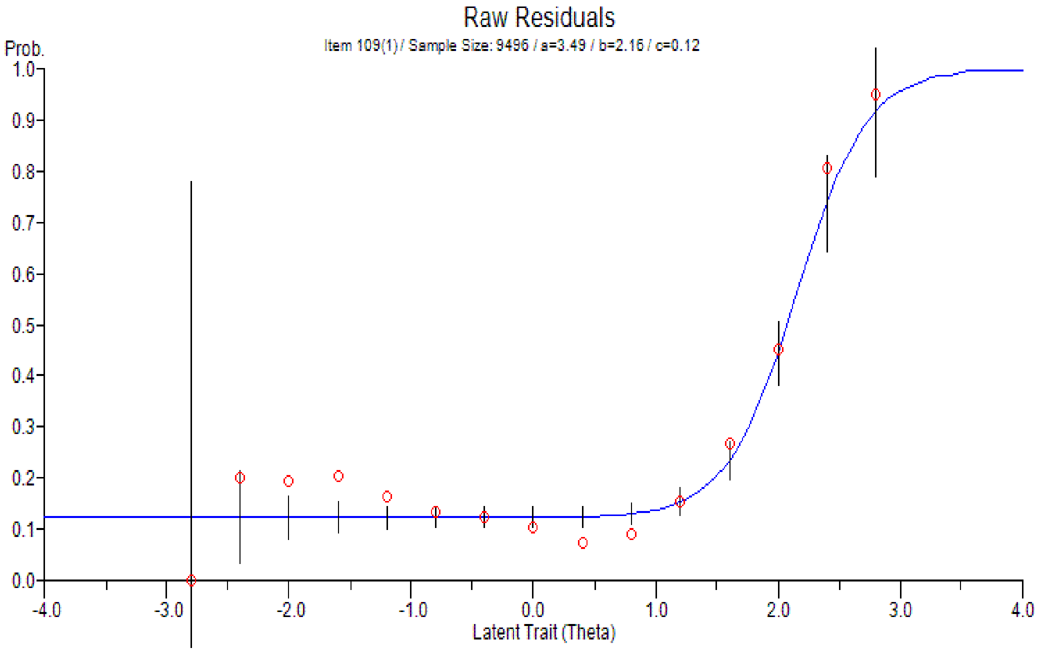

Raw Residuals
Item 93(1) / Sample Size: 9496 / a=1.13 / b=-0.56 / c=0.25
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)


Raw Residuals
Item 94(1) / Sample Size: 9496 / a=2.16 / b=0.59 / c=0.12
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)


Raw Residuals
Item 101(1) / Sample Size: 9496 / a=1.62 / b=-0.67 / c=0.25
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)


Raw Residuals
Item 102(1) / Sample Size: 9496 / a=0.61 / b=-0.74 / c=0.00
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)


Raw Residuals
Item 107(1) / Sample Size: 9496 / a=1.24 / b=0.83 / c=0.17
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)


Raw Residuals
Item 109(1) / Sample Size: 9496 / a=3.49 / b=2.16 / c=0.12
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

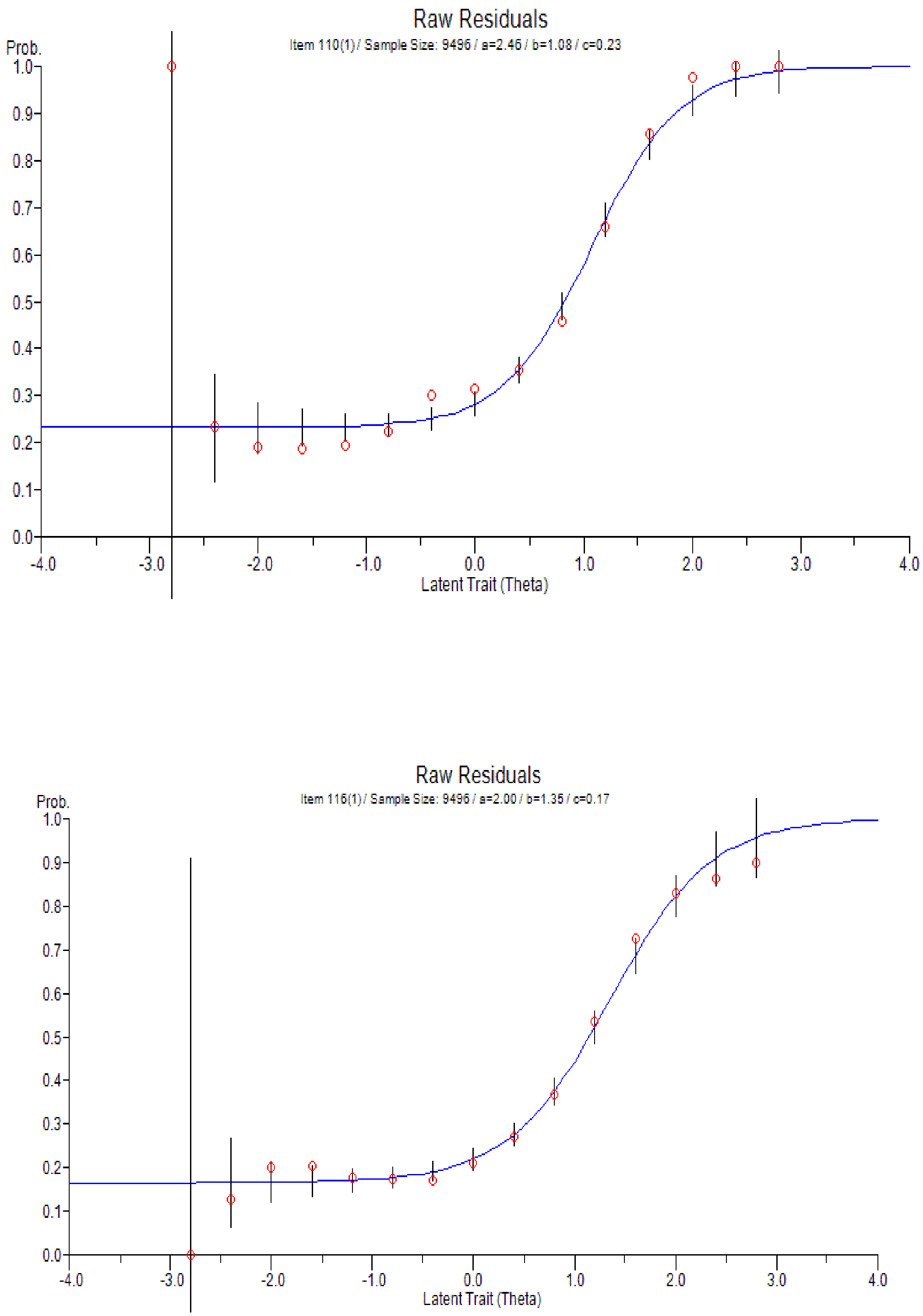
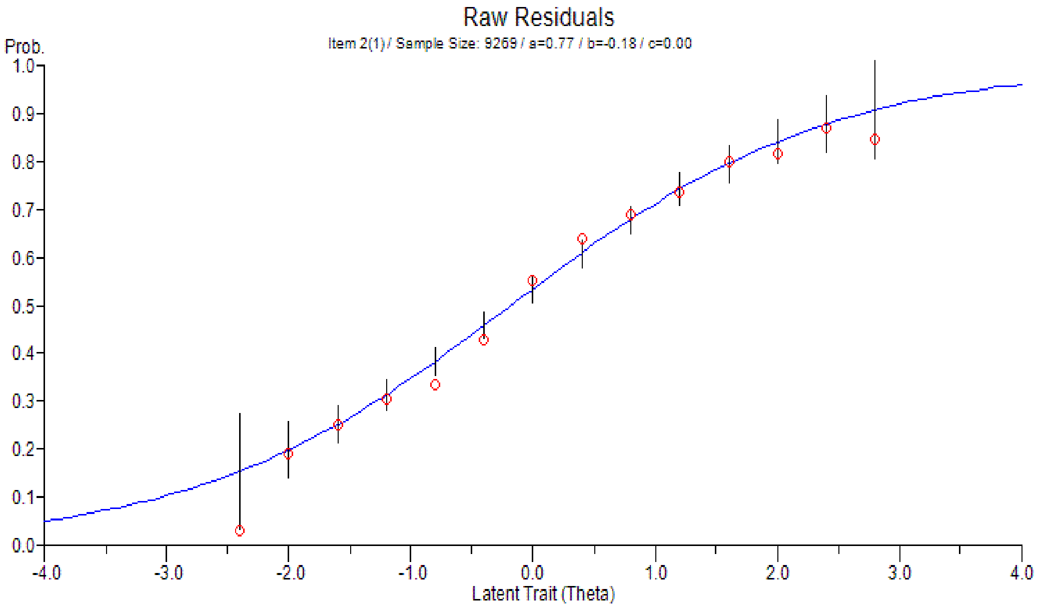
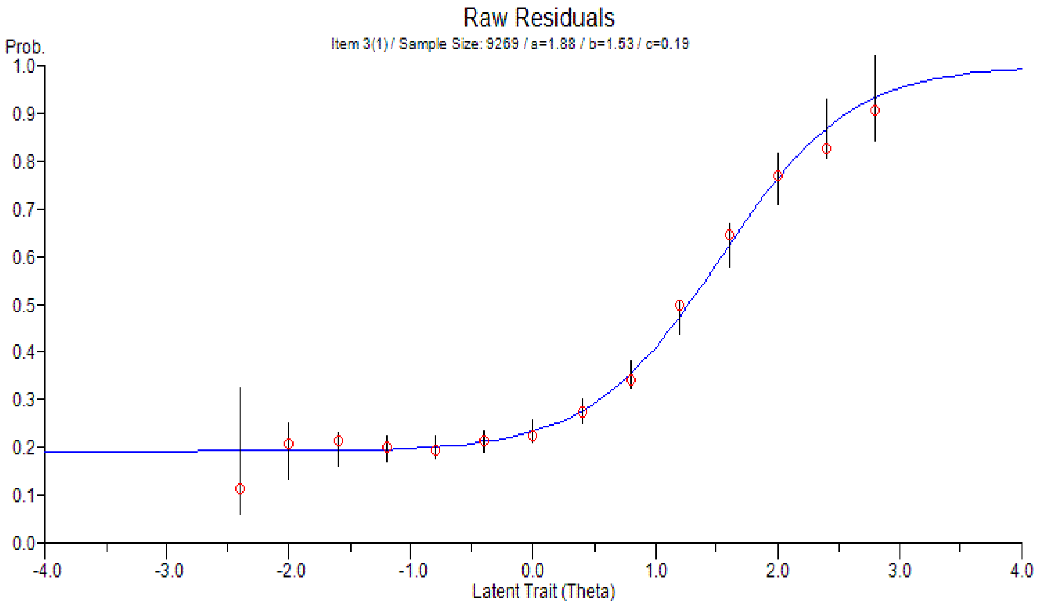
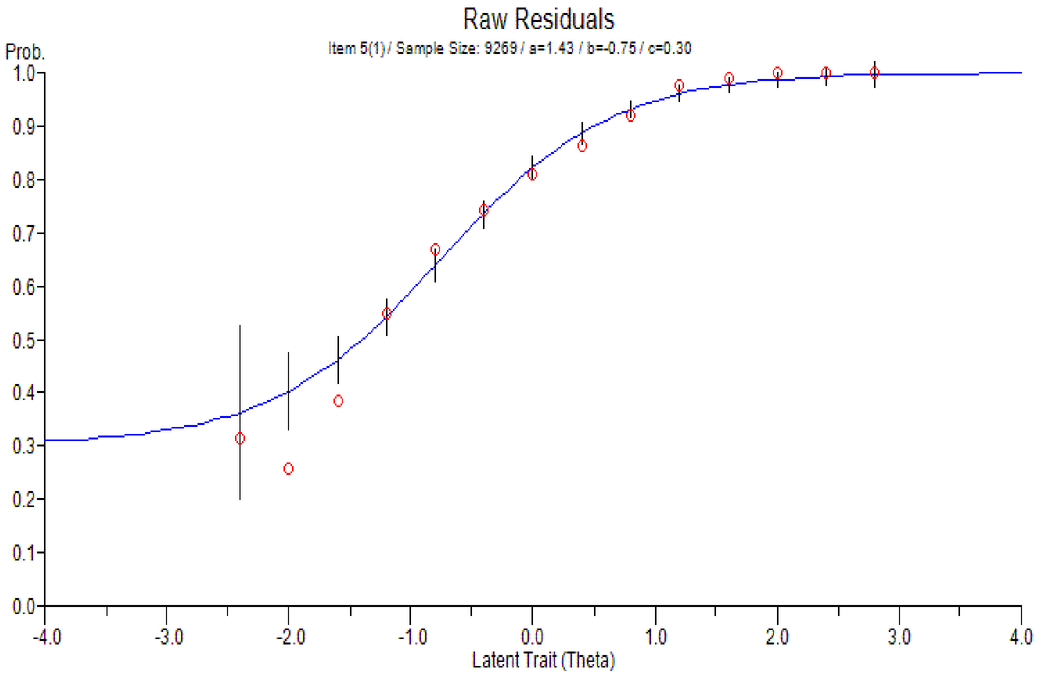
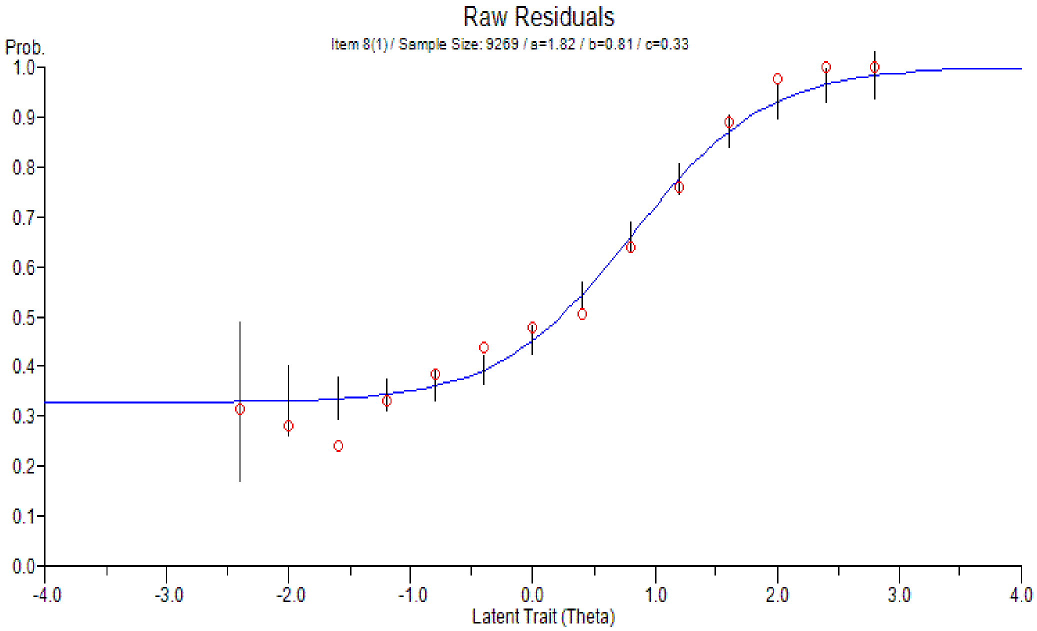


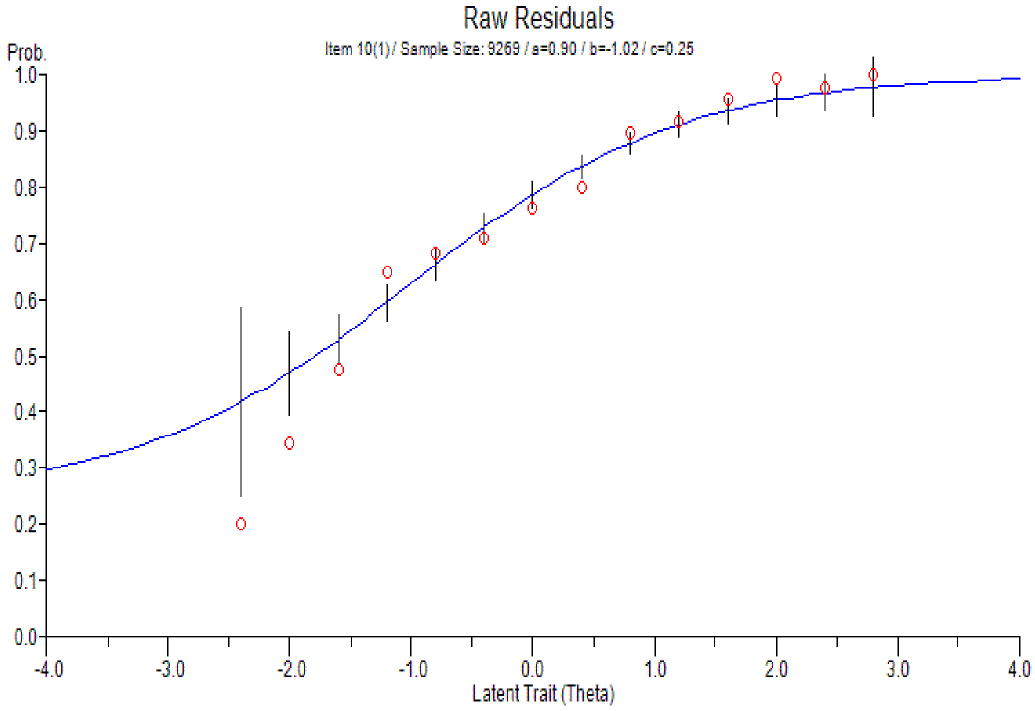
Figure D1. Plots of the Item Residuals for the Misfit Items for the 3PL model on Form A

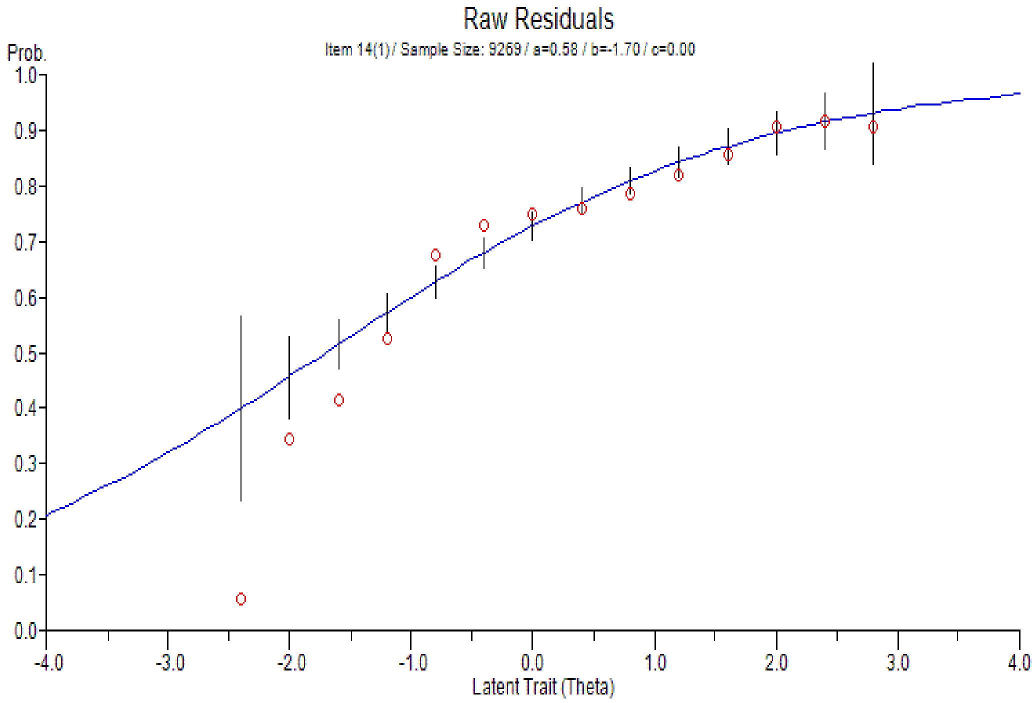

Raw Residuals
Item 2(1) / Sample Size: 9269 / a=0.77 / b=-0.18 / c=0.00
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

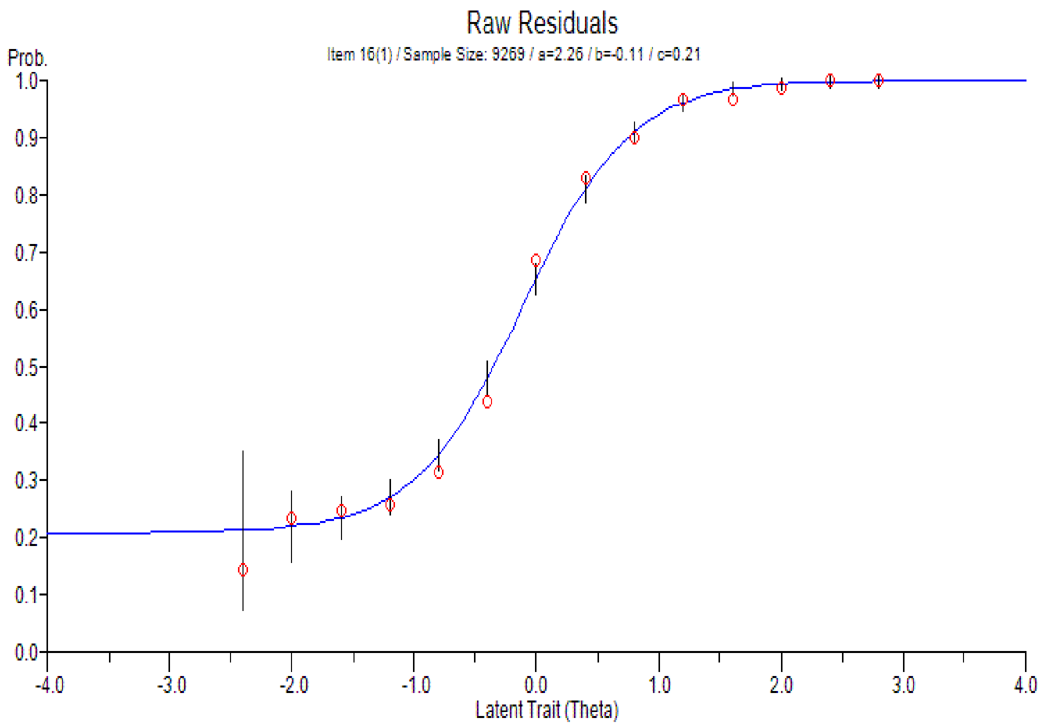

Raw Residuals
Item 3(1) / Sample Size: 9269 / a=1.88 / b=1.53 / c=0.19
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

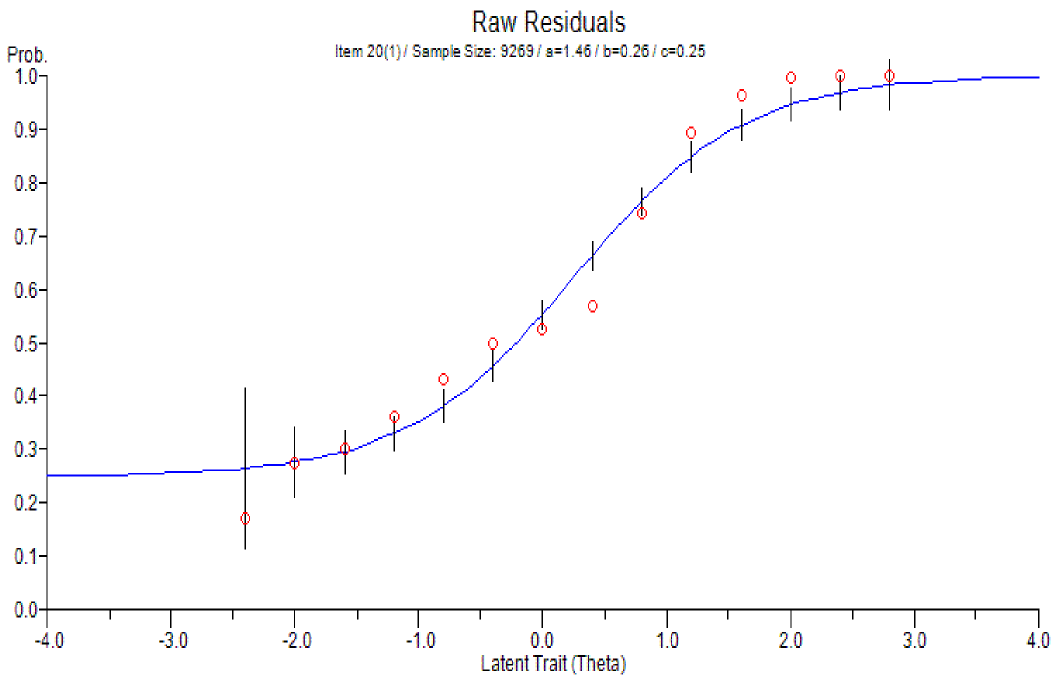

Raw Residuals
Item 5(1) / Sample Size: 9269 / a=1.43 / b=-0.75 / c=0.30
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

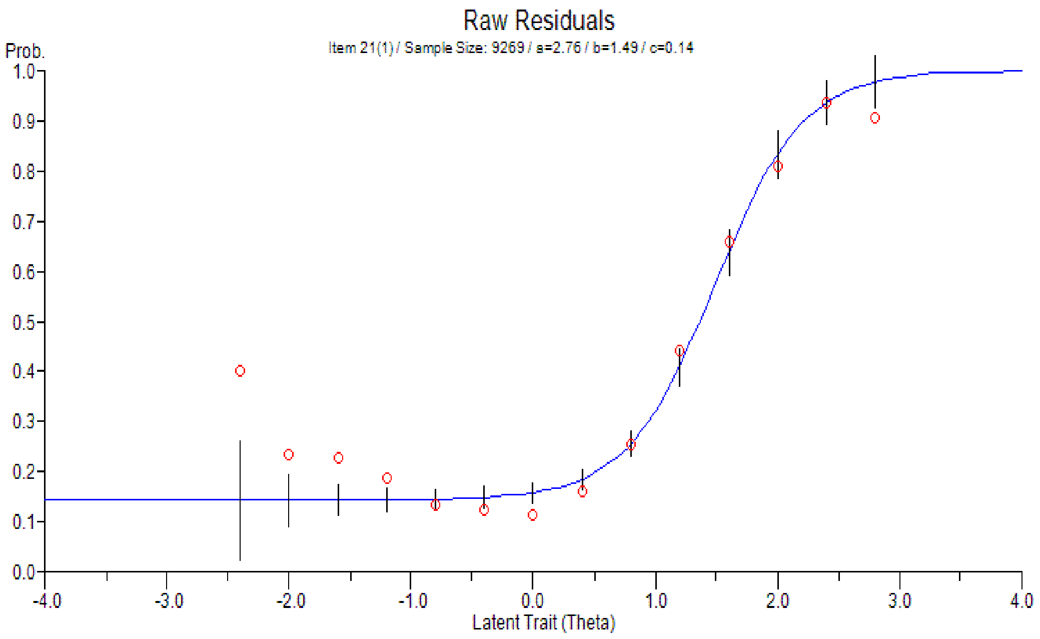

Raw Residuals
Item 8(1) / Sample Size: 9269 / a=1.82 / b=0.81 / c=0.33
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

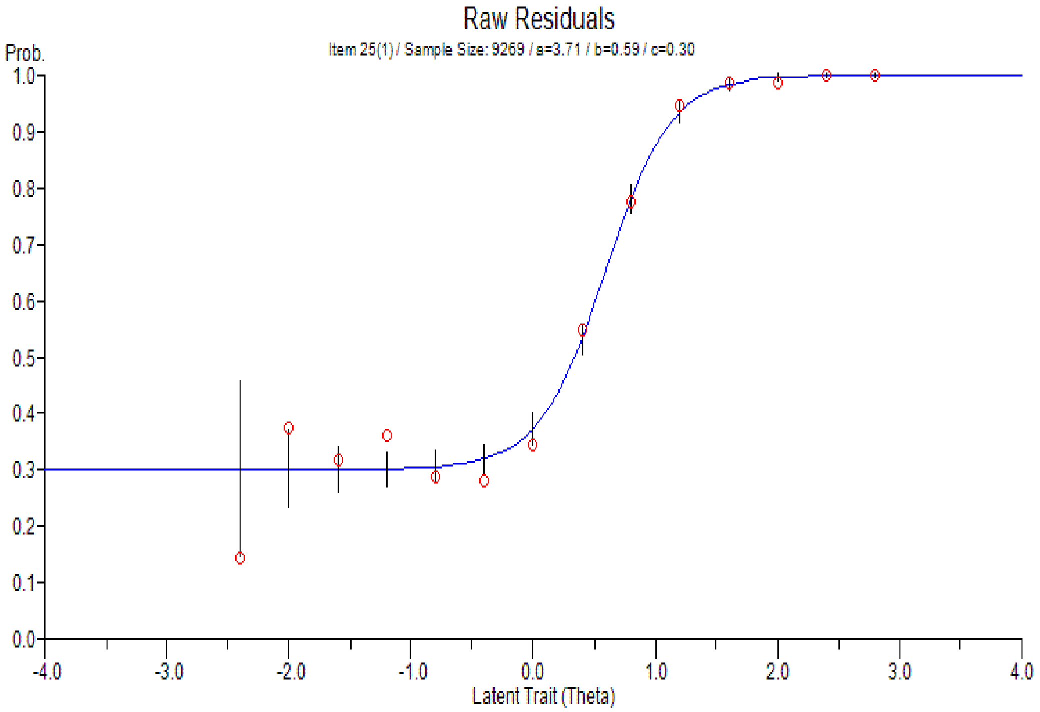

Raw Residuals
Item 10(1) / Sample Size: 9269 / a=0.90 / b=-1.02 / c=0.25
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

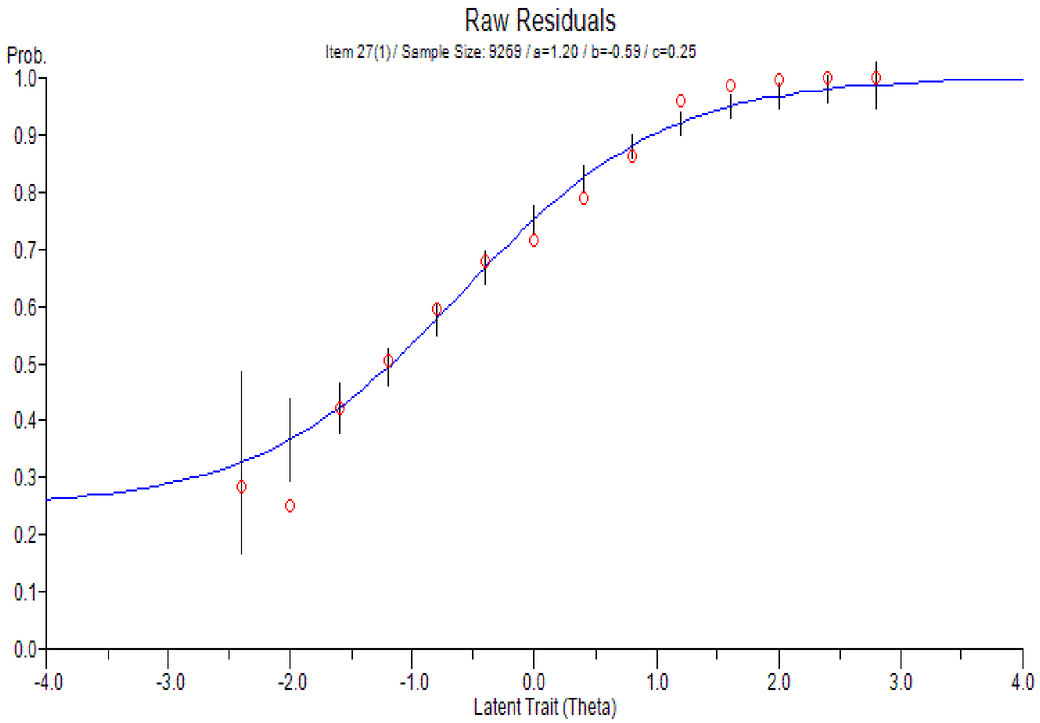

Raw Residuals
Item 14(1) / Sample Size: 9269 / a=0.58 / b=-1.70 / c=0.00
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

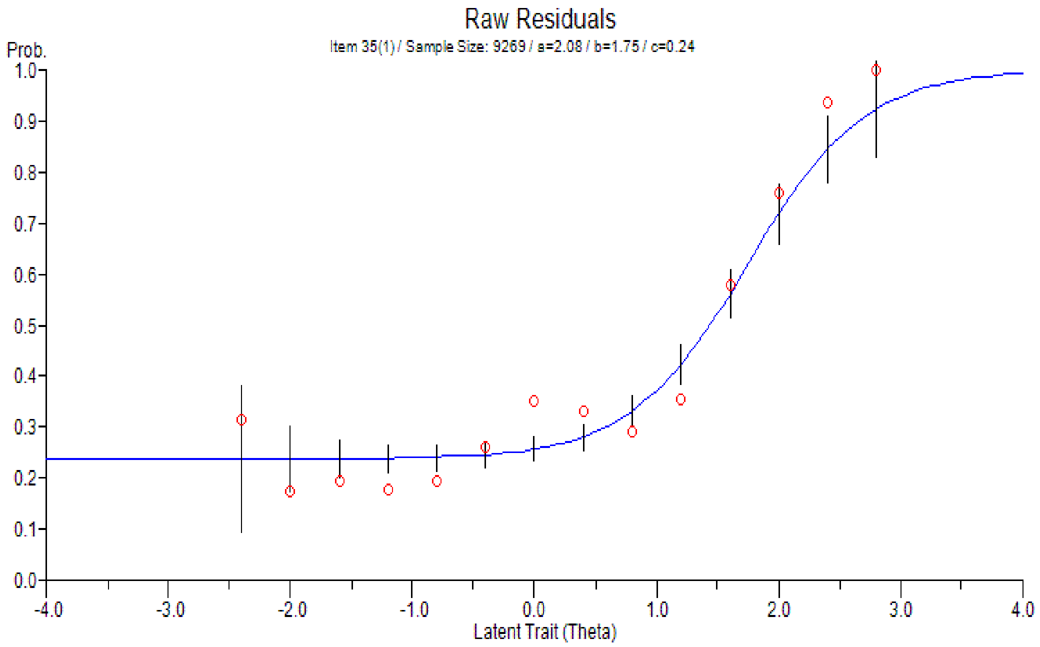

Raw Residuals
Item 16(1) / Sample Size: 9269 / a=2.26 / b=-0.11 / c=0.21
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

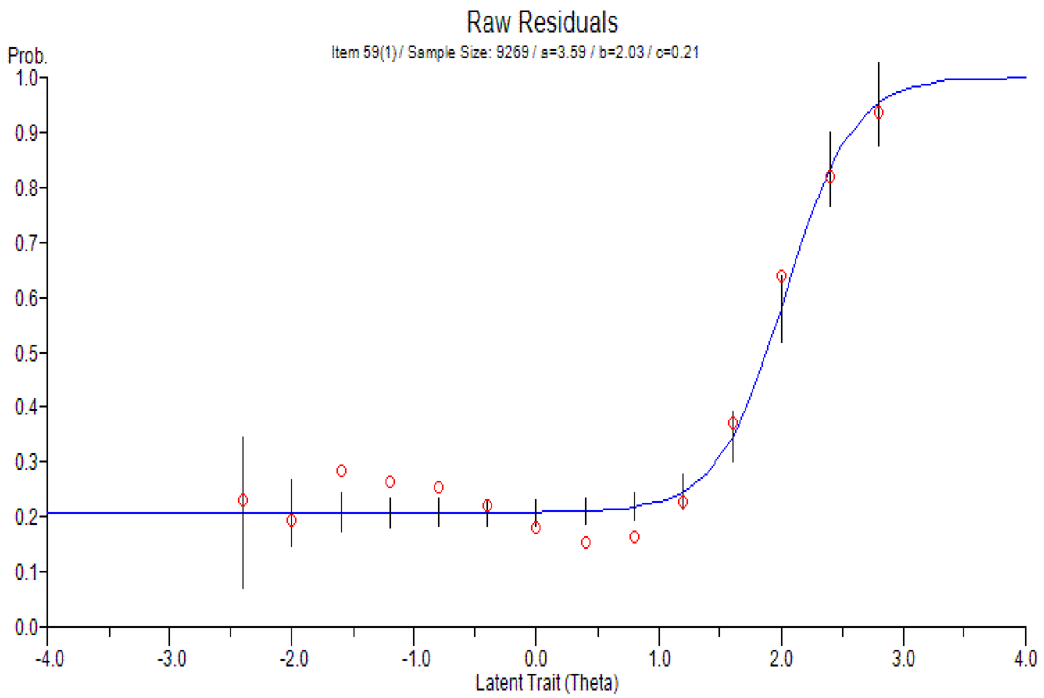

Raw Residuals
Item 20(1) / Sample Size: 9269 / a=1.46 / b=0.26 / c=0.25
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

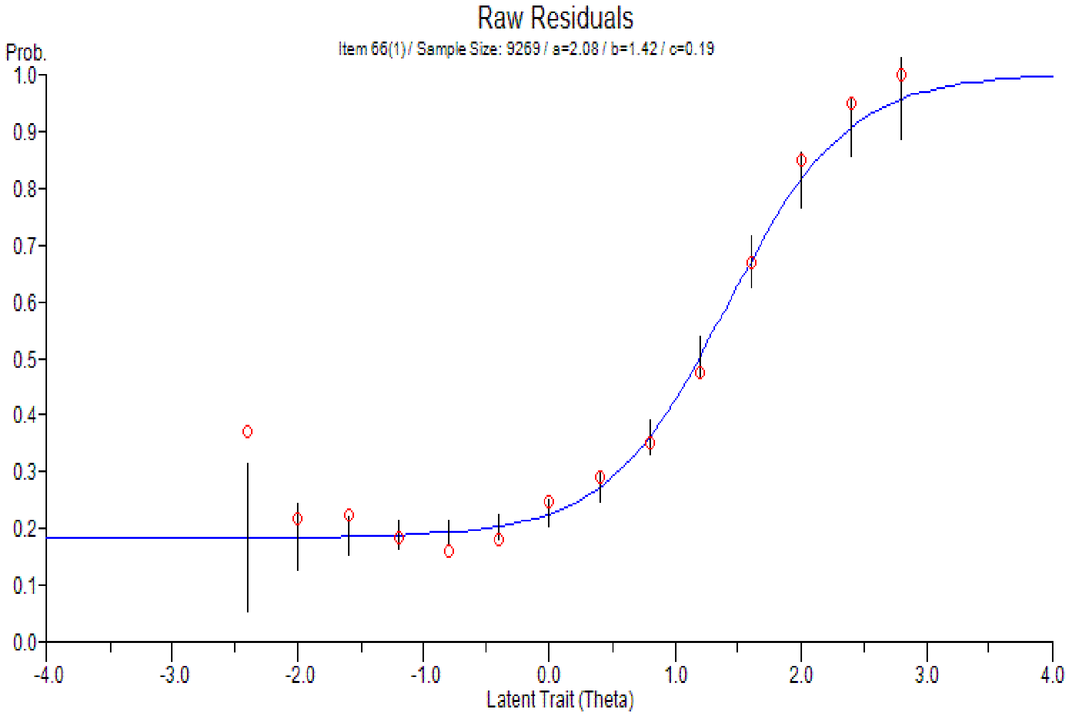

Raw Residuals
Item 21(1) / Sample Size: 9269 / a=2.76 / b=1.49 / c=0.14
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

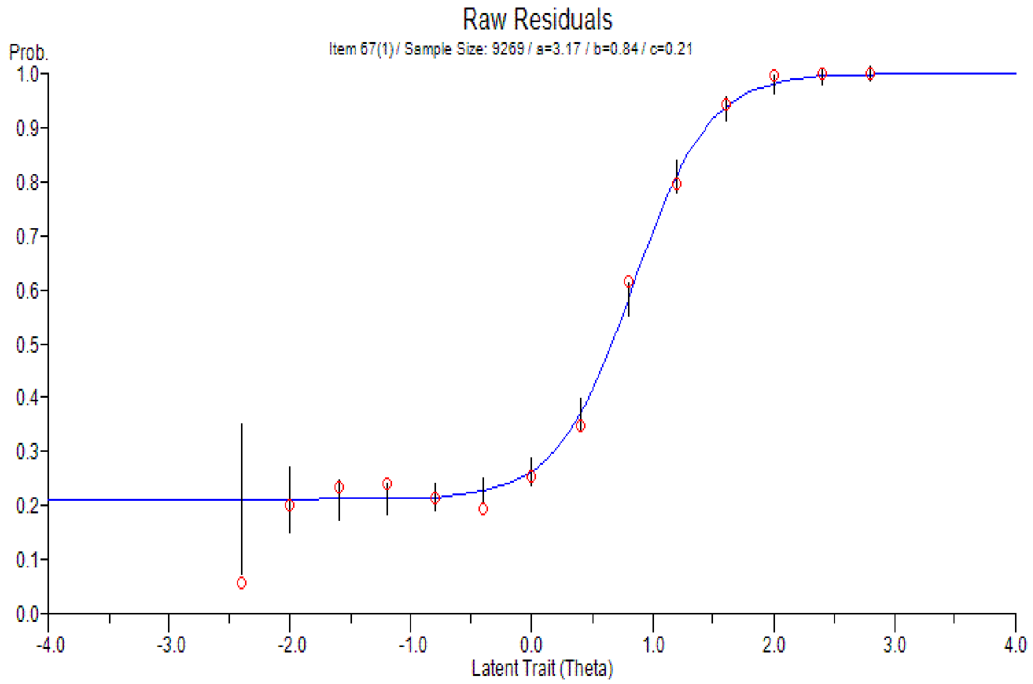

Raw Residuals
Item 25(1) / Sample Size: 9269 / a=3.71 / b=0.59 / c=0.30
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

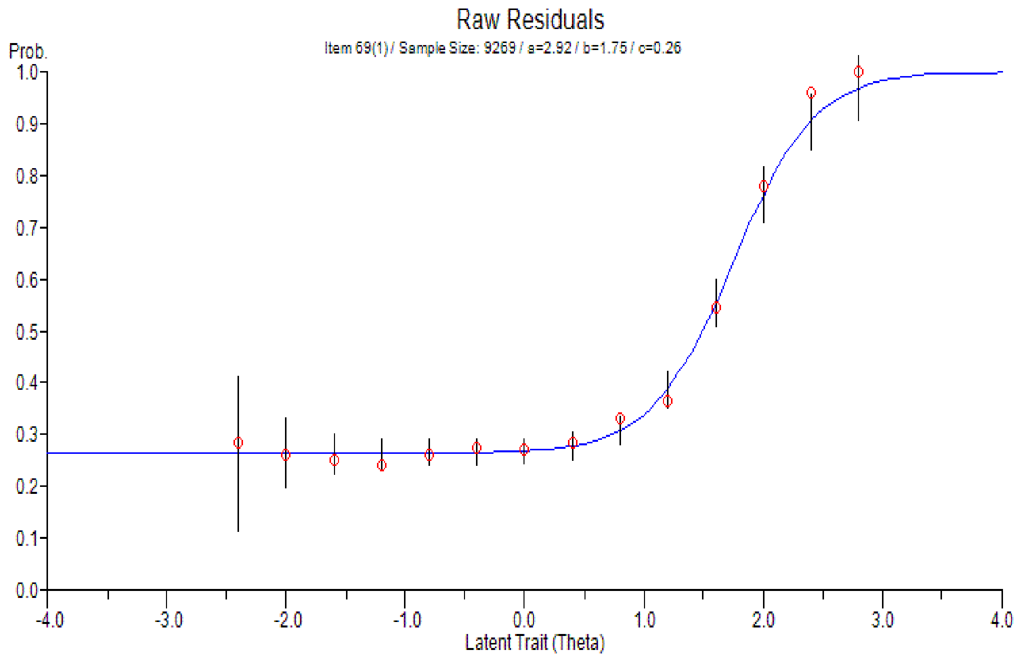

Raw Residuals
Item 27(1) / Sample Size: 9269 / a=1.20 / b=-0.59 / c=0.25
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

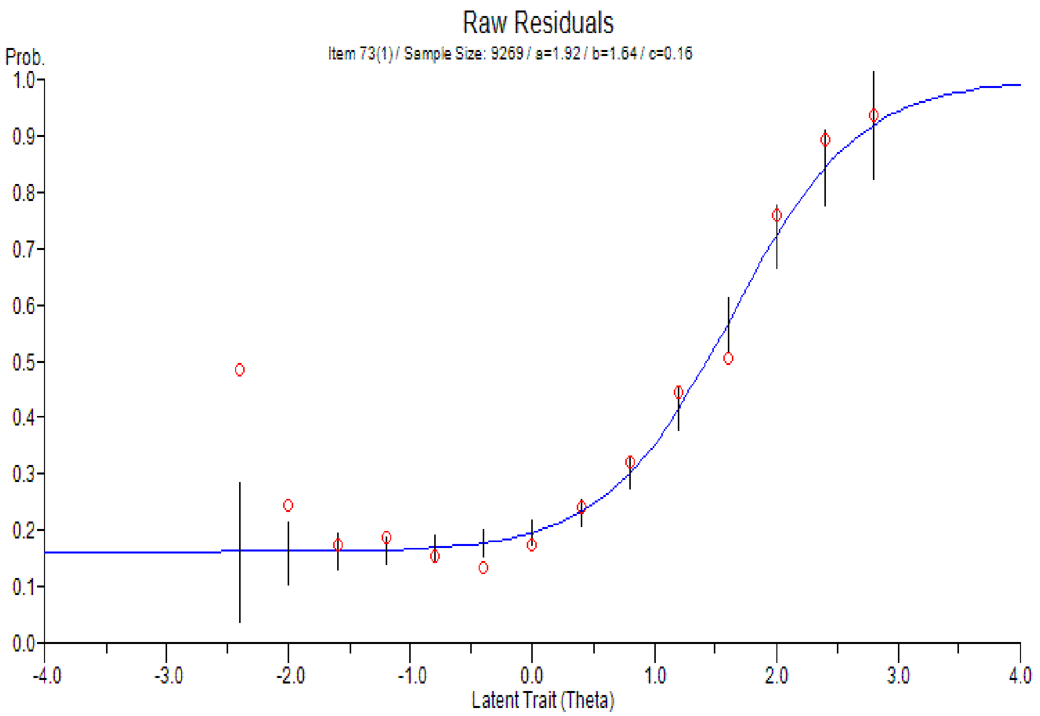

Raw Residuals
Item 35(1) / Sample Size: 9269 / a=2.08 / b=1.75 / c=0.24
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

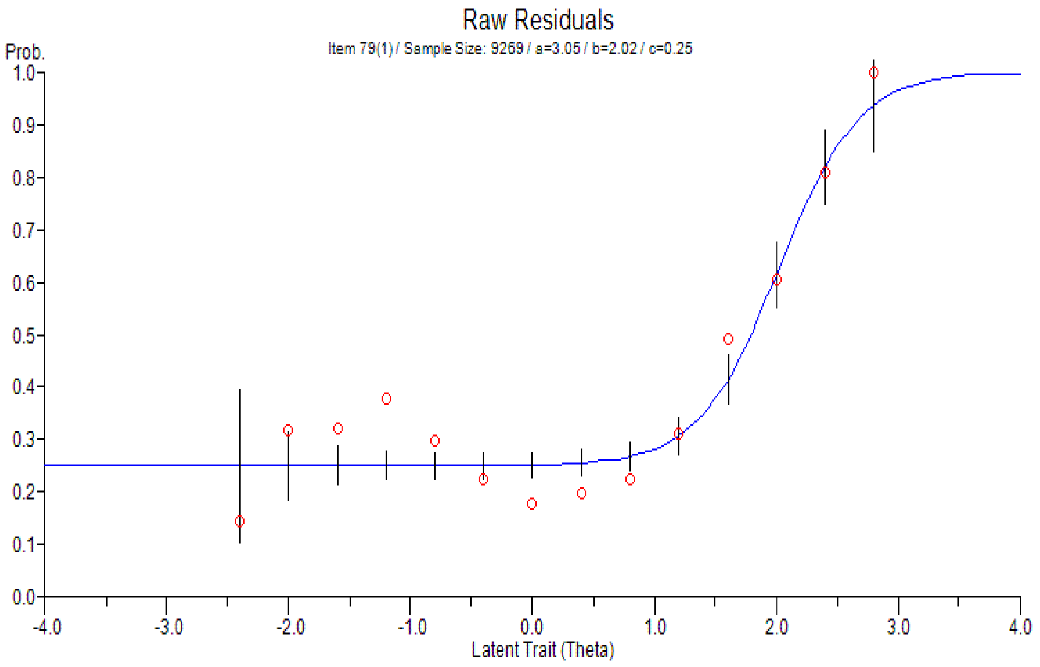

Raw Residuals
Item 59(1)/ Sample Size: 9269 / a=3.59 / b=2.03 / c=0.21
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

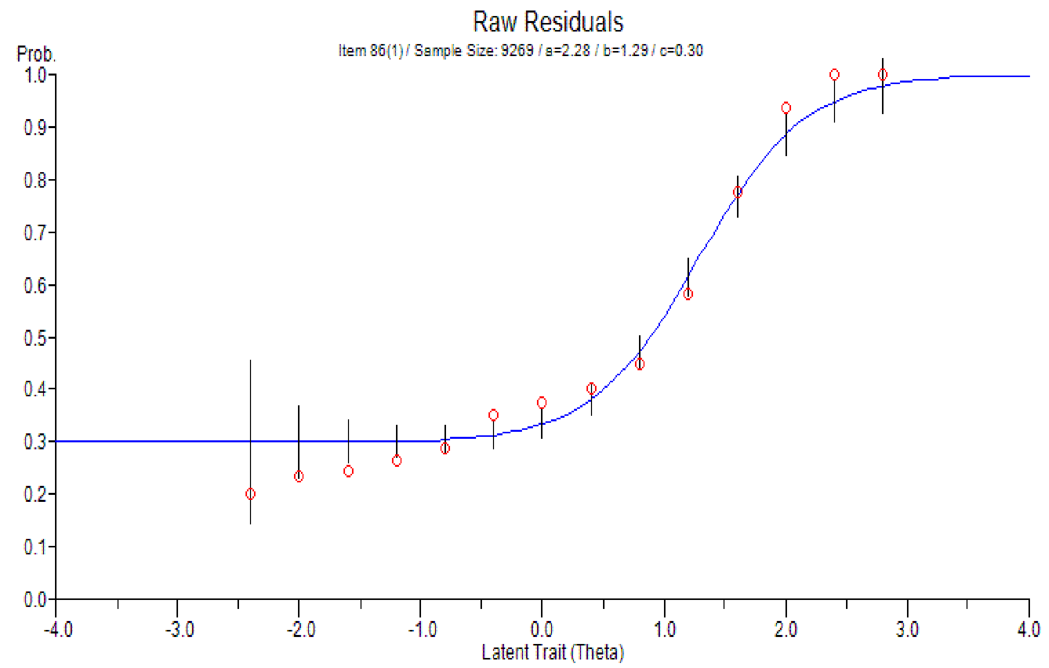

Raw Residuals
Item 66(1)/ Sample Size: 9269 / a=2.08 / b=1.42 / c=0.19
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

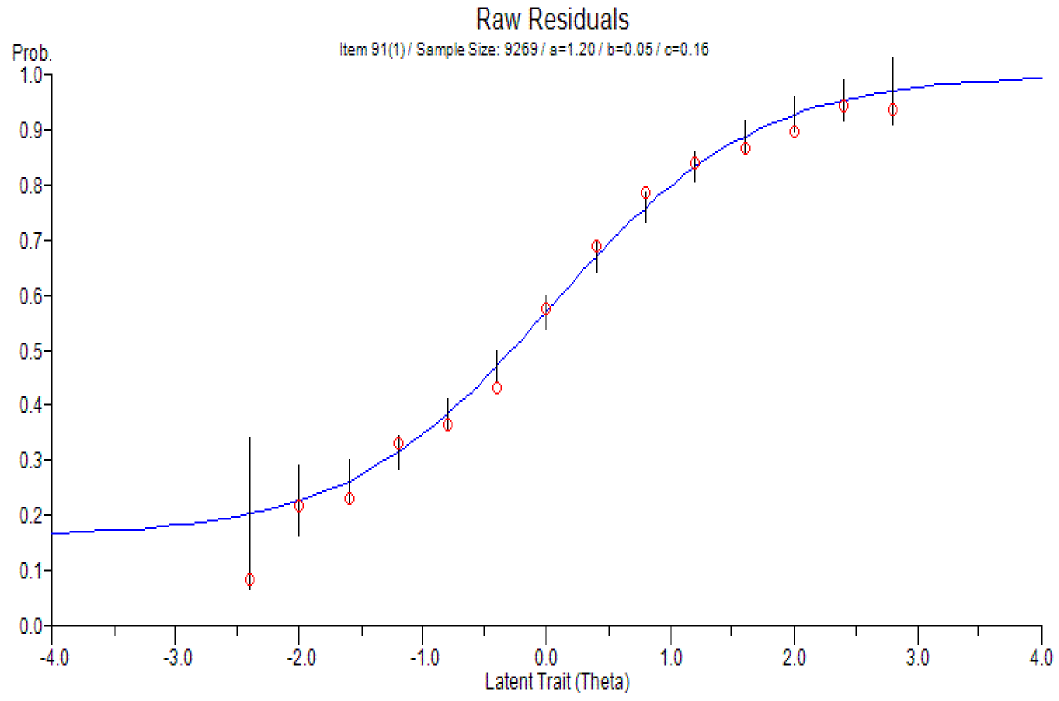

Raw Residuals
Item 67(1) / Sample Size: 9269 / a=3.17 / b=0.84 / c=0.21
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

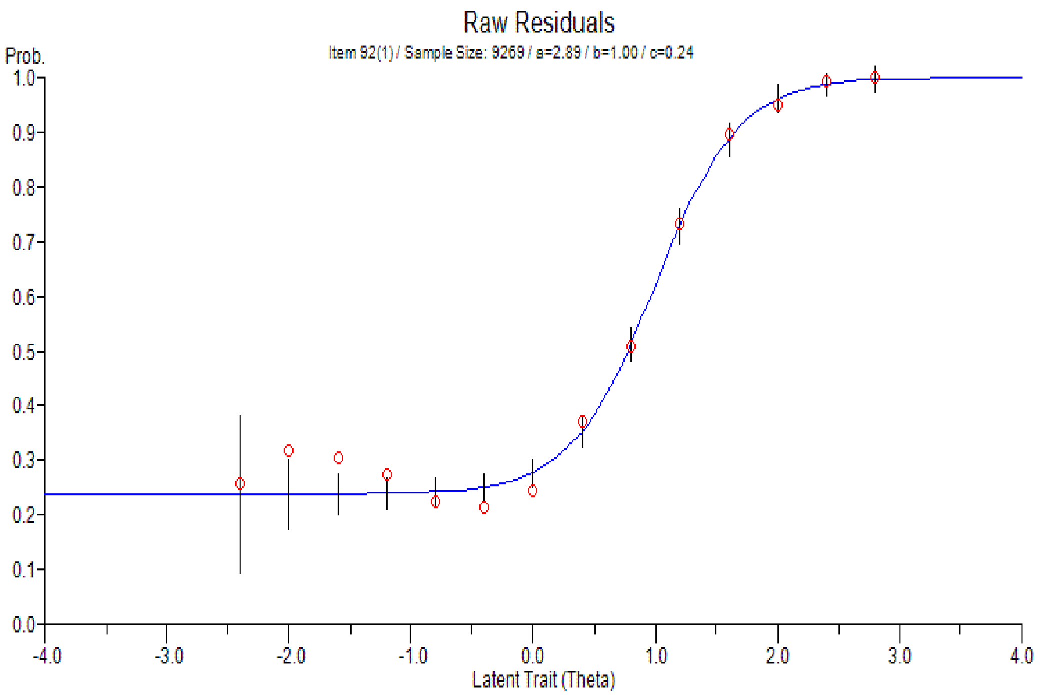

Raw Residuals
Item 69(1) / Sample Size: 9269 / a=2.92 / b=1.75 / c=0.26
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

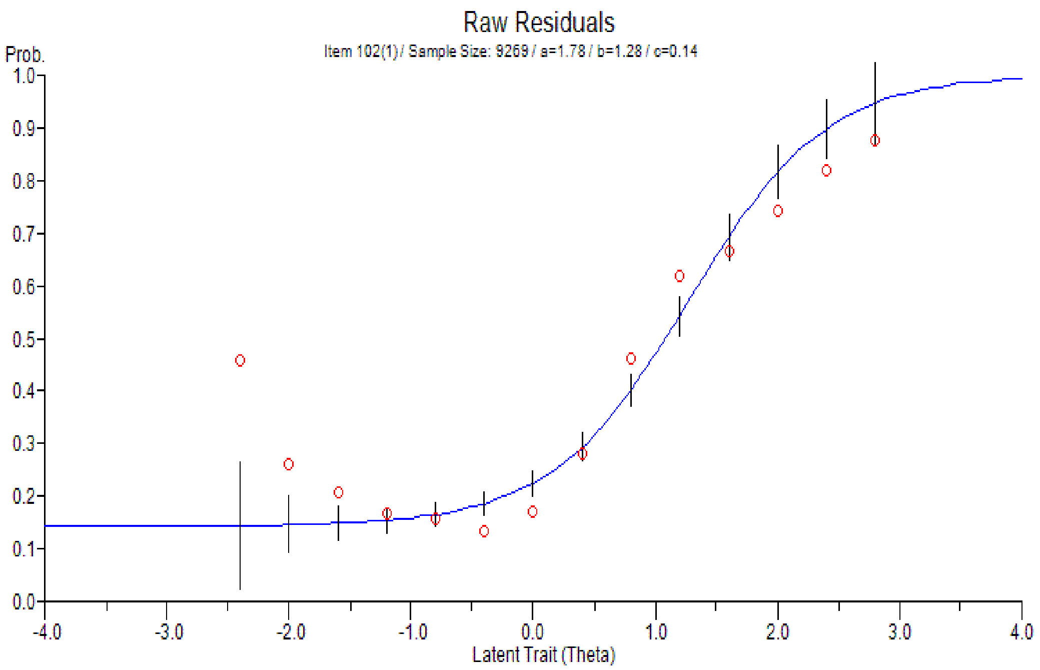

Raw Residuals
Item 73(1) / Sample Size: 9269 / a=1.92 / b=1.64 / c=0.16
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

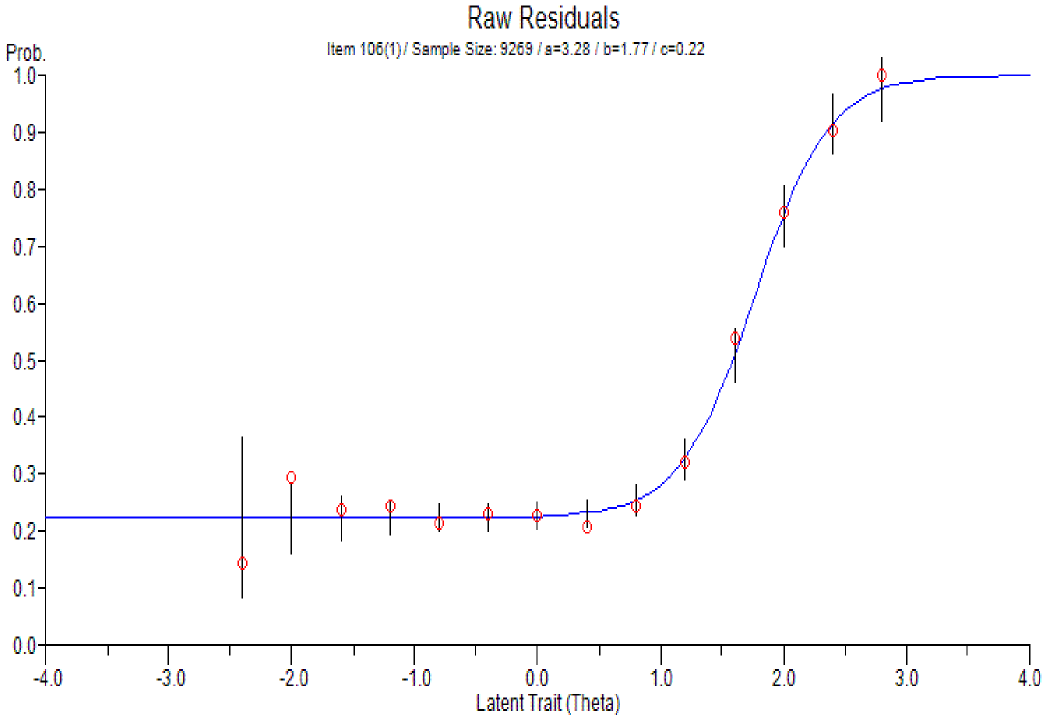

Raw Residuals
Item 79(1) / Sample Size: 9269 / a=3.05 / b=2.02 / c=0.25
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

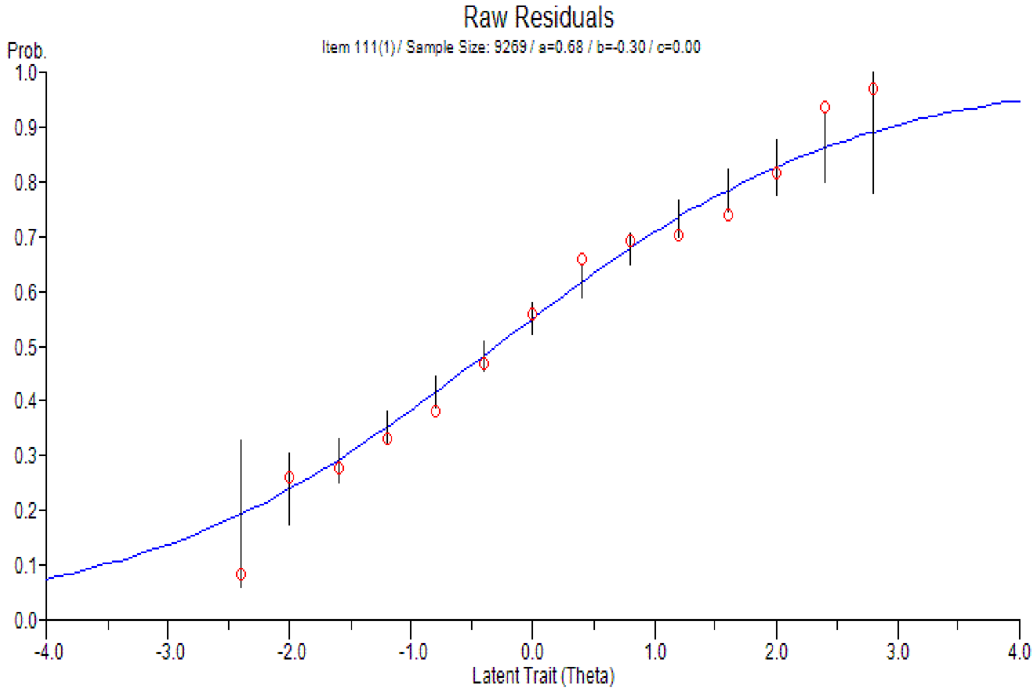

Raw Residuals
Item 86(1) / Sample Size: 9269 / a=2.28 / b=1.29 / c=0.30
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)


Raw Residuals
Item 91(1) / Sample Size: 9269 / a=1.20 / b=0.05 / c=0.16
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)


Raw Residuals
Item 92(1) / Sample Size: 9269 / a=2.89 / b=1.00 / c=0.24
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)


Raw Residuals
Item 102(1) / Sample Size: 9269 / a=1.78 / b=1.28 / c=0.14
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)


Raw Residuals
Item 106(1) / Sample Size: 9269 / a=3.28 / b=1.77 / c=0.22
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)


Raw Residuals
Item 111(1) / Sample Size: 9269 / a=0.68 / b=-0.30 / c=0.00
Prob.
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

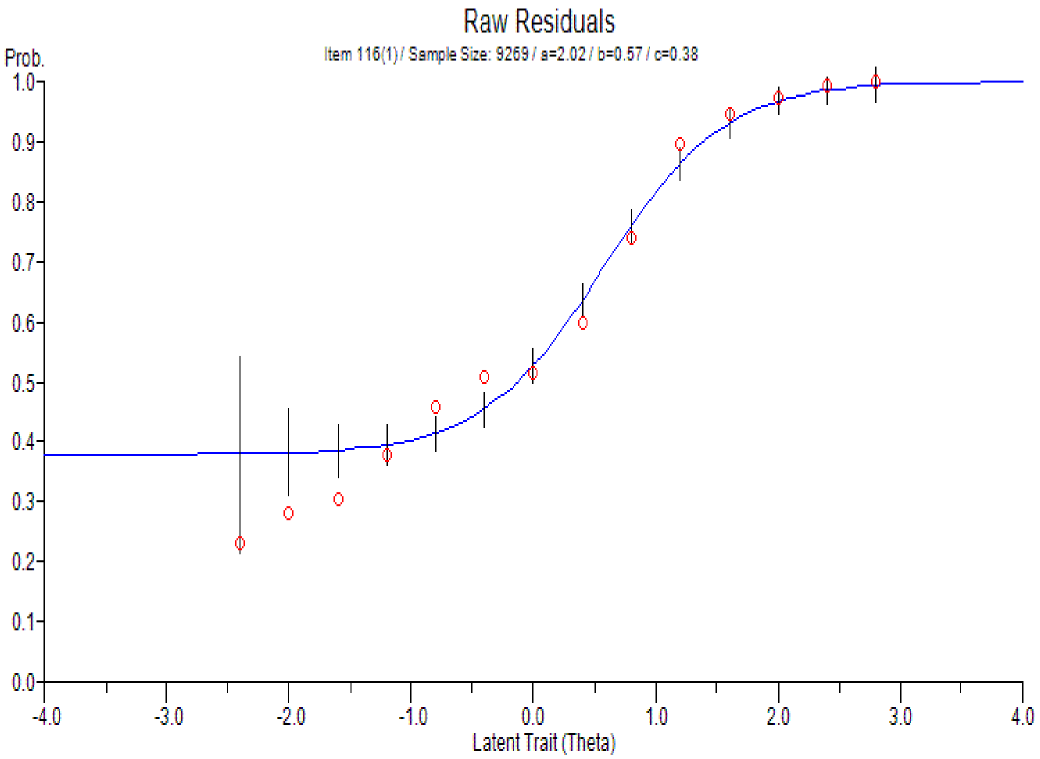


Figure D2. Plots of the Item Residuals for the Misfit Items for the 3PL model on Form B

APPENDIX E

EQUIVALENT SCORES FOR EQUATED FORMS A AND B USING THE POSTSMOOTHING WITH *S* = 0.01

Table E1. Equivalent Scores for Equated Forms A and B using the Postsmoothing with *S* = 0.01

Form B Score	Form A Equivalent Score
0	0.00
1	0.99
2	1.98
3	2.97
4	3.96
5	4.95
6	5.94
7	6.93
8	7.92
9	8.91
10	9.90
11	10.89
12	11.89
13	12.88
14	13.87
15	14.86
16	15.85
17	16.84
18	17.83
19	18.82
20	19.81
21	20.80
22	21.79
23	22.78
24	23.64

Table E1 (continued)

Form B Score	Form A Equivalent Score
25	24.57
26	25.62
27	26.71
28	27.84
29	29.08
30	30.42
31	31.81
32	33.18
33	34.48
34	35.69
35	36.87
36	38.06
37	39.26
38	40.45
39	41.67
40	42.91
41	44.17
42	45.41
43	46.65
44	47.85
45	48.97
46	50.07
47	51.25
48	52.51
49	53.79
50	54.97
51	56.05

Table E1 (continued)

Form B Score	Form A Equivalent Score
52	57.09
53	58.13
54	59.18
55	60.23
56	61.28
57	62.32
58	63.34
59	64.33
60	65.29
61	66.24
62	67.21
63	68.19
64	69.19
65	70.21
66	71.24
67	72.24
68	73.17
69	74.06
70	74.94
71	75.84
72	76.77
73	77.76
74	78.77
75	79.79
76	80.77
77	81.68
78	82.54

Table E1 (continued)

Form B Score	Form A Equivalent Score
79	83.39
80	84.30
81	85.25
82	86.23
83	87.19
84	88.16
85	89.14
86	90.11
87	91.05
88	91.97
89	92.86
90	93.70
91	94.48
92	95.21
93	95.92
94	96.64
95	97.41
96	98.25
97	99.10
98	99.92
99	100.67
100	101.40
101	102.09
102	102.76
103	103.40
104	104.02
105	104.65

Table E1 (continued)

Form B Score	Form A Equivalent Score
106	105.31
107	105.99
108	106.72
109	107.46
110	108.19
111	108.90
112	109.61
113	110.35
114	111.16
115	112.08
116	113.73
117	115.38
118	117.03
119	118.68

APPENDIX F

EQUATING FORMS B AND A OF THE CEPA-ENGLISH TEST USING THE POSTSMOOTHING WITH *S* = 0.01

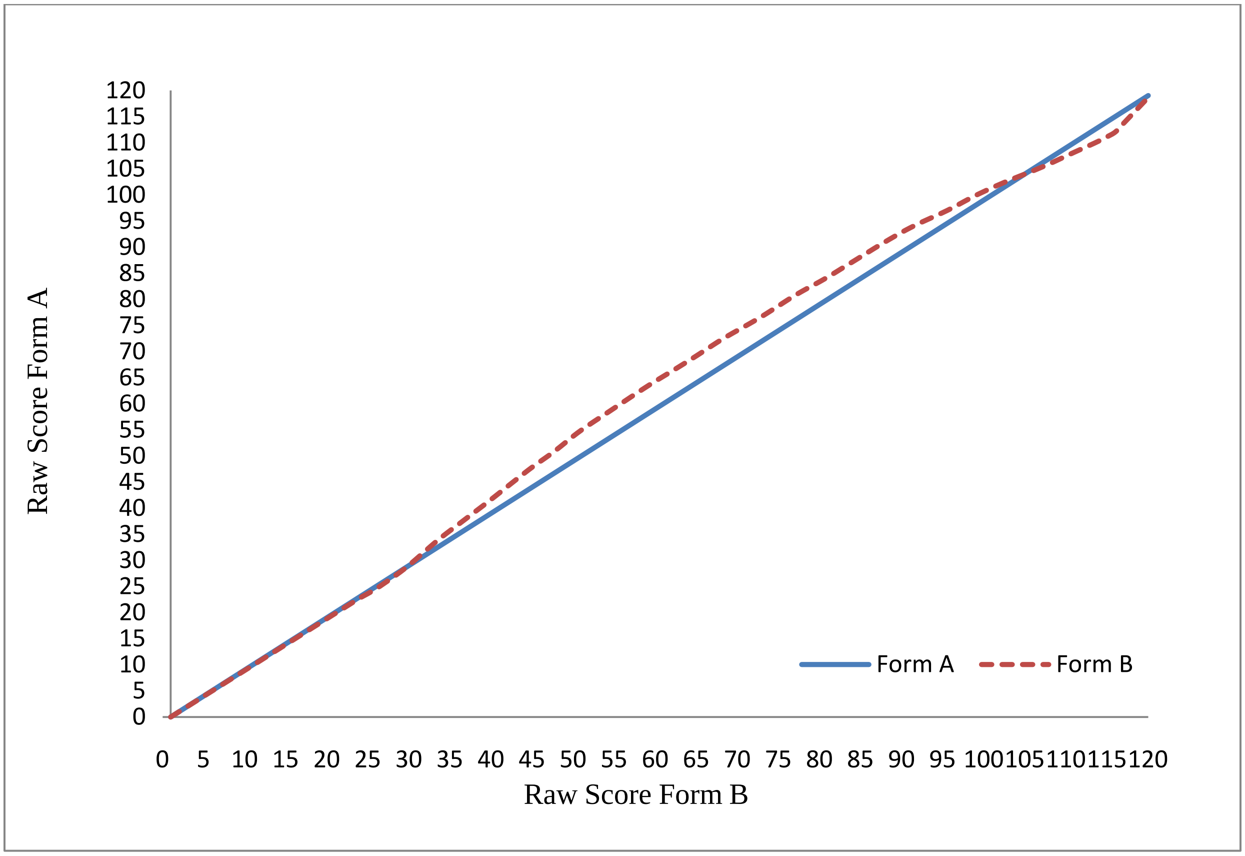


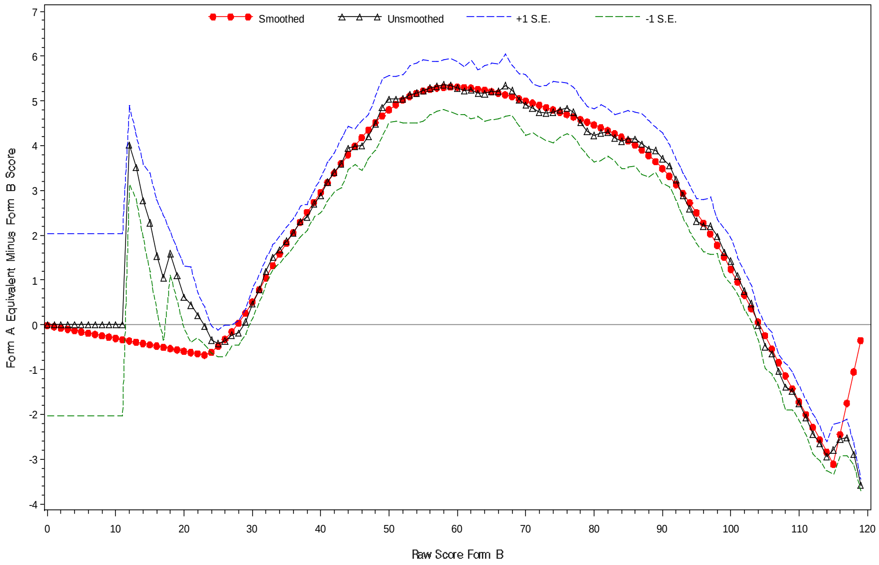
Figure F1. Equating Forms B and A of the CEPA-English Test using the Postsmoothing with

S = 0.01

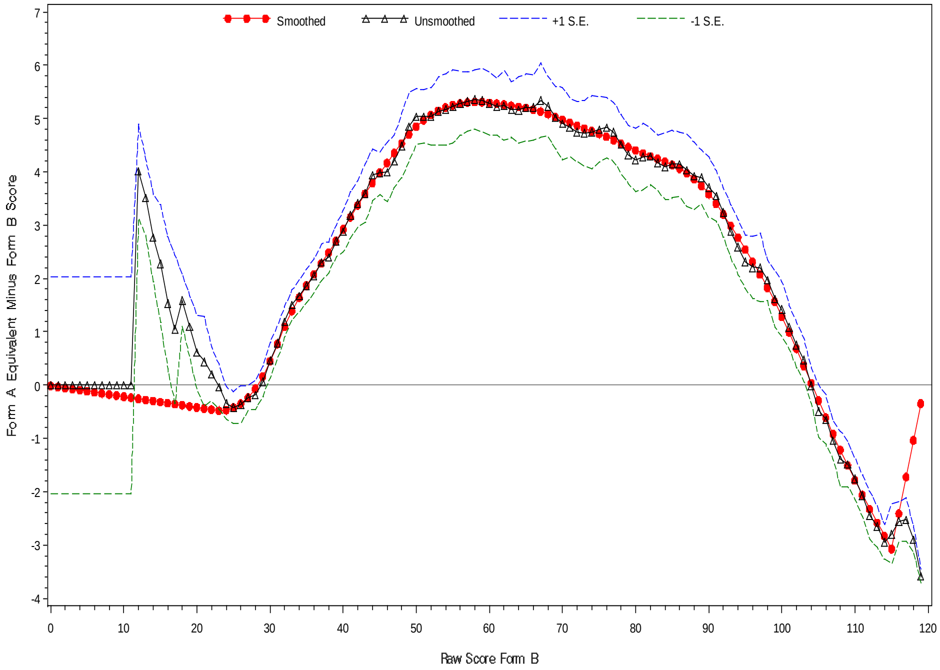
APPENDIX G

RAW-TO-RAW SCORE EQUIVALENTS FOR CUBIC SPLINE POSTSMOOTHING AT NINE DIFFERENT VALUES OF *S*

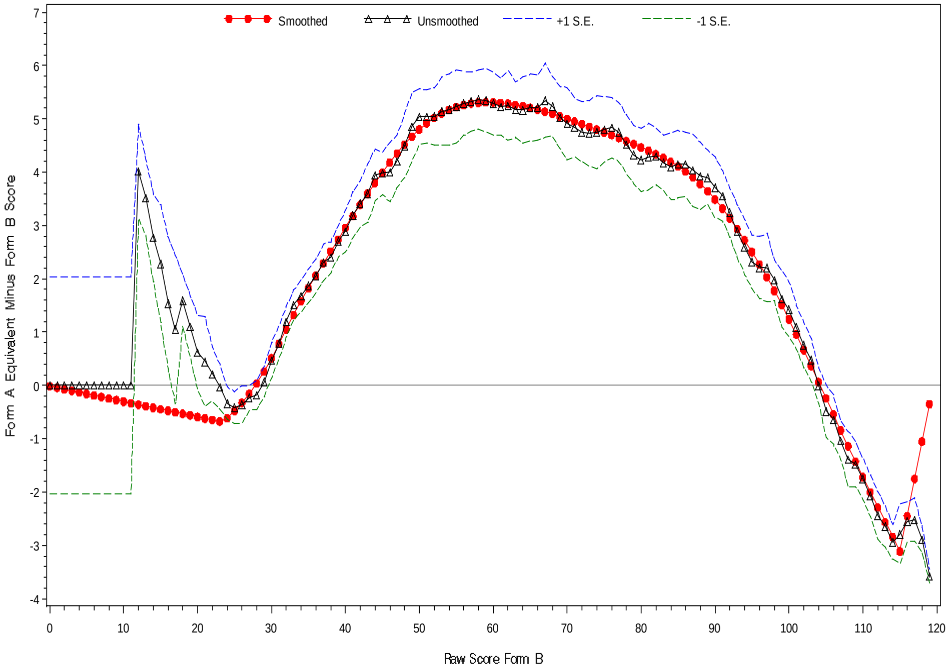
Raw-to-raw equivalents for postsmoothing, S = 0.1



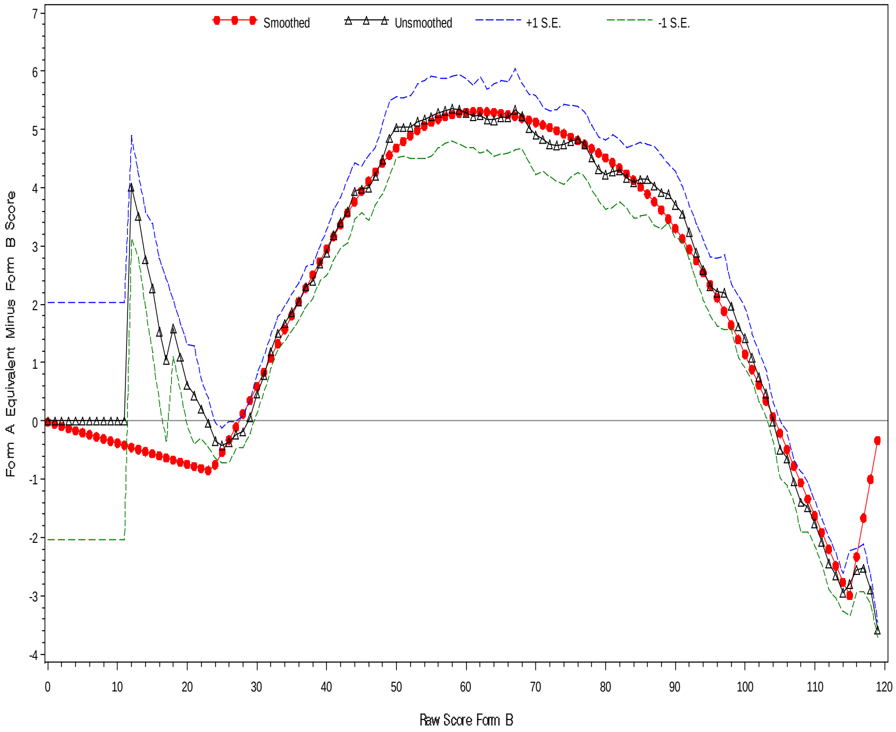
Raw-to-raw equivalents for postsmoothing, S = 0.05



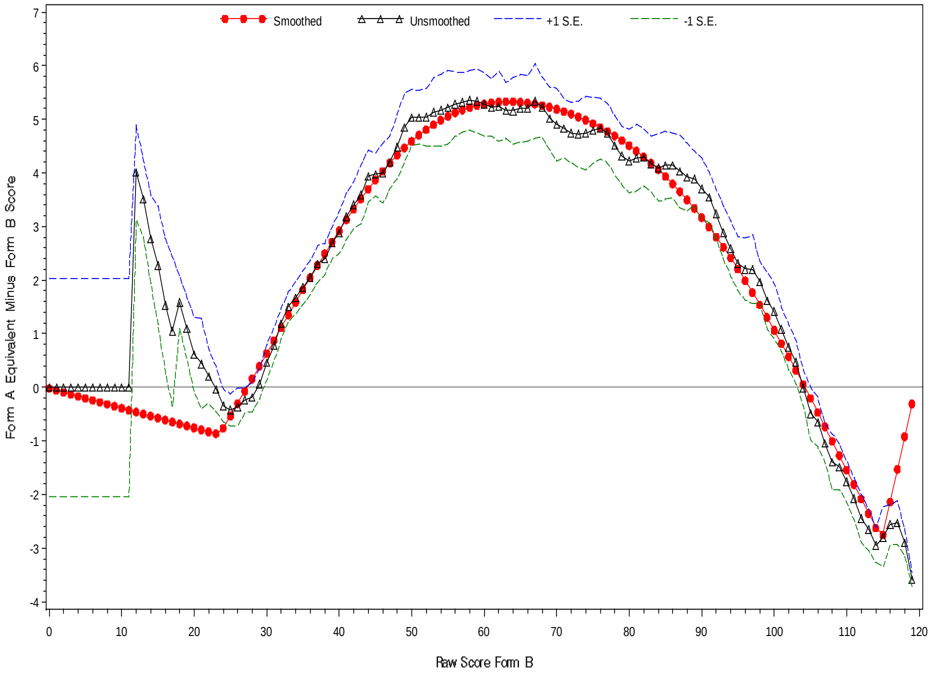
Raw-to-raw equivalents for postsmoothing, S = 0.1



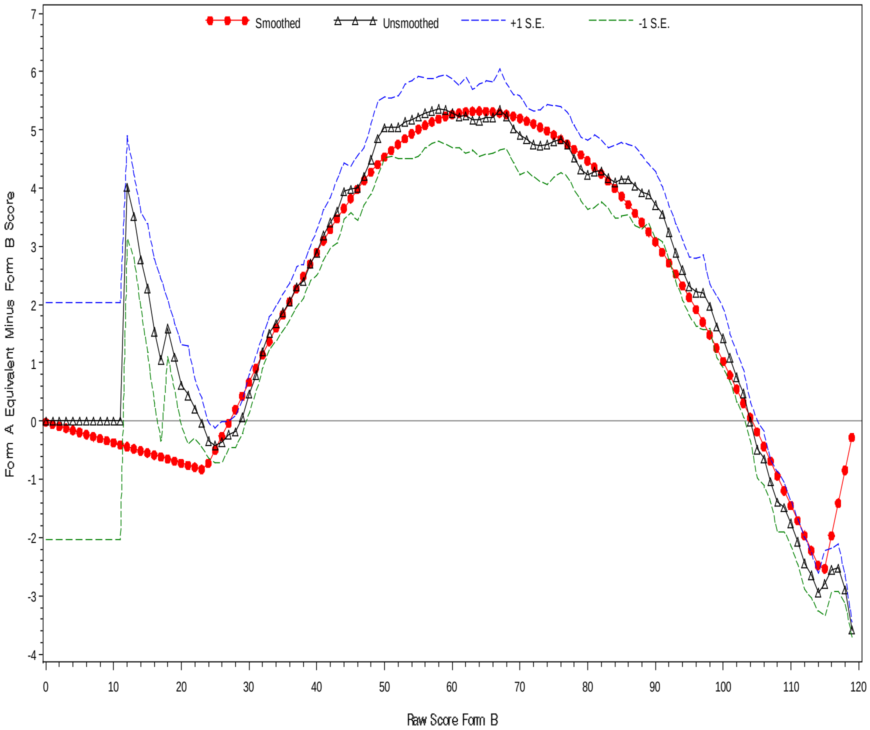
Raw-to-raw equivalents for postsmoothing, S = 0.2



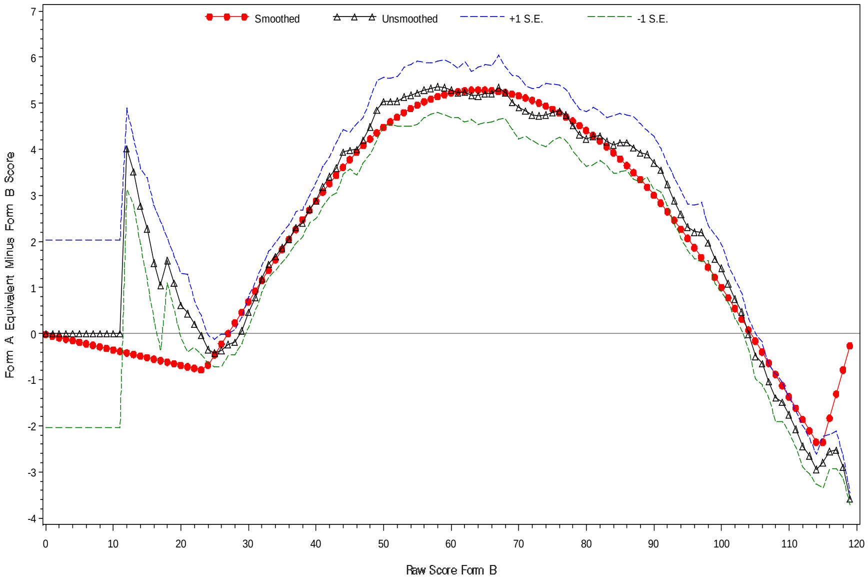
Raw-to-raw equivalents for postsmoothing, S = 0.3



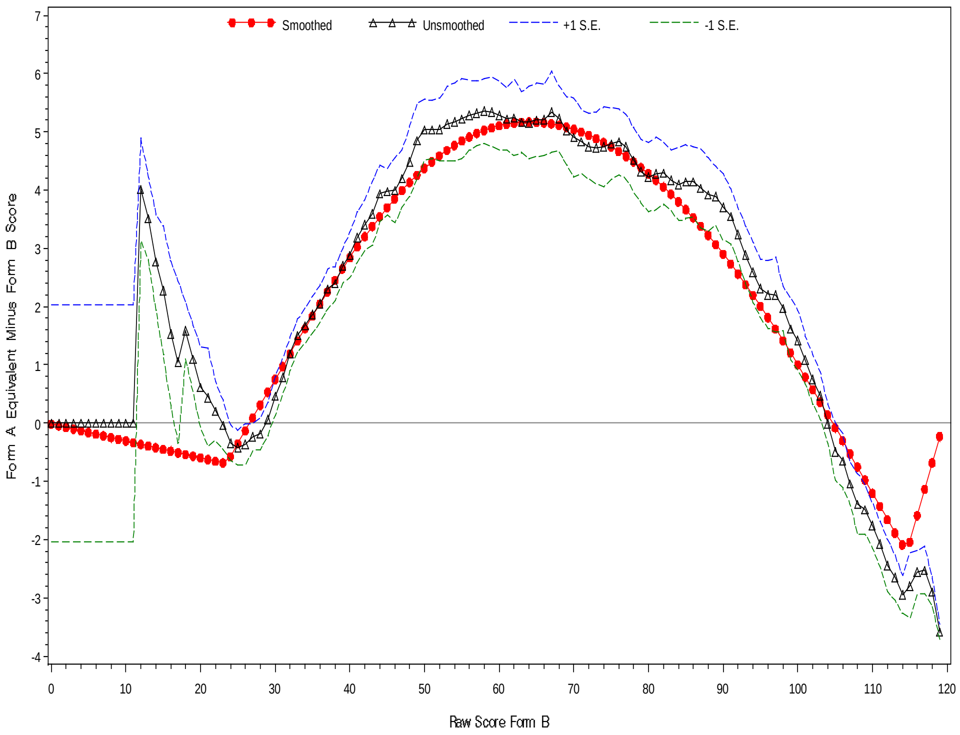
Raw-to-raw equivalents for postsmoothing, S = 0.4



Raw-to-raw equivalents for postsmoothing, S = 0.5



Raw-to-raw equivalents for postsmoothing, S = 0.75



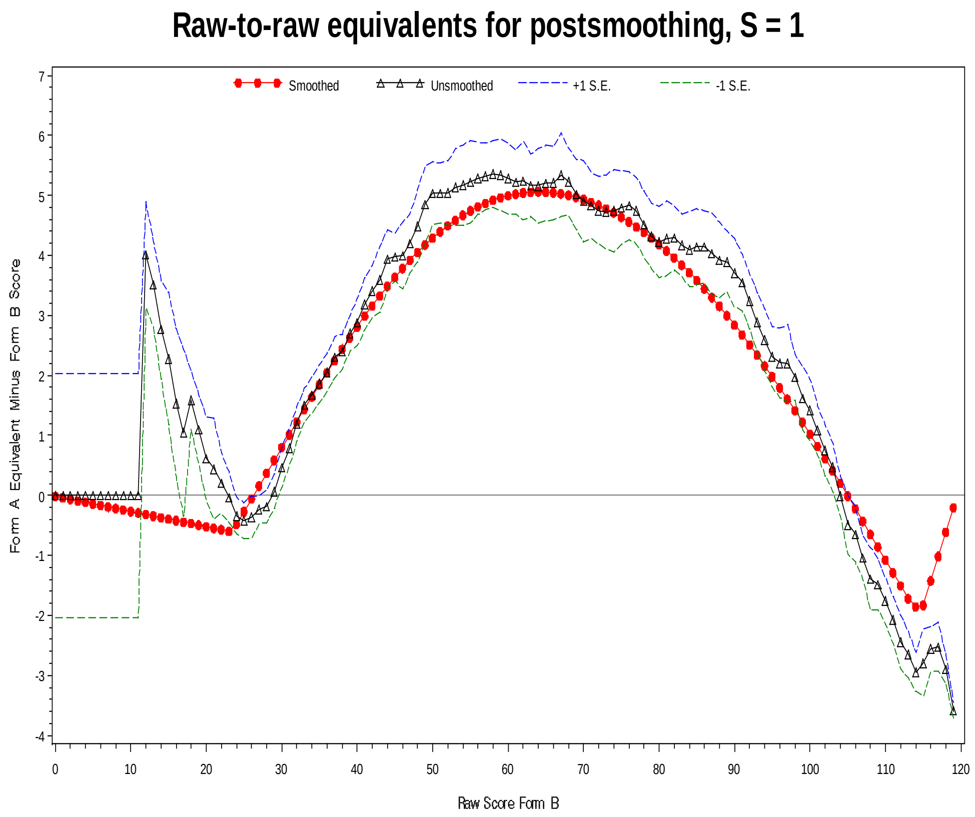


Figure G1. Raw-to-Raw Score Equivalents for Cubic Spline Postsmoothing (*S* =.01, .05, .1, .2, .3, .4, .5, .75, and 1)

BIBLIOGRAPHY

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29,* 67-91.

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). An NCME instructional module on using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, *22,* 37-53.

Alderson, Charles J.; Clapham, Caroline; Wall, Diane. 1995: *Language Test Construction and Evaluation.* Cambridge: Cambridge University Press. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Angoff, W.H. (1993). Perspectives on differential item functioning methodology. In P.W.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational *measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Assessment Systems Corporation (1995). *XCALIBRE: Marginal maximum-likelihood estimation for the 2- and 3- parameter logistic IRT model* (version 1.10). St. Paul, MN: Author.

Ayres, J. B., & Peters, R. M. (1977). Predictive validity of the test of English as a foreign language for Asian graduate students in engineering, chemistry, or mathematics. *Educational and Psychological Measurement, 37* (2), 461- 463.

Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement*, 28, 147-162.

Be´guin, A. A., & Glas, C. A. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika, 66,* 541-561.

Bejar, I. I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. *Journal of Educational Measurement, 17*, 283-296.

Bellingham, L. 1993. The relationship of Language Proficiency to Academic Success for International Students. *New Zealand Journal of Educational Studies*, 30 (2), 229-232.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29-51.

Bock, R.D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12,* 261-280.

Bock, R.D., & Lieberman, M. (1970). Fitting a response model for N dichotomously scored items. *Psychometrika, 26,* 347-372.

Bock, R.D., & Lieberman, M. (1970). Fitting a response model for N dichotomously scored items. *Psychometrika, 26,* 347-372.

Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 132-162). Newbury Park, CA: Sage.

Brown, A. (2008). *English language assessment for the preparation and selection of students for higher education in the UAE*. Unpublished manuscript.

Brown, A. (2007). *CEPA 2007 Examiners report*. UAE: NAPO, Ministry of Higher Education and Scientific Research.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), Testing structural equation models (pp. 136–162). Beverly Hills, CA: Sage.

Burgess, T. C., & Greis, N. B. (1984). English language proficiency and academicachievement among students of English as a second language at the college level. (ERIC Document Reproduction Service No. ED 074 812)

Camilli, G., & Shepard L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Carroll, J. B. (1945). The effect of difficulty and chance success on correlation between items and between tests. *Psychometrika, 26,* 347-372.

Cattell, R.B. (1966). The meaning and the strategic use of factor analysis, in R.B. Cattell(ed), *Handbook of Multivariate Experimental Psychology*, Chicago: Rand McNally.

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22,* 265-289.

Christoffersson, A. (1975). Factor analysis of dichotomized variables.*Psychometrika, 40*(1), 5-32.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items: An NCME instructional module. *Educational Measurement: Issues and Practice*, 17(1), 3-44.

Clauser, B., Mazor, K., & Hambleton, R. K. (1993). The effects of purification for the matching criterion on the identification of DIF using the MH procedure. *Applied Measurement in Education*, *6*(4), 269-279.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cotton, F. and Conrow, F. 1998. An Investigation of the Predictive Validity of IELTS amongst a Group of International Students studying at the University of Tasmania. *English Language Testing System Research Reports*, 1, 72-115.

Crocker, L. & Algina, J. (1986). *Introduction to classical & modern test theory*. Belmont, CA: Wadsworth Group.

De Champlain, A. F. (1999). *An overview of nonlinear factor analysis and its relationship to item response theory*: Law School Admission Council.

De Champlain, A. F.,&Tang, K. L. (1997). CHIDIM: A FORTRAN program for assessing the dimensionality of binary item responses based on Mc-Donald's nonlinear factor analytic model. *Educational and Psychological Measurement, 57*, 174-178.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39*, 1-38.

Donoghue, J. R., & Allen, N. L. (1993). Thin versus thick matching in the Mantel-Haenszel procedure for detecting DIF. *Journal of Educational Statistics*, *18*(2), 131-154.

Dooey, P.1999. An investigation into the predictive validity of the IELTS Test as an indicator of future academic success. In Martin, Stanley and Davison (eds) *Teaching in the Disciplines/Learning in Context*, 114-118, Proceedings of the 8th Annual Teaching Learning Forum, University of Western Australia, February 1999.

Dorans, N.J., & Holland, P.W. (1993). DIF detection and description: Mantel-Haenszel and Standardization. In P. W. Holland & H. Thayer (Eds.), *Differential item functioning* (pp. 35-66), Hillsdale, NJ: Lawrence Erlbaum Associates.

Dorans, N.J., & Kulick, E. (1983). *Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977*: *an application of* *the standardization approach* (Research Rep. No. 83-9). Princeton, NJ: Educational Testing Service.

Duran, B., & Weffer, R. (1992). Immigrants' aspirations, high school process, and academic outcomes. *American Educational Research Journal, 29*(1), 163-181.

Douglas, J., Kim, H. R., Habing, B., & Gao, F. (1998). Investigating local dependence with conditional covariance functions. *Journal of Educational and Behavioral Statistics, 23*, 129-151.

Educational Testing Service (2009). Retrieved February 4, 2009, from: http://www.ets.org/toefl/

Educational Testing Service (2009). *The TOEFL test and score manual supplement: 1994-1995 Edition.* Retrieved February 4, 2009, from: http://www.ets.org/Media/Research/pdf/TOEFL-SUM-9495.pdf

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Elder, C. 1993. Language proficiency as a predictor of performance in teacher education. *Melbourne Papers in Language Testing,* 2 (1), 68-87.

Fava, J. L., & Velicer, W. F. (1992). The effects of over-extraction on factor and component analysis. *Multivariate Behavioral Research, 27,* 387-415.

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for differential item functioning detection. *Educational and Psychological Measurement, 67,* 373-393.

Ford, J. K., MacCallum, R. C., & Tait, M. (1986). The application of exploratory factor analysis in applied psychology: A critical review and analysis. *Personnel Psychology, 39*, 291-314.

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research, 23,* 267-269.

Gauch, R.R. (2000). *Statistical methods for researchers made very simple*. Lanham, MD: University Press of America, Inc.

Gessaroli, M. E. & De Champlain, A. F. (1996). Using an approximate chi-square statistic to test the number of dimensions underlying the response to a set of items. *Journal of Educational Measurement, 33*, 157-179.

Gibson, C. and Rusek, W. 1992. The validity of an overall band score of 6.0 on the IELTS test as a predictor of adequate English language level appropriate for successful academic study. Unpublished Masters of Arts (Applied Linguistics) thesis, Macquarie University, New South Wales.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information bi-factor analysis. *Psychometrika, 57*, 423-436.

Green, S. B. (1983). Identifiability of spurious factors with linear factor analysis with binary items. *Applied Psychological Measurement*, *7*, 3-13.

Gorsuch, R. L. (1983) *Factor Analysis*. Hillsdale, NJ: Erlbaum

Gosz, J. K. &Walker, C. M. (2002, April). *An Empirical Comparison of Multidimensional Item Response Data Using TESTFACT and NOHARM.* Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, LA.

Graham, J.G. 1987. English Language Proficiency and the Prediction of Academic Success. *TESOL Quarterly*,21 (3), 505-521.

Green, S. B. (1983). Identifiability of spurious factors with linear factor analysis with binary items. *Applied Psychological Measurement*, *7*, 3-13.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, *22,* 144-149

Hale, C. Standsfield, and R. Duran (Eds.), *Summaries of Studies Involving the Test of English as a Foreign Language* (pp.163-1982). Princeton, N. J.: Educational Testing Service.

Hair, J. E., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. 2006. *Multivariate data analysis* (6th ed.). Upper Saddle River, NJ: Pearson Prentice

Hall.Hair, J.F., Anderson, R.E., Tatham, R.L., and Black, W.C., 1992. *Multivariate Data Analysis* (3rd ed.). Macmillan: New York.

Hambleton, R. K., Jones, R. W., & Rogers, H. J. (1993). Influence of item parameter estimation errors in test development. *Journal of Educational Measurement, 30*(2), 143-155.

Hambleton, R. K., & Rogers, J. H. (1986, April). *Promising directions for assessing item response model fit to test data.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Hambleton, R. K., & Rovinelli, R. J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement, 10*(3), 287-302.

Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991). *Fundametals of Item Response Theory*. USA: SAGE publications, Inc.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Boston: Kluwer-Nijhoff.

Hambleton, R.K. & Zaal, J.N. (1991). Advances in Educational and Psychological Testing: Theory and Application, Kluwer Academic Publisher, Boston.

Harman, H. H. (1976).*Modern factor analysis* (3rd ed.). Chicago: University of Chicago Press.

Hattie, J.A. (1981). A four stage factor analytic approach to studying behavioral domains. *Applied Psychological Measurement, 5,* 77-88.

Hattie, J. (1984).An empirical study of various indices for determining unidimensionality. *Multivariate Behavioural Research, 19*(1), 49-78.

Hattie, J., Krakowski, K., Rogers, H. J., & Swaminathan, H. (1996). An assessment of Stout's index of essential unidimensionality. *Applied Psychological Measurement, 20*(1), 1-14.

HCT. (2009). Retrieved January 12, 2009, from: http://www.hct.ac.ae/

Heil, D., & Aleamoni, L. (1974). Assessment of the proficiency in the use and understanding of English by foreign students as measured by the Test of English as a Foreign Language. In G. Hale, C. Standsfield, and R. Duran (Eds.), *Summaries of Studies Involving the Test of English as a Foreign Language* (pp.163-1982). Princeton, N. J.: Educational Testing Service.

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.

Hill, K. and Lynch, B. (1999), A Comparison of IELTS and TOEFL as Predictors of Academic Success, in Robyn Tulloh's (Ed) *International English Language Testing System Research Reports 1999*, vol2, IELTS Australia Pty Ltd,

Hills, J. R. (1989). Screening for potentially biased items in testing programs. *Educational Measurement*: *Issues and Practice, 8,* 5–11.

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning 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. (Eds.). (1993). *Differential item functioning*. Hillsdale, J:Lawrence Erlbaum Associates.

Ho, D.Y.F. and Spinks, J.A. 1985. Multivariate Prediction of Academic Performance by Hong Kong University Students. *Contemporary Educational Psychology*, 10, 249-259.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*, 179-185.Hughes, A. (1989): *Testing for language teachers*. Cambridge University Press, Cambridge.

Hughey, W. A., & Hinson, D. (1993). Assessing the efficacy of the test of English as aforeign language. Psychological Reports, 73, 187-193.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). Item response theory: Applications to psychological measurement. Homewood, IL: Dow Jones–Irwin.

Hulin, C. L., Drasgow, F., & Parsons, C. K. 1983. *Item Response Theory: Applications to Psychological Measurement*. Homewood, IL: Dow Jones Irwin.

Humphreys, L. G.,&Montanelli, R. G., Jr. (1975).An investigation of the parallel analysis criterion for determining the number of common factors. *Multivariate Behavioral Research*, *10*,193-205.

Hung, P., Wu, Y., & Chen, Y. (1991). *IRT item parameter linking: Relevant issues for the purpose of item banking.* Paper presented at the International Academic Symposium on psychological Measurement, Tainan, Taiwan.

Hwang, K., & Dizney, H. F. (1970). Predictive validity of the test of English as a foreign language for Chinese graduate students at an American university. Educational and Psychological Measurement, 30, 475-477.

International English Language Testing System (2009). Retrieved February 4, 2009, from: http://www.ielts.org/

Jodoin, M. G., & Gierl, M.J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14,* 329-349.

Johnson, P. (1988). English language proficiency and academic performance of undergraduate international students. *TESOL Quarterly, 22(1),* 164-168.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141-151.

Kerstjens, M and Nery, C. 2000. Predictive Validity in the IELTS Test: A Study of the Relationship Between IELTS Scores and Students' Subsequent Academic Performance. *English Language Testing System Research Reports*, 3, 85-108.

Kim, J.O., & Mueller, C.W. (1978). *Factor analysis: Statistical methods and practical issues.* Sage University paper series on quantitative applications in the social sciences, series no. 07-014. Beverly Hills and London: Sage Publications.

Kim, S. -H., & Cohen, A. S. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement, 29,* 51-66.

Knol, D. L., & Berger, M. P. F. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research, 26*(3), 457-477.

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

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: methods and practices*. Springer-Verlag New York, Inc.

Kuncel, N. R., Campbell, J. P., & Ones, D. S. (1998). GRE validity: Estimated or tacitly known. *American Psychologist, 53,* 567-568.

Lewis, C. (1993). A note on the value of including the studied item in the test score when analyzing test items for DIF. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 317-319). Hillsdale, NJ: Lawrence Erlbaum Associates.

Light, R., Xu, M., & Mossop, J. (1987). English proficiency and academic performance of international students. *TESOL Quarterly, 21(2),* 251-260.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord. F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

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.

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., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement, 52*, 443-451.

McDonald, R. P. (1997). Normal-ogive multidimensional model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*. New York: Springer.

McDonald, R. P. (1994). Testing for approximate dimensionality. In D. Laveault, B. Zumbo, M. E. Gessaroli, & M. W. Boss (Eds.), *Modern theories of measurement: Problems and Issues* (pp.63-86). Ottawa: Edumetrics Research Group.

McDonald, R. P. (1991). The dimensionality of tests and test items. *British Journal of Mathematical and Statistical Psychology, 34,* 100-117.McDonald, R.P. (1981). The dimensionality of tests and items. *The British Journal of Mathematical and Statistical Psychology, 34,* 100-117.

McDonald, R. P. (1967). *Nonlinear factor analysis* (Psychometric Monographs, No. 15). The Psychometric Society.

McKinley, R., & Mills, C. (1 985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement, 9,* 49-57.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-104). New York: Macmillan.

Miller, T. R. (1991). Empirical estimation of standard errors of compensatory MIRT model parameters obtained from the NOHARM program (Research report 91-2). Iowa City, IA: American College Testing.

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics, 11*(1), 3-31.

Muraki, E., & Engelhard, G., Jr., (1985). Full-information item factor analysis: Applications of EAP scores. *Applied Psychological Measurement, 9,* 417-430.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika, 43*(4), 551-560.

Muthén, B. O., & Muthén, L. K. (2001). *Mplus: Statistical analysis with latent variables*. Los Angeles: Statmodel.

Nandakumar, R. (1994). Assessing dimensionality of a set of item responses-Comparison of different approaches. *Journal of Educational Measurement, 31*(1), 17-35.

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. *Journal of Educational Measurement, 28*, 99-117.

Nandakumar, R., & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics, 18*, 41–68.

National Admissions and Placement Office (2009). Retrieved January 13, 2009, from: http://www.napo.ae/cepa/

Narayanan, P. & Swaminathan H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement*, 20(3), 257-274.

Neter, J., Kutner, M., Nachtsheim, C., and Wasserman, W. (1996), .*Applied Linear Statistical Models*., McGraw-Hill Companies, Inc., NY.

Nunnally, J. C. (1978). *Psychometric Theory*. New York, McGraw-Hill.

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers, 32* (3), 396-402.

Oshima, T.C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement, 31*, 200-219.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24, 50-64.*

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

Pang, X. L. (1999). *Assessing the performance of the approximate chi-square and Stout's T statistics with different test structures*. Unpublished Doctoral dissertation, University of Ottawa.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum.

Pyo, K. H. (2000, April). *Assessing dimensionality of a set of language test data*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. Reckase, M. D. (1981). *The formation of homogeneous item sets when guessing is a factor in item responses* (Office of Naval Research Report 81-85). Columbia: University of Missouri, Department of Educational Psychology. (ERIC Document Reproduction Service No. ED20935)

Reckase, M.D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4,* 207-230.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. Applied Psychological Measurement, 17, 105-116.

Schmitt, A. P., Holland, P. W., & Dorans, N. J. (1993). Evaluating hypotheses about differential item functioning. In P. W. Holland & H. Thayer (Eds.), *Differential item functioning* (pp. 281-315), Hillsdale, NJ: Lawrence Erlbaum Associates.

Shealy, R. T., & Stout, W. F. (1993a). 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*, 159-194.

Shealy, R. T., & Stout, W. F. (1993b). An item response theory model for test bias and differential item functioning. In P. Holland & H. Wainer (Eds.), *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Stevens, J.P. (1992). *Applied Multivariate Statistics for the Social Sciences* (2nd edition). Hillsdale, NJ: Erlbaum.

Stocking, M. L. & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*(2), 201-210.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika, 52*, 589-617.

Stone, C. A. & Yeh, C.-C. (2006). Assessing the dimensionality and factor structure of multiple-choice exams: An empirical comparison of methods using the Multistate Bar Examination. *Educational and Psychology Measurement*, *66*(2), 193-214.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensional assessment and ability estimation. *Psychometrika, 55,* 293-326.

Swaminathan, H. & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27,* 361-370.

Sweeney, K. P. (1996). *A Monte Carlo investigation of the likelihood-ratio procedure in the detection of differential item functioning.* Unpublished doctoral dissertation, Fordham University, New York, NY.

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of response to test items. *Applied Psychological Measurement, 27*(3), 159-203.

Thissen, D. (1991). MULTILOG: Multiple, categorical item analysis and test scoring using item response theory (Version 6.0). Mooresville, IN: Scientific Software.

Tsutakawa, R. K., & Johnson, J. (1990). The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika, 55*(2), 371-390.

UAEU. (2009). Retrieved January 12, 2009, from: http://www.uaeu.ac.ae/

Uaeinteract (2009). Retrieved January 8, 2009, from: http://www.uaeinteract.com/education/

UAE Yearbook. (2009). Retrieved January 7, 2009, from: http://www.uaeyearbook.com/uaeint_misc/pdf_2009/

Von Davier, A. A., Holland, P. W., Thayer, D. T. (2004). *The Kernel Method of Test Equating*. New York: Springer-Verlag. Wainer, H., & Braun, H. L. (1988) *Test validity*. Hillsdale, NJ: Erlbaum.

Waller, N.G. (1998). EZDIF: Detection of uniform and nonuniform differential item functioning with the Mantel-Haenszel and logistic regression procedures. *Applied Psychological Measurement*, 22(4), 391.

Way, W. D., & Tang, K. L. (1991). *A comparison of four logistic model equating methods.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Wilson, D. T., & Wood, R., & Gibbons, R. D.(1987). *TESTFACT: Test scoring, items statistics, and full-information item factor analysis*. Mooresville, IN: Scientific Software International.

Wilson, D., Wood, R., Gibbons, R., Schilling, S., Muraki, E., & Bock, R. D. (2003). *TESTFACT: Test scoring and full information item factor analysis* (Version 4.0). Lincolnwood, IL: Scientific Software International.

Wood, J. M., Tataryn, D. J., & Gorsuch, R. L. (1996). Effects of under- and over-extraction on principal axis factor analysis with varimax rotation. *Psychological Methods, 1*, 345-365.

Yen, W. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. Journal of Educational Measurement, 30(3), 187-213.

Yen, W.M., & Fitzpatrick, A.R. (2006). Item response theory. In R.L. Brennan (Ed.). *Educational measurement*, 4th Ed. Westport, CT: American Council on Education and Praeger Publishers.

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 [Electronic version]. *Educational and Psychological Measurement,63*(1), 51-64.

Zhang, B., & Stone, C. (2004, April). *Direct and indirect estimation of three parameter compensatory multidimensional item response models*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Zhang, J. & Stout, W. (1999). The theoretical detect index of dimensionality and its application to approximate simple structure. *Psychometrika, 64*, 231-249.

Zhu, W. (1998). Test equating: What, why, how? Research Quarterly for Exercise and Sport. 69, 11-23.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Thayer (Eds.), *Differential item functioning* (pp. 337-347), Hillsdale, NJ: Lawrence Erlbaum Associates.

ZU. (2009). Retrieved January 12, 2009, from:
http://www.zu.ac.ae/

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning: Logistic regression modeling as a unitary framework for binary and Likert-type item scores. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

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.

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement, 26*(1), 55-66.