ABSTRACT

Title of Dissertation: INVESTIGATING DIFFERENTIAL ITEM FUNCTION AMPLIFICATION AND CANCELLATION IN APPLICATION OF ITEM RESPONSE TESTLET MODELS

HAN BAO, Doctor of Philosophy, 2007

Dissertation directed by: C. Mitchell Dayton
Department of Measurement, Statistics and Evaluation

Many educational tests use testlets as a way of providing context, instead of presenting only discrete multiple-choice items, where items are grouped into testlets (Wainer & Kiely, 1987) or item bundles (Rosenbaum, 1988) marked by shared common stimulus materials. One might doubt the usual assumption of standard item response theory of local independence among items in these cases. Plausible causes of local dependence might be test takers' different levels of background knowledge necessary to understand the common passage, as a considerable amount of mental processing may be required to read and understand the stimulus, and different persons' learning experiences. Here, the local dependence can be viewed as additional dimensions other than the latent

traits. Furthermore, from the multidimensional differential item functioning (DIF) point of view, different distributions of testlet dimensions among different examinee subpopulations (race, gender, etc) could be the cognitive cause of individual differences in test performance. When testlet effect and item idiosyncratic features of individual items are both considered to be the reasons of DIF, it is interesting to investigate the phenomena of DIF amplification and cancellation resulting from the interactive effects of these two factors.

This dissertation presented a study based on a multiple-group testlet item response theory model developed by Li et al. (2006) to examine in detail different situations of DIF amplification and cancellation at the item and testlet level using testlet characteristic curve procedures with signed/ unsigned area indices and logistic regression procedure. The testlet DIF model was estimated using a hierarchical Bayesian framework with the Markov Chain Monte Carlo (MCMC) method implemented in the computer software WINBUGS. The simulation study investigated all of the possible conditions of DIF amplification and cancellation attributed to person-testlet interaction effect and individual item characteristics. Real data analysis indicated the existence of testlet effect and its magnitudes of difference on the means and/or variance of testlet distribution between manifest groups imputed to the different contexts or natures of the passages as well as its interaction with the manifest groups of examinees such as gender or ethnicity.

INVESTIGATING DIFFERENTIAL ITEM FUNCTION AMPLIFICATION AND CANCELLATION IN APPLICATION OF ITEM RESPONSE TESTLET MODELS

by

Han Bao

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2007

Advisory Committee:

Professor C. Mitchell Dayton, Chair
Professor Paul J. Smith
Professor Amy Hendrickson
Professor George Macready
Professor Robert Lissitz

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Differential item functioning (DIF) is present when individuals of the same ability but from different subpopulations of examinees have different probabilities of success on a given item (Hambleton, Swaminathan & Rogers, 1991). To date, a variety of DIF analysis procedures with strong theoretical bases have been developed (Clauser & Mazor, 1998) and many DIF related studies have been published (Haladyna & Downing, 2004). However, DIF analysis approaches have been limited by the lack of statistical confirmation of hypotheses about DIF-causing dimensions (Roussos & Stout, 1996) and even fewer studies have investigated DIF with testlets. (Thissen, Steinberg, & Mooney, 1989; Wainer, 1995; Wainer & Lewis, 1990; Wainer, Sireci, & Thissen, 1991). This dissertation presents a method based on a testlet item response theory model from a multidimensional DIF analysis framework. Because of its multidimensional modeling approach, the testlet model offers an opportunity to understand the cognitive causes of individual difference in test performance and item functioning and further helps to investigate simultaneous DIF amplification and cancellation at both item level and testlet level. Stout (2002, p. 498) wrote:

> "……the explicitly multidimensional nature of the model allows, and exhorts, us to rigorously study and understand the necessary role of secondary dimensions in causing DIF (Differential item functioning), DBF (Differential bundle functioning), or DTF (Differential testlet functioning). In particular, when several items (perhaps forming a dimensionally homogeneous and substantively interpretable bundle, such as a set of items on a geography test that each require map-reading skills) each depend on the same secondary dimension, the possibility of a large amount of DBF experienced by the focal group at the bundle subset score level caused by individual item *DIF amplification* (see Nandakumar, 1993) becomes an issue. Or, when the influences of multiple secondary dimensions interact, the possibility of *DIF cancellation* (see Nandakumar, 1993, again) of the influence of DBF-producing bundles at the test score level (such as a reading comprehension test where the paragraphs are carefully balanced by content, based on

explicit consideration of gender) or cancellation of the influence of DIF-producing items at the bundle score level, becomes important. Since people are cognitively heterogeneous and since items can not, and should not, be context-free, the notion of DIF cancellation and DBF cancellation is perhaps more important than casual thought first suggests."

Generally, *DIF amplification* means that items within a testlet or bundle (a subset of items sharing common stimulus materials, common item stems, or common item structures) that show no detectable item DIF could show significant DIF when aggregated at the testlet or bundle level. Testlet (or bundle)-level DIF analysis increases the sensitivity of detecting DIF. *DIF cancellation* means that significant item DIF in different directions could be cancelled out within the testlet or bundle (Wainer, 1995).

Beyond Stout's perspectives, DIF amplification and cancellation could occur both at the item level and testlet level because the possible causes of DIF might be the secondary dimensions and also idiosyncratic features of individual items functioning homogeneously or heterogeneously among different groups. When the secondary dimensions and item difficulty attributes all favor one of the groups across items within a testlet, more significant DIF should be detected at the item level and could be even more obvious at the testlet level; when the secondary dimensions and item difficulty attributes favor different groups, DIF could be cancelled at the item level but might be significant when cumulated at the testlet level; when the secondary dimensions and item attributes favor the same group for some of the items within testlet but function on the contrary for the rest of items within testlet, DIF could be amplified at the individual item level but cancelled out at the testlet level. The accumulated DIF amplification and cancellation at the testlet level is highly related to the situation of simultaneous DIF amplification and cancellation at the item level. An aggregate DIF effect at the testlet level is of more interest.

Study of DIF amplification and cancellation can be very useful for test construction purposes. Undetectable item DIF accumulated at the testlet level would increase the sensitivity of detection which is especially useful for those focal groups that are relatively rare in the examinee population. A certain amount of item DIF cancelled out at the testlet level provides a solution to yield a perfectly acceptable test construction unit which is especially important in adaptive testing where the test is usually built out of testlets (Wainer, 1995).

Recent trends in test construction toward focusing tests on a level larger than individual items indicate a favorable future for the use of testlets (Wainer, Sireci and Thissen, 1991). These context-dependent items are often regarded as more realistic and possibly even better for measuring problem-solving in a context that is difficult to develop in a single item. However, these situations call into question the assumption of local independence of item response theory. Under a fixed-effects approach, the local dependence due to shared variation among items in a testlet can be expressed in terms of response *patterns* within an item bundle (Wang & Wilson, 2005; Wilson & Hoskens, 2001; Hoskens & De Boeck, 1997). Under a random-effects approach, Bradlow, Wainer and Wang (1999) extended Birnbaum's two-parameter model to include an additional random effect for modeling local dependence within each given testlet. Plausible causes of local dependence might be test takers' different levels of background knowledge necessary to understand the common passage as a considerable amount of mental processing may be required to read and understand the stimulus and different persons' learning experiences. Here, the local dependence can be viewed as an additional dimension other than the latent trait. A random testlet effect captures the interaction

between examinee and testlets beyond the latent trait of interest and individual item parameters. From the multidimensional DIF point of view, the multi-group testlet model helps us to differentiate between DIF and *impact*, where the former is due to both the different distributions of testlet factors for different examinee subpopulations and idiosyncratic features of individual items and the latter is due to the actual ability differences between groups in proficiency intended to be measured. Moreover, the testlet effect provides reasons for group differences on a set of items found within the test specifications that might prove more useful in explaining why a bundle of items functions differentially between two groups matched on abilities (Douglas, Roussos & Stout, 1996).

The possible sources of nuisance dimensions related to the testlet factor could be the content and cognitive dimensions associated with passages and the possible sources of item attributes could be the item type or format, negatively worded item stems, and the presence of pictures or other reference materials such as tables, charts and diagrams. For example, the literature background knowledge underlying the passage might be advantageous to the female group relative to the male group. On the other hand, several items within the testlet might require a specific form of inferential reasoning skills, which are in favor of the male group instead. Both of the factors can be present and function together and it provides us a great opportunity to study DIF amplification and cancellation at the item and testlet level. By searching out possible sources of or patterns to the occurrence of DIF over items within a testlet, hypotheses as to the sources of DIF can be checked with more confidence because of the presence of more items for analysis.

By communicating those results to item writers, any patterns or trends detected can be used to assist in developing a protocol for creating items less likely to be inappropriate.

This study was intended to investigate the interactive effects of secondary testlet dimension and item attributes on the phenomena of DIF amplification and cancellation at both of item and testlet levels in application of multiple-group Testlet Item Response Theory Model developed by Li et al. (2006). Instead of Li's approach of estimating a multi-group testlet model using the MML-EM algorithm and detecting DIF using the Item Response Theory (IRT) likelihood ratio test, the testlet DIF model was estimated using a hierarchical Bayesian framework with the MCMC method implemented in the computer software WINBUGS1.4 (Spiegelhalter, et al., 2000). The purpose of this study is to examinate in detail different situations of DIF amplifications and cancellations at the item and testlet level using Testlet Characteristic Curve procedure with Signed/ Unsigned area indices and Logistic Regression procedure and to present policy implications based on our findings. There are two ways of thinking about testlet DIF that relate to the "amplification and cancellation" phenomenon. They have to do with whether DIF is modeled (1) at the level of any given testlet as a DIF parameter applying constantly or non-constantly to all items in the testlet, or (2) at the level of individual items, each one having its own item DIF parameters. The study has been conducted using real and simulated datasets.

Overview of later chapters

Chapter 2 provids some background on DIF, DIF amplification and cancellation, including definitions and the types of DIF, reviews of item response testlet models and

reviews of the statistical methods used for detecting DIF and their extensions for explaining DIF. Chapter 3 briefly introduces the MCMC method of Bayesian parameter estimation based on Gibbs sampling and presents analysis designs for simulation study and real data analysis. Chapter 4 discusses the estimation of parameters with WinBUGS, shows some simulation results, applies the proposed models and methods to real test data and summarizes our methods and results. Chapter 5 contains conclusions of results from simulation study and real data analysis, and discusses limitations and possible future work.

Chapter Two: Literature Review

In studies of construct validity, the assessment of DIF is an important issue in evaluating the inferences from educational and psychological testing. DIF occurs when test items display different statistical or psychometric properties in different groups of examinees after matching on the same intended-to-be-measured underlying proficiency, θ. In other words, the absence of differential item functioning occurs when the item responses and group variables are independent after conditioning on ability (Millsap & Everson, 1993). Although the assessment of test fairness is often based on the single-item DIF analysis, DIF can occur at the subtest and test score level (Stout, 2002). As a natural extension of DIF, differential testlet/bundle functioning and differential test functioning can be defined when the expected subscores on testlets or interpretable bundles of items (test scores) differ across groups of examinees with the same intended-to-be-measured latent traits.

Along with the differential testlet/bundle functioning and differential test functioning analysis, the issue of DIF amplification and cancellation was of interest. Nandakumar (1993) has argued that the possibilities of DIF amplification and cancellation should be investigated in any DIF analysis. At the test score level, several empirical studies of DIF cancellation have been done by Drasgow (1987), Roznowski (1987), and Reith and Roznowski (1991). They concluded that when the sources of DIF are diverse, it might cancel out the cumulative DIF effects across groups might canceled out at the test score level. Nandakumar (1993) provided a systematic study of the phenomena of simultaneous DIF amplification and cancellation at the test score level using SIBTEST (Shealy & Stout, 1993). In application of the two-dimensional three-

parameter logistic model with compensatory abilities (Reckase & McKinley, 1983), a variety of simulation studies have been done to investigate DIF amplification when assuming items with only one nuisance ability and DIF cancellation when assuming items with two different nuisance abilities. As demonstrated in simulation and real data analysis, the cumulative effect of DIF could either be amplified or cancelled out partially or completely because of the multidimensionality nature of the test.

Comparatively few studies of simultaneous DIF have been done at the testlet level. Wainer et al. (1991) developed the definition of differential testlet functioning (DTF), stated the advantages of simultaneous DIF at the testlet level under conceptual framework, and detected testlet DIF using Bock's (1972) model. Later, empirical studies of the differential performance of items within testlets were investigated by Wainer (1995) using Mantel-Haenszel methods. However, fitting a polytomous item response model to testlets restricted its applications in detecting DIF at the test level by throwing away the information of individual items. Shealy and Stout (1993) proposed a nonparametric procedure to study the systematical nuisance dimension using differential bundle functioning (DBF) at the item as well as at the testlet level. A limitation of their approach is that it could display DIF of items in a testlet on average but no detailed information about individual item functioning within a testlet. Recently, Li et al. (2006) posed a multi-group testlet model using a marginal maximum likelihood estimation method based on the EM algorithm and provided a stepwise mechanism to detect two kinds of sources of DIF at the testlet level: testlet factors and item characteristics.

In contrast to those previous studies in this area, the significance of this research is as follows: First, here the testlet DIF model is used to specify the testlet effect as a second

dimension intended /not intended to be measured, which is partially or completely responsible for the DIF. Second, compared to Nandakumar's (1993) belief that multiple nuisance dimensions contribute to DIF amplification and cancellation within a set of items, we consider that testlet effects might function consistently or non-consistently and item characteristics that can be another source of DIF besides the nuisance dimension. Through taking both of the testlet factor and item-specific factor into consideration, the multiple-group Testlet model helps to detect DIF amplification and cancellation at the item as well as testlet level.

Identifying the causes or substantive themes that characterize items exhibiting DIF used to be a fundamental problem in the study of group differences using DIF methods. Roussos and Stout (1996) proposed a *multidimensionality-based DIF analysis paradigm* to bridge the gap between statistical and substantive analyses by linking both to the Shealy-Stout multidimensional model for DIF (Shealy & Stout, 1993). It has been recognized and accepted that the general cause of DIF is the presence of multidimensionality in items displaying DIF; that is, such items measure at least one dimension in addition to the primary dimension(s) the item is intended to measure (Lord, 1980). Each second dimension is further categorized as either an *auxiliary* dimension if the secondary dimension is intended to be measured or as a *nuisance* dimension if the secondary dimension is not intended to be measured. It is referred to as *benign* DIF when it is caused by an *auxiliary* dimension; or it is referred to as *adverse* DIF when it is caused by a *nuisance* dimension (Roussos & Stout, 1996). For example, in accordance with the testlet effect, content knowledge is of interest in some reading comprehension tests and moreover, it can be judged to favor one manifest group over the other (e.g.,

males vs. females) on specific subject domain. However, sometimes the passage effect causing the dependence among items can be considered a nuisance dimension in the sense that it intrudes on the intended focus of the test. Therefore, once the evidence that local item dependence reflects *benign* DIF or *adverse* DIF has been investigated, the testlet DIF analysis could give us evidence about the sources of DIF and also the amount of DIF by checking testlet effect.

In the remainder of this section, statistical definitions of DIF, multidimensionality-based DIF and testlet DIF were introduced; next, the history and development of item response theory testlet models were overviewed and statistical methods for detecting DIF in literature are briefly reviewed and compared. Then, the methods used for this study of DIF amplification and cancellation were introduced in more detail.

Statistical Definitions of DIF

The first formal statistical definition of DIF (Definition 1) is stated at a general level to include the form of multidimensionality-based DIF (Definition 2) and can be extended to cumulative DIF at the testlet/bundle level (Definition 3).

Definition 1: General DIF

$$E[Y \mid W = w, G = R] \neq E[Y \mid W = w, G = F] \qquad \forall w, \tag{2.1}$$

where

Y is the observed score on a single item or the observed scores on a set of items; Y can be a dichotomous or a polytomous score;

W is the latent proficiency variable for which Y is the observed indicators and W can be univariate (when there is only one latent trait intended to be measured) or multivariate (when where are several latent traits intended to be measured);
G is a group variable indicating demographic information, such as ethnicity, gender, or age, or a latent class indicator.

Generally, two groups are defined where R represents a reference group and F represents a focal group. Under the item response theory framework, this definition is equivalent to non-overlapping item response functions or item characteristic curves of the two groups.

Definition 2: Multidimensionality-based DIF

$$E[Y(W_2 = \vec{\eta}) \mid W_1 = \vec{\theta}, G = R_1] \neq E[Y(W_2 = \vec{\eta}) \mid W_1 = \vec{\theta}, G = F] \quad \forall W_1, W_2, \qquad (2.2)$$

where

W_1 represents a vector of main latent proficiencies, $\vec{\theta}$, intended to be measured;

W_2 represents a vector of secondary latent proficiencies, $\vec{\eta}$.

If η is a *nuisance* dimension not intended to be measured, then the difference on observed value Y between focal group and reference group is called *Adverse* DIF; If η is an *auxiliary* dimension intended to be measured, then the difference on observed value Y between focal group and reference group is called *Benign* DIF. This definition illustrates conditioning on the same distributions of main latent proficiencies; it is the secondary dimensions that lead to differential performances between the focal group and reference group. Moreover, it is important to differentiate *Impact* from DIF when the different

performances of the two groups are attributed to the different distributions of main latent proficiencies: $E[Y(W_1 = \vec{\theta}) \mid G = R] \neq E[Y(W_1 = \vec{\theta}) \mid G = F] \quad \forall W_1$.

Definition 3: Testlet DIF Effects

Let $S = \{Y_{i_1}, Y_{i_2}, \cdots, Y_{i_T}\}$ be any subtest of items to be studied for differential testlet/bundle functioning and define

$$B = \sum_{t=1}^{T} (E[Y_{i_t} \mid W = w, G = R] - E[Y_{i_t} \mid W = w, G = F]) \quad \forall w, \tag{2.3}$$

where

B is the cumulative DIF effects of all of the items within a testlet/bundle;

B could be positive, zero or negative. Here, $\sum_{i=1}^{T} (E[Y_{i_T} \mid W = w, G = R])$ and $\sum_{i=1}^{T} (E[Y_{i_T} \mid W = w, G = F])$ are equivalent to the *testlet characteristic curves* of the reference and focal groups in item response theory respectively. And thus, B is equivalent to the signed-area index of measuring the areas between the two testlet characteristic curves of the reference and focal groups.

Item Response Theory as Applied to Differential Item Functioning

Item Response Theory includes a family of mathematical models that specify probabilistic relationships between a person's item response and the person's underlying proficiency levels and item characteristics. It is useful for detection of DIF because DIF can be modeled through the use of estimated item parameters and latent traits, and different item functions between two groups can be described in a precise and graphical manner (Hambleton, Swaminathan & Rogers, 1991).

Item response theory models can vary in the number of dimensions representing the underlying proficiencies of interest, the dichotomous or polytomous scoring of the item response, and the number of item parameters and the normal ogive/logistic formats of the model. There are three standard unidimensional models: one-parameter, two-parameter, and three-parameter logistic models. General testlet models have been developed based on these standard unidimensional models (e.g., two-parameter and three-parameter models) by adding an item-testlet interaction effect parameter. Extending from the general testlet model, the multiple-group testlet model may offer particular advantages in the study of DIF.

I. One-parameter Logistic Model (I-PL model or Rasch model)

The Rasch model (Rasch, 1960) is the simplest of unidimensional models, and can be given by the formula:

$$P(y_{ij}=1)=\frac{e^{(\theta_j-b_i)}}{1+e^{(\theta_j-b_i)}} \quad , \tag{2.4}$$

where

y_{ij} is examinee j's response category to item i;

P (y_{ij}=1) is the probability that examinee j answers item i correctly;

θ_j is examinee j's proficiency level;

b_i is the difficulty parameter of item i, which indicates the point on the ability continuum when an examinee has a 50% probability of answering item i correctly.

An assumption that is implicit in the model is that it assumes that all items have the same discrimination value.

II. Two-parameter Logistic Model (2-PL model)

The 2-PL model proposed by Birnbaum (1957, 1958a, 1958b) extends the 1PL model by adding item discrimination parameter, and is given by the formula:

$$P(y_{ij}=1\ )=\frac{e^{(a_i(\theta_j-b_i))}}{1+e^{(a_i(\theta_j-b_i))}}, \tag{2.5}$$

where a_i is a discrimination parameter of item i and all other terms in the formula have the same interpretations as in (2.4).

With the discrimination parameter, the 2-PL model allows for variation in item discrimination parameter across items in a test. Both the 1-PL model and 2-PL model assume there is no impact of guessing on item responses, which will usually be violated for a test composed of multiple-choice items.

III. Three-parameter Logistic Model (3-PL model)

To take into account the impact of guessing, Birnbaum (1968) proposed the 3-PL model, which is given by

$$P(y_{ij}=1\ )=c_i+(1-c_i)\frac{e^{(a_i(\theta_j-b_i))}}{1+e^{(a_i(\theta_j-b_i))}}\ , \tag{2.6}$$

where

c_i is the lower asymptote or pseudo-chance parameter for item j indicating the probability of answering item j correctly for persons having no ability. All other terms in the model have the same interpretations as in (2.5).

IV. Testlet Model

The cornerstone of item response theory is the assumption of *local independence*. Local independence posits that an examinee's response to a given test item depends on an unobservable examinee parameter, θ, but not on the identity of or responses to other items that may have been presented to the examinee (Lord, 1980). More formally, it is

asserted that responses to test items are conditionally independent, given item parameters and θ. Local independence can be violated when a test consists of items nested in testlets, where groups of items share a common stimulus.

The 3PL testlet (3PL-t) model proposed by Wainer, Bradlow, and Du (2000) and Du (1998) is an extension of Birnbaum's (1968) 3PL model in which local dependence is specifically modeled by adding a random effect parameter, γ. This dependency is assumed to be unique to a testlet and is considered a second dimension in the sense that it is different from the intended focus of the test. The probability that examinee j answers item i correctly in the 3PL-t model is given by,

$$P_{ij} = c_i + (1 - c_i)\frac{e^{(a_i(\theta_j - b_i - \gamma_{jd(i)})}}{1 + e^{(a_i(\theta_j - b_i - \gamma_{jd(i)})}}, \qquad (2.7)$$

where

$\gamma_{jd(i)}$ is a random effect representing the interaction of person j with testlet $d(i)$ (i.e., the testlet that contains item i).

All other terms in the model have the same interpretations as in (2.6). The addition of the γ parameter reflects the effect of this nuisance dimension. The value of $\gamma_{jd(i)}$ is constant within a testlet for person j, but the value of $\gamma_{jd(i)}$ differs for each person. The variances of γ are allowed to vary across testlets and indicate the amount of local dependence in each testlet. The items within the testlet can be considered conditionally independent if the variance of γ is zero. The amount of local dependence increases as the variance of γ increases.

V. Multiple-group Testlet Model

As an extension and application of the random-effects approach using the testlet model, the main interest of this dissertation lies in detecting whether and how testlets function differently for individuals with different group membership. That is, different genders, ethnic groups, etc. may have different mental processes, levels of background knowledge, or learning experiences, which cause the amount of local dependence between items within the testlets to differ across these groups. The multiple-group testlet model is given in the following formula:

$$P(Y_{ijg}=1\,|\,\Omega_{ijg}) = c_{ig} + (1-c_{ig})\frac{e^{(a_{ig}(\theta_j - b_{ig} + \gamma_{jd(i)g}))}}{1+e^{(a_{ig}(\theta_j - b_{ig} + \gamma_{jd(i)g}))}}, \tag{2.8}$$

where

$P(Y_{ijg}=1)$ denotes the probability that examinee j =1, … , J of group g receives score 1 on item i; generally, there are two groups: the focal group and the reference group;

Ω_{ijg} is the vector of parameters $(a_{ig}, b_{ig}, c_{ig}, \theta_j, \gamma_{jd(i)g})$;

a_{ig}, b_{ig}, c_{ig} are the item slope parameter, item difficulty parameter and "guessing" parameter of group g;

θ_j represents the proficiency of examinee j;

$\gamma_{jd(i)g}$ is the interaction of person j in group g with item i nested in the testlet d(i).

A special case of Model (2.8) has $c_{ig}=0$ for all groups, and then it is the multiple-group 2-PL testlet model proposed by Bradlow, Wainer and Wang (1999). The 2-PL multiple-group testlet model is given by,

$$P(Y_{ijg}=1\,|\,\Omega_{ijg}) = \frac{e^{(a_{ig}(\theta_j - b_{ig} + \gamma_{jd(i)g}))}}{1+e^{(a_{ig}(\theta_j - b_{ig} + \gamma_{jd(i)g}))}}, \tag{2.9}$$

Glas et al. (2000) discuss three alternative ways of model formulation regarding to the testlet parameter: (1). as part of ability, assuming the item parameters are constant across all examinees; (2). as part of difficulty, by grouping testlet effect as part of item difficulty; (3). as an independent entity, by separating the testlet parameter from both ability and difficulty. Treating the testlet parameter as part of item difficulty, Wang and Wilson (2005) presented a procedure for detecting differential item functioning in testlet-based tests, where DIF was taken into account by adding DIF parameters into the Rasch testlet model. Here from the multidimensionality-based DIF point of view, the testlet parameter is treated as a second dimension other than the primary proficiency of interest. Therefore, the model is defined differently by adding testlet parameter instead of subtracting the testlet parameter.

From the mathematical definition of the model, there are two potential sources of DIF: (1) the random person-testlet interaction effect γ and (2) item characteristic parameters (a, b). Considering DIF at the testlet level, the difference caused by γ of each item might be amplified at the testlet level keeping the item parameters the same for both the focal group and reference group across all items in the testlet. On the other hand, because of its own characteristics, each item in the testlet might not function consistently for the two groups although they have the same γ. These two sources might function simultaneously. Larger γ values and smaller b values for one of the group than those for the other group or smaller γ values and larger b values for one of the group than those of the other group would lead to DIF amplification at the individual item level; on the contrary, larger γ values for one of the group and larger b values for the same group or smaller γ values and smaller b values for the same group of examinees would lead to DIF

cancellation at the individual item level. Items with small but systematic DIF may go statistically unnoticed, but when combined into a testlet, DIF may be detected at the testlet level. This is referred to as amplification at the testlet level; Items within testlet with large and un-systematic DIF could be statistically noticed, but when combined, DIF may be cancelled at the testlet level. This is referred to as cancellation at the testlet level. This dissertation is to study the pattern of DIF amplification and cancellation of items in testlets modeled by the multiple-group 2-PL testlet model.

Statistical Methods for Detecting DIF

Currently there are numerous methods for conducting DIF assessment for dichotomously and polytomously scored items (see Millsap & Everson, 1993, and Potenza & Dorans, 1995, for reviews). Some techniques are based on IRT such as the area between two item response functions (Rudner, 1977; Rudner, Getson, & Knight, 1980), Lord's (1980) χ^2 test, Thissen, Steinberg, & Wainer's (1988) likelihood ratio test and Shealy and Stout's (1993) simultaneous item bias test (SIBTEST); others do not use IRT, such as the Mantel-Haenszel (MH) method (Holland & Thayer, 1988) and the logistic regression procedure (Swaminathen & Rogers, 1990).

Within the item response theory framework, item characteristic curves provide a means of comparing the response of two different groups matched on ability. In other words, DIF may be investigated whenever the conditional probabilities of correct response differ for the two groups. Item Characteristic Curves (ICCs) determined by their discrimination parameter, difficulty parameter or guessing parameter can be graphed to broaden our understanding of items showing DIF. Two categories of DIF: *Uniform* and

Nonuniform (or Crossing DIF) can be described graphically by ICCs. *Uniform DIF* exists when the ICCs for the two groups do not cross over the entire ability range. Thus one group performs better than the other group at all ability levels. Nonuniform DIF exists when the ICCs for the two groups cross at some point on the θ scale. Thus DIF for and against a group might cancel out to a certain amount. The same procedure can be applied to the *testlet characteristic curve* (the expected true score curves) obtained by summing ICCs across items in a testlet within groups, and comparing these testlet characteristic curves across groups.

Among the several approaches in the IRT framework, the signed-area/unsigned-area procedures provide an index that quantifies the difference between two ICCs, which can be applied to testlet characteristic curves. SIBTEST is a non-parametric multidimensional-based IRT approach, which can be used to test the hypotheses of uniform DIF/ nonuniform DIF (unidirectional DIF/crossing DIF using Li and Stout's terminology). Lord's chi-square test is used to test the equality of the parameters of the ICCs. The likelihood ratio test is used to test the model fit. For the statistical methods not using IRT, Mantel-Haenszel is a nonparametric statistical approach using an estimated constant odds ratio to provide a measure of effect size for evaluating the amount of DIF, which is designed to detect uniform DIF. Logistic regression is a parametric approach used to detect both uniform and non-uniform DIF. In the current study of DIF, since appropriate for both uniform DIF and crossing DIF, signed/unsigned area procedures and logistic regression approach are to be used, and these two approaches are reviewed briefly.

I. Signed-area/Unsigned-area indices

Rudner (1977; Rudner, Getson & Knight, 1980) proposed that DIF can be defined mathematically through the following formulas:

$$SIGNED-AREA=\int[P_R(\theta)-P_F(\theta)]d\theta, \tag{2.9}$$

$$UNSIGNED-AREA=\sqrt{\int[P_R(\theta)-P_F(\theta)]^2 d\theta}\ . \tag{2.10}$$

Note that the probability of a correct response for the focal group is subtracted from that of the reference group. DIF effect size based on areas between item response functions (Raju, 1988) is set to 0.4 to reflect moderate DIF and 0.8 to reflect large DIF. The signed-area index is appropriate for uniform DIF and unsigned-area index is appropriate to detect nonuniform DIF. The advantage of the simple area indices is that they can be easily graphed and visualized; the disadvantages are that they are not accurate when the highest density of examinees are located at the extreme region of the ability scale, and are not appropriate when the guessing parameters for the two groups are unequal. Additionally, there are no associated tests of significance.

II. Logistic Regression

Swaminathan and Rogers (1990) proposed the use of logistic regression for DIF detection through introducing estimated coefficients for group, total score, and the interaction of the total score and group and testing for significance with a model comparison strategy. The general logistic regression model may be written as:

$$P(Y=1)=\frac{e^{(\psi)}}{1+e^{(\psi)}}, \tag{2.11}$$

where

$$\psi=\tau_0+\tau_1\theta+\tau_2 G+\tau_3(\theta G),$$

Y is the examinee's item response score coded as 1 (right) or 0 (wrong);

θ is the estimated examinee's latent trait value;

G is the group index, coded as 1 (Focal group) or 2 (Reference group);

τ_0 represents the weight associated with the intercept;

τ_1 indicates the ability differences between subgroups of examinees in the propensity to get the item right; when τ_1 is statistically significant, it means that the examinees with higher ability levels have better odds of getting the item right.

τ_2 is the combined odds ratio; when τ_2 is statistically different from zero, it means that the odds of getting an item right are different for the two groups.

τ_3 is the interaction of group and estimated latent trait score; and when τ_3 is statistically significant, it means that the item shows larger differences in group performance at some ability levels than at others.

The direction of each regression coefficient (τ) could provide the information about whether the focal group or the reference group is favored**.** Zumbo (1999) suggested three steps for hypothesis testing of uniform DIF and non-uniform DIF and provided an index to measure the amount of DIF by computing the difference of the squared multiple correlations (R^2). Regarding flexibility in specification of the regression equation, this approach can incorporate more than one ability estimate into a single regression analysis to obtain more accurate matching criteria and to differentiate multidimensional item impact from DIF (Mazor, Kanjee, & Clauser, 1996).

Extending from Zumbo's (1990) three steps of hypothesis testing, here a five-step process is recommended to accommodate the four parameters (e.g., $(a_{ig}, b_{ig}, \theta_j, \gamma_{jd(i)g})$) of the multiple-group 2-PL testlet model: $\log it(Y=1|\theta,\gamma) = a_{ig}\theta_{jg} + a_{ig}\gamma_{jd(i)g} - a_{ig}b_{ig}$.

Step1: The matching or conditioning variable (e.g. the estimated examinee's latent trait score) is entered into the regression equation,

Model 1: $\psi = \tau_0 + \tau_1\theta$

This serves as the baseline model.

Step 2: The testlet parameter is entered into the regression,

Model 2: $\psi = \tau_0 + \tau_1\theta + \tau_2\gamma$

The effect of testlet parameter can be investigated by checking the improvement in R-squared based effect size against model 1; that is, Model 2 is compared to Model 1.

Step 3: The group variable is entered into the regression equation,

Model 3: $\psi = \tau_0 + \tau_1\theta + \tau_2\gamma + \tau_3 G$

The presence of uniform DIF can be tested by examining the improvement in R-squared based effect size associated with adding a term for group membership (G) against model 2. That is, Model 3 is compared to Model 2.

Step 4: The interaction term based on the main dimension of θ is added,

Model 4: $\psi = \tau_0 + \tau_1\theta + \tau_2\gamma + \tau_3 G + \tau_4(\theta G)$

The presence of crossing DIF occurring on the θ scale can be tested by examining the improvement in R-squared based effect size associated with adding a term for the interaction between the estimated latent trait score and group membership ($\theta * G$) against Model 3. In other words, Model 4 is compared to Model 3.

Step 5: The interaction term based on the nuisance dimension is finally added,

Model 5: $\psi = \tau_0 + \tau_1\theta + \tau_2\gamma + \tau_3 G + \tau_4(\theta G) + \tau_5(\gamma G)$

The presence of crossing DIF occurring on the γ scale can be tested by examining the improvement in R-squared based effect size associated with adding an additional term

from Model 3 for the interaction between estimated nuisance latent trait scores and group membership ($\gamma * G$).

Additionally, Zumbo (1999) provided a measure of DIF effect size, called ΔR^2, which is the difference in the R-squared values at each step of DIF modeling. ΔR^2 is given as:

$$\Delta R^2 = R_2^2 - R_1^2 \ , \tag{2.13}$$

where

R_2^2 and R_1^2 are the sums of the products of the standardized regression coefficients for each explanatory variable and the correlation between the response and each explanatory variable for the augmented and baseline models, respectively.

Jodoin and Gierl (2000) recently presented guidelines for measurement of magnitude of overall DIF. Negligible DIF: Null hypothesis is retained or null hypothesis is rejected and $\Delta R^2 < 0.035$; Moderate DIF: Null hypothesis is rejected and $0.035 \leq \Delta R^2 \leq 0.070$; large DIF: Null hypothesis is rejected and $\Delta R^2 \geq 0.070$. These guidelines are applicable to both uniform and non-uniform DIF.

Chapter 3 Methodology and Research Design

In item response models, the item parameters and the person parameters are often estimated simultaneously. Generally, the ability parameters are assumed to be incidental parameters and the item parameters as the structural parameters. (Neyman & Scott, 1948**;** Andersen, 1972). In general, the incidental parameters are estimated to have consequences that are increased to yield stable estimates of the structural parameters. The basic problem pointed out by Neyman and Scott is that the maximum likelihood estimators of the structural parameters are not consistent in the presence of incidental parameters. Andersen (1972) demonstrated this in the case of the Rasch model when he showed that with a fixed number of items, maximum likelihood estimation fails when a perfect score or a zero score (score of examinee approaching infinity) is encountered. He further showed that consistent maximum likelihood estimators can be obtained by conditioning the likelihood function on the number correct score, the minimal sufficient statistic for the ability parameters in the Rasch model. Unfortunately, sufficient statistics for the ability parameters are not available in the two-parameter (or the three-parameter) logistic model. Hence, conditional maximum likelihood estimators of the item parameters cannot be obtained for the two- and three-parameter models. Bock and Lieberman (1970) derived "marginal maximum likelihood" estimators of the item parameters in the normal ogive model by integrating out the ability parameters. Since this procedure requires numerical integration and the evaluation of the likelihood function over 2^n response patterns, where n is the number of items, the procedure becomes tedious whenever n is even moderately large. Bock and Aitkin (1981), employing the

expectation-maximization (E-M) algorithm, obtained marginal maximum likelihood estimators of the item parameters more efficiently.

When several parameters have to be estimated simultaneously, and when, as in item response models, structural as well as incidental parameters have to be estimated, a Bayesian approach may be appropriate (Zellner, 1971**,** p. 112-114). This is particularly true when prior information about the parameters is available, since the incorporation of such information will certainly increase the meaningfulness and the "accuracy" of the estimates.

In order to estimate the parameters of the 2-PL Testlet model and the 3-PL Testlet model, Wainer, et al. (1999, 2000) used the Markov Chain Monte Carlo (MCMC) procedure in the full Bayesian framework. Glas et al. (2000) compared marginal maximum likelihood (MML) and expected a posteriori (EAP) estimation with MCMC and pointed out that MML and MCMC estimates for the testlet model were highly correlated but MCMC provided more accurate interval and point estimation results. Moreover, MCMC is more promising for evaluating complex IRT models, such as the testlet model, from the multidimensional perspective although the tradeoff is its intensive computation Wainer, et al. (2000). Since the testlet models were developed under the Bayesian hierarchical framework, the specifications of prior distributions of item and person parameters have been testified. Moreover, since our interest lies in investigating the phenomena of DIF cancellation and amplification attributed to both the testlet factor and each item's idiosyncratic features, including item difficulty and item discrimination parameters and thus a certain degree of accuracy in estimation of item parameters was in demand, the MCMC method was chosen for estimating parameters for the multiple group

2-PL testlet model. First, the model estimation procedure is introduced and then the research design for both simulated and real data is introduced.

Model Estimation under Bayesian Framework

The use of MCMC estimation for IRT models was introduced by Patz and Junker (1990a), and has since been used to estimate a variety of models. Embedding the multiple-group 2-PL testlet model into the larger Bayesian hierarchical framework allows for more precise and meaningful parameter estimates to be obtained (Swaminathan & Gifford, 1985). Under the 2-PL testlet model, the probability that examinee j answers a dichotomous item i correctly or incorrectly can be expressed as:

$$P(Y_{ijg} = 1 \mid \Omega_{ijg}) = \frac{e^{(a_{ig}(\theta_{jg} - b_{ig} + \gamma_{jd(i)g}))}}{1 + e^{(a_{ig}(\theta_{jg} - b_{ig} + \gamma_{jd(i)g}))}}, \tag{3.1}$$

or

$$P(Y_{ijg} = 0 \mid \Omega_{ijg}) = \frac{1}{1 + e^{(a_{ig}(\theta_{j} - b_{ig} + \gamma_{jd(i)g}))}}. \tag{3.2}$$

Suppose that J examinees each take an examination consisting of I items; the I items are nested within K testlets, which is, $d(i) \in \{1, \ldots, K\}$, where $k_{d(i)}$ and $\mathrm{K}_{d(i)}$ represent the number of items nested within testlet d(i) and number of testlets. We can get the likelihood for an observed test score matrix $Y = (y_{ijg})$ as:

For the dichotomous case,

$$P(Y \mid \Delta_1) = \prod_{j=1}^{J} \{ \prod_{i \in I} [\frac{e^{a_{ig}(\theta_{jg} - b_{ig} + \gamma_{jd(i)g})}}{1 + e^{a_{ig}(\theta_{jg} - b_{ig} + \gamma_{jd(i)g})}}]^{y_{ijg}} [\frac{1}{1 + e^{a_{ig}(\theta_{jg} - b_{ig} + \gamma_{jd(i)g})}}]^{1 - y_{ijg}} \}, \tag{3.3}$$

where $\Delta_1 = \{\vec{\theta}_g, \vec{a}_g, \vec{b}_g, \vec{\gamma}_g\}$ is the set of likelihood parameters. Note that we assume local independence exists after conditioning on the main dimension θ and the nuisance dimension, testlet effect, γ.

The likelihood parameters Δ_1 in the model are assumed to have normal prior distributions in order to take advantages of conjugate features and they are specified as:

$$\theta_{jg} \sim Normal(0,1), \tag{3.4}$$

$$a_{if} \sim \log(Normal(\mu_{af}, \sigma_a^2)), \tag{3.5}$$

$$a_{ir} \sim \log(Normal(\mu_{ar}, \sigma_a^2)), \tag{3.6}$$

$$b_{if} \sim Normal(\mu_{bf}, \sigma_b^2), \tag{3.7}$$

$$b_{ir} \sim Normal(\mu_{br}, \sigma_b^2), \tag{3.8}$$

$$\gamma_{jd(i)f} \sim Normal(\mu_{\gamma(d)f}, \sigma_{\gamma(d)f}^2), \text{ d=1,…, K} \tag{3.9}$$

$$\gamma_{jd(i)r} \sim Normal(\mu_{\gamma(d)r}, \sigma_{\gamma(d)r}^2), \text{ d=1,…, K} \tag{3.10}$$

Assuming DIF is caused by effects other than the main proficiency of interest, we set the ability distribution of θ_g as a standardized normal distribution across the focal group and reference group to identify the model. Unlike Wainer, et al.'s testlet models, the item discrimination parameter and item difficulty parameters will be manipulated to have different means but the same variances for the two groups; the prior distribution of testlet effect γ has different means and different variances for the two groups.

The joint posterior density function given the data (Y) is the product of the likelihood function and the prior distribution functions of all the unknown parameters (Δ_1).

$$p(\Delta_1 \mid Y) \propto \{\prod_{I=1}^{I}\prod_{j=1}^{J}\left[\frac{e^{a_{ig}(\theta_j - b_{ig} + \gamma_{jd(i)g})}}{1+e^{a_{ig}(\theta_j - b_{ig} + \gamma_{jd(i)g})}}\right]^{y_{ij}}\left[\frac{1}{1+e^{a_{ig}(\theta_j - b_{ig} + \gamma_{id(i)g})}}\right]^{1-y_{ij}}\} *$$

$$\{\prod_{j=1}^{J} e^{-\frac{1}{2}\theta_j^2}\}\{\prod_{i=1}^{I} e^{-\frac{1}{2\sigma_a^2}(a_{ig}-\mu_{ag})}\}\{\prod_{i=1}^{I} e^{-\frac{1}{2\sigma_b^2}(b_{ig}-\mu_{bg})}\}\{\prod_{d=1}^{K}\prod_{i=1}^{I} e^{-\frac{1}{2\sigma_{\gamma(d)g}^2}(\gamma_{jd(i)g}-\mu_{\gamma(d)g})^2}\}.$$

(3.11)

The goal of Bayesian modeling is to define a posterior distribution as opposed to arriving at a point estimate and its confidence interval for each of the unknown (Δ_1). After specification of all of the required prior distributions, an MCMC computation approach, the Gibbs sampler, is used to estimate the unknown parameters by sampling from their posterior distributions, conditioning on the previously drawn values of all other parameters and the data. Assuming certain regularity conditions hold (see Tiemey, 1994**)** these chains of values for each parameter will eventually converge to a target distribution, which is the posterior distribution of the parameters in the model. After stationauty is attained, future draws will be distributed like draws from the posterior distribution, and parameter estimates can be obtained by sampling from that distribution.

The Gibbs sampler is a Markovian updating scheme that proceeds as follows (Gelfand & Smith, 1990). Generally, let $\theta = \theta_1, \theta_2, ..., \theta_R$ denote the R parameters in the model and let Y denote the observed data. The posterior distribution of interest, from which we would like to sample, is then $P(\theta \mid Y)$. Let $P(\theta_r \mid Y, \theta_{-r})$ denote the full conditional distribution of the r^{th} model parameter, the conditional distribution of the parameter given the data (Y) and all other model parameters (θ_{-r}). It can be shown that a joint distribution may be defined by the complete set of such full conditional distributions. Thus in the Bayesian case, the joint posterior distribution of model

parameters may be defined as the complete set of full conditional posterior distributions. That is, the joint posterior $P(\theta \mid Y)$ may be defined by,

$P(\theta_1 \mid Y, \theta_{-1}), P(\theta_2 \mid Y, \theta_{-2}), ..., P(\theta_R \mid Y, \theta_{-R})$. Sampling from the joint posterior then reduces to sampling from the full conditional distributions.

Let θ_r^k denote the value of model parameter r at iteration k. The first iteration of a Gibbs sampler consists of proceeding to the following steps:

1. Initialize the parameters by assigning values for $\theta_1^0, \theta_2^0, ..., \theta_R^0$;
2. Draw values for the first model parameter conditional on current values of all other model parameters and the observed data. That is, draw θ_1^1 from

   $P(\theta_1 \mid Y, \theta_2^0, ..., \theta_R^0)$;

3. For r=2 ,…, R, draw values of parameters for r conditional on current values of all other model parameters and the observed data. Draw a value from each of the following

   $P(\theta_2 \mid Y, \theta_1^1, \theta_3^0, \theta_4^0, ..., \theta_R^0)$

   $P(\theta_3 \mid Y, \theta_1^1, \theta_2^0, \theta_4^0, ..., \theta_R^0)$

   ……

   $P(\theta_r \mid Y, \theta_1^1, ..., \theta_{r-1}^0, \theta_{r+1}^0, \theta_R^0)$

   ……

   $P(\theta_R \mid Y, \theta_1^1, ..., \theta_{R-1}^1)$

Note that each draw is subsequently used in the full conditionals for the remaining

parameters. For example, the value θ_1 drawn (in step 2) is used in the full conditionals for r=2… R (in step 3). Similarly the value of θ_2 drawn (in step 3) is used in the full conditionals for r=3,…,R (in step 3), and so on.

To subsequently iterate the Gibbs sampler, the same general principle is followed: values for a parameter are obtained by sampling from its full conditional distribution given the data and most recent values of the other parameters. More formally, for iteration k in the chain, the full condition for parameter r from which to sample is $P(\theta_r \mid Y, \theta_{<r}^{k}, \theta_{>r}^{k-1})$. In cases where one can be constructed, WinBUGS1.4 (Spiegelhalter, et al., 2000) employs the Gibbs sampler to obtain a solution.

The set of the posterior conditional densities which are required to implement the Gibbs sampler for the 2-PL testlet model in (2.9) can be derived from the joint posterior density in (3.11).

The posterior conditional distributions for the unknown parameters $\Delta_1 = ((a_{1g},\ldots,a_{ig}),(b_{ig},\ldots,b_{jg}),(\gamma_{jd(i)g},\ldots,\gamma_{jd(i)g}))$ are as follows:

$$[a_{ig} \mid Y, \Delta_{-a_{ig}}] \propto \{\prod_{j=1}^{J_g}[\frac{e^{a_{ig}(\theta_{jg}+\gamma_{id(i)g}-b_{ig})}}{1+e^{a_{ig}(\theta_{jg}+\gamma_{jd(i)g}-b_{ig})}}]^{y_{ijg}}[\frac{1}{1+e^{a_{ig}(\theta_{jg}+\gamma_{jd(i)k}-b_{ig})}}]^{1-y_{ijg}}\} * e^{-\frac{1}{2\sigma_a^2}(a_{ig}-\mu_{ag})},$$

$$\text{j=1, …, } J_g \qquad (3.12)$$

$$[b_{ig} \mid Y, \Delta_{-b_{ig}}] \propto \{\prod_{j=1}^{J}[\frac{e^{a_{ig}(\theta_{jg}+\gamma_{id(i)g}-b_{ig})}}{1+e^{a_{ig}(\theta_{jg}+\gamma_{jd(i)g}-b_{ig})}}]^{y_{ijg}}[\frac{1}{1+e^{a_{ig}(\theta_{jg}+\gamma_{jd(i)g}-b_{ig})}}]^{1-y_{ijg}}\} * e^{-\frac{1}{2\sigma_b^2}(b_{ig}-\mu_{bg})},$$

$$\text{j=1, … , } J_g \qquad (3.13)$$

$$[\theta_j \mid Y, \Delta_{-\theta_j}] \propto \{\prod_{i=1}^{I}[\frac{e^{a_{ig}(\theta_{jg}+\gamma_{id(i)g}-b_{ig})}}{1+e^{a_{ig}(\theta_{jg}+\gamma_{jd(i)g}-b_{ig})}}]^{y_{ijg}}[\frac{1}{1+e^{a_{ig}(\theta_{jg}+\gamma_{jd(i)g}-b_{ig})}}]^{1-y_{ijg}}\} * e^{-\frac{1}{2}\theta_j^2},$$

$$\text{i=1, …, I} \qquad (3.14)$$

$$[\gamma_{jdg} \mid Y, \Delta_{-\gamma_{jdg}}] \propto \{\prod_{i=d(1)}^{d(k)} [\frac{e^{a_{ig}(\theta_{jg}+\gamma_{jd(i)g}-b_{ig})}}{1+e^{a_{ig}(\theta_{jg}+\gamma_{jd(i)g}-b_{ig})}}]^{y_{ijg}} [\frac{1}{1+e^{a_{ig}(\theta_{jg}+\gamma_{jd(i)g}-b_{ig})}}]^{1-y_{ijg}}\} * e^{-\frac{1}{2\sigma^2_{\gamma(d)g}}(\gamma_{jd(i)g}-\mu_{\gamma(d)g})^2},$$

$$\text{id} = 11,\ldots,\text{IK}. \tag{3.15}$$

In this research, the WinBUGS software was used in the analysis of the model using MCMC. When using WinBUGS, the number of iterations needed to reach convergence, known as burn-in, has to be determined at first. Iterations from the MCMC estimation procedure prior to convergence may lead to false or imprecise information for inferences. In WinBUGS, several graphical methods are utilized to assess the convergence of the MCMC algorithms. Once convergence is reached, the history plots (trace plots) shows random sampling within the same part of the same space for all chains, Brooks-Gelman-Rubin (BGR) plots show the convergence of both the pooled and within interval widths to stability, and the plot of the auto-correlation function shows the autocorrelation has decreased to zero. After the convergence is achieved, there are a number of guidelines for discovering how much iteration are "enough" to best represent the posterior. One method of assessing whether enough iteration has been completed is to examine the density plots. Once the posterior distribution is sampled fully one would expect to see smooth curves. A rule of thumb often applied in practice is that enough iterations are run so that the error due to the nature of MCMC being an empirical approximation to the posterior is less than 5% of the estimated posterior standard deviation (Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D., 2003). Note that this is easily monitored in WinBUGS1.4.

Analysis Design

In order to investigate the phenomena of DIF amplification and cancellation, at first different simulation conditions were studied, using a data generation computer program in SAS 8.2 (SAS Institute, 2002) for this purpose and then operational data from the ACT (American College Testing) reading test made up of testlets was analyzed.

I. Simulation study

To measure the DIF effect on the testlet parameter and item characteristic parameters and to mimic the real world DIF situations, two simulation designs were performer: (1) Even split DIF with 50/50 manifest split and (2) Uneven split DIF with 80/20 manifest split. Three models were studied: (1) 2-parameter logistic model assuming local independence; (2) 2-parameter logistic testlet model assuming local dependence and testlet effect function homogeneous across groups; (3) 2-parameter logistic testlet DIF model assuming local dependence and testlet effect function heterogeneous across groups. For each model in the first example, a total of seven favorability conditions were defined by counting combinations of difference in preference of focal group and reference group on item difficulty parameter and item discrimination parameters. For each model in the second example, there were a total of five combinations of conditions. A detailed description of these conditions was presented later. For each condition, a common test structure was set by fixing the test composed of first 10 independent dichotomous items and latter two testlets composed of 10 dichotomous items, among which 10 individual items and the first testlet were set as DIF free. This test design was utilized to mimic many current operational tests in which independent binary multiple-

choice items were followed by passage/portfolio testlets. Ten replications were conducted under each situation. The θ distribution for a total of 5000 simulees was assumed to be distributed as $Normal(0,1)$, same for focal and reference subgroups, consistent with the assumption of multidimensional DIF.

A factor that was varied across the simulation study was the testlet variance σ_γ^2. The variances of testlet effects, indicating the degree of within-testlet dependence, were chosen to be 0.2 and 1.5. It was important to note that the larger σ_γ^2, the greater the proportion of total variance in test scores that was attributable to the given testlet and it was interesting to investigate its impact on DIF.

To make the simulated test data as similar as possible to real-world applications, the parameters of the 10 independent dichotomous items were adopted from Lord's (1968) study of the Verbal Scholastic Aptitude Test. These parameters are listed in Table 1. A two-parameter logistic (2PL) model was used to generate items responses for the dichotomous items.

TABLE 1:
Item Parameters for Items 1-10 of Simulation Study

Dichotomous Item Parameters		
Item Number	a_i	b_i
1	1.1	-1.0
2	1.0	-0.9
3	1.3	0.1
4	0.7	1.1
5	1.4	0.4
6	1.2	0.1
7	1.4	0.7
8	0.9	0.6
9	0.8	0.8
10	0.6	1.0

To be consistent with the studies of testlet models by Wainer, the population distributions for the parameters of the two testlets used to generate the data were those corresponding to previous analyses of the Scholastic Aptitude Test (SAT). Specifically, we set $a_i \sim \log Normal(\mu_a, 0.1581^2)$, $b_i \sim Normal(\mu_b, 0.3162^2)$. In order to focus on the difference of means and simplify the simulation conditions of items embedded in the testlet of interest, the variances of the a-parameter and b-parameter were fixed to be very small. To measure the DIF in the second testlet, the difference on the item discrimination parameter and item difficulty parameter were set to be 0.3 and 0.5, respectively. (See Table 3 for descriptions of conditions used in the simulation study.) The 0.3 and 0.5 differences on the discrimination parameter and difficulty parameter were selected to coincide with other DIF studies. For example, Kim and Cohen (1992) used a difference of 0.16 or 0.32; Wang and Wilson (2005) chose a difference of 0.4 or 0.6 to represent a moderate DIF effect. The random effects γ_d for the focal group were generated from $Normal(-1.2247, \sigma_{\gamma f}^2)$, and the $\gamma_{\gamma r}$ for the reference group were generated from

$Normal(0, \sigma_{\gamma r}^2)$. We assume that the mean difference between the focal group and reference group is -1.2247, one standard deviation of the large testlet effect, which is the same across the 10 items embedded in the testlet and that the reference group or focal group was favored because of the content of the passage.

DTF analysis of testlet DIF amplification and cancellation were examined at both the testlet level and the individual item level and our analysis were focused on the second testlet. Descriptions of the simulation design were listed in Table 2 and Table 3, where Table 2 described the models used and Table 3 described all of the conditions. For the first design, there were three models: Model 1: presume that there was no testlet effect and items were local independent with each other; Model 2: presume that the testlet parameter apply constantly to all items in the testlet, that was the γ_d was DIF free and 2-PL testlet model was used. There were two situations related to this model, small variance of testlet effect (I) and large variance of testlet effect (II). Model 3: presume that the testlet parameter applies unequally to all items in the testlet, that was γ_d was sampled from $N(\mu_{\gamma r}, \sigma_{\gamma r}^2)$ and $N(\mu_{\gamma f}, \sigma_{\gamma f}^2)$ for the reference group and focal group, respectively, and that was multiple-group 2-PL testlet model was used. For the first design, the testlet parameter was distributed as $\gamma_{df} \sim N(-1.2247, 0.4472^2)$ and $\gamma_{dr} \sim N(0, 0.4472^2)$, where the mean difference between the two groups was -1.2247 and the variances of two groups were equal to indicate small effect. For each of these models, seven variations on the item characteristics with 1 DIF free situation as a baseline condition and 3 combinations of differences of favorability on item difficulty parameter (conditions of uniform DIF) and 3 combinations of differences of favorability on item discrimination parameter (conditions of non-uniform DIF) were specified: at first for the baseline condition (A), the

distributions of item difficulty parameters and item discrimination parameters were equal for all of the 10 items within the testlet; Second, for the Uniform DIF, (B1) the mean difference between the difficulty parameters was 0.5 and the reference group was favored for the first 5 items within the testlet and the focal group was favored for the rest of 5 items within testlet; (B2) the mean difference between the item difficulty parameters was 0.5 and the reference group was favored for all of the 10 items within the testlet; (B3) the mean difference between the item difficulty parameters was 0.5 and the focal group was favored for all of the 10 items within testlet; Third, as to the non-uniform DIF/crossing DIF, the three conditions were: (C1) the mean difference between the discrimination parameters was 0.3 and the reference group was favored for the first 5 items within testlet and the focal group was favored for the rest 5 items within testlet; (C2) the mean difference between the discrimination parameters was 0.3 and the reference group was favored for all of the 10 items within testlet. (C3) The mean difference between the discrimination parameters was 0.3 and the focal group was favored for all of the 10 items within testlet. By consideration of the various conditions of testlet distribution, there were 28 (4*7) data sets generated. Next, for the second design, three models were also applied. Taking testlet means and variances into consideration, as well as, keeping five of the total seven combinations as above, there were a total of 25 (5*5) simulated data sets. These three variations of testlet distributions were: (1) the mean of testlet distribution of reference group was 0 and the mean of testlet distribution of focal group was -1.2247 indicating the reference group was favored; the variances of testlet distributions of both of the reference group and focal group were 0.2, indicating small dependence; (2) the means of the testlet distributions of both groups were 0; the variance of testlet distribution

of reference group was 0.2 and the variance of testlet distribution of focal group was 1.5 indicating there were larger variability on the testlet dimension among focal group; (3) the mean of testlet distribution of reference group was 0 and the mean of testlet distribution of focal group was -1.2247 indicating the reference group was favored; the variance of testlet distribution of reference group was 0.2 and the variance of testlet distribution of focal group was 1.5 indicating there were larger dependence among focal group. The detailed descriptions of each situation by taking the combinations of models in Table 2 and conditions in the Table 3 were provided in the Appendix G.

Specifically, the second model was taken to study the DIF amplification and cancellation phenomenon attributed to the testlet parameter and item characteristic parameters simultaneously. DTF analysis of testlet DIF amplification and cancellation were based on both the item level and testlet level. On the one hand, at the individual item level, when both of the testlet parameter and item difficulty parameters favored the reference group, which means that the reference group has higher ability on the secondary dimension on average and also the item seems easier for the reference group, it was the test for simultaneous effects of amplification at the item level. However, when the testlet parameter and item difficulty parameter favored different groups, it was the test for simultaneous effect of cancellation at the item level; on the other hand, at the testlet level, summing up small or significant amount of item DIF favoring the same group for all items embedded in the testlet would reflect amplification; However, summing up small or significant amounts of item DIF favoring different groups for different items embedded in the testlet would reflect cancellation. The phenomena could be assessed by comparing the results from the second or third model with those from the first model.

TABLE 2:
The description of model for Differential Testlet Functioning Analysis in Simulation Design

Model	Equation	Testlet distribution for Design I and Design II
Model 1: (M1) Assuming local Independence; 2-Parameter Logistic Model	$P(Y_{ijg}=1\mid\Omega_{ijg})=\frac{\exp(a_{ig}(\theta_j-b_{ig}))}{1+\exp(a_{ig}(\theta_j-b_{ig}))}$	
Model 2: (M2) Assuming local dependence and testlet effect function homogeneously between groups; Testlet model with no DIF on testlet effect	$P(Y_{ijg}=1\mid\Omega_{ijg})=\frac{\exp(a_{ig}(\theta_j-b_{ig}+\gamma_{jd(i)}))}{1+\exp(a_{ig}(\theta_j-b_{ig}-\gamma_{jd(i)}))}$	(Design 1: 50/50 split) Where: I. $\gamma_{jd(i)f}=\gamma_{jd(i)r}\sim N(0,0.4472^2)$; II. $\gamma_{jd(i)f}=\gamma_{jd(i)r}\sim N(0,1.2247^2)$; (Design 2: 80/20 split) Where: I. $\gamma_{jd(i)f}=\gamma_{jd(i)r}\sim N(0,0.4472^2)$;
Model 3: (M3) Assuming local dependence and testlet effect function heterogeneously between groups; Testlet model with DIF on testlet effect	$P(Y_{ijg}=1\mid\Omega_{ijg})=\frac{\exp(a_{ig}(\theta_j-b_{ig}+\gamma_{jd(i)g}))}{1+\exp(a_{ig}(\theta_j-b_{ig}-\gamma_{jd(i)g}))}$	(Design 1: 50/50 split) Where: I. $\gamma_{jd(i)f}\sim N(-1.2247,0.7071^2)$ & $\gamma_{jd(i)r}\sim N(0,0.7071^2)$ (Design 2: 80/20 split) Where: I. $\gamma_{jd(i)f}\sim N(-1.2247,0.4472^2)$ & $\gamma_{jd(i)r}\sim N(0,0.4472^2)$; II. $\gamma_{jd(i)f}\sim N(0,1.2247^2)$ & $\gamma_{jd(i)r}\sim N(0,0.4472^2)$; III. $\gamma_{jd(i)f}\sim N(-1.2247,1.2247^2)$ & $\gamma_{jd(i)r}\sim N(0,0.4472^2)$

TABLE 3:
The description of condition for Differential Testlet Functioning Analysis in Simulation Design

Conditions		
A.Baseline Condition	For i=1,…,10, $b_{if} = b_{ir} \sim N(0,0.3162^2)$ & $a_{if} = a_{ir} \sim \log\ normal\ \ (0,0.1581^{\ 2})$	For design I and II
B1. Uniform DIF	For i=1,…,5, $b_{if} \sim N(0,0.3162^2) \& b_{ir} \sim N(-0.5,0.3162^2)$; $a_{if} = a_{ir} \sim \log\ normal\ \ (0,0.1581^{\ 2})$; For i=6,…,10, $b_{if} \sim N(-0.5,0.3162^2) \& b_{ir} \sim N(0,0.3162^2)$; $a_{if} = a_{ir} \sim \log\ normal\ \ (0,0.1581^{\ 2})$;	For design I and II
B2. Uniform DIF	For i=1,…,10, $a_{if} = a_{ir} \sim \log\ normal\ \ (0,0.1581^{\ 2})$ $b_{if} \sim N(0,0.3162^2) \& b_{ir} \sim N(-0.5,0.3162^2)$;	For design I and II
B3. Uniform DIF	For i=1,…,10, $a_{if} = a_{ir} \sim \log\ normal\ \ (0,0.1581^{\ 2})$ $b_{if} \sim N(-0.5,0.3162^2) \& b_{ir} \sim N(0,0.3162^2)$;	For design I only
C1. Non-uniform DIF	For i=1,…,5, $a_{if} \sim \log normal(0,0.1581^2)$ $a_{ir} \sim \log normal(-0.3,0.1581^2)$; $b_{if} = b_{ir} \sim N(0,0.3162^2)$; For i=6,…,10, $a_{if} \sim \log normal(-0.3,0.1581^2)$ $a_{ir} \sim \log normal(0,0.1581^2)$; $b_{if} = b_{ir} \sim N(0,0.3162^2)$	For design I and II
C2. Non-uniform DIF	For i=,…,10, $a_{if} \sim \log normal(0,0.1581^2)$ $a_{ir} \sim \log normal(-0.3,0.1581^2)$ $b_{if} = b_{ir} \sim N(0,0.3162^2)$;	For design I and II
C3. Non-uniform DIF	For i=,…,10, $a_{if} \sim \log normal(-0.3,0.1581^2)$ $a_{ir} \sim \log normal(0,0.1581^2)$ $b_{if} = b_{ir} \sim N(0,0.3162^2)$;	For design I only

In order to measure the magnitude of the DIF effects across all of 53 (4*7+5*5) simulation conditions, item characteristic curves and testlet characteristic curves provided visual descriptions of DIF and the signed or unsigned indices served as the basis for a statistical characterization of DIF. Moreover, the logistic regression procedure gives information on both the significance and direction of the DIF parameter. All of the three DIF detecting procedures were programmed in MATLAB 7.2 (The Mathworks Inc., 2004).

II. Real Data Analysis

The computer estimation programs WinBUGS1.4 (Spiegelhalter, et al., 2000) and MATLAB7.2 2 (The Mathworks Inc., 2004) were used in analyzing one set of real data. The data set was obtained from released form of the American College Testing (ACT) in Reading (1995). This test was chosen for analysis due to its structure and content of testlets. The Reading section of ACT was composed of 40 test items nested within 4 testlets. The Reading Test consisted of four passages: Prose Fiction, Social Science, Humanities, and Natural Science. All four passages were given equal weight in scoring. There were a total of 3078 females and 2875 males for the analysis of gender DIF and a total of 1271 minority students and 3171 Caucasian students for the analysis of ethnic DIF. A cross-validation procedure was used by randomly partitioning the data into two subsets, one with 1528 females/ 1432 males and the other one with 1550 females/ 1443 males, and two samples each with 652 minority / 1550 Caucasians, such that the analysis was initially performed on the first subset, while the other subset was retained for subsequent use in confirming and validating the initial analysis.

Previous to the study of the DIF amplification and cancellation, an important first step was to assess the model fit. The deviance information criterion (DIC) was used as a model selection index. First, the 2-PL testlet and 2-PL models were fit to the data to investigate whether there was local dependence due to testlet by comparing the model fit. Next, the multiple-group 2-PLM testlet model or multiple- group 2-PLM model were fit to the data to detect whether there was DIF at the whole test level. Finally, a detailed examination of DIF at the item level and testlet level was constructed using ICC/TCC, Signed/Unsigned Area Indices, and a Logistic Regression procedure. The cross-validation sample was then used for confirmatory analysis of those three steps.

Spiegelhalter et al. (1998) described the Deviance Information Criterion (DIC) as a model selection index under Bayesian theoretical framework for arbitrarily complex models. The DIC was defined as

$$\text{DIC} = mean[-2\log L(\omega^t \mid y)] - \{mean[-2\log L(\omega^t \mid y) - 2\log L(\omega^* \mid y)]\}, \qquad (3.14)$$

where the first term was called deviance indicating the MCMC average of the log-likelihoods calculated at the end of an iteration of the Gibbs sampler. The log-likelihood in the second term was calculated using the posterior means of parameters ω. The second term was the penalty function for increasing model complexity.

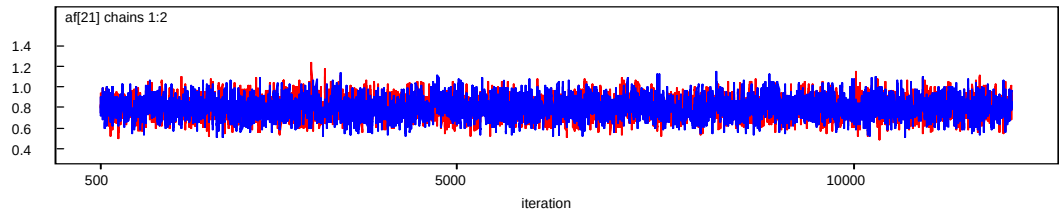
Chapter 4 Results and Discussion

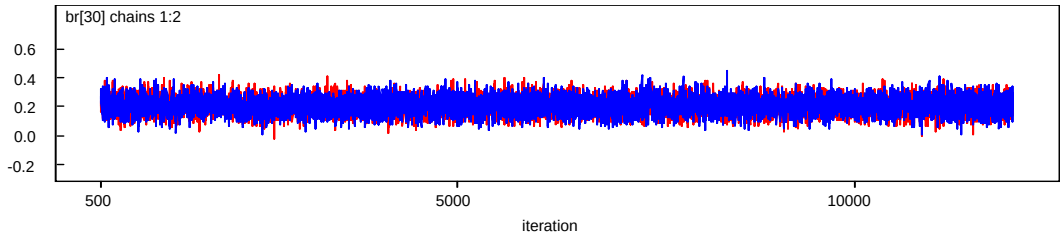
Results of Simulation Study

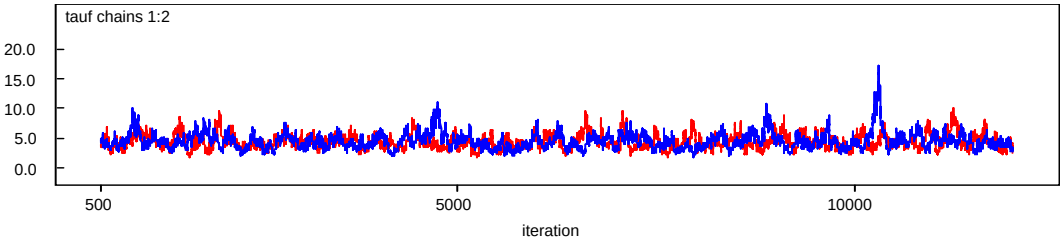
I. Model Convergence and Parameter Recovery

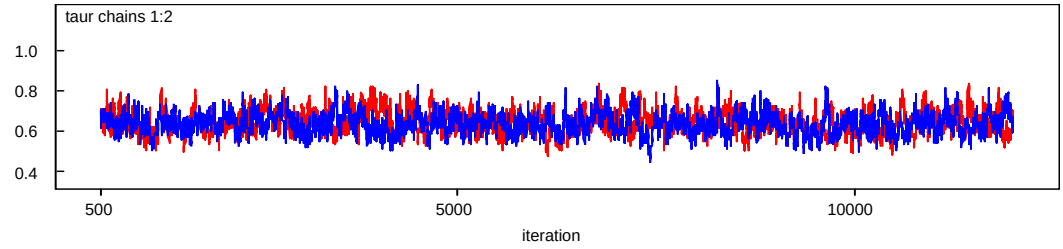
Prior distributions for Bayesian estimation of the model parameters were assumed to be normally distributed with mean of zero and a variance of one for the latent trait θ In order to deal with the model indeterminacy issue, the testlet parameters were assumed to be normally distributed with fixed means in the simulation design (e.g., -1.2247 or 0, according to different simulation conditions) and a common precision (1/variance) for all of the situations with a hyperprior of a Gamma distribution, $Gamma(1,1)$. This was a weakly informative prior distribution; the posterior distribution of such parameters was driven by data as well as proper priors which are required in the WinBUGS1.4 computations. This avoided technical problems arising from improper posterior distributions. In each situation that was studied, two chains were run for 4,000 iterations with a burn-in of 1,500 iterations. Although all of the parameters including the hyperprior parameter, tau, converge after 1,500 iterations, more iterations were necessary to achieve stable posterior values. Figure[1] 1 showes representative history plots, BGR diagnostic plots, and graphs of autocorrelations for several representative parameters.

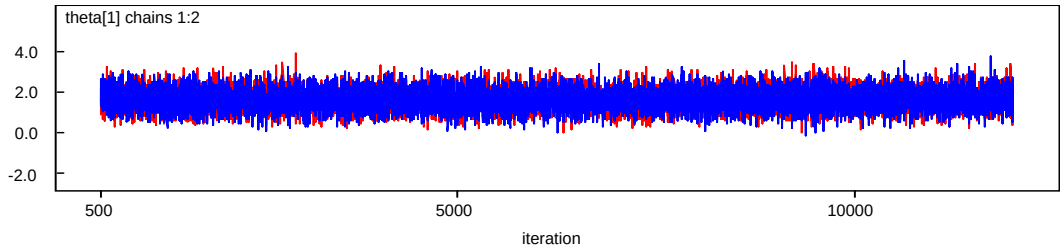
[1] The blue and red color of the graphs represent for the two chains.

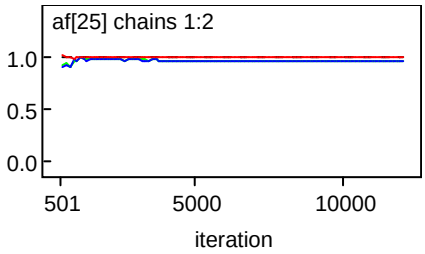

af[21] chains 1:2
1.4
1.2
1.0
0.8
0.6
0.4
500
5000
10000
iteration

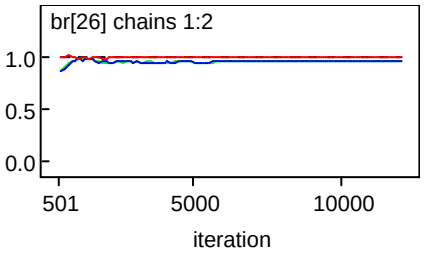

br[30] chains 1:2
0.6
0.4
0.2
0.0
-0.2
500
5000
10000
iteration

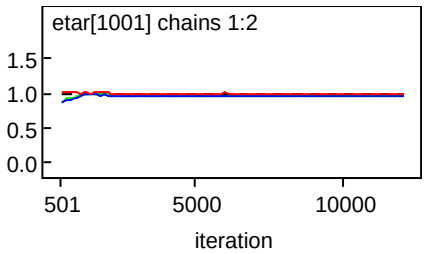

tauf chains 1:2
20.0
15.0
10.0
5.0
0.0
500
5000
10000
iteration

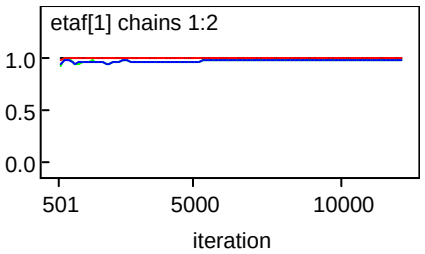

taur chains 1:2
1.0
0.8
0.6
0.4
500
5000
10000
iteration

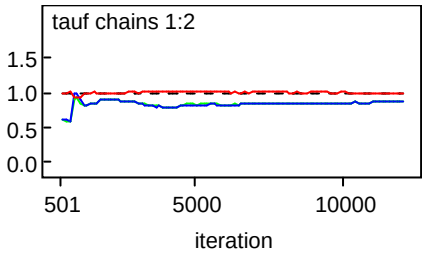

theta[1] chains 1:2
4.0
2.0
0.0
-2.0
500
5000
10000
iteration

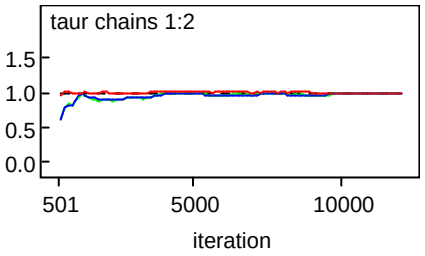

af[25] chains 1:2
1.0
0.5
0.0
501
5000
10000
iteration

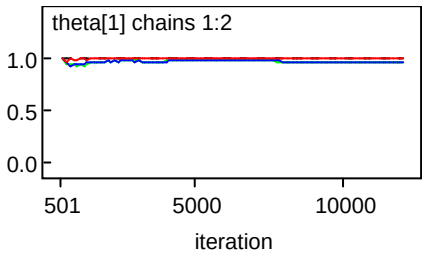

br[26] chains 1:2
1.0
0.5
0.0
501
5000
10000
iteration


etar[1001] chains 1:2
1.5
1.0
0.5
0.0
501
5000
10000
iteration


etaf[1] chains 1:2
1.0
0.5
0.0
501
5000
10000
iteration


tauf chains 1:2
1.5
1.0
0.5
0.0
501
5000
10000
iteration


taur chains 1:2
1.5
1.0
0.5
0.0
501
5000
10000
iteration


theta[1] chains 1:2
1.0
0.5
0.0
501
5000
10000
iteration

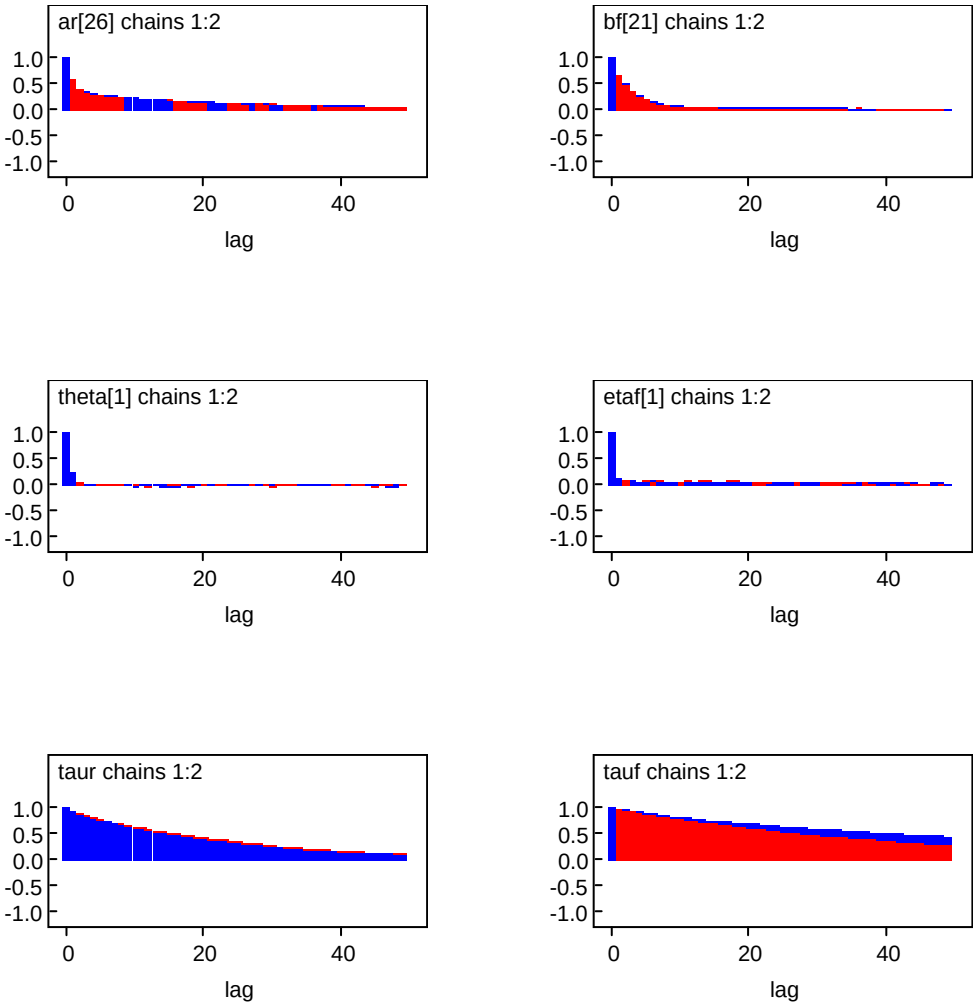


FIGURE 1: Gibbs sampling history plots, BGR diagnostic plots, autocorrelation plots of representative a-parameter, b-parameter, theta-parameter, testlet-parameter (etaf and etar) and testlet precision parameter (tauf and taur)

Once the estimations were done, the results of the WinBUGS runs illustrated the simulation conditions under which those models could recover the parameters used to generate the data, given the model that generated the data matched the model used. For one representative example, the correlations of true and estimated a-parameter, b-parameter, theta-parameter and testlet-parameter are listed in Table 4.

TABLE 4:
Correlations of True and Estimated Item and Person Parameters Obtained by MCMC Estimation

Examples	Parameter Estimates							
	a		b		θ		γ	
	Focal Group	Reference Group	Focal Group	Reference Group	Focal Group	Reference Group	Focal Group	Reference Group
Gender Example	0.992 (Sig:0.00) (N:2500)	0.981 (Sig:0.00) (N:2500)	0.924 (Sig:0.00) (N:2500)	0.990 (Sig:0.00) (N:2500)	0.891 (Sig:0.00) (N:5000)		0.546 (Sig:0.00) (N:2500)	0.614 (Sig:0.00) (N:2500)
Ethnic Example	0.983 (Sig:0.00) (N:1000)	0.988 (Sig:0.00) (N:4000)	0.914 (Sig:0.00) (N:1000)	0.994 (Sig:0.00) (N:4000)	0.882 (Sig:0.00) (N:5000)		0.397 (Sig:0.00) (N:1000)	0.774 (Sig:0.00) (N:4000)

According to the information given in the above table, the correlations of true and estimated item and person parameters were around 0.8~0.9, which were higher than those of item-person interaction parameter γ. Therefore, the parameter recovery of testlet parameter γ seemed to be not as good as those of a-parameter, b-parameter and θ-parameter, which might be due to the facts that every examinee provided information to estimate a relatively small number of item parameter a and b, ever item provided information to estimate ability parameters θ, but each testlet provided relatively little information to estimate its person-testlet interaction parameter γ since items nested within testlet had the same testlet parameter.

II. Logistic regression modeling results and signed_area/unsigned_area indices of simulation study

For each situation in the simulation design, Table 5 and Table 8 show the R^2 based effect size results of the DIF analysis of one (or two) of the representative item(s) of the total of ten of interest for the even split design and uneven split design based on the five-step logistic regression modeling process. The values in bolded face were significant R^2 based effect size based on the Jodoin and Gierl (2000) classification criterion. The column with the title "$R_2^2 - R_1^2$" displays an increased portion of R^2 after adding the testlet variable conditioning on the main latent trait θ; the column with the title "$R_3^2 - R_2^2$" indicated the increase of R^2 after entering the dummy group variable into the two dimensional (θ and γ) regression model; the column with the title "$R_4^2 - R_3^2$" displays the increase portion of R^2 after adding the interaction of the group variable and the main latent trait θ; the column with the title "$R_5^2 - R_3^2$" indicats an increase portion of R^2 after adding the interaction of the group variable and the testlet variable. Since the same item discrimination parameters[2] had been defined in the current testlet model to capture the interaction of item with the main dimension θ and testlet dimension γ, the results of $R_4^2 - R_3^2$ and $R_5^2 - R_3^2$ were supposed to be similar. The results of regression coefficients of logistic regression model for one (or two) of the representative items were listed in Table 6 for the even split design and Table 9 for the uneven split design. The columns with $\tau_0 \sim \tau_4$ represent the regression coefficients for constant variable, θ-parameter, γ-

[2] See Li, Bolt & Fu, "A Comparison of Alternative Models for Testlet ", *Applied Psychological Measurement*, Vol. 30 No.1, p.3-21 for testlet models with different item discrimination parameters for θ and γ.

parameter, group variable, interaction term of θ and group variable, interaction term of γ and group variable respectively. The values significantly different from zero were bolded, with the "+"sign of the values indicating reference group was favored and the "-" sign indicating the focal group was favored. The results of signed-area/unsigned-area indices are listed in Table 7 and Table 10 for these two designs. The item characteristic curves (ICC) of representative item(s) and testlet characteristic curves of each situation in simulation design are attached in Appendix B and C.

From the R^2 based effect size results in Table 5 and Table 8, we could conclude that, in general, a moderate amount of uniform DIF was detected in situations B1, B2 and B3 of each model for the first design and situations B1 and B2 of each model for the second design, and a moderate amount of nonuniform DIF was detected in situation C1, C2 and C3 of each model for first design and situations C1 and C2 of each model for second design. The relatively smaller effect size for the second case might be due to the unequal sample sizes of the two subgroups. Additionally, the columns of $R_2^2 - R_1^2$ in Table 5 and Table 8 indicate the contributions of the testlet variable. The values of model 3 were, in general, larger than those of model 2 due to the distributions of testlet parameters of two subgroups. Specifically comparing results of condition II with condition I of model 2 for the first design in Table 5, the larger values of $R_2^2 - R_1^2$ indicated the larger contribution because of the larger variance of testlet distribution. Whereas the even larger values in model 3 suggested the different means of the distributions of the testlet parameter between the two subgroups. As to the three conditions of model 3 for the second design, the relatively smaller values of $R_2^2 - R_1^2$ of condition II suggested the different variances of testlet parameter between two subgroups

could not be easily detected comparing to the mean difference. However, the relatively larger values of $R_2^2 - R_1^2$ of condition II of model 3 with reference group having small testlet variance and focal group having large testlet variance, than those of condition I of model 2 with both of the two subgroups having small testlet variances, indicated the difference of variance of testlet distributions between two subgroups. The influence of the different variances of testlet distributions was also reflected in the results of signed-area/unsigned-area indices of situation C1 and C2 of model 1, model 2 (both of groups had small testlet effect) and condition II of model 3 (focal group had large testlet effect and reference group had small testlet effect) (bolded in green) of Table 10. Regarding model 1 as baseline condition, those values of situation C1 and C2, especially the signed-area indices, of condition II of model 3 were much smaller than those of model 2, and again they were all smaller than those of model 1, which might indicate that the larger variance of testlet distribution of focal group paid off its unfavorability of small sample size, and DIF at the item level was cancelled out more completely than those of Model 2 where both subgroups had the same small testlet distributions.

Still in Table 5 and Table 8, we also found especially exaggerated large/small $R_2^2 - R_1^2$ values and exaggerated small $R_3^2 - R_2^2$ values (boded in green) for situation B1, B2 and B3 of model 3 in the even split design and situation B1 and B2 of condition C1 and C3 of model 3 in the uneven split design. The reason might be that DIF amplification and cancellation happened at the item level with interactive effects of item difficulty parameter and testlet parameter and testlet parameter entered into the regression model before item difficulty parameter.

Consistent with the simulation design, the favorability of one of the two subgroups was indicated by the signs of regression coefficients in Table 6 and Table 9. For situation B1, B2 and B3 that were associated with different distributions of the item difficulty parameters in the even split example, the positive regression coefficients suggested half of the 10 items under situation B1 and all of the items under situation B2 favored the reference group, and the negative regression coefficients reflected the rest half of the 10 the items under situation B1 and all of the items under situation B3 favored the focal group. Similar results had been received for the situation B1 and B2 for the uneven split example. For situation C1, C2 and C3 that were associated with different distributions of item discrimination parameters for the even split design, the positive regression coefficients suggested that half of the 10 items under situation C1 and all of the items under situation C2 had smaller item discrimination parameters for the reference group than for the focal group; on the contrary, the negative regression coefficients suggested the rest half of items under condition C1 and all of the items under condition C3 had larger item discrimination parameters for the reference group than that for the focal group. Similar results had been found for situation C1 and C2 of the uneven split example.

The most important results of DIF amplification and cancellation had been categorized into the following seven points and reflected in results of signed-area /unsigned-area indices of exemplified item(s) and testlets in Table 7 and Table 10. These seven categories were:

1. No DIF amplification and cancellation at both item and testlet levels;
2. No DIF amplification and DIF cancellation at the item level but DIF

amplification at the testlet level;

3. No DIF amplification and DIF cancellation at the item level but DIF cancellation at the testlet level;
4. DIF amplification at the item level and DIF amplification at the testlet level;
5. DIF amplification at the item level and DIF cancellation at the testlet level;
6. DIF cancellation at the item level and DIF amplification at the testlet level;
7. DIF cancellation at the item level and DIF cancellation at the testlet level.

All of the conditions under model 1 and situation A of model 2 in Table 7 and Table 10 served as baseline conditions with no testlet effect for model 1 and constant testlet effect for the two subgroups for situation A. Therefore, there was basically no DIF amplification and cancellation although there was an ignorable amount of DIF cumulated at the testlet level for situation A of model 2 due to the existence of a testlet effect.

The phenomena of the second point was reflected on situation B2 and B3 of model 2 of even split example and situation B2 of model 2 and situation B2 of condition II of model 3 of uneven split example, where there was uniform DIF due to different item difficulty parameters of two subgroups but no difference on the means of testlet parameters. For example, for situation B3 of condition I of model 2 in Table 7 where there was small testlet effect and it functioned homogenously between the reference group and focal group and item difficulty parameter favored focal group, the signed-area and unsigned-area index of one of the items embedded in the testlet were -0.4082 and 0.1977 and the two indices for the whole testlet were -4.0816 and 1.9307. Whereas the phenomena of the third point was reflected on situation B1 of model 2 of even split example and situation B1 of model 2 and condition II of model 3 of uneven split

example, where there were uniform DIF due to different item difficulty parameters with half of the items within the testlet favoring the reference group and half of them on the contrary, and still no difference on the means of the testlet parameters. Therefore, there were no DIF amplification and cancellation at the individual item level but DIF cancellation happened at the testlet level. For example, as to the situation B1 of condition II of model 3 in Table 9, we could see that the signed-area and unsigned-area of one of the first five items with items difficulty parameters favoring reference group were 0.4044 and 0.1947 and those for one of the latter five items with item difficulty parameters favoring focal group are -0.4188 and 0.2022, and then those two indices of the whole testlet were -0.0723 and 0.1511.

Regarding to the fourth point of DIF amplification at the item level and testlet level, it was exemplified on situation B2 of model 3 of the first design and situation B2 of condition I and III of model 3 of second design, where reference group was more capable than focal group on the testlet dimension and what was more, items were more easier for reference group than for focal group. For example, as to the situation B2 of model 3 of even split design, the signed-area and unsigned-area indices of one of the representative item were 0.5563 and 0.2635, which were larger than those of baseline situation B2 in model 1: 0.4484 and 0.2131, and they were even larger at the testlet level of 5.5629 and 2.6126. And then, the fifth point of DIF amplification at the item level but cancellation at the testlet level was reflected on situation C1 of condition I of model 3 of even split example and situation C1 of condition I and III of model 3 of uneven split example, where the amplification occurred because the reference group was more capable than the focal group on average on the testlet dimension, and cancellation at the testlet level was

because of the different item discrimination parameters of items within the testlet. Taking situation C1 of condition III of model 3 of the uneven split design as an example, the signed-area and unsigned-area indices of two representative items with item discrimination parameter estimates favoring reference group and focal group respectively, were 0.1451 and 0.1356, and 0.1248 and 0.1474, which were slightly larger than those of baseline condition C1 of model 1 with signed-area and unsigned-area indices of two representative items of 0.0076 and 0.1302 for one item and 0.0850 and 0.1314 for the other. The signed-area and unsigned-area indices for the testlet of situation C1 of condition III of model 3 were 1.3498 and 0.6446, which were close to those of situation A of condition III of model 3. This indicates that DIF at the testlet level was due to the different distributions of testlet parameter and DIF caused by item discrimination parameter had been cancelled out when half of the item discrimination parameters favored the reference group and half of them favored the focal group.

For the sixth category of DIF cancellation at the item level but amplification at the testlet level, there were two reasons for DIF cancellation at the item level: one was because of the different item discrimination parameters of the two subgroups although it still could be detected by the unsigned-area index, and the other one was because the testlet effect and item difficulty parameter function differently between two subgroups. The situation C2 and C3 of model 2 of even split design and situation C2 of condition I of model 2 and condition II of model 3 of the uneven split design gave expression to the first reason. One example of it was the situation C2 of condition II of model 3 of uneven split design, where the signed-area and unsigned-area indices for one of the representative items were -0.0023 and 0.1156, and those for the testlet were -0.0232 and

1.0891. The situation B3 of condition I of model 3 of the first design gave expression to the second reason of this category, where the reference group had higher ability on the testlet dimension but item were more difficult to them than to the focal group. Signed-area and unsigned-area indices were |-0.3016|[3] and 0.1505, which were smaller than those of the baseline situation B3 of model 1 of |-0.4130| and 0.1947, and those for the testlet were much larger: |-3.0160| and 1.4428. One special case was situation B1 of condition I of model 3 of the first design and situation B1 of condition I and III of model 3 of the second design, where half of items within testlet had DIF amplification and half of them have DIF cancellation at the item level because of the interaction of item difficulty parameter and testlet effect but DIF amplification all happened at the testlet level because of the existence of the testlet effect. For example, the signed-area and unsigned-area of two representative items under situation B1 of condition I of model 3 in Table 10 were 0.5414 and 0.2580, -0.2504 and 0.1223 respectively, and those for the testlet were 1.4555 and 0.6901.

Finally, the DIF cancellation at both item and testlet levels was reflected in situation C1 of condition I of model 2 and situation C1 of condition II of model 3 of uneven split design. Taking the former one as an example, the signed-area and unsigned-area indices of two representative items were -0.0213 and 0.1250, and 0.0235 and 0.1474, and those for the testlet were 0.0113 and 0.1219.

[3] Absolute values of the signed-area indices were considered here to compare the magnitude of DIF of different situations.

TABLE 5:
The R^2 based Effect Size of Simulation Study of Even Split Design

		R_1^2 (θ)	R_2^2 (G)	R_3^2 (θ*G)	R_4^2	R_5^2	$R_2^2-R_1^2$ Uniform DIF	$R_3^2-R_2^2$ Non uniform DIF	$R_4^2-R_3^2$	$R_5^2-R_3^2$
Model 1 (no testlet effect)	A	0.9948	0.9993	1.0000			0.0045	7.40E-04		
	B1	0.9410 0.9464	0.9998 0.9998	1.0000 1.0000			**0.0534** **0.0589**	0.0002 0.0002		
	B2	0.9225	0.9992	1.0000			**0.0767**	0.0001		
	B3	0.9310	0.9992	1.0000			**0.0682**	0.0001		
	C1	0.9725 0.9748	0.9752 0.9778	1.0000 1.0000			0.0027 0.0030	**0.0248** **0.0222**		
	C2	0.9700	0.9739	1.0000			0.0039	**0.0261**		
	C3	0.9762	0.9807	1.0000			0.0045	**0.0193**		
		R_1^2 (θ)	R_2^2 (γ)	R_3^2 (G)	R_4^2 (θ*G)	R_5^2 (γ*G)	$R_2^2-R_1^2$ DIF on γ	$R_3^2-R_2^2$ Uniform DIF	$R_4^2-R_3^2$ Non uniform DIF	$R_5^2-R_3^2$ Non uniform DIF
Model 2 Condition I (equal small testlet effect)	A	0.9920	0.9925	0.9973	1.0000	1.0000	0.0005[4]	0.0048	0.0027	0.0027
	B1	0.9296 0.9327	0.9332 0.9300	0.9972 0.9967	1.0000 1.0000	1.0000 1.0000	0.0004 0.0004	**0.0640** **0.0668**	0.0028 0.0033	0.0028 0.0033
	B2	0.9157	0.9163	0.9977	1.0000	1.0000	0.0006	**0.0814**	0.0023	0.0023
	B3	0.9276	0.9280	0.9973	1.0000	1.0000	0.0004	**0.0694**	0.0027	0.0027
	C1	0.9652 0.9641	0.9654 0.9646	0.9691 0.9676	1.0000 1.0000	1.0000 1.0000	0.0001 0.0005	0.0037 0.0040	**0.0309** **0.0324**	**0.0309** **0.0324**
	C2	0.9707	0.9716	0.9763	1.0000	1.0000	0.0008	0.0048	**0.0237**	**0.0237**
	C3	0.9694	0.9706	0.9743	1.0000	1.0000	0.0012	0.0037	**0.0257**	**0.0257**
Model 2 Condition II **(**equal large testlet effect**)**	A	0.9831	0.9942	0.9978	1.0000	1.0000	0.0111	0.0037	0.0022	0.0022
	B1	0.9245 0.9213	0.9352 0.9317	0.9973 0.9984	1.0000 1.0000	1.0000 1.0000	0.0107 0.0104	**0.0621** **0.0668**	0.0027 0.0016	0.0027 0.0016
	B2	0.9173	0.9276	0.9979	1.0000	1.0000	0.0103	**0.0703**	0.0021	0.0021
	B3	0.9087	0.9189	0.9979	1.0000	1.0000	0.0102	**0.0790**	0.0021	0.0022
	C1	0.9578 0.9564	0.9675 0.9664	0.9724 0.9699	0.9997 0.9997	1.0000 1.0000	0.0097 0.0100	0.0049 0.0035	**0.0273** **0.0298**	**0.0276** **0.0301**
	C2	0.9641	0.9748	0.9788	0.9998	1.0000	0.0106	0.0040	**0.0210**	**0.0212**
	C3	0.9577	0.9697	0.9737	0.9997	1.0000	0.0122	0.0000	**0.0284**	**0.0287**
Model 3 Condition I (unequal median testlet effect)	A	0.9842	0.9967	0.9982	1.0000	1.0000	**0.0125**	0.0016	0.0018	0.0018
	B1	0.8895 0.9668	0.9748 0.9795	0.9980 0.9985	1.0000 1.0000	1.0000 1.0000	**0.0853** **0.0127**	**0.0232** **0.0190**	0.0020 0.0015	0.0020 0.0015
	B2	0.8742	0.9724	0.9983	1.0000	1.0000	**0.0982**	**0.0259**	0.0017	0.0017
	B3	0.9519	0.9737	0.9974	1.0000	1.0000	**0.0218**	**0.0237**	0.0026	0.0026
	C1	0.9611 0.9583	0.9700 0.9697	0.9722 0.9708	0.9999 0.9999	1.0000 1.0000	**0.0096** **0.0114**	0.0015 0.0011	**0.0278** **0.0292**	**0.0278** **0.0292**
	C2	0.9635	0.9767	0.9783	1.0000	1.0000	**0.0132**	0.0016	**0.0217**	**0.0217**
	C3	0.9623	0.9742	0.9758	0.9996	1.0000	**0.0119**	0.0016	**0.0241**	**0.0242**

Note: The results listed were averaged over 10 replications and averaged over 10 items for condition A, B2, B3, C2 and C3; the two sets of values for condition B1 and C1 represented for the results averaged over the former and latter five items respectively.

[4] The different depth of colors represent different amount of numbers related to different models. (Same for the results in other tables.)

TABLE 6:
Logistic Regression Coefficients of Simulation Study of Even Split Design

		τ_0	τ_1 (θ)	τ_2 (G)	τ_3 (θ*G)	τ_4	τ_5
Model 1 (no testlet effect)	A	-0.0334	1.0424	0.0271	-0.0349		
	B1	-0.0262 0.4632	1.0427 1.0134	**0.4669** **-0.4247**	-0.0085 -0.0107		
	B2	-0.0333	1.0427	**0.5320**	-0.0250		
	B3	0.4877	1.0460	**-0.4940**	-0.0388		
	C1	-0.0261 -0.0286	1.0491 0.7539	-0.0304 0.0672	**-0.2879** **0.2460**		
	C2	-0.0331	1.0425	0.0230	**-0.2926**		
	C3	-0.0255	0.7597	0.0106	**0.2457**		
		τ_0	τ_1 (θ)	τ_2 (γ)	τ_3 (G)	τ_4 (θ*G)	τ_5 (γ*G)
Model 2 Condition I (equal small testlet effect)	A	0.0104	1.0065	1.0065	0.0240	0.0034	0.0033
	B1	0.0229 0.5042	1.0404 0.9528	1.0404 0.9528	**0.4758** **-0.4558**	-0.0340 0.0657	-0.0340 0.0657
	B2	0.0106	1.0059	1.0059	**0.5280**	0.0091	0.0091
	B3	0.5040	0.9971	0.9971	**-0.4810**	0.0121	0.0121
	C1	0.0233 0.0084	1.0434 0.7010	1.0434 0.7010	-0.0262 0.0409	**-0.3052** **0.3144**	**-0.3052** **0.3144**
	C2	0.0106	1.0064	1.0064	0.0114	**-0.2588**	**-0.2588**
	C3	0.0087	0.7392	0.7392	0.0148	**0.2706**	**0.2706**
Model 2 Condition II **(**equal large testlet effect**)**	A	0.0482	0.9872	0.9872	-0.0102	0.0242	0.0242
	B1	0.0457 0.5365	1.0141 0.9617	1.0141 0.9617	**0.4515** **-0.4688**	-0.0160 0.0632	-0.0160 0.0632
	B2	0.0489	0.9869	0.9869	**0.4843**	0.0361	0.0361
	B3	0.5485	0.9825	0.9824	**-0.5114**	0.0272	0.0272
	C1	0.0451 0.0462	1.0180 0.7068	1.0180 0.7068	-0.0408 0.0211	**-0.2830** **0.3153**	**-0.2830** **0.3153**
	C2	0.0481	0.9874	0.9874	-0.0214	**-0.2417**	**-0.2417**
	C3	0.0378	0.7333	0.7333	-1E-05	**0.2775**	**0.2775**
Model 3 Condition I (unequal median testlet effect)	A	-0.0026	0.9965	0.9965	0.0294	0.0104	0.0104
	B1	0.0118 0.4742	1.0298 0.9609	1.0298 0.9609	**0.4868** **-0.4150**	-0.0229 0.0600	-0.0229 0.0600
	B2	-0.00410	0.9954	0.9954	**0.5385**	0.0226	0.0226
	B3	0.5095	0.9944	0.9944	**-0.4821**	0.0121	0.0121
	C1	0.0102 -0.0083	1.0275 0.7120	1.0275 0.7120	-0.0197 0.0692	**-0.2913** **0.3058**	**-0.2913** **0.3058**
	C2	-0.0020	0.9972	0.9972	0.0180	**-0.2513**	**-0.2513**
	C3	0.0039	0.7383	0.7383	0.0232	**0.2685**	**0.2685**

Note: The results listed were averaged over 10 replications and averaged over 10 items for condition A, B2, B3, C2 and C3; the two sets of values for condition B1 and C1 represented for the results averaged over the former and latter five items respectively.

TABLE 7:
The Signed-area and Unsigned-area Indices of Simulation Study of
Even Split Design

		Signed-Area	Unsigned-Area	Signed-Area For TCC	Unsigned-Area For TCC
Model 1 (no testlet effect)	A	0.0208	0.0452		
	B1	0.3913 -0.3625	0.1851 0.1704		
	B2	0.4484	0.2131		
	B3	-0.4130	0.1947		
	C1	-0.0497 0.0737	0.1383 0.1230		
	C2	0.0425	0.1772		
	C3	0.0238	0.1225		
Model 2 Condition I (equal small testlet effect)	A	**0.0187**	**0.0615**	**0.1872**	**0.0919**
	B1	0.4009 -0.3983	0.1967 0.1945	0.0133	0.0747
	B2	0.4406	0.2124	4.4063	2.0845
	B3	-0.4082	0.1977	-4.0816	1.9307
	C1	**-0.0224** **0.0339**	**0.1453** **0.1517**	**0.0577**	**0.0546**
	C2	0.0147	0.1282	0.1466	1.1841
	C3	0.0137	0.1301	0.1366	1.2354
Model 2 Condition II (equal large testlet effect)	A	**-0.0086**	**0.0518**	**-0.0859**	**0.1037**
	B1	0.3756 -0.3970	0.1856 0.1952	-0.1070	0.1163
	B2	0.3959	0.1923	3.9586	1.8911
	B3	-0.4279	0.2093	-4.2791	2.0573
	C1	**-0.0371** **0.0158**	**0.1374** **0.0925**	**-0.1067**	**0.0925**
	C2	-0.0207	0.1185	-0.2071	1.1030
	C3	0.0020	0.1311	0.0197	1.2514
Model 3 Condition I (unequal median testlet effect)	A	0.1291	0.0824	1.2909	0.6108
	B1	0.5127 -0.2586	0.2464 0.1279	1.2704	0.6067
	B2	0.5563	0.2635	5.5629	2.6126
	B3	-0.3016	0.1505	-3.0160	1.4428
	C1	0.0750 0.1610	0.1476 0.1644	1.1799	0.5499
	C2	0.1135	0.1366	1.1349	1.2756
	C3	0.1230	0.1452	1.2297	1.3603

Note: The results listed were averaged over 10 replications and averaged over 10 items for condition A, B2, B3, C2 and C3; the two sets of values for condition B1 and C1 represented for the results averaged over the former and latter five items respectively.

TABLE 8:
The R^2 based Effect Size of Simulation Study of Uneven Split Design

		R_1^2 (θ)	R_2^2 (G)	R_3^2 (θ*G)	R_4^2	R_5^2	$R_2^2 - R_1^2$ Uniform DIF	$R_3^2 - R_2^2$ Non uniform DIF	$R_4^2 - R_3^2$	$R_5^2 - R_3^2$
Model 1 (no testlet effect)	A	0.9942	0.9988	1.0000			0.0046	0.0012		
	B1	0.9484 0.9673	0.9996 0.9989	1.0000 1.0000			**0.0512** **0.0316**	0.0004 0.0011		
	B2	0.9415	0.9989	1.0000			**0.0574**	0.0011		
	C1	0.9679 0.9848	0.9756 0.9867	1.0000 1.0000			0.0077 0.0019	**0.0244** **0.0133**		
	C2	0.9738	0.9793	1.0000			0.0055	**0.0207**		
		R_1^2 (θ)	R_2^2 (γ)	R_3^2 (G)	R_4^2 (θ*G)	R_5^2 (γ*G)	$R_2^2 - R_1^2$ DIF on γ	$R_3^2 - R_2^2$ Uniform DIF	$R_4^2 - R_3^2$ Non uniform DIF	$R_5^2 - R_3^2$ Non uniform DIF
Model 2 Condition I (equal testlet effect)	A	0.9959	0.9964	0.9989	1.0000	1.0000	0.0005	0.0025	0.0011	0.0011
	B1	0.9508 0.9593	0.9512 0.9598	0.9986 0.9989	1.0000 1.0000	1.0000 1.0000	0.0004 0.0005	**0.0474** **0.0391**	0.0014 0.0011	0.0014 0.0011
	B2	0.9452	0.9456	0.9988	1.0000	1.0000	0.0004	**0.0532**	0.0012	0.0012
	C1	0.9758 0.9838	0.9763 0.9842	0.9794 0.9854	0.9998 1.0000	1.0000 1.0000	0.0005 0.0004	0.0031 0.0017	**0.0206** **0.0146**	**0.0206** **0.0146**
	C2	0.9807	0.9811	0.9834	1.0000	1.0000	0.0005	0.0023	**0.0116**	**0.0116**
Model 3 Condition I (Different testlet effect for means)	A	0.9878	0.9982	0.9989	1.0000	1.0000	**0.0104**	0.0007	0.0011	0.0011
	B1	0.9217 0.9807	0.9912 0.9944	0.9985 0.9994	1.0000 1.0000	1.0000 1.0000	**0.0695** **0.0137**	**0.0073** **0.0051**	0.0015 0.0006	0.0015 0.0006
	B2	0.9153	0.9915	0.9992	1.0000	1.0000	**0.0761**	**0.0077**	0.00081	0.0000
	C1	0.9685 0.9801	0.9802 0.9863	0.9811 0.9866	1.0000 1.0000	1.0000 1.0000	**0.0117** **0.0062**	0.0008 0.0002	**0.0189** **0.0134**	**0.0189** **0.0134**
	C2	0.9688	0.9825	0.9832	1.0000	1.0000	**0.0137**	0.0008	**0.0168**	**0.0168**
Model 3 Condition II (different testlet effect for variances)	A	0.9950	0.9973	0.9993	1.0000	1.0000	0.0023	0.0020	0.0007	0.0007
	B1	0.9519 0.9518	0.9538 0.9539	0.9992 0.9990	1.0000 1.0000	1.0000 1.0000	0.0018 0.0021	**0.0454** **0.0451**	0.0008 0.0010	0.0008 0.0010
	B2	0.9472	0.9489	0.9989	1.0000	1.0000	0.0016	**0.0500**	0.0011	0.0011
	C1	0.9744 0.9844	0.9803 0.9851	0.9840 0.9860	0.9998 1.0000	1.0000 1.0000	0.0058 0.0008	0.0037 0.0009	**0.0160** **0.0140**	**0.0160** **0.0140**
	C2	0.9792	0.9839	0.9863	0.9999	1.0000	0.0047	0.0024	**0.0137**	**0.0137**
Model 3 Condition III (different testlet effect for means and variances)	A	0.9897	0.9977	0.9991	1.0000	1.0000	**0.0080**	0.0014	0.0009	0.0009
	B1	0.9171 0.9760	0.9735 0.9818	0.9990 0.9995	1.0000 1.0000	1.0000 1.0000	**0.0565** **0.0058**	**0.0254** **0.0177**	0.0010 0.0005	0.0010 0.0005
	B2	0.9143	0.9737	0.9991	1.0000	1.0000	**0.0594**	**0.0254**	0.0009	0.0009
	C1	0.9656 0.9800	0.9836 0.9853	0.9860 0.9867	0.9998 1.0000	1.0000 1.0000	**0.0176** **0.0053**	0.0025 0.0014	**0.0139** **0.0133**	**0.0139** **0.0133**
	C2	0.9682	0.9837	0.9854	0.9999	1.0000	**0.0156**	0.0017	**0.0146**	**0.0146**

Note: The results listed were averaged over 10 replications and averaged over 10 items for condition A, B2, B3, C2 and C3; the two sets of values for condition B1 and C1 represented for the results averaged over the former and latter five items respectively.

TABLE 9:
Logistic Regression Coefficients of Simulation Study of Ethnic Example

		τ_0	τ_1 (θ)	τ_2 (G)	τ_3 (θ*G)	τ_4	τ_5
Model 1 (no testlet effect)	A	-0.0580	1.0349	0.0593	-0.0201		
		-0.0577	1.0633	**0.5234**	-0.0341		
	B1	0.4473	1.0207	**-0.4088**	-0.0053		
	B2	-0.0583	1.0347	**0.5617**	-0.0142		
		-0.0586	1.0692	0.0264	**-0.3173**		
	C1	-0.0483	0.7512	0.0867	**0.2635**		
	C2	-0.0579	1.0345	0.0558	**-0.2811**		
		τ_0	τ_1 (θ)	τ_2 (γ)	τ_3 (G)	τ_4 (θ*G)	τ_5 (γ*G)
Model 2 Condition I (equal testlet effect)	A	-0.0005	0.9982	0.9982	0.0342	0.0150	0.0150
		-0.0043	1.0252	1.0252	**0.5086**	-0.0002	-0.0002
	B1	0.4945	0.9518	0.9518	**-0.4427**	0.0591	0.0591
	B2	-0.0021	0.9910	0.9910	**0.5322**	0.0283	0.0283
		-0.0063	1.0317	1.0317	0.0171	**-0.2696**	**-0.2690**
	C1	0.0191	0.7172	0.7172	0.0304	**0.2849**	**0.2849**
	C2	-0.0026	0.9902	0.9902	0.0174	**-0.2423**	**-0.2424**
Model 3 Condition I (Different testlet effect for means)	A	-0.0160	1.0068	1.0068	0.0499	0.0071	0.0071
		-0.0177	1.0184	1.0184	**0.5168**	0.0070	0.0070
	B1	0.4571	0.9663	0.9663	**-0.4057**	0.0453	0.0453
	B2	-0.0170	0.9980	0.9980	**0.5472**	0.0209	0.0209
		-0.0238	1.0150	1.0150	0.0168	**-0.2623**	**-0.2623**
	C1	0.0017	0.7304	0.7304	0.0511	**0.2778**	**0.2778**
	C2	-0.0170	0.9979	0.9979	0.0320	**-0.2501**	**-0.2501**
Model 3 Condition II (different testlet effect for variances)	A	0.0415	0.9845	0.9845	-0.0291	0.0268	0.0268
		0.0125	1.0056	1.0056	**0.4874**	0.0187	0.0187
	B1	0.5311	0.9457	0.9457	**-0.4793**	0.0645	0.0645
	B2	0.0121	0.9831	0.9831	**0.5106**	0.0507	0.0507
		0.0116	1.0102	1.0102	-0.0006	**-0.2485**	**-0.2485**
	C1	0.0492	0.7164	0.7164	0.0007	**0.2854**	**0.2854**
	C2	0.0217	0.9789	0.9790	-0.0069	**-0.2272**	**-0.2272**
Model 3 Condition III (different testlet effect for means and variances)	A	-0.0120	0.9760	0.9760	0.0246	0.0350	0.0350
	B1	-0.0396	0.9796	0.9796	**0.5388**	0.0472	0.0472
		0.4802	0.9660	0.9660	**-0.4286**	0.0470	0.0470
	B2	-0.0222	0.9718	0.9718	**0.5524**	0.0481	0.0481
	C1	-0.0353	0.9913	0.9913	0.0462	**-0.2293**	**-0.2293**
		0.0025	0.7231	0.7232	0.0471	**0.2786**	**0.2786**
	C2	-0.0155	0.9729	0.9729	0.0390	**-0.2255**	**-0.2255**

Note: The results listed were averaged over 10 replications and averaged over 10 items for condition A, B2, B3, C2 and C3; the two sets of values for condition B1 and C1 represented for the results averaged over the former and latter five items respectively.

TABLE 10:
The Signed-area and Unsigned-area Indices of Simulation Study of Ethnic Example

		Signed-area	Unsigned-area	Signed-area For TCC	Unsigned-area For TCC
Model 1 (no testlet effect)	A	0.0476	0.0650		
		0.4373	0.2086		
	B1	-0.3469	0.1671		
	B2	0.4730	0.2245		
		0.0076	**0.1302**		
	C1	**0.0850**	**0.1314**		
	C2	**0.0404**	**0.1396**		
Model 2 Condition I (equal small testlet effect)	**A**	**0.0271**	**0.0507**	**0.2713**	**0.1419**
		0.4193	**0.2028**		
	B1	**-0.3858**	**0.1870**	**0.1675**	**0.1494**
	B2	**0.4446**	**0.2125**	**4.4460**	**2.0975**
		0.0158	**0.1291**		
	C1	**0.0204**	**0.1362**	**0.1813**	**0.1045**
	C2	**0.0192**	**0.1167**	**0.1915**	**1.1208**
Model 3 Condition I (Different testlet effect for means)	A	0.1481	0.0919	1.4814	0.6954
		0.5414	0.2580		
	B1	-0.2504	0.1223	1.4555	0.6901
	B2	0.5732	0.2681	5.7322	2.6616
		0.1095	0.1466		
	C1	0.1492	0.1541	1.2934	0.5967
	C2	0.1262	0.1424	1.2616	1.2984
Model 3 Condition II (different testlet effect for variances)	**A**	**-0.0275**	**0.0399**	**-0.2748**	**0.1504**
		0.4044	**0.1947**		
	B1	**-0.4188**	**0.2022**	**-0.0723**	**0.1511**
	B2	**0.4207**	**0.2022**	**4.2071**	**1.9891**
		0.0026	**0.1269**		
	C1	**-0.0087**	**0.1320**	**-0.0310**	**0.0679**
	C2	**-0.0023**	**0.1156**	**-0.0232**	**1.0891**
Model 3 Condition III (different testlet effect for means and variances)	A	0.1211	0.0740	1.2109	0.5921
		0.5525	0.2627		
	B1	-0.2701	0.1312	1.4120	0.6866
	B2	0.5636	0.2671	5.6355	2.6545
		0.1451	0.1356		
	C1	0.1248	0.1474	1.3498	0.6446
	C2	0.1364	0.1279	1.3637	1.2280

Note: The results listed were averaged over 10 replications and averaged over 10 items for condition A, B2, B3, C2 and C3; the two sets of values for condition B1 and C1 represented for the results averaged over the former and latter five items respectively.

Results of Real Data Analysis

Analysis of the ACT reading data yielded some interesting findings about differential item functioning at the item and testlet levels. The first finding was that the person-testlet interaction effect did exist in real data and its magnitude varied among different subjects of content; second, within the same examination, magnitude of person-testlet interaction varied among different subgroups of examinees; third, the item characteristics did interact with testlet effect to yield DIF amplification and DIF cancellation at the item and /or testlet level.

I. Results of Model Comparisons

An important first step in the data analysis of DIF was to assess the relative fit of four models, 2-parameter logistic model for one group, 2-parameter logistic model for two groups, and 2-parameter testlet model for one group and 2-parameter testlet model for two groups, to the observed data. If the testlet model fit better than the 2-PL model by taking the person–testlet interaction into consideration, it indicates that there is a testlet effect to capture the local dependence among items nested within testlets; If the two-group 2-PL/2PL testlet models fit the data better than one-group 2-PL/2PL testlet models, it suggests that the test functioned differently between the reference group and focal group. Regarding the model identification issue, constraints were set as the 2-PL testlet model by fixing the difficulty of the last item as the negative sum of the item difficulties of the rest of items in the test and for the convenience of model convergence, the prior distributions of testlet parameters were all set as $Normal(0,1)$.

The DIC results of those four models were shown in Table 11 and Table 12. For both of the two gender samples, the DICs of 2-PLM testlet models were smaller than those of 2-PLM models and additionally, the DIC of 2-group 2-PL testlet model were decreased by about 197 for sample 1 and by 242 for sample 2, compared with those of the one-group 2-PL testlet model. For both ethnic samples, similarly, the DICs of 2-PLM testlet models were smaller than those of 2-PL models and the DIC of 2-group 2-PL testlet model decreased by about 68 for sample 1 and by 63 for sample 2 compared with those of one-group 2-PL testlet model. Thus there was evidence that testlet effect did exist in the test and the test functioned differently between males and females and between minorities and Caucasians.

Additionally, we made a detailed investigation of the DIC difference between one-group and two-group models. Since the 2-PL model ignored the testlet effect, the DIC difference between the one-group 2PL model and two-group 2PL model reflected the difference of item attributes between two subgroups. However, by considering the testlet parameter, the DIC difference between the one-group 2PL testlet model and two-group 2PL testlet model might reflect the difference of combination effect of two sources of DIF: item attributes and testlet distribution. For the first gender sample, DIC of two-group 2PL model decreased about 230 from that of one-group 2PL model, and the difference of DIC values between two-group 2PL testlet model and one-group 2PL testlet model was 197, which might suggest that the different performance between male group and female group could be attributed more to the item characteristics than the testlet effect. These results were confirmed by the second gender sample. For the first ethnic sample, the DIC decreased only one point from two group 2PL model to one-group 2PL

model, but the reduction of DIC value of two-group 2PL testlet model and one-group 2PL testlet model was about 68, which might suggest that the different performance between minorities and Caucasians could be attributed more to testlet effect rather than the item characteristics. Similar results could be confirmed by the second ethnic sample.

TABLE 11:
DIC of 2-PL Model and 2-PL testlet Model of Gender Example

DIC	One- Group 2-PLM	Two-Group 2-PLM	One- Group 2-PLTM	Two-Group 2-PLTM
Sample 1	137869.000	137639.000	136828.000	136631.000
Sample 2	138704.000	138439.000	137711.000	137469.000

TABLE 12:
DIC of 2-PL Model and 2-PL testlet Model of Ethnic Example

DIC	One- Group 2-PLM	Two-Group 2-PLM	One- Group 2-PLTM	Two-Group 2-PLTM
Sample 1	102691.000	102690.000	101855.000	101787.000
Sample 2	102698.000	102684.000	101852.000	101789.000

II. Magnitudes of Differences in Testlet Effect and Item Characteristics

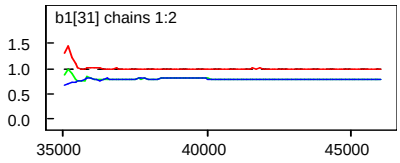
Evidence was provided above that there were testlet effects and item idiosyncratic features that functioned differently between reference group and focal group. (For the gender sample, males were referred as reference group and females were referred to as focal group; for the race sample, Caucasians were referred as reference group and minorities were referred to as focal group). However, what was the real difference in means and variances of testlet parameter between two subgroups? In this investigation, testlet models were constructed by setting much more uninformative prior distributions to the testlet parameter with its means distributed as $Normal(0,1)$ and its variance distributed as Gamma (1,1), and also, in order to deal with the indeterminacy problem, by

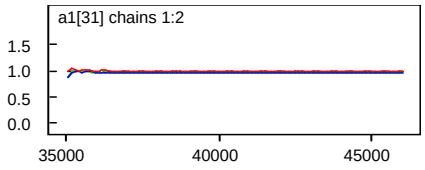
setting four constraints to the item difficulty parameters with the item difficulty of the last item of each testlet as the negative sum of those of the rest of 9 items. The distributions of the main latent proficiency θ were set as $Normal(0,1)$, same for the two subgroups to meet our assumption that the causation of DIF was not depend on the main dimension. The second sample of gender data and race data were used for this study.

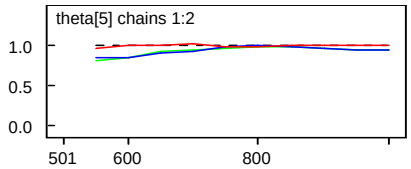
1. Convergence of Models

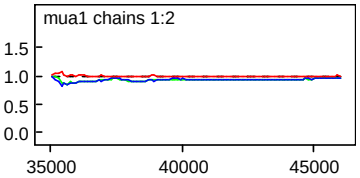
Not surprisingly, the means and precisions (1/ variance) of the testlet distributions proved to be much more challenging to estimate than the item discrimination parameters and latent proficiency parameter θ. Item difficulty parameters were also quite difficult to estimate because the weak priors of testlet parameters caused the problem of model identification. Based on the BGR diagnostic plots, it was shown that these means and precisions required a burn-in of approximately 35,000 iterations. The one noteworthy indicator was that the BGR diagnostics had stabilized around one. See Figure 2 for typical diagnostic plots of gender sample and Figure 5 for typical diagnostic plots of ethnicity sample. The autocorrelations of means and precisions of testlet distributions were relatively higher than those of item discrimination parameters, item difficulty parameters and latent proficiency parameters, albeit the evidence of convergence had been provided by BGR statistics, which probably due to the cross-correlations among four testlet effects in the model. See Figure 3 for autocorrelation plots for some parameters of gender sample and Figure 6 for Autocorrelation plots for some parameters of the ethnic sample. Finally it seemed prudent to end up with a sample of approximately 60,000 iterations (around 20,000 extra iterations after burn-in) in order to be comfortable

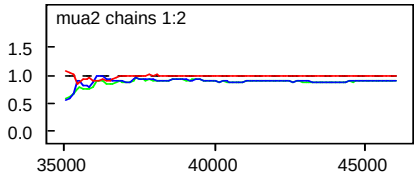
making inferences regarding the posterior distributions. When that was done the density plots were smooth (See Figure 4: for gender sample and Figure 7: for ethnicity sample) and the standard deviations to the MC-error ratios were less than the recommended ratio of 0.05 (See Table 13: for gender sample and Table 14: for ethnic sample of testlet mean and precision parameters and See Appendix A for item characteristics parameters).

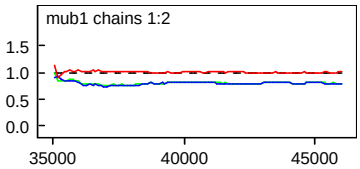

b1[31] chains 1:2
1.5
1.0
0.5
0.0
35000
40000
45000
iteration

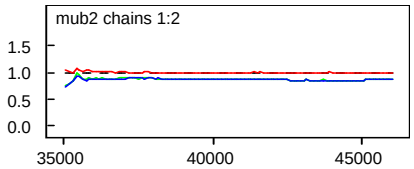

a1[31] chains 1:2
1.5
1.0
0.5
0.0
35000
40000
45000
iteration

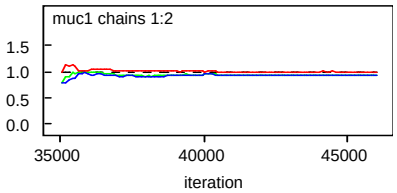

theta[5] chains 1:2
1.0
0.5
0.0
501
600
800
iteration

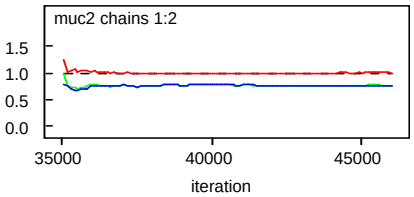

mua1 chains 1:2
1.5
1.0
0.5
0.0
35000
40000
45000
iteration

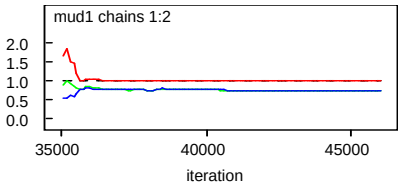

mua2 chains 1:2
1.5
1.0
0.5
0.0
35000
40000
45000
iteration

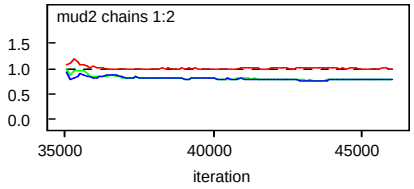

mub1 chains 1:2
1.5
1.0
0.5
0.0
35000
40000
45000
iteration

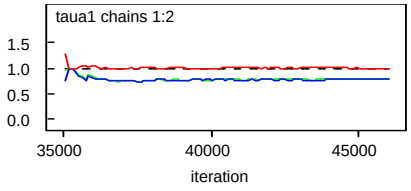

mub2 chains 1:2
1.5
1.0
0.5
0.0
35000
40000
45000
iteration

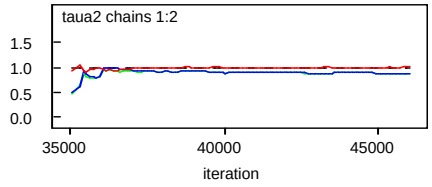

muc1 chains 1:2
1.5
1.0
0.5
0.0
35000
40000
45000
iteration


muc2 chains 1:2
1.5
1.0
0.5
0.0
35000
40000
45000
iteration


mud1 chains 1:2
2.0
1.5
1.0
0.5
0.0
35000
40000
45000
iteration


mud2 chains 1:2
1.5
1.0
0.5
0.0
35000
40000
45000
iteration


taua1 chains 1:2
1.5
1.0
0.5
0.0
35000
40000
45000
iteration


taua2 chains 1:2
1.5
1.0
0.5
0.0
35000
40000
45000
iteration

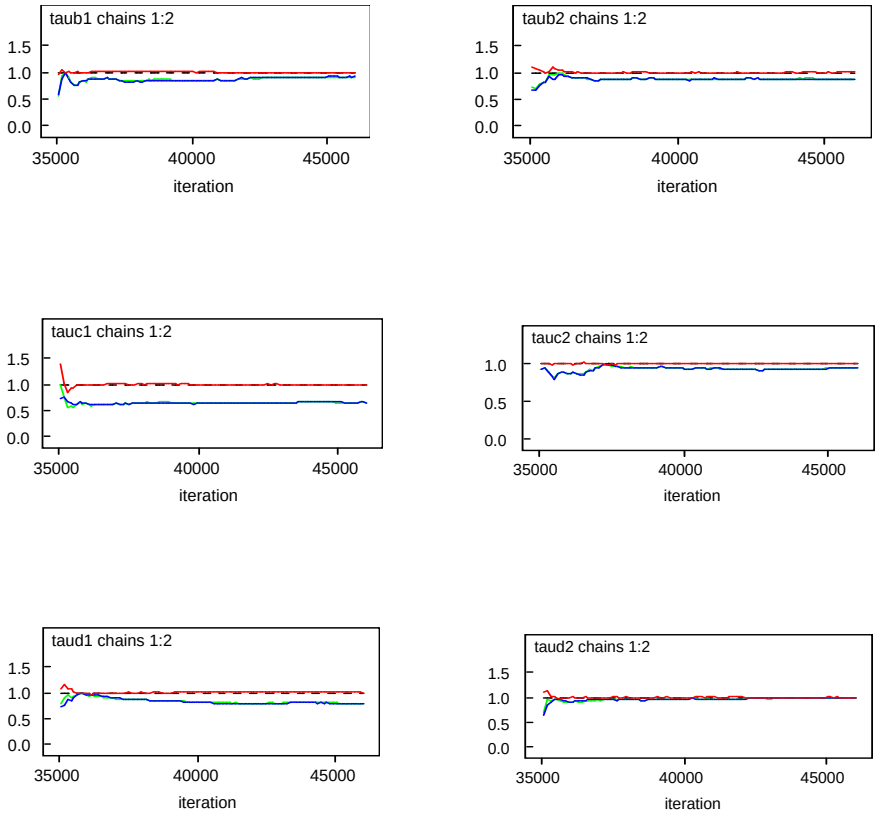
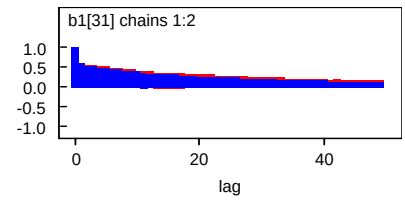
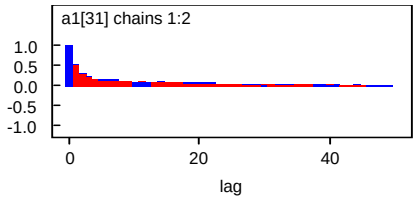
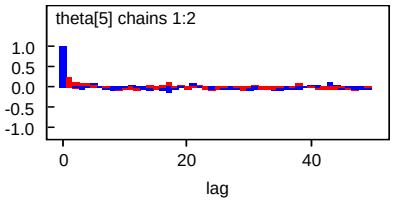
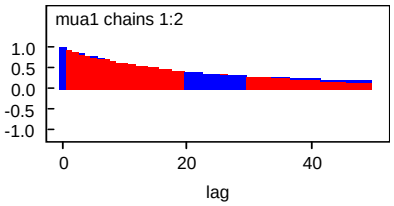


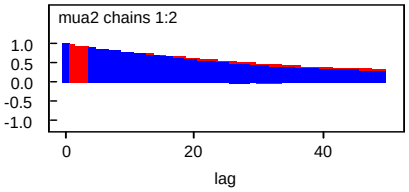
FIGURE 2: Gibbs sampling BGR diagnostic plots of several representative parameter of Gender sample (a1, b1, theta, mua1, mua2, mub1, mub2, muc1, muc2, mud1, mud2 (means of four testlet distributions of each subgroup) and taua1, taua2, taub1, taub2, tauc1, tauc2, taud1, taud2 (precisions of four testlet distributions of each subgroup)

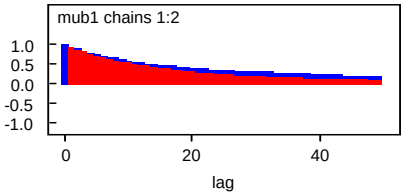

b1[31] chains 1:2
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag

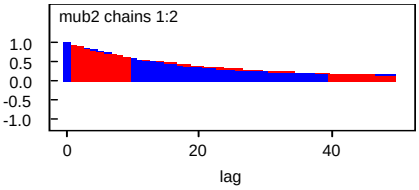

a1[31] chains 1:2
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag

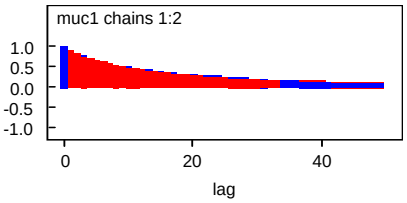

theta[5] chains 1:2
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag

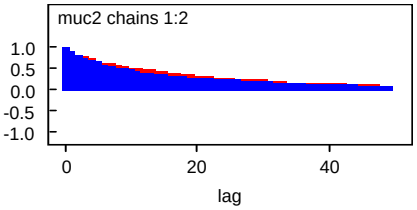

mua1 chains 1:2
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag

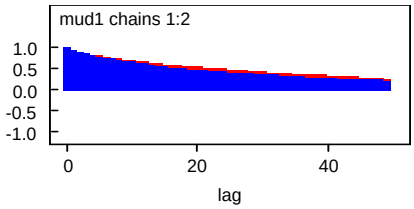

mua2 chains 1:2
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag

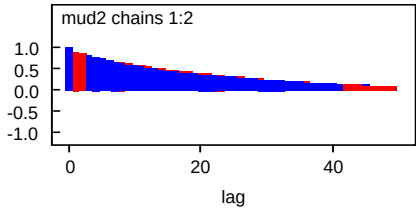

mub1 chains 1:2
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag

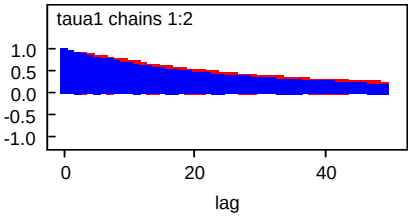

mub2 chains 1:2
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag

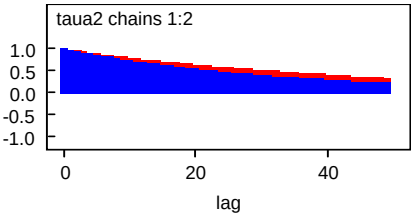

muc1 chains 1:2
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag


muc2 chains 1:2
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag


mud1 chains 1:2
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag


mud2 chains 1:2
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag


taua1 chains 1:2
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag


taua2 chains 1:2
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag

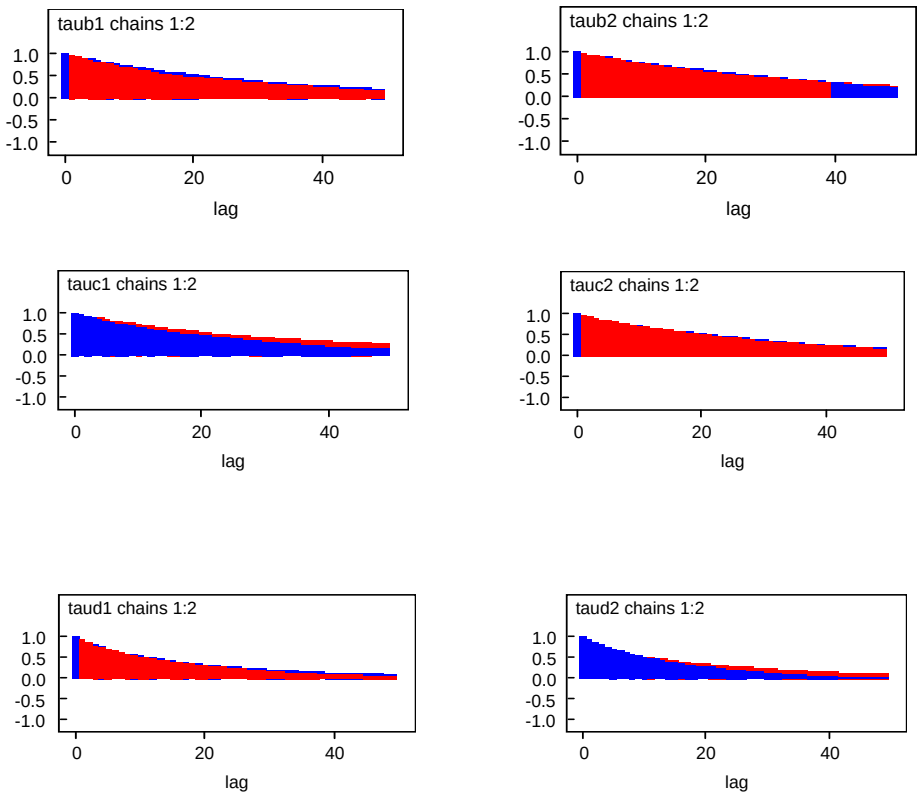
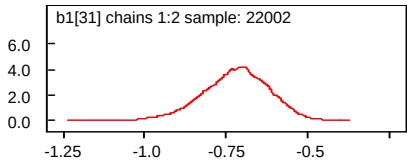
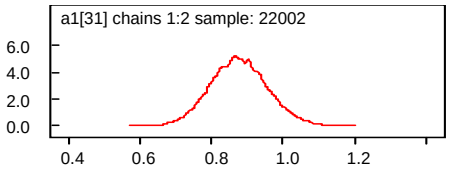
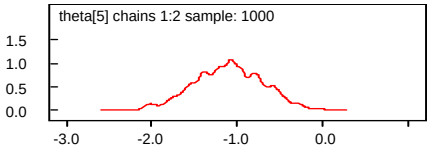
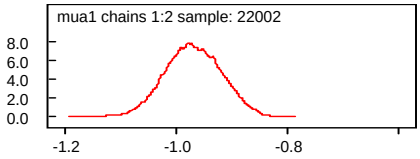


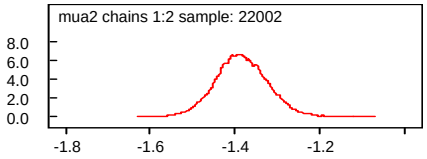
FIGURE 3: Gibbs sampling autocorrelation plots of several representative parameter of Gender sample (a1, b1, theta, mua1, mua2, mub1, mub2, muc1, muc2, mud1, mud2 (means of four testlet distributions of each subgroup) and taua1, taua2, taub1, taub2, tauc1, tauc2, taud1, taud2 (precisions of four testlet distributions of each subgroup

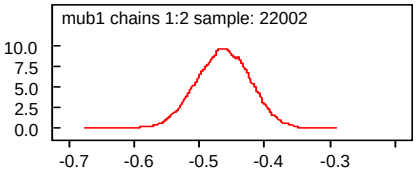

b1[31] chains 1:2 sample: 22002
6.0
4.0
2.0
0.0
-1.25
-1.0
-0.75
-0.5

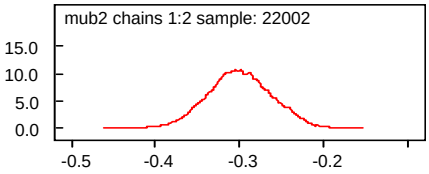

a1[31] chains 1:2 sample: 22002
6.0
4.0
2.0
0.0
0.4
0.6
0.8
1.0
1.2

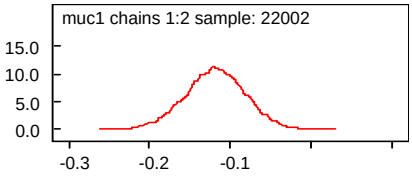

theta[5] chains 1:2 sample: 1000
1.5
1.0
0.5
0.0
-3.0
-2.0
-1.0
0.0

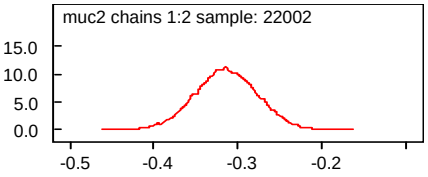

mua1 chains 1:2 sample: 22002
8.0
6.0
4.0
2.0
0.0
-1.2
-1.0
-0.8

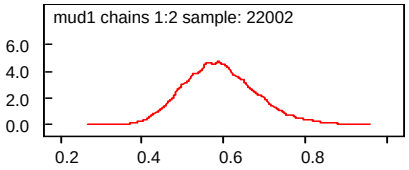

mua2 chains 1:2 sample: 22002
8.0
6.0
4.0
2.0
0.0
-1.8
-1.6
-1.4
-1.2

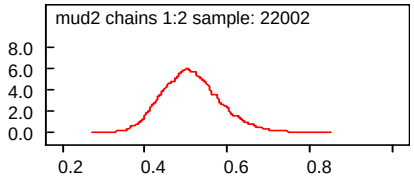

mub1 chains 1:2 sample: 22002
10.0
7.5
5.0
2.5
0.0
-0.7
-0.6
-0.5
-0.4
-0.3


mub2 chains 1:2 sample: 22002
15.0
10.0
5.0
0.0
-0.5
-0.4
-0.3
-0.2


muc1 chains 1:2 sample: 22002
15.0
10.0
5.0
0.0
-0.3
-0.2
-0.1


muc2 chains 1:2 sample: 22002
15.0
10.0
5.0
0.0
-0.5
-0.4
-0.3
-0.2


mud1 chains 1:2 sample: 22002
6.0
4.0
2.0
0.0
0.2
0.4
0.6
0.8


mud2 chains 1:2 sample: 22002
8.0
6.0
4.0
2.0
0.0
0.2
0.4
0.6
0.8

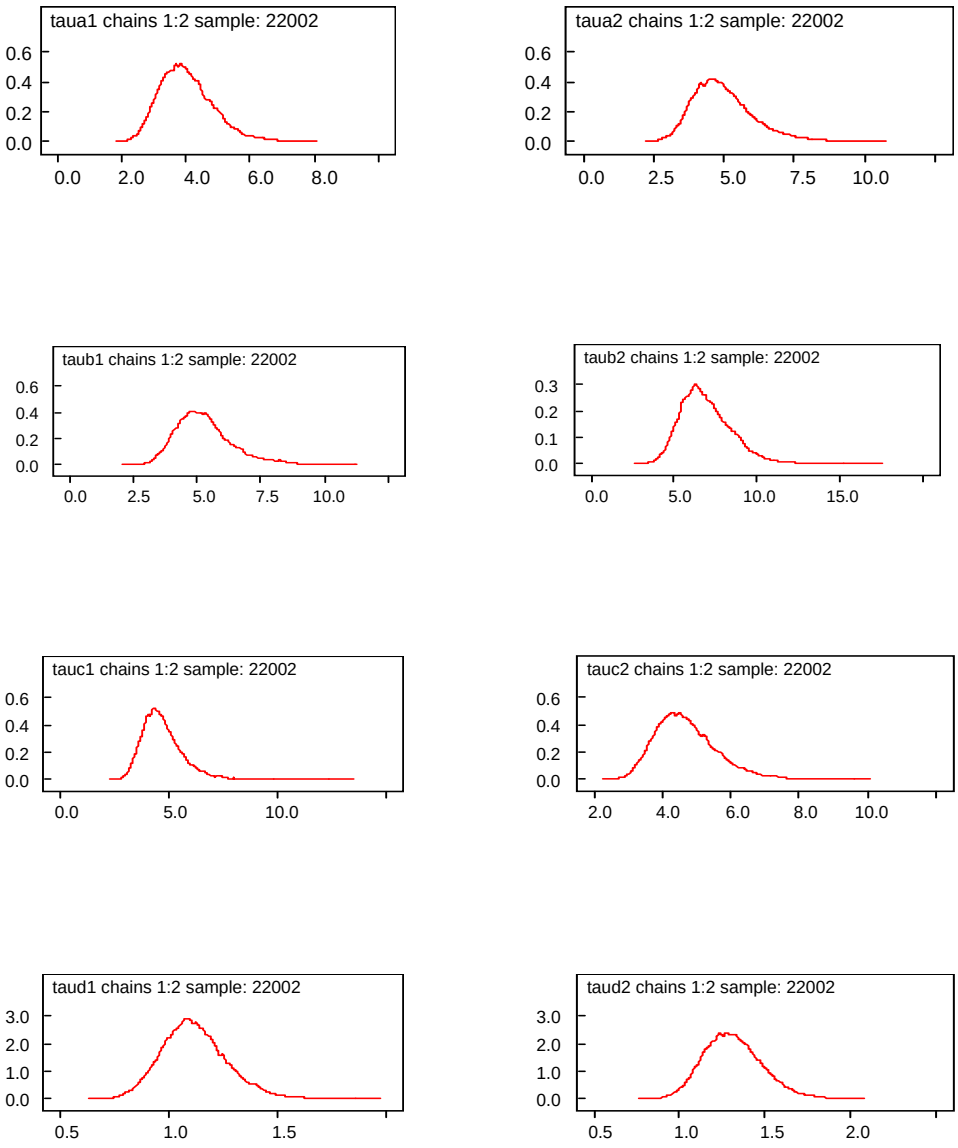
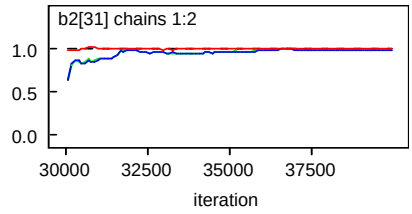
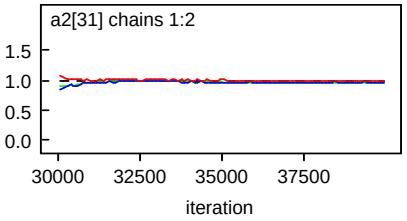
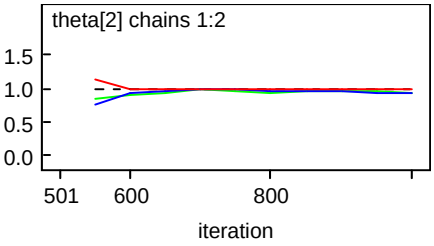
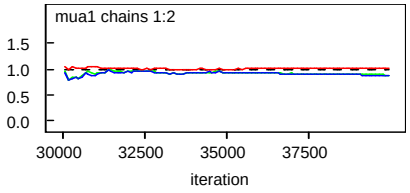


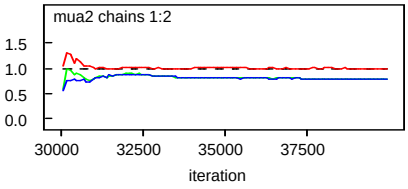
FIGURE 4: Gibbs sampling density function plots of several representative parameter of Gender sample (a1, b1, theta, mua1, mua2, mub1, mub2, muc1, muc2, mud1, mud2 (means of four testlet distributions of each subgroup) and taua1, taua2, taub1, taub2, tauc1, tauc2, taud1, taud2 (precisions of four testlet distributions of each subgroup)

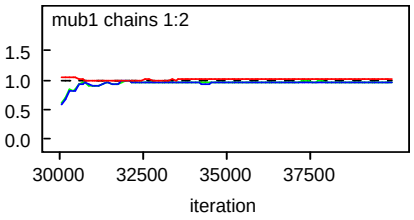

b2[31] chains 1:2
1.0
0.5
0.0
30000
32500
35000
37500
iteration

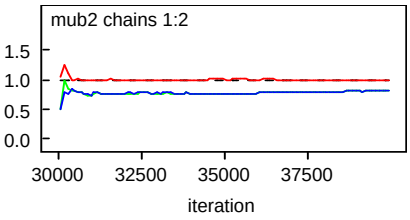

a2[31] chains 1:2
1.5
1.0
0.5
0.0
30000
32500
35000
37500
iteration

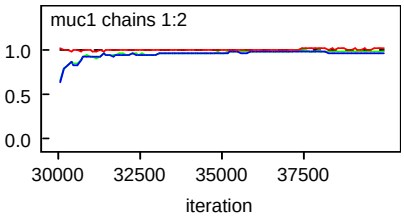

theta[2] chains 1:2
1.5
1.0
0.5
0.0
501
600
800
iteration

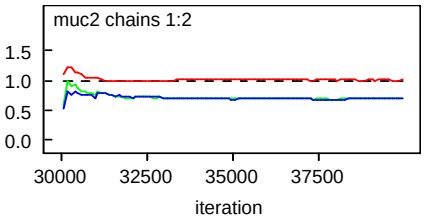

mua1 chains 1:2
1.5
1.0
0.5
0.0
30000
32500
35000
37500
iteration

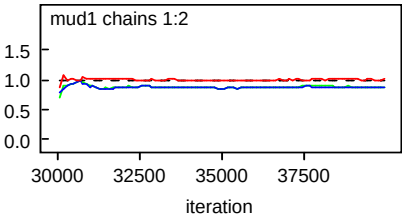

mua2 chains 1:2
1.5
1.0
0.5
0.0
30000
32500
35000
37500
iteration

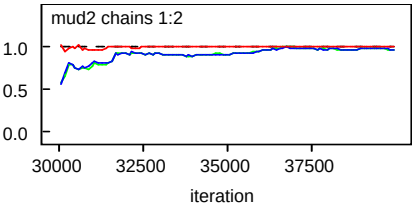

mub1 chains 1:2
1.5
1.0
0.5
0.0
30000
32500
35000
37500
iteration


mub2 chains 1:2
1.5
1.0
0.5
0.0
30000
32500
35000
37500
iteration


muc1 chains 1:2
1.0
0.5
0.0
30000
32500
35000
37500
iteration


muc2 chains 1:2
1.5
1.0
0.5
0.0
30000
32500
35000
37500
iteration


mud1 chains 1:2
1.5
1.0
0.5
0.0
30000
32500
35000
37500
iteration


mud2 chains 1:2
1.0
0.5
0.0
30000
32500
35000
37500
iteration

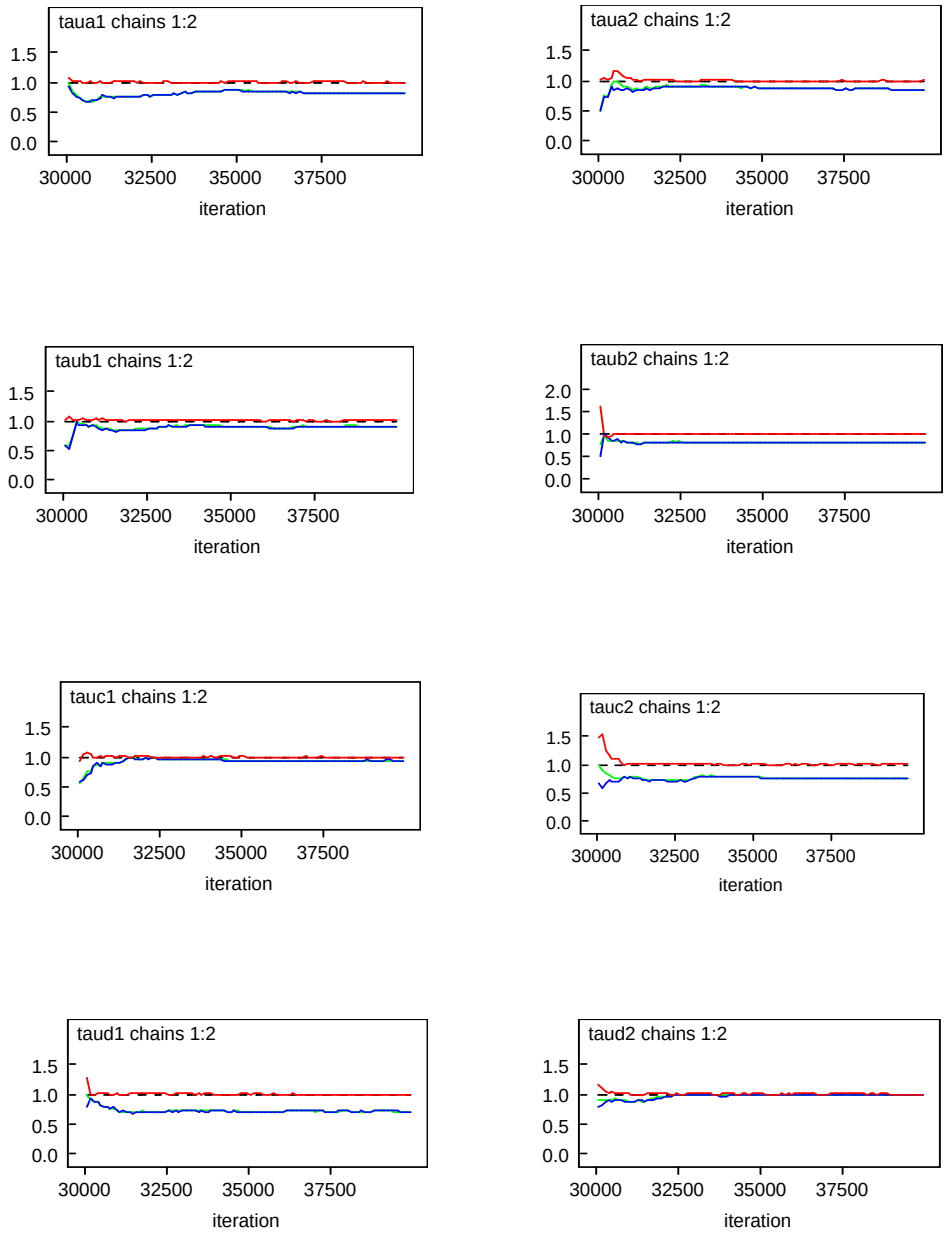
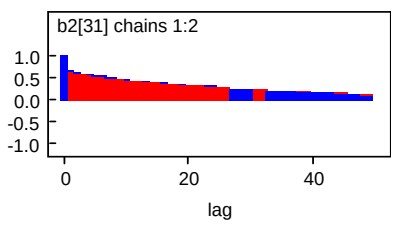
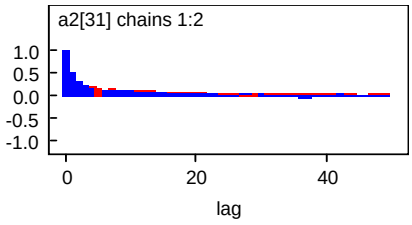
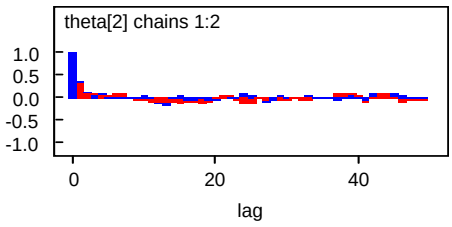
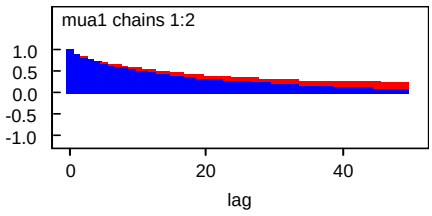


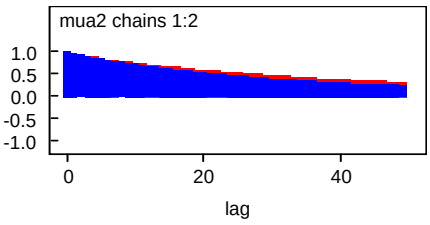
FIGURE 5: Gibbs sampling BGR diagnostic plots of several representative parameter of Ethnic sample (a1, b1, theta, mua1, mua2, mub1, mub2, muc1, muc2, mud1, mud2 (means of four testlet distributions of each subgroup) and taua1, taua2, taub1, taub2, tauc1, tauc2, taud1, taud2 (precisions of four testlet distributions of each subgroup)

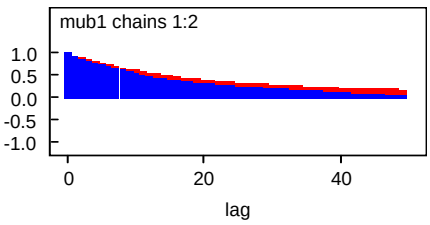

b2[31] chains 1:2
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag

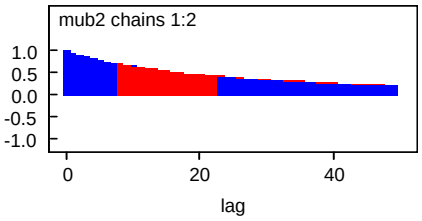

a2[31] chains 1:2
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag

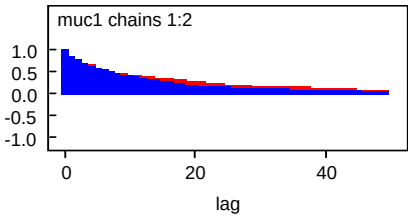

theta[2] chains 1:2
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag

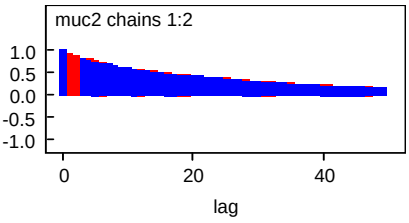

mua1 chains 1:2
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag


mua2 chains 1:2
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag


mub1 chains 1:2
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag


mub2 chains 1:2
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag


muc1 chains 1:2
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag


muc2 chains 1:2
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag

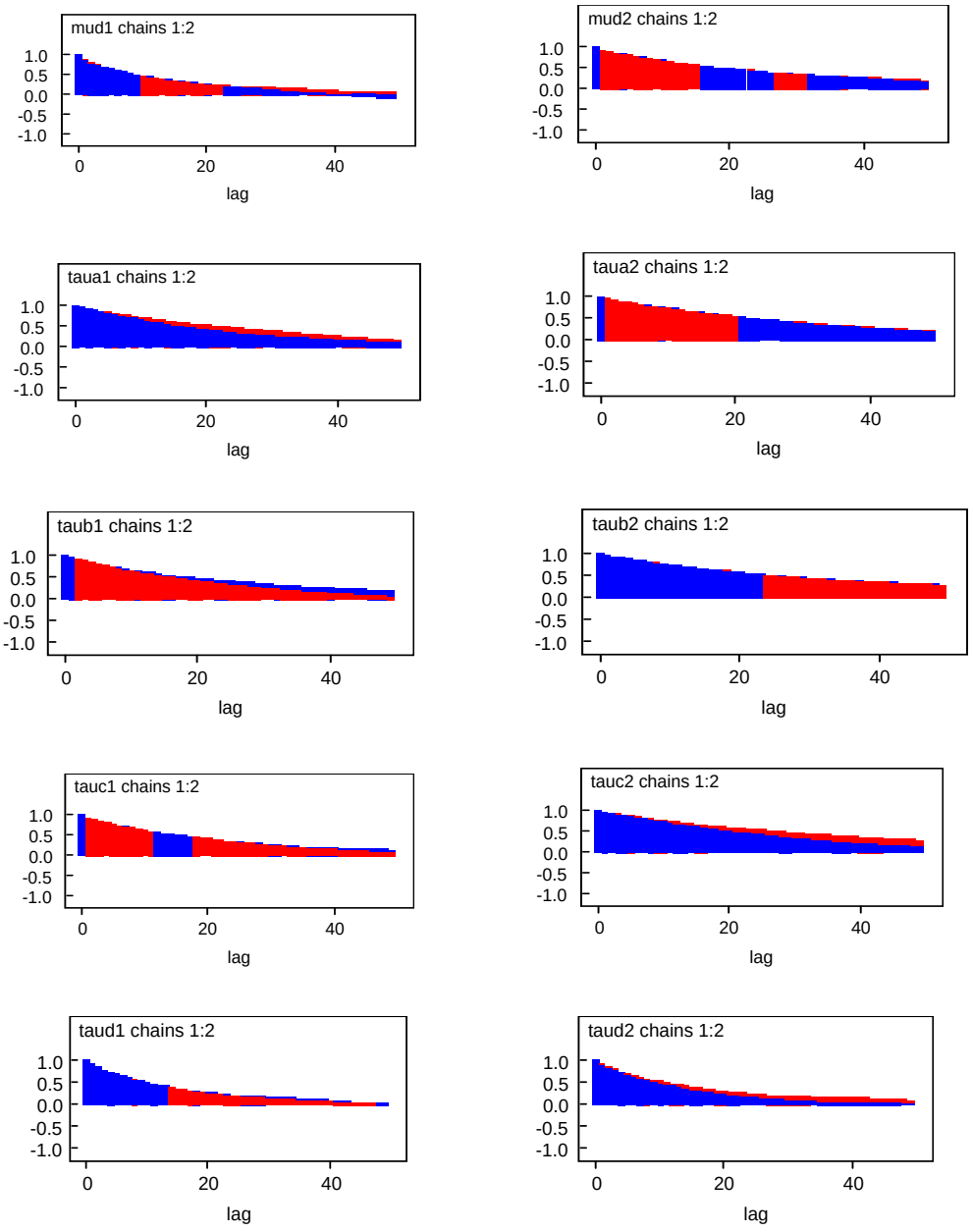
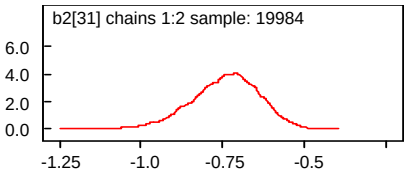
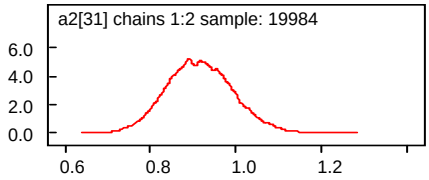
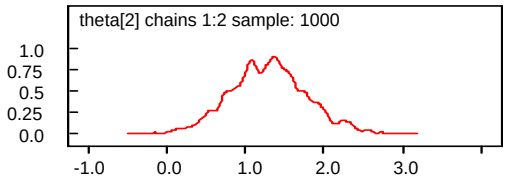
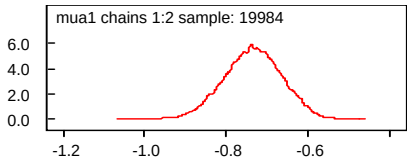


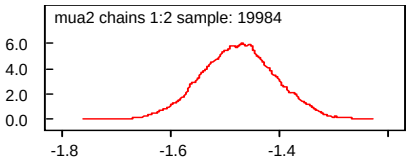
FIGURE 6: Gibbs sampling autocorrelation plots of several representative parameter of Gender sample (a1, b1, theta, mua1, mua2, mub1, mub2, muc1, muc2, mud1, mud2 (means of four testlet distributions of each subgroup) and taua1, taua2, taub1, taub2, tauc1, tauc2, taud1, taud2 (precisions of four testlet distributions of each subgroup

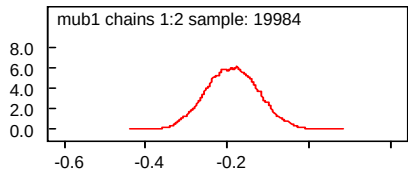

b2[31] chains 1:2 sample: 19984
6.0
4.0
2.0
0.0
-1.25
-1.0
-0.75
-0.5

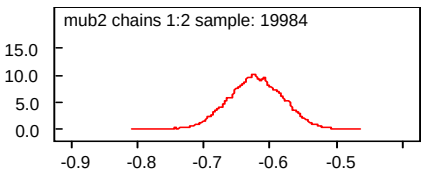

a2[31] chains 1:2 sample: 19984
6.0
4.0
2.0
0.0
0.6
0.8
1.0
1.2

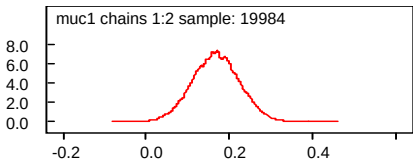

theta[2] chains 1:2 sample: 1000
1.0
0.75
0.5
0.25
0.0
-1.0
0.0
1.0
2.0
3.0

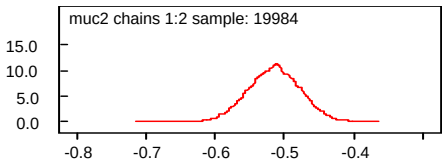

mua1 chains 1:2 sample: 19984
6.0
4.0
2.0
0.0
-1.2
-1.0
-0.8
-0.6

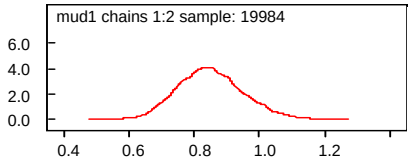

mua2 chains 1:2 sample: 19984
6.0
4.0
2.0
0.0
-1.8
-1.6
-1.4

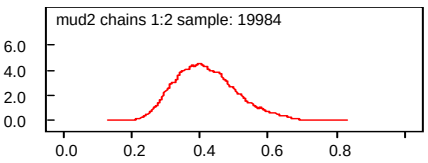

mub1 chains 1:2 sample: 19984
8.0
6.0
4.0
2.0
0.0
-0.6
-0.4
-0.2


mub2 chains 1:2 sample: 19984
15.0
10.0
5.0
0.0
-0.9
-0.8
-0.7
-0.6
-0.5


muc1 chains 1:2 sample: 19984
8.0
6.0
4.0
2.0
0.0
-0.2
0.0
0.2
0.4


muc2 chains 1:2 sample: 19984
15.0
10.0
5.0
0.0
-0.8
-0.7
-0.6
-0.5
-0.4


mud1 chains 1:2 sample: 19984
6.0
4.0
2.0
0.0
0.4
0.6
0.8
1.0
1.2


mud2 chains 1:2 sample: 19984
6.0
4.0
2.0
0.0
0.0
0.2
0.4
0.6
0.8

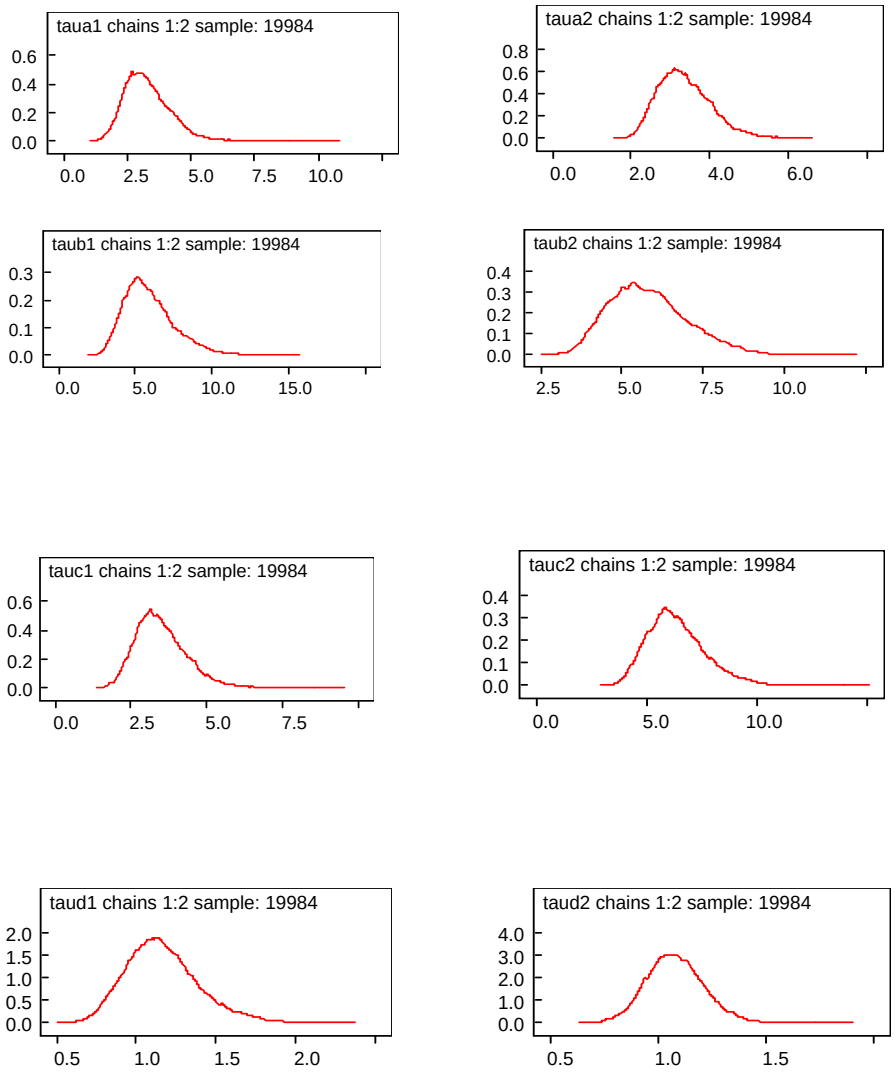


FIGURE 7: Gibbs sampling density plots of several representative parameter of Ethnic sample (a1, b1, theta, mua1, mua2, mub1, mub2, muc1, muc2, mud1, mud2 (means of four testlet distributions of each subgroup) and taua1, taua2, taub1, taub2, tauc1, tauc2, taud1, taud2 (precisions of four testlet distributions of each subgroup)

2. Magnitudes of Differences on Testlet Parameters

When people mention "the content of the ACT Reading Test", they are actually referring two different things. The first type of content refers to the subject matter of the passage, which is, in my understanding, the testlet effect. The second type refers to the sorts of questions asked about the passage, which is, in my understanding, the reading comprehension ability that the test is mainly testing.

Regardless of different subjects of the passages, the essential reading comprehension skills, such as, 1. identify specific details and facts; 2. determine the meaning of words through context; 3. draw inferences from given evidence; 4. understand character and character motivation; 5. identify the main idea of a section or the whole passage; 6. identify the author's point of view or tone; 7. identify cause-effect relationships; 8. make comparisons and analogies, could be acknowledged by every student through certain amount of training in class.

However, different subjects of knowledge of passages could mean differently to minorities and majorities because of the different levels of familiarity with the cultures and also could mean different things to females and males because of certain different cognitive attributes between gender such as motivation or interests, etc.

Appearing in order, the 1996 ACT Reading Test consists of four passages: Prose Fiction adapted from Carol Shiels, "Invitations", about a girl's reactions to invitations to different parties; Social Science talking about the story of a politician, Humanities about the history of Victorian houses in California, and Natural Science about uniqueness of the creatures in Biosphere. Actually, the different distributions of acknowledgement of these four subjects of contents, in other words, the different distributions of Testlet parameters

associated with the four passages, have been found between minorities and Caucasians, females and males in this study.

The results of statistics of testlet parameters were listed in Table 13 for gender sample and Table 14 for the ethnic sample.

TABLE 13:
Statistics of Testlet Parameters from Gender Example

Node	Mean	SD	MC error	2.50%	Median	97.50%
mua1	**0.9685**	0.05110	0.002116	1.0690	0.9689	0.8706
mua2	**1.3840**	0.06280	0.002991	1.5040	1.3860	1.2560
mub1	**0.4634**	0.04176	0.001655	0.5459	0.4629	0.3837
mub2	**0.2995**	0.03866	0.001512	0.3743	0.3001	0.2237
muc1	**0.1187**	0.03726	0.001290	0.1938	0.1184	0.0468
muc2	**0.3140**	0.03706	0.001366	0.3874	0.3141	0.2422
mud1	**-0.5930**	0.09016	0.004051	-0.4330	-0.5872	-0.7906
mud2	**-0.5135**	0.07160	0.002746	-0.3859	-0.5090	-0.6685
taua1	**4.0350**	0.83180	0.036700	2.6920	3.9410	5.9640
taua2	**4.9820**	1.08600	0.052550	3.2970	4.8290	7.5690
taub1	**5.3040**	1.11700	0.049730	3.5520	5.1530	8.0850
taub2	**6.9450**	1.55600	0.071700	4.4940	6.7300	10.470
tauc1	**4.7220**	0.92810	0.041490	3.3190	4.5810	6.8890
tauc2	**4.7040**	0.90860	0.039910	3.2630	4.5850	6.8230
taud1	**1.1210**	0.15080	0.007540	0.005547	0.001993	0.8535
taud2	**1.3140**	0.17190	0.008595	0.006397	0.002198	1.0130

Notes: mua1, mua2 represent for the means of the distributions of four testlets of Males group; taua1, taua2 represent for the precisions of the distributions of four testlets of Males group; mub1, mub2 represent for the means of the distributions of four testlets of Females group; taub1, taub2 represent for the precisions of the distributions of four testlets of Females group.

TABLE 14:
Statistics of Testlet Parameters from Ethnic Example

Node	Mean	SD	MC error	2.50%	Median	97.50%
mua1	**0.7379**	0.07250	0.002945	0.8833	0.7355	0.6034
mua2	**1.4800**	0.07018	0.003543	1.6230	1.4770	1.3490
mub1	**0.1852**	0.06670	0.002656	0.3143	0.1864	0.0529
mub2	**0.6231**	0.04199	0.001879	0.7045	0.6225	0.5417
muc1	**-0.1649**	0.05807	0.002140	-0.0504	-0.1649	-0.2771
muc2	**0.5159**	0.03730	0.001622	0.5891	0.5159	0.4435
mud1	**-0.8510**	0.10620	0.003947	-0.6608	-0.8440	-1.0780
mud2	**-0.4187**	0.08394	0.003636	-0.2735	-0.4131	-0.5959
taua1	**3.2210**	0.85200	0.037540	1.9070	3.1000	5.2380
taua2	**3.3110**	0.66250	0.031420	2.2800	3.2230	4.7960
taub1	**5.6780**	1.54300	0.067390	3.2770	5.4950	9.2490
taub2	**5.9180**	1.27600	0.063850	3.9980	5.7240	8.9740
tauc1	**3.6100**	0.93510	0.039550	2.1920	3.4820	5.8330
tauc2	**6.3600**	1.29300	0.062990	4.1940	6.2190	9.2690
taud1	**1.1580**	0.22390	0.007656	0.7876	1.1350	1.6570
taud2	**1.0800**	0.13530	0.004629	0.8362	1.0750	1.3720

Notes: mua1, mua2 represent for the means of the distributions of four testlets of Minority group; taua1, taua2 represent for the precisions of the distributions of four testlets of Caucasians group; mub1, mub2 represent for the means of the distributions of four testlets of Minority group; taub1, taub2 represent for the precisions of the distributions of four testlets of Caucasians group.

For the first passage, males scored 0.9685 on average, which was less than females' scores since girls were more interested in and more familiar with the topics about foods, dresses, etc. related to parties; not surprisingly, minorities scored lower on average than Caucasians by about 0.7 because of unfamiliarity with the Western culture. The variances of first testlet parameter were about 0.2 to 0.3 and were similar between these two sets of subgroups.

For the second passage, males scored about 0.2 higher than females on average. It seemed make sense that boys were usually more interested in Politics and Economics. Again, minorities scored about 0.5 lower than Caucasians. The variances of the second

testlet were about 0.1 to 0.2. They were similar between race subgroups. And the variability of the females group was slightly smaller than that of males group.

For the third passage, girls scored about 0.2 higher than boys on average. The reason maybe that girls were more interested in the arts of Architecture. Minorities scored -0.1649 on average, which was much lower than Caucasians' mean score: 0.5159. The variability of this testlet was around 0.46, same for males and females. The variance of minority group was 0.5263 and that of Caucasian group was 0.3965, which might indicate that the background knowledge varied more among the group of students from different foreign countries, albeit its small sample size.

For the last passage, boys and girls, minorities and Caucasian were all scored lower than zero on average, minorities were especially lower. The complexity and unfamiliarity of this topic might be the reason of lower scores. The variance of this testlet was around 1, which was the highest among those of four testlets. It suggested that certain level of Natural Science background knowledge was required to understand the content of this passage.

In contrast with the R^2 based effect size indices of logistic regression procedure in Table 17 and Table 21, for the gender sample in Table 17, the average R^2 based effect size of the four testlets were 0.0155, 0.0156, 0.0167 and 0.1503 respectively, as for the ethnic sample in Table 21, the average R^2 based effect size of the four testlets were 0.1286, 0.0731, 0.0676 and 0.3104. The indices reflected the mean and variance differences of the four tesltet distributions for the two samples. There was local dependence among the four testlets, especially for the last one. The different

performances of two subgroups of two samples were more obvious on the last testlet and the first testlet than those of the other two.

3. Magnitudes of Difference on Item Characteristic Features

For the gender sample, the results of estimates of item difficulty parameters and item discrimination parameters are listed in the Appendix A. Table 15 and Table 16 showed that items identified as functioning differentially from a gender perspective, including the magnitude of the differential item functioning on item difficulty parameters (shown as the mean for bdif) and on item discrimination parameters (shown as the mean for adif) separately, for each of the 40 items in this test. Items that are bolded are those for which the confidence interval for the difference between the item difficulties and item discriminations in the two subgroups did not contain zero. Table 17 listes the R^2 based effect sizes of logistic regression procedure of detecting DIF. Items that were bolded in the table were those for which the magnitudes of DIF were relatively larger than others. Table 18 listes the regression coefficients. Again, the coefficients showing statistically significant different from zero are bolded in the Table. The "+"sign of the values of regression coefficients indicated females group was favored and "-" sign indicated that the males group was favored.

For the item difficulty parameters, the largest DIF between two subgroups had been found for Item 20 with the mean bdif of -0.8027, and Item 7 with the mean bdif of 0.8836. The result of Item 20 was consistent with the values (including sign) of regression coefficients τ_3 (if the values were significantly differed from zero, it denoted items displaying uniform DIF), where Item 20 seemed much easier for Males than for

Females. The result was also confirmed by the R^2 based effect sizes in Table 17, where $R_3^2 - R_2^2$ suggesting the magnitude of DIF due to item difficulty after conditioning on the effects of testlet and main latent trait. However, for some reason, the large amount of DIF on Item 7 had been detected by the model but not by R^2 based effect sizes albeit satisfactorily indicated by the value and sign of the regression coefficient in Table 18. It might be still attributed to the influence of exaggerated contributions of testlet effect because of its order of entering the regression model. It also could be the reason of the dispersion of large magnitude of DIF on item discrimination parameters. Then moderate amount of DIF around 0.4 ~ 0.5 had been found on item 16, 23, 6 and 9. They were confirmed by the results in Table 17 and Table 18 except the DIF on item 9 could not be detected by R^2 based effect sizes. Finally, negligible amount of DIF on item difficulty parameters have been detected for Item 3, 5, 13, 14, 17, 18, 19, 22, 25, 26, 32 and 37.

As to the item discrimination parameter, the large amount of DIF had been detected for Item 7 with the mean of adif of 0.3199, indicating that item discriminated more highly among males than females. The result was consistent with those of logistic regression procedure, where $R_4^2 - R_3^2$ and $R_5^2 - R_3^2$ indicated the interaction term of item with the main dimension θ and with the secondary dimension γ separately. And again, it was consistent with the value and sign of regression coefficients τ_4 and τ_5 (if the values were significantly different from zero, it denoted items displaying nonuniform DIF because of the integrations between subgroups and the main dimension θ , for τ_4 , and interaction between subgroups and the testlet dimension, for τ_5, respectively). The

evidences of small magnitude of DIF had been detected on Item 12 and item 26, and evidences of negligible magnitudes of DIF had been detected on Item 10, 23 and 29.

TABLE 15:
DIF Analysis of Item Difficulty Parameters of Gender Example

Node	Mean	SD	MC error	2.50%	Median	97.50%
Testlet A						
difb[1]	-0.1027	0.20660	0.004978	-0.5054	-0.10680	0.3176
difb[2]	-0.07048	0.17980	0.004094	-0.4165	-0.07379	0.2910
difb[3]	**-0.2163**	**0.12440**	**0.002524**	**-0.4605**	**-0.21610**	**0.02824**
difb[4]	0.1609	0.18090	0.003842	-0.1831	0.15740	0.5267
difb[5]	**-0.1561**	**0.10840**	**0.002581**	**-0.3680**	**-0.15620**	**0.05826**
difb[6]	**-0.3430**	**0.21360**	**0.004160**	**-0.7436**	**-0.35120**	**0.09494**
difb[7]	**0.8836**	**0.33190**	**0.011960**	**0.2800**	**0.86390**	**1.5840**
difb[8]	0.08549	0.11770	0.002637	-0.1444	0.08531	0.3164
difb[9]	**-0.4054**	**0.12650**	**0.002657**	**-0.6538**	**-0.40570**	**-0.1544**
difb[10]	0.1641	0.16700	0.002404	-0.1457	0.16000	0.5060
Testlet B						
difb[11]	0.01597	0.16680	0.003601	-0.3215	0.01744	0.3374
difb[12]	-0.0468	0.14830	0.003248	-0.3527	-0.04119	0.2312
difb[13]	**0.3100**	**0.08970**	**0.001404**	**0.1366**	**0.30830**	**0.4891**
difb[14]	**-0.3080**	**0.09294**	**0.001199**	**-0.4952**	**-0.30700**	**-0.1304**
difb[15]	0.1350	0.12880	0.002048	-0.1253	0.13710	0.3851
difb[16]	**0.4034**	**0.09899**	**0.001296**	**0.2109**	**0.40210**	**0.6001**
difb[17]	**0.2566**	**0.08075**	**0.001200**	**0.1007**	**0.25690**	**0.4162**
difb[18]	**-0.1928**	**0.08174**	**0.001069**	**-0.3526**	**-0.19280**	**-0.03259**
difb[19]	**0.2292**	**0.09441**	**0.001482**	**0.0445**	**0.22920**	**0.4140**
difb[20]	**-0.8027**	**0.16410**	**0.002941**	**-1.1430**	**-0.79610**	**-0.4992**
Testlet C						
difb[21]	-0.08023	0.09739	0.001630	-0.2720	-0.08039	0.1122
difb[22]	**0.2353**	**0.09491**	**0.001625**	**0.05234**	**0.23400**	**0.4273**
difb[23]	**0.5834**	**0.15050**	**0.003246**	**0.3031**	**0.57800**	**0.8927**
difb[24]	-0.05519	0.07604	9.42E-04	-0.2027	-0.05511	0.09284
difb[25]	**-0.2171**	**0.07815**	**9.01E-04**	**-0.3704**	**-0.21620**	**-0.06645**
difb[26]	**-0.3228**	**0.07504**	**0.001007**	**-0.4690**	**-0.32270**	**-0.1762**
difb[27]	-0.1241	0.06634	7.76E-04	-0.2557	-0.12410	0.005894
difb[28]	-0.1600	0.08541	0.001410	-0.3266	-0.16010	0.007737
difb[29]	7.97E-04	0.13400	0.002310	-0.2555	-9.59E-04	0.2714
difb[30]	0.1399	0.09844	0.001241	-0.05149	0.13880	0.3368
Testlet D						
difb[31]	-0.08681	0.13550	0.004271	-0.3605	-0.08547	0.1771
difb[32]	**0.2845**	**0.12520**	**0.004209**	**0.0340**	**0.28510**	**0.5304**
difb[33]	-0.1676	0.13670	0.004695	-0.4404	-0.16550	0.09622
difb[34]	0.0917	0.13960	0.004725	-0.1877	0.09262	0.3639
difb[35]	0.01212	0.13890	0.003788	-0.2671	0.01253	0.2839
difb[36]	-0.1205	0.15710	0.003927	-0.4323	-0.11950	0.1877
difb[37]	**-0.3371**	**0.15930**	**0.004111**	**-0.6528**	**-0.33580**	**-0.02637**
difb[38]	0.07977	0.17350	0.003646	-0.2650	0.07946	0.4209
difb[39]	-0.3908	0.21130	0.003529	-0.8121	-0.38780	0.01833
difb[40]	0.6348	0.83860	0.034170	-0.9690	0.60770	2.3740

TABLE 16:
DIF Analysis of Item Discrimination Parameters of Gender Example

Node	Mean	SD	MC error	2.50%	Median	97.50%
Testlet A						
difa[1]	0.07967	0.11610	0.002883	-0.1494	0.07881	0.3097
difa[2]	-0.01535	0.14530	0.003642	-0.3009	-0.01630	0.2707
difa[3]	**0.2174**	**0.14540**	**0.002714**	**-0.06702**	**0.21750**	**0.5036**
difa[4]	-0.0617	0.10640	0.002250	-0.2713	-0.06188	0.1473
difa[5]	-0.1392	0.13307	0.002244	-0.4006	-0.13810	0.1222
difa[6]	0.1245	0.08405	0.001538	-0.04106	0.12470	0.2913
difa[7]	**0.3199**	**0.09535**	**0.002690**	**0.1311**	**0.32090**	**0.5030**
difa[8]	-0.0905	0.11020	0.001511	-0.3059	-0.09091	0.1257
difa[9]	-0.07703	0.09534	0.001304	-0.2650	-0.07645	0.1082
difa[10]	**0.2074**	**0.12170**	**0.002171**	**-0.03176**	**0.20720**	**0.4461**
Testlet B						
difa[11]	-0.08951	0.11040	0.002319	-0.3043	-0.09008	0.1300
difa[12]	**-0.3029**	**0.16100**	**0.003433**	**-0.6210**	**-0.30110**	**0.01032**
difa[13]	0.06761	0.15360	0.002659	-0.2296	0.06654	0.3704
difa[14]	-0.00111	0.12130	0.001851	-0.2392	-0.00173	0.2380
difa[15]	-0.1544	0.12020	0.001972	-0.3915	-0.15370	0.08113
difa[16]	-0.00907	0.10190	0.001300	-0.2079	-0.00938	0.1909
difa[17]	0.1447	0.12120	0.001637	-0.08955	0.14400	0.3864
difa[18]	0.1955	0.12009	0.001632	-0.03826	0.19510	0.4338
difa[19]	0.0506	0.11050	0.001496	-0.1642	0.05007	0.2693
difa[20]	0.08035	0.10140	0.001704	-0.1187	0.08043	0.2790
Testlet C						
difa[21]	-0.00249	0.14001	0.002385	-0.2782	-0.00107	0.2734
difa[22]	-0.01042	0.14170	0.002422	-0.2899	-0.00996	0.2681
difa[23]	0.03811	0.10690	0.002038	-0.1693	0.03726	0.2476
difa[24]	0.1045	0.13610	0.001879	-0.1600	0.10430	0.3722
difa[25]	-0.08945	0.12430	0.001860	-0.3333	-0.08915	0.1532
difa[26]	**0.3495**	**0.12970**	**0.001898**	**0.09668**	**0.34870**	**0.6082**
difa[27]	-0.05898	0.14190	0.002006	-0.3373	-0.05965	0.2188
difa[28]	-0.01824	0.12850	0.001975	-0.2714	-0.01882	0.2295
difa[29]	**-0.1712**	**0.08929**	**0.001328**	**-0.3451**	**-0.17110**	**0.004395**
difa[30]	0.08928	0.10840	0.001599	-0.1249	0.08962	0.2992
Testlet D						
difa[31]	0.01698	0.10940	0.002103	-0.1966	0.01607	0.2347
difa[32]	-0.05541	0.13210	0.002285	-0.3138	-0.05515	0.2021
difa[33]	-0.1670	0.12580	0.002305	-0.4113	-0.16800	0.08451
difa[34]	0.1000	0.11650	0.002019	-0.1262	0.09931	0.3296
difa[35]	-0.01602	0.10820	0.002039	-0.2279	-0.01531	0.1946
difa[36]	-0.1364	0.08587	0.001421	-0.3049	-0.13620	0.03172
difa[37]	-0.1268	0.08430	0.001301	-0.2939	-0.12690	0.03852
difa[38]	-0.03055	0.09003	0.001691	-0.2007	-0.03038	0.1450
difa[39]	0.03172	0.07344	0.001189	-0.1104	0.03122	0.1762
difa[40]	-0.0649	0.06345	0.002378	-0.1888	-0.06501	0.06116

TABLE 17:
Results of R^2 based Effect Size of Logistic Regression of Gender Example

Item	R_1^2	R_2^2	R_3^2	R_4^2	R_5^2	$R_2^2 - R_1^2$ DIF on γ	$R_3^2 - R_2^2$ Uniform DIF	$R_4^2 - R_3^2$ Non uniform DIF	$R_5^2 - R_3^2$ Non uniform DIF
Testlet A						0.01549			
1	0.8107	0.8154	0.8174	0.8182	0.8183	0.0047	0.0019	0.0008	0.0009
2	0.6636	0.6797	0.6848	0.7111	0.7112	0.0161	0.0051	0.0264	0.0264
3	0.6207	0.6224	0.6325	0.6632	0.6634	0.0017	**0.0101**	**0.0308**	**0.0310**
4	0.6966	0.7569	0.7795	0.8400	0.8401	0.0603	0.0226	0.0606	0.0607
5	0.5756	0.5937	0.6036	0.6609	0.6610	0.0181	0.0099	0.0574	0.0575
6	0.8438	0.8441	0.8826	0.9217	0.9224	0.0003	**0.0385**	0.0391	0.0397
7	0.7718	0.7740	0.7866	0.8643	0.8648	0.0022	**0.0126**	**0.0777**	**0.0782**
8	0.6531	0.6929	0.7099	0.7746	0.7746	0.0398	0.0170	0.0647	0.0647
9	0.8310	0.8390	0.8599	0.8670	0.8671	0.0080	**0.0209**	0.0071	0.0072
10	0.7609	0.7646	0.7660	0.7725	0.7726	0.0037	0.0013	**0.0065**	**0.0066**
Testlet B						0.01555			
11	0.7879	0.7939	0.8000	0.8068	0.8068	0.0059	0.0061	0.0069	0.0069
12	0.4963	0.4967	0.5202	0.5921	0.5923	0.0004	0.0235	**0.0719**	**0.0721**
13	0.6036	0.6077	0.6092	0.6092	0.6092	0.0040	**0.0015**	0.0000	0.0000
14	0.6572	0.6879	0.7017	0.7380	0.7387	0.0307	**0.0138**	0.0363	0.0370
15	0.6748	0.6753	0.6944	0.7282	0.7283	0.0004	0.0192	0.0338	0.0339
16	0.7732	0.7754	0.7927	0.8010	0.8011	0.0021	**0.0173**	0.0083	0.0083
17	0.6758	0.6845	0.6849	0.6898	0.6899	0.0086	0.0004	0.0050	0.0051
18	0.5656	0.5981	0.6168	0.6992	0.7002	0.0325	0.0187	0.0823	0.0833
19	0.7342	0.7413	0.7439	0.7439	0.7440	0.0071	0.0026	0.0001	0.0000
20	0.5995	0.6634	0.7213	0.7228	0.7245	0.0638	**0.0580**	0.0015	0.0032
Testlet C						0.01674			
21	0.6280	0.6391	0.6391	0.6439	0.6440	0.0111	0.0000	0.0048	0.0048
22	0.5624	0.5831	0.5935	0.6380	0.6384	0.0207	**0.0104**	0.0445	0.0450
23	0.6979	0.7469	0.7789	0.8202	0.8211	0.0490	**0.0320**	0.0413	0.0422
24	0.6309	0.6381	0.6402	0.6408	0.6408	0.0072	0.0021	0.0006	0.0006
25	0.6545	0.6662	0.6663	0.6734	0.6735	0.0118	0.0001	0.0071	0.0071
26	0.5143	0.5145	0.5458	0.6215	0.6223	0.0002	**0.0313**	**0.0757**	**0.0765**
27	0.5586	0.5674	0.5675	0.5769	0.5769	0.0088	0.0001	0.0094	0.0094
28	0.6260	0.6347	0.6353	0.6368	0.6368	0.0087	0.0006	0.0016	0.0016
29	0.7769	0.8033	0.8037	0.8361	0.8365	0.0264	0.0004	**0.0324**	**0.0329**
30	0.7118	0.7353	0.7381	0.7458	0.7459	0.0235	0.0028	0.0077	0.0078
Testlet D						0.15025			
31	0.5774	0.7039	0.7054	0.7054	0.7054	0.1265	0.0015	0.0000	0.0000
32	0.4664	0.5578	0.5657	0.5896	0.5940	0.0913	**0.0079**	0.0239	0.0283
33	0.5582	0.6775	0.6776	0.6786	0.6788	0.1193	0.0000	0.0011	0.0013
34	0.5382	0.6499	0.6500	0.6509	0.6510	0.1117	0.0001	0.0009	0.0010
35	0.5707	0.6950	0.6950	0.6963	0.6965	0.1244	0.0000	0.0013	0.0015
36	0.6856	0.8664	0.8665	0.8667	0.8668	0.1808	0.0001	0.0002	0.0003
37	0.6816	0.8599	0.8696	0.8711	0.8715	0.1783	**0.0097**	0.0015	0.0019
38	0.6483	0.8087	0.8089	0.8104	0.8107	0.1604	0.0002	0.0015	0.0018
39	0.6876	0.8714	0.8903	0.8934	0.8943	0.1838	**0.0189**	0.0031	0.0040
40	0.7394	0.9653	0.9696	0.9697	0.9697	0.2260	0.0043	0.0001	0.0001

TABLE 18:
Regression Coefficients of Gender Example

Item	τ_0	τ_1 (for θ)	τ_2 (for γ)	τ_3 (for G)	τ_4 (for θ*G)	τ_5 (for γ*G)
Testlet A						
1	0.4234	0.8203	0.8203	-0.1172	-0.0797	-0.0797
2	0.6055	1.0280	1.0280	-0.0642	0.0160	0.0160
3	0.1257	1.2570	1.2570	**-0.2466**	**-0.2170**	**-0.2170**
4	0.0359	0.6934	0.6934	0.1246	0.0616	0.0616
5	-0.3032	1.0440	1.0440	-0.2250	0.1390	0.1390
6	-0.0082	0.5339	0.5339	**-0.1385**	-0.1245	-0.1245
7	0.2137	0.7352	0.7352	**0.2738**	**-0.3199**	**-0.3199**
8	-0.5048	0.8248	0.8248	0.0229	0.0905	0.0905
9	-0.4399	0.6487	0.6487	**-0.3467**	0.0770	0.0770
10	0.0477	0.9887	0.9887	0.1182	**-0.2074**	**-0.2074**
Testlet B						
11	0.6911	0.7390	0.7390	0.0970	0.0895	0.0895
12	1.2275	1.0500	1.0500	**0.2906**	**0.3030**	**0.3030**
13	0.4668	1.3350	1.3350	**0.3690**	-0.0680	-0.0680
14	0.3516	1.0140	1.0140	**-0.3123**	0.0010	0.0010
15	0.6071	0.8670	0.8670	0.2458	0.1540	0.1540
16	-0.5532	0.8107	0.8107	**0.3245**	0.0091	0.0091
17	-0.6780	1.1700	1.1700	**0.3471**	-0.1450	-0.1450
18	-0.2056	1.1750	1.1750	-0.1544	-0.1958	-0.1958
19	-0.8653	0.9883	0.9883	0.2592	-0.0506	-0.0506
20	-1.0113	0.8513	0.8513	**-0.5237**	-0.0803	-0.0803
Testlet C						
21	0.9861	1.1970	1.1970	-0.0939	0.0030	0.0030
22	0.7469	1.2090	1.2090	**0.2939**	0.0110	0.0110
23	0.4984	0.8118	0.8118	**0.4277**	-0.0381	-0.0381
24	0.4186	1.2720	1.2720	-0.0990	-0.1050	-0.1050
25	0.2543	1.0620	1.0620	**-0.2286**	0.0890	0.0890
26	-0.2071	1.3910	1.3910	**-0.2839**	**-0.3500**	**-0.3500**
27	-0.0773	1.3320	1.3320	-0.1760	0.0590	0.0590
28	-0.8487	1.2080	1.2080	-0.2088	0.0180	0.0180
29	-0.5016	0.5845	0.5845	-0.1464	**0.1713**	**0.1713**
30	-0.8642	1.0090	1.0090	0.2049	-0.0889	-0.0889
Testlet D						
31	0.6327	0.8779	0.8779	-0.0870	0.0000	0.0000
32	0.4344	1.0990	1.0990	**0.3501**	0.0000	0.0000
33	1.0087	0.9280	0.9280	-0.0025	0.0000	0.0000
34	0.9577	0.9848	0.9848	-0.0163	0.0000	0.0000
35	-0.0830	0.8921	0.8921	0.0095	0.0000	0.0000
36	0.3028	0.5423	0.5423	-0.0056	0.0000	0.0000
37	0.3743	0.5355	0.5355	**-0.1346**	0.0000	0.0000
38	-0.3402	0.6625	0.6625	0.0396	0.0000	0.0000
39	-0.0806	0.4835	0.4835	**-0.1713**	0.0000	0.0000
40	-0.9107	0.2489	0.2489	-0.0382	0.0000	0.0000

For the ethnic sample, the results of estimates of item difficulty parameters and item discrimination parameters were listed in the Appendix A. Table 19 and Table 20 showed that items identified as functioning differentially from a ethnic perspective, including the magnitude of the differential item functioning on item difficulty parameters (shown as the mean for bdif) and on item discrimination parameters (shown as the mean for adif) separately, for each of the 40 items in this test. Table 21 listed the R^2 based effect sizes of logistic regression procedure of detecting DIF. Table 22 listed the regression coefficients.

Large magnitudes of DIF on item difficulty parameters have been detected by our model on Item 1, 4, and 34, where Caucasians were favored. Unfortunately, due to the reason mentioned above, DIF on item difficulty parameters of item 1 and 4 has not been detected by R^2 based effect size indices. Evidence of relatively moderate magnitude of DIF has been found on Item 8 and 9, where they seemed unusually easy to minority group. A detailed study of these two items revealed that the simplest reading skill: identify specific details and facts were to be tested and Caucasian students seemed to be distracted from the correct answer based on their own empirical understandings. For example, item 9 was to ask student to infer a sentence in the passage: "usual spun-out wastes of time that had to be scratched endlessly for substance" The correct answer to simply identify the fact was "bored and lacking in interesting things to do." However, a lot of Caucasian students selected one of the distractions: "somewhat festive but socially insincere." And thus, relatively more minority students gave correct answers to these two items. Small or negligible amount of DIF on item difficulty parameters have been detected on Item 12 and 26. Moderate amount of DIF on item discrimination parameters

have been detected on Item 1; negligible amount of DIF have been detected on Item 12, 16, 25 and 28.

TABLE 19:
DIF Analysis of Item Difficulty Parameters of Ethnic Example

Node	Mean	SD	MC error	2.50%	Median	97.50%
Testlet A						
difb[1]	**0.9498**	**0.3001**	**0.011230**	**0.4162**	**0.92890**	**1.5930**
difb[2]	-0.08906	0.1750	0.003675	-0.4287	-0.08808	0.2559
difb[3]	-0.03996	0.1413	0.002768	-0.3186	-0.04037	0.2413
difb[4]	**0.7140**	**0.2137**	**0.005366**	**0.3261**	**0.70430**	**1.1590**
difb[5]	-0.07201	0.1362	0.002701	-0.3385	-0.07184	0.1911
difb[6]	0.002077	0.1746	0.002986	-0.3350	-0.00136	0.3548
difb[7]	-0.3222	0.3198	0.009072	-0.9698	-0.31510	0.2685
difb[8]	**-0.4887**	**0.1317**	**0.002667**	**-0.7540**	**-0.48720**	**-0.2366**
difb[9]	**-0.4877**	**0.1549**	**0.003251**	**-0.7948**	**-0.48550**	**-0.1843**
difb[10]	-0.1663	0.1992	0.002514	-0.5655	-0.16520	0.2194
Testlet B						
difb[11]	-0.03778	0.1968	0.003747	-0.4417	-0.03154	0.3379
difb[12]	**-0.2921**	**0.2138**	**0.004772**	**-0.7502**	**-0.28020**	**0.09283**
difb[13]	-0.1086	0.1242	0.002150	-0.3604	-0.10630	0.1277
difb[14]	-0.1059	0.1117	0.001522	-0.3314	-0.10570	0.1125
difb[15]	0.09901	0.1933	0.003864	-0.2974	0.10340	0.4658
difb[16]	-0.00137	0.1227	0.001786	-0.2374	-0.00434	0.2427
difb[17]	0.02145	0.1196	0.001852	-0.2055	0.01868	0.2650
difb[18]	-0.06604	0.1101	0.001642	-0.2784	-0.06671	0.1518
difb[19]	0.1251	0.1456	0.002385	-0.1452	0.11860	0.4292
difb[20]	0.3662	0.2767	0.005379	-0.1059	0.34210	0.9782
Testlet C						
difb[21]	0.05679	0.1298	0.002525	-0.1979	0.05749	0.3107
difb[22]	-0.02023	0.1151	0.002046	-0.2468	-0.01959	0.2017
difb[23]	0.1020	0.1742	0.004429	-0.2367	0.10110	0.4526
difb[24]	-0.09023	0.09745	0.001268	-0.2800	-0.09094	0.1036
difb[25]	0.08529	0.1073	0.001242	-0.1195	0.08392	0.2999
difb[26]	**-0.1856**	**0.09121**	**0.001274**	**-0.3626**	**-0.18620**	**-0.00285**
difb[27]	-0.07009	0.09135	0.001334	-0.2463	-0.07095	0.1119
difb[28]	-0.05849	0.1244	0.002441	-0.2897	-0.06350	0.2014
difb[29]	0.2001	0.1739	0.003309	-0.1133	0.19010	0.5755
difb[30]	-0.01958	0.1291	0.001730	-0.2594	-0.02384	0.2506
Testlet D						
difb[31]	0.1635	0.1643	0.004512	-0.1546	0.16410	0.4864
difb[32]	0.1722	0.1484	0.004273	-0.1227	0.17220	0.4625
difb[33]	0.09963	0.1654	0.004896	-0.2265	0.10040	0.4231
difb[34]	**0.4743**	**0.1720**	**0.004935**	**0.1319**	**0.47410**	**0.8146**
difb[35]	0.2311	0.1885	0.004078	-0.1259	0.22690	0.6137
difb[36]	0.3314	0.1773	0.004319	-0.01078	0.32990	0.6831
difb[37]	0.02936	0.2050	0.004337	-0.3671	0.02721	0.4397
difb[38]	0.1974	0.2200	0.004296	-0.2112	0.19160	0.6447
difb[39]	-0.1354	0.2301	0.003916	-0.5726	-0.14060	0.3326
difb[40]	-1.5640	0.9251	0.035030	-3.3580	-1.56400	0.2954

TABLE 20:
DIF Analysis of Item Discrimination Parameters of Ethnic Example

Node	Mean	SD	MC error	2.50%	Median	97.50%
Testlet A						
difa[1]	**0.3990**	**0.14160**	**0.003524**	**0.1344**	**0.39490**	**0.6906**
difa[2]	0.1125	0.18750	0.004065	-0.2414	0.10680	0.4932
difa[3]	0.1667	0.19150	0.004114	-0.1987	0.16080	0.5519
difa[4]	0.1328	0.13000	0.002850	-0.1169	0.13060	0.3949
difa[5]	0.004188	0.15010	0.002514	-0.2844	8.64E-04	0.3047
difa[6]	0.03734	0.11470	0.001909	-0.1815	0.03619	0.2674
difa[7]	-0.0279	0.10770	0.002821	-0.2320	-0.03023	0.1843
difa[8]	0.02016	0.15450	0.002538	-0.2759	0.01740	0.3283
difa[9]	-0.1151	0.12210	0.002001	-0.3505	-0.11760	0.1254
difa[10]	0.003754	0.13740	0.002292	-0.2598	0.001183	0.2777
Testlet B						
difa[11]	-0.04498	0.13840	0.002763	-0.3114	-0.04596	0.2356
difa[12]	**-0.3833**	**0.17120**	**0.003960**	**-0.7170**	**-0.38510**	**-0.03942**
difa[13]	-0.06591	0.19230	0.003306	-0.4341	-0.07019	0.3187
difa[14]	-0.1602	0.15640	0.002466	-0.4601	-0.16310	0.1522
difa[15]	-0.1180	0.13760	0.002802	-0.3834	-0.11990	0.1567
difa[16]	**0.3142**	**0.13900**	**0.001874**	**0.05551**	**0.30970**	**0.5974**
difa[17]	-0.1896	0.14200	0.001916	-0.4642	-0.19180	0.09171
difa[18]	-0.1892	0.15240	0.002233	-0.4799	-0.19350	0.1174
difa[19]	-0.07618	0.12900	0.001852	-0.3251	-0.07699	0.1821
difa[20]	-0.1935	0.11830	0.002106	-0.4231	-0.19550	0.04458
Testlet C						
difa[21]	0.01375	0.16260	0.003117	-0.2998	0.01060	0.3369
difa[22]	-0.1669	0.18030	0.003254	-0.5119	-0.16990	0.1973
difa[23]	0.1193	0.13790	0.002958	-0.1441	0.11590	0.3958
difa[24]	0.04945	0.16760	0.002514	-0.2704	0.04534	0.3873
difa[25]	**-0.3074**	**0.15100**	**0.002287**	**-0.6019**	**-0.30910**	**-0.0103**
difa[26]	0.08698	0.18150	0.002829	-0.2635	0.08249	0.4538
difa[27]	-0.08755	0.17650	0.002740	-0.4252	-0.09159	0.2664
difa[28]	**-0.3184**	**0.15450**	**0.002867**	**-0.6136**	**-0.32180**	**-0.01224**
difa[29]	0.002131	0.11640	0.002054	-0.2205	7.70E-05	0.2364
difa[30]	0.003511	0.14570	0.002455	-0.2757	0.002433	0.2963
Testlet D						
difa[31]	-0.06947	0.13730	0.002603	-0.3327	-0.07260	0.2027
difa[32]	0.1943	0.17980	0.003076	-0.1428	0.18730	0.5657
difa[33]	-0.0429	0.14670	0.002605	-0.3218	-0.04696	0.2579
difa[34]	-0.1352	0.13490	0.002284	-0.3959	-0.13580	0.1336
difa[35]	-0.1107	0.12610	0.002297	-0.3505	-0.11210	0.1390
difa[36]	0.1419	0.12290	0.002250	-0.08962	0.13840	0.3894
difa[37]	-0.08644	0.10010	0.001574	-0.2781	-0.08710	0.1135
difa[38]	0.04689	0.11550	0.002294	-0.1711	0.04401	0.2791
difa[39]	0.1827	0.10680	0.001914	-0.01977	0.18060	0.3993
difa[40]	0.09759	0.07353	0.002411	-0.04192	0.09635	0.2448

TABLE 21:
Results of R^2 based Effect Size of Logistic Regression of Ethnic Example

Item	R_1^2	R_2^2	R_3^2	R_4^2	R_5^2	$R_2^2 - R_1^2$ (DIF on γ)	$R_3^2 - R_2^2$ (Uniform DIF)	$R_4^2 - R_3^2$ (Non uniform DIF)	$R_5^2 - R_3^2$ (Non uniform DIF)
Testlet A						0.1286			
1	0.7695	0.9344	0.9345	0.9956	1.0000	0.1649	**1E-04**	**0.0611**	**0.0655**
2	0.8893	0.9938	0.9979	0.9999	1.0000	0.1045	0.0041	0.0020	0.0021
3	0.8799	0.993	0.9958	0.9997	1.0000	0.1131	0.0028	0.0039	0.0042
4	0.6991	0.9785	0.9942	0.9996	1.0000	0.2794	**0.0157**	0.0054	0.0058
5	0.8612	0.9996	1.0000	1.0000	1.0000	0.1384	0.0004	0.0000	0.0000
6	0.8491	0.9992	0.9993	1.0000	1.0000	0.1501	1E-04	0.0007	0.0007
7	0.8924	0.9955	0.9995	1.0000	1.0000	0.1031	0.0040	0.0005	0.0005
8	0.9272	0.9807	0.9999	1.0000	1.0000	0.0535	**0.0192**	1E-04	1E-04
9	0.9186	0.9807	0.9951	0.9997	1.0000	0.0621	**0.0144**	0.0046	0.0049
10	0.8810	0.9978	1.0000	1.0000	1.0000	0.1168	0.0022	0.0000	0.0000
Testlet B						0.0731			
11	0.9291	0.9993	0.9994	1.0000	1.0000	0.0702	1E-04	0.0006	0.0006
12	0.8668	0.9745	0.9789	0.9995	1.0000	0.1077	**0.0044**	**0.0206**	**0.0211**
13	0.9442	0.9992	0.9995	1.0000	1.0000	0.0550	0.0003	0.0005	0.0005
14	0.9384	0.9956	0.9958	0.9999	1.0000	0.0572	0.0002	0.0041	0.0042
15	0.8763	0.9906	0.9963	0.9999	1.0000	0.1143	0.0057	0.0036	0.0037
16	0.9046	0.9700	0.9700	0.9992	1.0000	0.0654	0.0000	**0.0292**	**0.0300**
17	0.9305	0.9935	0.9935	0.9998	1.0000	0.0630	0.0000	0.0063	0.0065
18	0.9378	0.9941	0.9943	0.9999	1.0000	0.0563	0.0002	0.0056	0.0057
19	0.9194	0.9980	0.9985	1.0000	1.0000	0.0786	0.0005	0.0015	0.0015
20	0.9242	0.9872	0.9873	0.9997	1.0000	0.0630	1E-04	0.0124	0.0127
Testlet C						0.0676			
21	0.9223	0.9998	1.0000	1.0000	1.0000	0.0775	0.0002	0.0000	0.0000
22	0.9112	0.9948	0.9966	0.9998	1.0000	0.0836	0.0018	0.0032	0.0034
23	0.932	0.994	0.9952	0.9998	1.0000	0.0620	0.0012	0.0046	0.0048
24	0.9434	0.9981	0.9996	1.0000	1.0000	0.0547	0.0015	0.0004	0.0004
25	0.8976	0.9828	0.9866	0.9994	1.0000	0.0852	0.0038	**0.0128**	**0.0134**
26	0.9493	0.9949	0.9991	1.0000	1.0000	0.0456	**0.0042**	0.0009	0.0009
27	0.9399	0.9988	0.9992	1.0000	1.0000	0.0589	0.0004	0.0008	0.0008
28	0.9475	0.9844	0.9870	0.9994	1.0000	0.0369	0.0026	**0.0124**	**0.0130**
29	0.8925	0.9959	1.0000	1.0000	1.0000	0.1034	0.0041	0.0000	0.0000
30	0.9322	1.0000	1.0000	1.0000	1.0000	0.0678	0.0000	0.0000	0.0000
Testlet D						0.3104			
31	0.6762	0.9937	1.0000	1.0000	1.0000	0.3175	0.0063	0.0000	0.0000
32	0.6913	0.9991	1.0000	1.0000	1.0000	0.3078	0.0009	0.0000	0.0000
33	0.6846	0.9971	1.0000	1.0000	1.0000	0.3125	0.0029	0.0000	0.0000
34	0.6102	0.9480	1.0000	1.0000	1.0000	0.3378	**0.0520**	0.0000	0.0000
35	0.6750	0.9931	1.0000	1.0000	1.0000	0.3181	0.0069	0.0000	0.0000
36	0.6791	0.9949	1.0000	1.0000	1.0000	0.3158	0.0051	0.0000	0.0000
37	0.6870	0.9979	1.0000	1.0000	1.0000	0.3109	0.0021	0.0000	0.0000
38	0.6762	0.9937	1.0000	1.0000	1.0000	0.3175	0.0063	0.0000	0.0000
39	0.7012	0.9999	1.0000	1.0000	1.0000	0.2987	1E-04	0.0000	0.0000
40	0.7166	0.9836	1.0000	1.0000	1.0000	0.2670	0.0164	0.0000	0.0000

TABLE 22:
Regression Coefficients of Ethnic Example

Item	τ_0	τ_1 (for θ)	τ_2 (for γ)	τ_3 (for G)	τ_4 (for θ*G)	τ_5 (for γ*G)
Testlet A						
1	0.1582	0.9277	0.9277	**0.4341**	**-0.3989**	**-0.3989**
2	0.5182	1.1660	1.1660	-0.1436	-0.1120	-0.1120
3	0.1032	1.2640	1.2640	-0.0575	-0.1670	-0.1670
4	-0.1915	0.7838	0.7838	**0.4973**	-0.1328	-0.1328
5	-0.3219	0.9614	0.9614	-0.0675	-0.0042	-0.0042
6	-0.3436	0.6376	0.6376	0.0214	-0.0373	-0.0373
7	0.4481	0.5314	0.5314	-0.1567	0.0279	0.0279
8	-0.1646	1.0090	1.0090	**-0.4801**	-0.0200	-0.0200
9	-0.3606	0.6624	0.6624	**-0.4419**	0.1152	0.1152
10	0.2264	0.7926	0.7926	-0.1323	-0.0038	-0.0038
Testlet B						
11	0.7425	0.7832	0.7832	0.0114	0.0450	0.0450
12	1.1536	0.8628	0.8628	**0.1473**	**0.3832**	**0.3832**
13	0.8185	1.2330	1.2330	-0.0978	0.0650	0.0650
14	0.1194	0.9912	0.9912	-0.1026	0.1598	0.1598
15	0.6297	0.7544	0.7544	0.1849	0.1180	0.1180
16	-0.3905	1.0180	1.0180	0.1195	**-0.3142**	**-0.3142**
17	-0.4614	0.9108	0.9108	-0.0723	0.1892	0.1892
18	-0.2179	0.9894	0.9894	-0.1196	0.1896	0.1896
19	-0.7236	0.8141	0.8141	0.0436	0.0762	0.0762
20	-1.1868	0.6233	0.6233	-0.0695	0.1935	0.1935
Testlet C						
21	0.8345	1.0780	1.0780	0.0496	-0.0140	-0.0140
22	0.8883	1.1600	1.1600	0.1011	0.1670	0.1670
23	0.7988	0.8626	0.8626	-0.0347	-0.1193	-0.1193
24	0.4392	1.2060	1.2060	-0.1222	-0.0490	-0.0490
25	-0.0132	0.9633	0.9633	0.1042	**0.3077**	**0.3077**
26	-0.2456	1.3680	1.3680	**-0.2221**	-0.0870	-0.0870
27	-0.1793	1.3060	1.3060	-0.1098	0.0880	0.0880
28	-0.8026	1.0460	1.0460	-0.3238	**0.3180**	**0.3180**
29	-0.7060	0.7390	0.7390	0.1495	-0.0022	-0.0022
30	-0.8187	1.0540	1.0540	-0.0175	-0.0040	-0.0040
Testlet D						
31	0.4900	0.8478	0.8478	0.1901	0.0000	0.0000
32	0.4808	1.2530	1.2530	0.1074	0.0000	0.0000
33	0.8675	0.8996	0.8996	0.1353	0.0000	0.0000
34	0.5977	0.8099	0.8099	**0.5478**	0.0000	0.0000
35	-0.1642	0.7828	0.7828	0.1833	0.0000	0.0000
36	0.3216	0.7938	0.7938	0.1586	0.0000	0.0000
37	0.3009	0.5255	0.5255	0.0675	0.0000	0.0000
38	-0.4614	0.7463	0.7463	0.1671	0.0000	0.0000
39	-0.1756	0.6650	0.6650	-0.0170	0.0000	0.0000
40	-0.8443	0.3311	0.3311	-0.1161	0.0000	0.0000

III. Phenomena of DIF Amplification and Cancellation at Item and Testlet Levels

The evidence of DIF amplification and cancellation at item and testlet levels has been investigated using signed-area / unsigned-area indices by calculating the areas between item characteristic curves of two subgroups of each item and the areas between testlet characteristic curves of two subgroups of each testlet. Table 23 and Table 24 listed the results of the indices of the two examples. The item characteristic curve of each item and testlet characteristic curve of each testlet was provided in the Appendix D and E.

TABLE 23:
Results of Signed-area/ Unsigned-area Indices of Gender Example

Item	Signed-Area	Unsigned-Area
Testlet A		
1	0.1392	0.0948
2	0.2456	0.1361
3	0.1099	0.1125
4	0.4059	0.1901
5	0.2429	0.1215
6	-0.0745	0.0942
7	0.2508	0.2453
8	0.3999	0.1877
9	0.0244	0.0341
10	0.2959	0.1993
Testlet Level	2.0401	1.1656
Testlet B		
11	-0.0324	0.0426
12	-0.0442	0.0916
13	0.1203	0.0730
14	-0.3714	0.1803
15	0.0500	0.0533
16	0.1784	0.0827
17	0.0768	0.0699
18	-0.3051	0.1643
19	0.0542	0.0365
20	-0.6327	0.3168
Testlet Level	-0.9063	0.4344
Testlet C		
21	0.0928	0.0494
22	0.3463	0.1822
23	0.4970	0.2458
24	0.1050	0.0679
25	-0.0105	0.0309
26	-0.1109	0.1264
27	0.0604	0.0344
28	0.0258	0.0147
29	**0.0714**	**0.1062**
30	0.2726	0.1321
Testlet Level	1.3498	0.6608
Testlet D		
31	**-0.0164**	**0.0140**
32	**0.2865**	**0.1476**
33	**0.0563**	**0.0315**
34	**0.0465**	**0.0270**
35	**0.0488**	**0.0253**
36	**0.0342**	**0.0190**
37	**-0.0971**	**0.0451**
38	**0.0707**	**0.0331**
39	**-0.1441**	**0.0673**
40	**-0.0215**	**0.0117**
Testlet Level	**0.2640**	**0.1872**

TABLE 24
Results of Signed-area/ Unsigned-area Indices of Ethnic Example

Item	Signed-Area	Unsigned-Area
Testlet A		
1	0.5777	0.3907
2	0.4019	0.2442
3	0.4706	0.2801
4	**0.7713**	**0.3982**
5	0.4754	0.2396
6	**0.4126**	**0.1986**
7	0.2246	0.1066
8	**0.1742**	**0.0934**
9	0.1782	0.0958
10	0.3464	0.1770
Testlet Level	4.0331	2.1227
Testlet B		
11	0.2703	0.1312
12	0.2591	0.1446
13	0.2637	0.1394
14	0.2767	0.1425
15	0.4012	0.1923
16	0.2995	0.2024
17	0.3509	0.1874
18	0.3016	0.1611
19	0.3898	0.1933
20	0.3506	0.2138
Testlet Level	3.1635	1.5206
Testlet C		
21	0.3605	0.1972
22	0.3623	0.1915
23	0.2700	0.1686
24	0.2670	0.1485
25	0.4302	0.2248
26	0.1881	0.1098
27	0.2938	0.1553
28	0.2412	0.1530
29	0.3935	0.1859
30	0.2899	0.1442
Testlet Level	3.0965	1.5163
Testlet D		
31	**0.4164**	**0.2041**
32	**0.3573**	**0.1945**
33	**0.3703**	**0.1846**
34	**0.6922**	**0.3381**
35	**0.4090**	**0.2021**
36	**0.3920**	**0.1904**
37	**0.2781**	**0.1299**
38	**0.3851**	**0.1936**
39	**0.2268**	**0.1114**
40	**0.0227**	**0.0132**
Testlet Level	**3.5499**	**1.7188**

1. DIF Amplification and Cancellation at the item level

Evidence of phenomena of DIF amplification at the item level has been found on one example of Item 4 nested within the first testlet of ethnic sample and phenomena of DIF cancellation at the item level has been found on one example of Item 8 of the same testlet of the same sample. As to this testlet, the mean difference of testlet effect between the minorities and Caucasians was about -0.7421. Taking a criteria item, Item 6, to make comparison, Item 6 reflected the DIF attributed only to the testlet effect, where there were no significantly statistical differences on item difficulty parameter and item discrimination parameter between two subgroups (see Figure 8). Then obviously, referring Item 4 to reflect DIF amplification at the item level, there were larger areas between the two ICCs because other than the mean difference of testlet effect, the difference on the item difficulty parameters between the two subgroups was 0.7140 and Caucasians were favored (See Figure 10). Referring Item 8 to reflect DIF cancellation at the item level, there were smaller areas between the two ICCs because other than the mean difference of testlet effect with Caucasians have higher abilities on testlet dimension, the difference on the item difficulty parameters between the two subgroups was 0.4887 and Minority group was favored (See Figure 11). The magnitudes of DIF of these three items measured by signed-area and unsigned-area indices were shown in Table 24. The other kind of DIF cancellation at the item level because of crossing of ICCs has been detected on Item 29 of the gender example (See Figure 9). The reason for DIF cancellation at the item level was because of the small difference of the item discrimination parameters between females and males groups. Since females have about 0.2 higher on the means of testlet distribution than males, the ICC of Females group

shifted a little bit to the left and thus two ICCs crossed at the lower left corner. The signed-area and unsigned-area indices of this item were 0.0714 and 0.1062. DIF could not be easily detected by signed-area index but still could by unsigned-area index (See Table 23).

2. DIF Amplification and Cancellation at the testlet level

Evidence of DIF amplification and cancellation at the testlet level has been found on examples of the last testlet of the two samples. Regarding the gender sample, although the testlet effect functioned approximately homogeneously between females and males, nearly half of the items nested within the testlet slightly favored males group (See Figure 13 as an example) and nearly half of them functioned on the opposite way (See Figure 12 as an example), and thus the cumulative effect of DIF cancelled out at the testlet level (See Figure 14). See Table 23 for magnitudes of DIF of those 10 items and the DIF at the testlet level. Regarding the ethnic sample, on the other hand, although item attributes functioned similarly between the minority group and Caucasian group (See Table 24 for evidence), the mean difference of the testlet distribution between the two subgroups was about 0.4323. Therefore, the cumulated effect of DIF amplified at the testlet level, albeit the small amount of DIF found on each item within the testlet (See Figure 12 and Figure 16 as two examples).

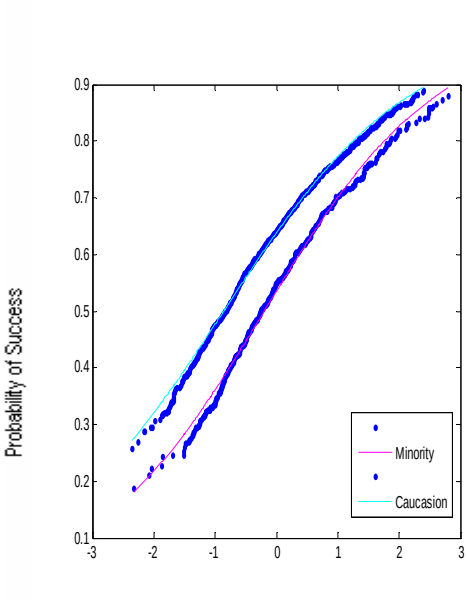


FIGURE 8: ICC of Item 6 of Ethnic Sample

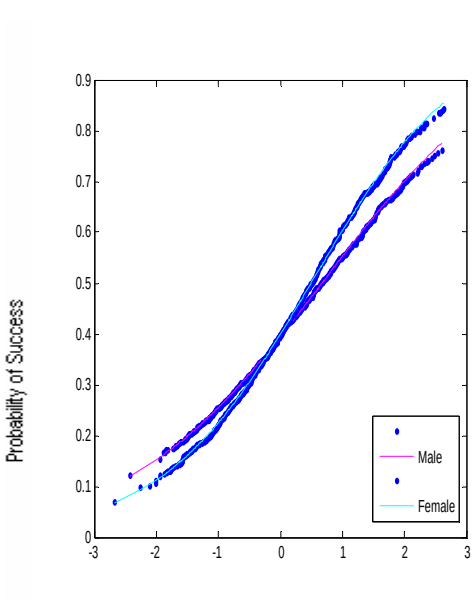


FIGURE 9: ICC of Item 29 of Gender Sample

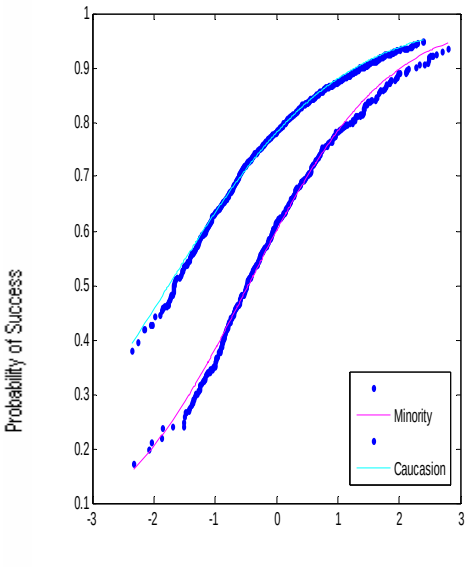


FIGURE 10: ICC of Item 4 of Ethnic Sample

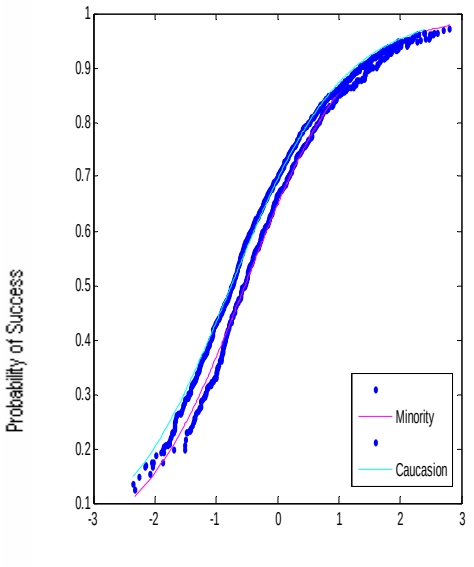


FIGURE 11: ICC of Item 8 of Ethnic Sample

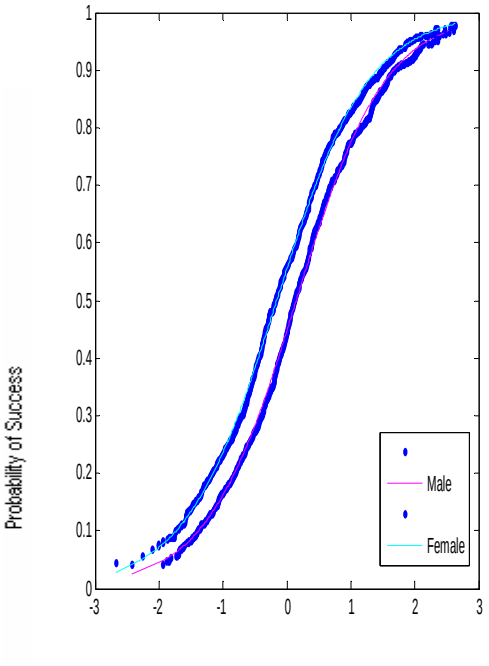


FIGURE 12: ICC of Item 32 of Gender Example

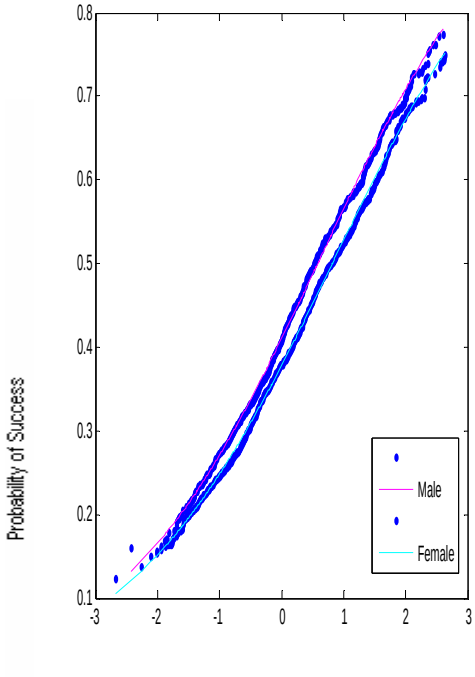


FIGURE 13: ICC of Item 39 of Gender Example

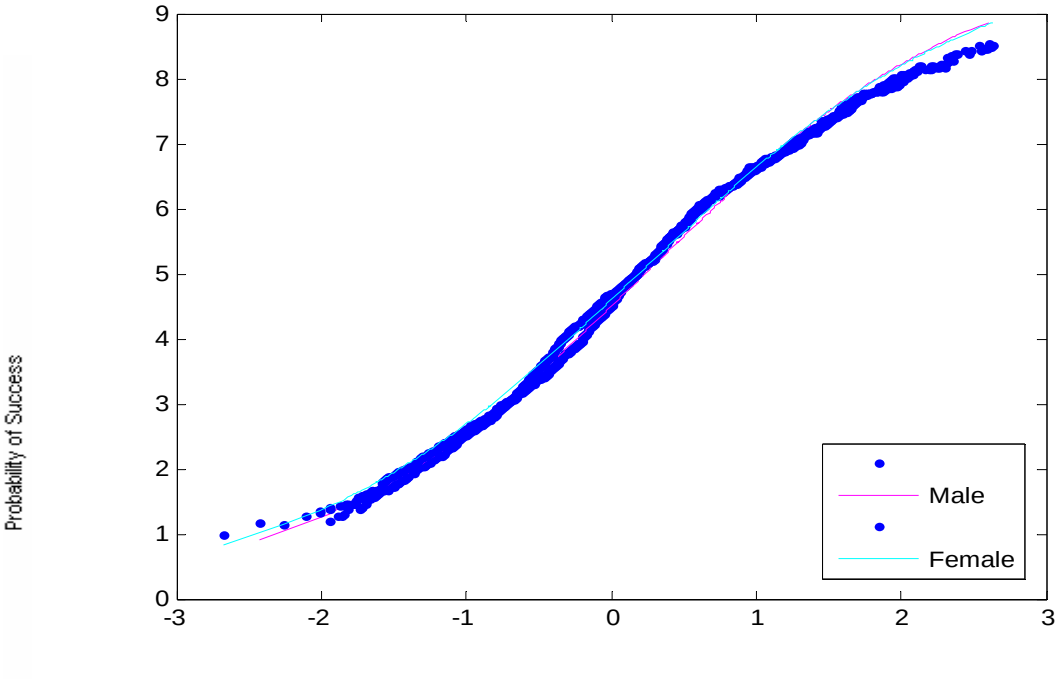


FIGURE 14: TCC of Testlet D of Gender Example

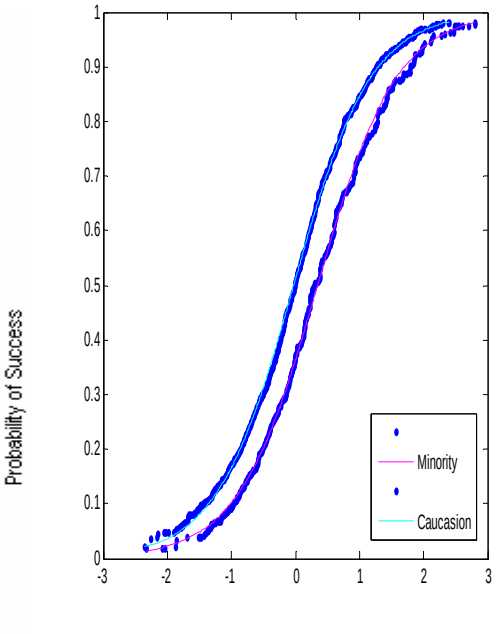


FIGURE 15: ICC of Item 32 of Ethnic Example

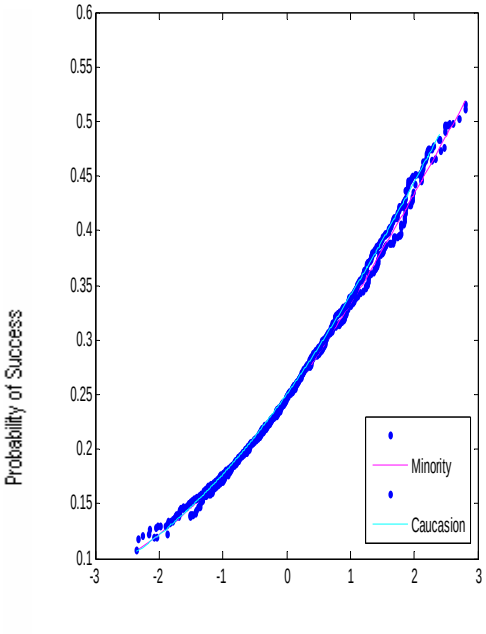


FIGURE 16: ICC of Item 40 of Ethnic Example

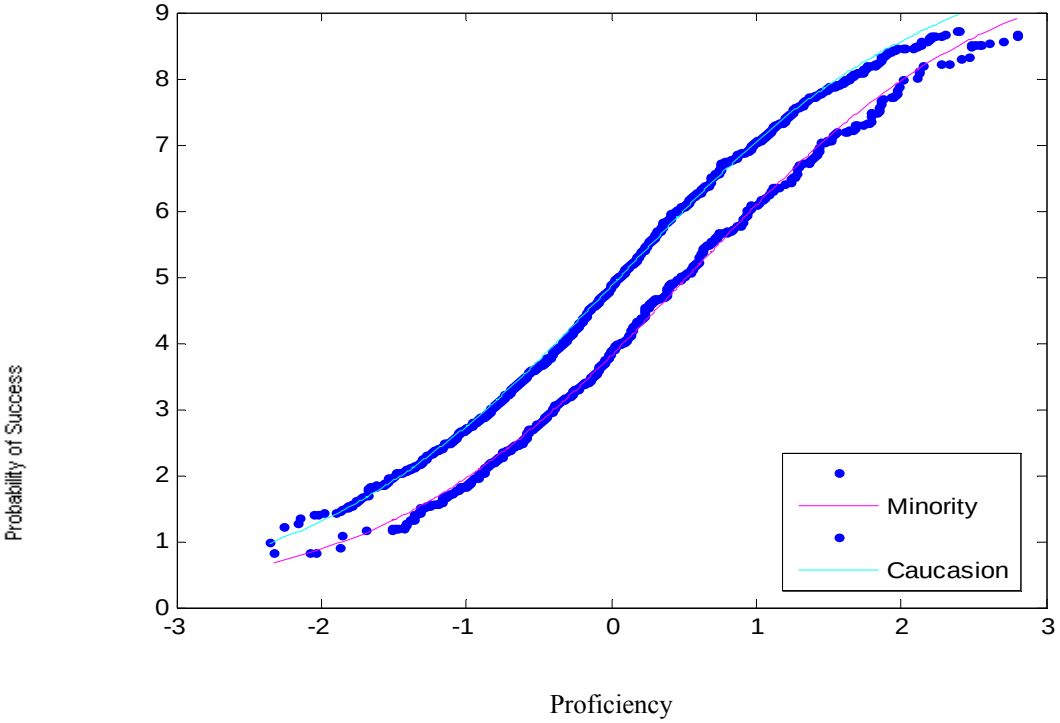


FIGURE 17: TCC of Testlet D of Ethnic Example

Summary

In summary, analyses of the simulated data and real data obtained from ACT reading test revealed that the person-testlet interaction effect did exist and the phenomena of DIF amplification and cancellation were attributed to comprehensive DIF effects of testlet distributions and idiosyncratic features of items within testlet. All of the possible situations of DIF due to the two factors were enumerated and a total of seven combination results of DIF amplification and cancellation at the item and testlet level were summarized in the simulation study. As indicated by the results of real data analysis, the magnitudes of person-testlet interaction effects, embodied on the means and/ or variances, were not the same, and they seemed to be attributed to the different contexts or natures of the passages as well as its interaction with the manifest groups of examinees such as gender or ethnicity. The effect was found to be larger in the passage about Natural Science than the other three topics related to Prose Fiction, Social Science, and Humanities. Additionally, larger magnitude of difference on the testlet effect was also found in ethnic example than that in gender example. The phenomena of DIF amplification and cancellation as examples of situations in the simulation study were also found in the real data analysis.

Chapter 5 Conclusion and Discussion

The focus of this study was to investigate DIF amplification and cancellation at the individual item level and testlet level. Based on simulation as well as real data analysis, logistic regression procedure and signed-area and unsigned-area indices on item response theory framework demonstrated their effectiveness to assess DIF at two levels. The signed-area and unsigned-area indices was useful to provide a magnitudes measure of DIF and the logistic regression was useful for identifying items with DIF and also for explaining the sources of the DIF. As demonstrated, at either item level or testlet level or both, the cumulative effect of DIF could either amplify or cancel out partially or completely.

The work conducted in this research used the advantages of the multiple-group item response testlet model proposed by Li, etal to investigate the sources of the DIF and the reason of DIF amplification and cancellation at the two levels. In this study, we used a Bayesian estimation method implemented by WinBUGS 1.4 software.

The results obtained from the simulation study and the analysis of real data led to the following conclusions:

- First, in general, the homogeneous functioning of testlet effect and item difficulty parameters between the two subgroups was the reason for DIF amplification at the item level. On the contrary, the heterogeneous functioning of the testlet effect and item difficulty parameters between the two subgroups was the reason for DIF cancellation at the item level. More usual reason for DIF cancellation at the item level was because of the different item discrimination parameters leading to the crossing of ICCs of two subgroups;

- Second, the reason for the DIF amplification at the testlet level was usually due to the existence of testlet effect and the reason for DIF cancellation at the testlet level was because of the heterogeneous functioning of individual items nested within the testlet;
- The difference in the variances of the testlet distributions between the reference group and focal group seems to have as significant influence as the difference on the means of testlet distributions on the phenomena of DIF amplification and cancellation;
- The person-testlet interaction effect did exist in real ACT test data. The magnitude of this effect varied from examination to examination and from testlet to testlet, depending on the nature of the test items included in the testlets and on the nature of the population to which the test was administered.

Roznowski (1988) has raised the issue that, because decisions are made at a level higher than the item, the study of DIF at the item level may only have limited importance. Since many current assessment are made up of testlets, it is very likely impossible to ignore its multidimensional nature. It is sensible to consider an aggregate measure of DIF at the testlet level by considering the interactive influence of testlet effect and the characteristic features of individual items within the testlet. DIF cancellation at the item and testlet level, under this argument, provided a graceful solution to yield a set of DIF-balanced test construction unit. However, it is hard to say whether or not it is beneficial for large-scale testing organizations to look for DIF and not find any due to the possibility of cancellation at the testlet level even though it really did exist at individual item level. Fortunately, at least at the testlet level, the multiple group testlet models could give us clues to locate the source of DIF. DIF amplification at item and testlet level, under this argument, provided a useful tool to ensure fairness through the increased

statistical power of detecting DIF for relatively rare focal groups in the examinee population. However, it was still important to assess whether the statistically significant amount of DIF present was of practical importance and also enough sample size was still necessary to ensure the power of certain statistical methods of detecting DIF.

Ideally speaking, to accomplish credibility of the DIF study, findings of DIF must be accompanied by a careful study with as large a sample size as could be found and it must also use the most efficient statistical model available to analyze data. One underlying assumption always exists albeit often overlooked was that IRT model that was assumed to underlie the individual item responses was appropriate. Fortunately, we considered these arguments in our development of the methodology presented here. The samples we have used were realistic for most practical situations leading to reliable detection of DIF and also appropriate to obtain reliable results from MCMC estimation of item response testlet models. Nonetheless, more elegant testlet models with different item discrimination parameters and with covariance to capture dependence between a set of testlets in the test would be useful and interesting for the future study. Moreover, although manifest groups such as genders and racial groups have been easily identified to be used in the traditional DIF study, regarding to the issues such as, the lack of homogeneity in manifest groups and possibilities that the groups being examined are not really the manifest groups affected, etc, a latent class approach using latent grouping variables to allow for the assessment of DIF without tying that DIF to any specific variable and set of variables could be possible to make a more definitive statement for investigation of the presence DIF.

Appendix A: Means and Standard Deviances of Estimates of Item Parameters of Gender Example and Ethnic Example

TABLE 1: Means and Standard Deviations of Estimates of Item Difficulty Parameters of Gender Example

Item	Mean of Item Difficulty Parameters for Male Group	Standard Deviation of Item Difficulty Parameters for Male Group	Mean of Item Difficulty Parameters for Female Group	Standard Deviation of Item Difficulty Parameters for Female Group
Testlet A				
1	-0.5161	0.12550	-0.4134	0.15940
2	-0.5890	0.11150	-0.5185	0.14010
3	-0.09998	0.07479	0.1163	0.09884
4	-0.05176	0.11000	-0.2126	0.14320
5	0.2904	0.07019	0.4465	0.08160
6	0.01532	0.13210	0.3583	0.17210
7	-0.2907	0.12170	**-1.1740**	**0.30970**
8	0.6120	0.07594	0.5265	0.08882
9	0.6782	0.08935	1.0840	0.08971
10	-0.04829	0.08491	-0.2124	0.14280
Testlet B				
11	-0.9352	0.12890	-0.9512	0.10620
12	-1.1690	0.12020	-1.1220	0.08719
13	-0.3497	0.06067	-0.6597	0.06683
14	-0.3467	0.07250	-0.0387	0.05915
15	-0.7002	0.09875	-0.8353	0.08386
16	0.6824	0.07401	0.2790	0.06691
17	0.5795	0.05720	0.3228	0.05745
18	0.1750	0.05560	0.3677	0.05949
19	0.8755	0.06841	0.6463	0.06523
20	1.1880	0.08681	1.9910	0.14150
Testlet C				
21	-0.8238	0.06788	-0.7435	0.07132
22	-0.6178	0.05956	-0.8531	0.07438
23	-0.6140	0.08060	-1.1970	0.12570
24	-0.3291	0.05144	-0.2739	0.05569
25	-0.2395	0.05712	-0.0224	0.05310
26	0.1489	0.04741	0.4717	0.05751
27	0.05803	0.04786	0.1821	0.04622
28	0.7026	0.06052	0.8626	0.06009
29	0.8582	0.10800	0.8574	0.08134
30	0.8565	0.07283	0.7166	0.06647
Testlet D				
31	-0.7207	0.10260	-0.6339	0.08759
32	-0.3953	0.09560	-0.6798	0.08003
33	-1.0870	0.10660	-0.9189	0.08464
34	-0.9725	0.10310	-1.0640	0.09342
35	0.09306	0.10600	0.08094	0.08942
36	-0.5583	0.12400	-0.4378	0.09621
37	-0.6990	0.12550	-0.3619	0.09737
38	0.5135	0.13210	0.4337	0.11190
39	0.1668	0.14520	0.5576	0.15440
40	3.6590	0.66740	3.0240	0.49820

TABLE 2:
Means and Standard Deviations of Estimates of Item Discrimination Parameters of Gender Example

Item	Mean of Item Discriminate Parameters for Male Group	Standard Deviation of Item Discriminate Parameters for Male Group	Mean of Item Discriminate Parameters for Female Group	Standard Deviation of Item Discriminate Parameters for Female Group
Testlet A				
1	0.8203	0.08408	0.7406	0.07730
2	1.0280	0.10070	1.0440	0.10400
3	1.2570	0.11150	1.0400	0.09446
4	0.6934	0.07176	0.7550	0.07766
5	1.0440	0.08901	1.1830	0.09952
6	0.5339	0.06385	0.4094	0.05574
7	0.7352	0.07702	0.4153	0.05641
8	0.8248	0.07502	0.9153	0.08060
9	0.6487	0.06619	0.7257	0.06873
10	0.9887	0.08990	0.7813	0.08031
Testlet B				
11	0.7390	0.07744	0.8285	0.07878
12	1.0500	0.10570	1.3530	0.12240
13	1.3350	0.11230	1.2670	0.10510
14	1.0140	0.08911	1.0150	0.08318
15	0.8670	0.08245	1.0210	0.08820
16	0.8107	0.07322	0.8198	0.07091
17	1.1700	0.09052	1.0250	0.08078
18	1.1750	0.09309	0.9792	0.07807
19	0.9883	0.08000	0.9377	0.07583
20	0.8513	0.07360	0.7710	0.07148
Testlet C				
21	1.1970	0.10050	1.2000	0.10030
22	1.2090	0.09808	1.2200	0.10200
23	0.8118	0.07538	0.7737	0.07540
24	1.2720	0.09965	1.1670	0.09070
25	1.0620	0.08539	1.1510	0.08937
26	1.3910	0.10230	1.0410	0.08058
27	1.3320	0.10000	1.3910	0.10260
28	1.2080	0.09087	1.2260	0.09092
29	0.5845	0.06061	0.7558	0.06555
30	1.0090	0.08037	0.9201	0.07352
Testlet D				
31	0.8779	0.07947	0.8609	0.07573
32	1.0990	0.09278	1.1540	0.09307
33	0.9280	0.08367	1.0950	0.0934
34	0.9848	0.08831	0.8848	0.07627
35	0.8921	0.07680	0.9081	0.07587
36	0.5423	0.05832	0.6787	0.06333
37	0.5355	0.05770	0.6623	0.06151
38	0.6625	0.06307	0.6931	0.06400
39	0.4835	0.05409	0.4518	0.05037
40	0.2489	0.04199	0.3138	0.04713

TABLE 3:
Means and Standard Deviations of Estimates of Item Difficulty Parameters of Ethnic Example

Item	Mean of Item Difficulty Parameters for Group	Standard Deviation of Item Difficulty Parameters for Male Group	Mean of Item Difficulty Parameters for Female Group	Standard Deviation of Item Difficulty Parameters for Group
Testlet A				
1	-0.1705	0.12090	**-1.1200**	**0.27690**
2	-0.4444	0.12100	-0.3554	0.12810
3	-0.08163	0.09650	-0.04168	0.10260
4	0.2443	0.11520	-0.4698	0.18020
5	0.3348	0.09929	0.4068	0.09257
6	0.5389	0.12850	0.5368	0.11940
7	-0.8433	0.24480	-0.5211	0.20740
8	0.1631	0.10020	0.6518	0.08545
9	0.5444	0.12500	1.0320	0.08974
10	-0.2856	0.14560	-0.1193	0.13700
Testlet B				
11	-0.9480	0.15890	-0.9102	0.11920
12	-1.3370	0.18940	-1.0440	0.10090
13	-0.6638	0.10200	-0.5552	0.07259
14	-0.1205	0.09510	-0.0146	0.05907
15	-0.8347	0.15380	-0.9337	0.11860
16	0.3836	0.09624	0.3850	0.07544
17	0.5066	0.10670	0.4852	0.05562
18	0.2202	0.09527	0.2862	0.05485
19	0.8888	0.13010	0.7637	0.06565
20	1.9040	0.26000	1.5380	0.09314
Testlet C				
21	-0.7741	0.09559	-0.8309	0.08662
22	-0.7658	0.09000	-0.7456	0.07287
23	-0.9260	0.11930	-1.0280	0.12970
24	-0.3642	0.07803	-0.2740	0.06035
25	0.01373	0.09314	-0.07156	0.05263
26	0.1795	0.07742	0.3651	0.04878
27	0.1373	0.07850	0.2074	0.04626
28	0.7673	0.11380	0.8258	0.05148
29	0.9553	0.15700	0.7553	0.07369
30	0.7768	0.11520	0.7964	0.05908
Testlet D				
31	-0.5780	0.13040	-0.7415	0.09941
32	-0.3837	0.11370	-0.5559	0.09351
33	-0.9643	0.12860	-1.0640	0.10340
34	-0.7380	0.13400	-1.2120	0.10760
35	0.2098	0.15980	-0.02135	0.09931
36	-0.4051	0.13590	-0.7366	0.11240
37	-0.5726	0.17050	-0.6019	0.11380
38	0.6182	0.18430	0.4208	0.11810
39	0.2641	0.18050	0.3994	0.14380
40	2.5500	0.62310	4.1130	0.67290

TABLE 4:
Means and Standard Deviations of Estimates of Item Discrimination Parameters of Ethnic Example

Item	Mean of Item Discriminate Parameters for Group	Standard Deviation of Item Discriminate Parameters for Male Group	Mean of Item Discriminate Parameters for Female Group	Standard Deviation of Item Discriminate Parameters for Group
Testlet A				
1	0.9277	0.12550	0.5288	0.06585
2	1.1660	0.15830	1.0540	0.10200
3	1.2640	0.16350	1.0970	0.09806
4	0.7838	0.10830	0.6510	0.07006
5	0.9614	0.12440	0.9572	0.08345
6	0.6376	0.09580	0.6003	0.06321
7	0.5314	0.08694	0.5593	0.06462
8	1.0090	0.12870	0.9890	0.08507
9	0.6624	0.09965	0.7776	0.07078
10	0.7926	0.11390	0.7888	0.07790
Testlet C				
11	0.7832	0.11550	0.8282	0.07765
12	0.8628	0.13000	1.2460	0.11310
13	1.2330	0.16030	1.2980	0.10850
14	0.9912	0.12680	1.1510	0.08987
15	0.7544	0.11020	0.8724	0.08192
16	1.0180	0.12320	0.7038	0.06516
17	0.9108	0.11430	1.1000	0.08258
18	0.9894	0.12500	1.1790	0.08752
19	0.8141	0.10590	0.8903	0.07247
20	0.6233	0.09543	0.8168	0.06963
Testlet C				
21	1.0780	0.13590	1.0640	0.08820
22	1.1600	0.14470	1.3270	0.10910
23	0.8626	0.11730	0.7433	0.07251
24	1.2060	0.14250	1.1570	0.08955
25	0.9633	0.11770	1.2710	0.09300
26	1.3680	0.15330	1.2810	0.09312
27	1.3060	0.14690	1.3940	0.09899
28	1.0460	0.12100	1.3640	0.09550
29	0.7390	0.09729	0.7368	0.06448
30	1.0540	0.12200	1.0500	0.07905
Testlet D				
31	0.8478	0.11260	0.9172	0.07979
32	1.2530	0.15730	1.0580	0.08553
33	0.8996	0.12180	0.9425	0.08092
34	0.8099	0.10760	0.9451	0.08104
35	0.7828	0.10240	0.8936	0.07375
36	0.7938	0.10670	0.6519	0.06140
37	0.5255	0.08122	0.6120	0.05815
38	0.7463	0.09802	0.6994	0.06153
39	0.6650	0.09423	0.4823	0.05102
40	0.3311	0.06291	0.2335	0.03725

Appendix B: Item Characteristic Curves of Representative Items in Simulation Study of Gender Example

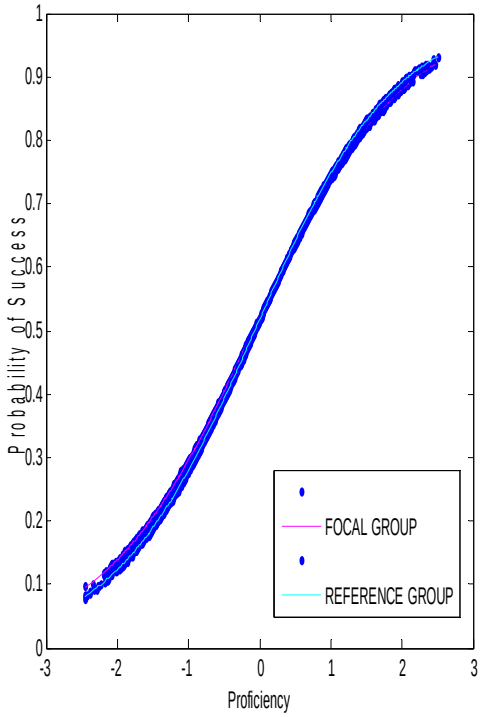


FIGURE 1: ICC of M2C1A

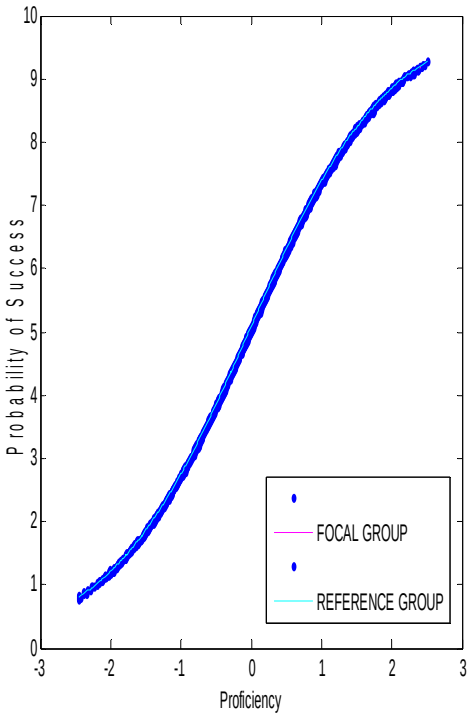


FIGURE 2: TCC of M2C1A

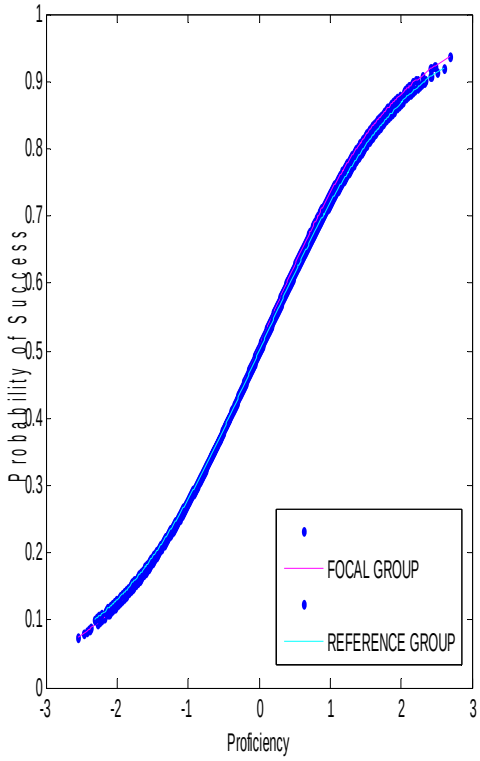


FIGURE 3: ICC of M1A

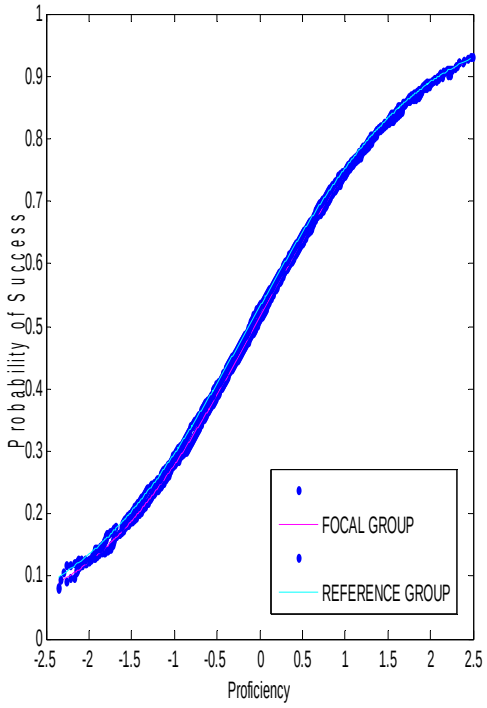


FIGURE 4: ICC of M2C2A

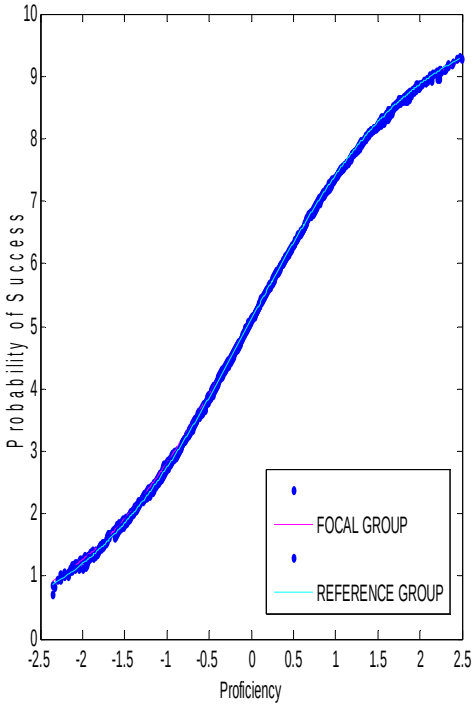


FIGURE 5: ICC of M2C2A

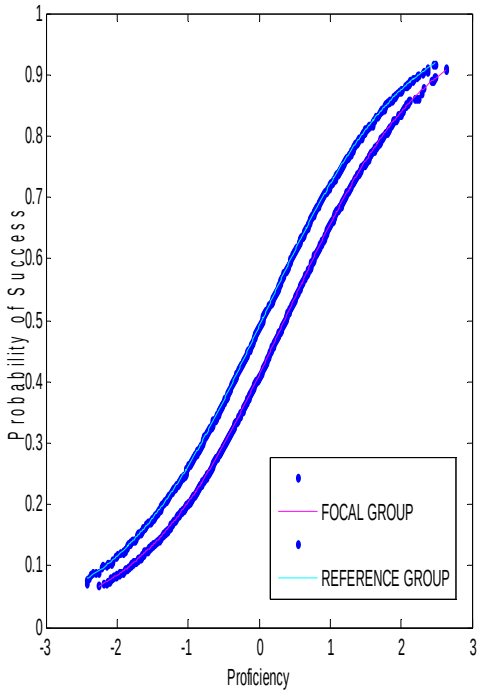


FIGURE 6: ICC of M3C1A

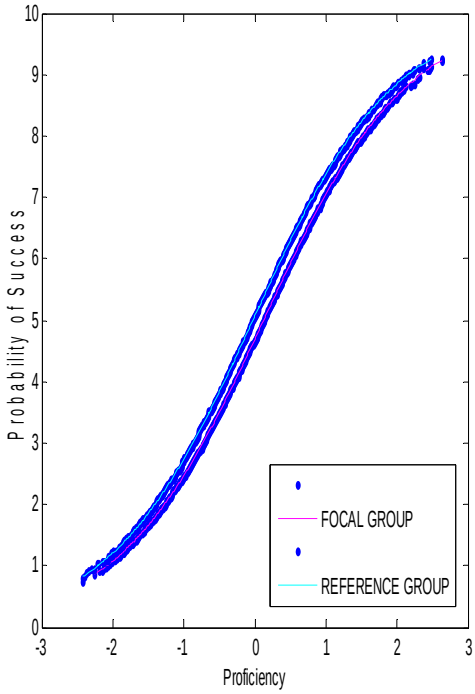
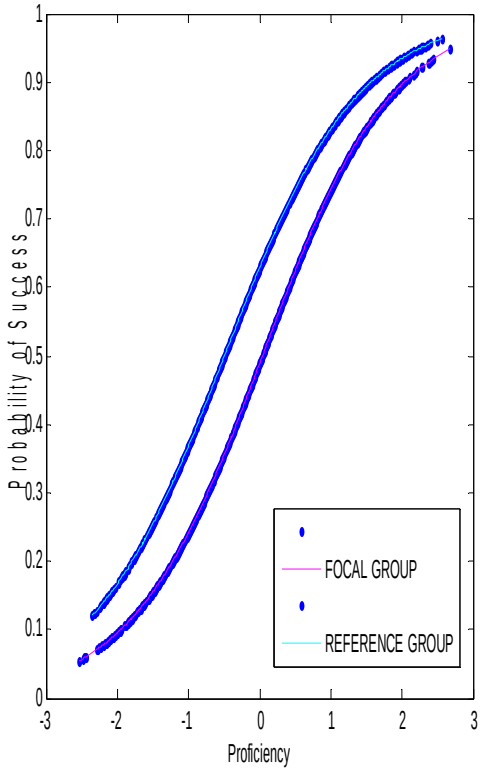


FIGURE 7: TCC of M2C1A



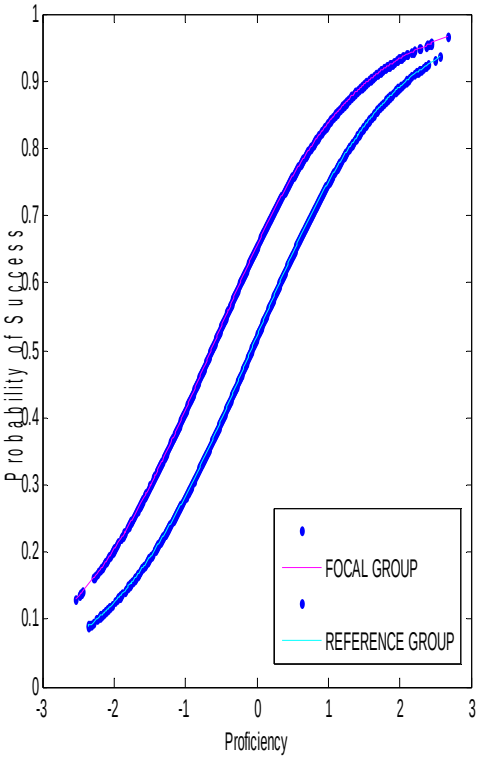


FIGURE 8: ICC of M1B1

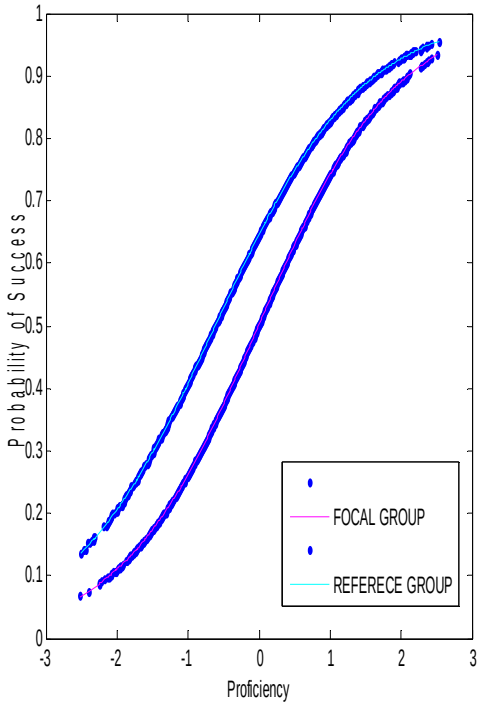


FIGURE 9: ICC of M1B1

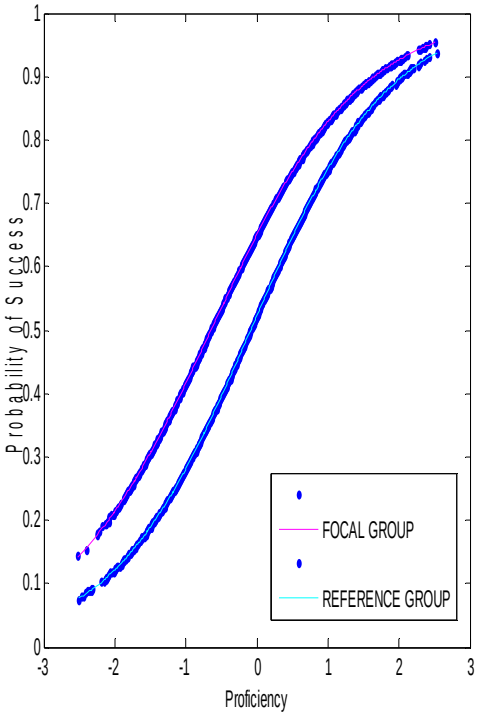


FIGURE 10: ICC of M2C1B1

FIGURE 11: ICC of M2C1B1

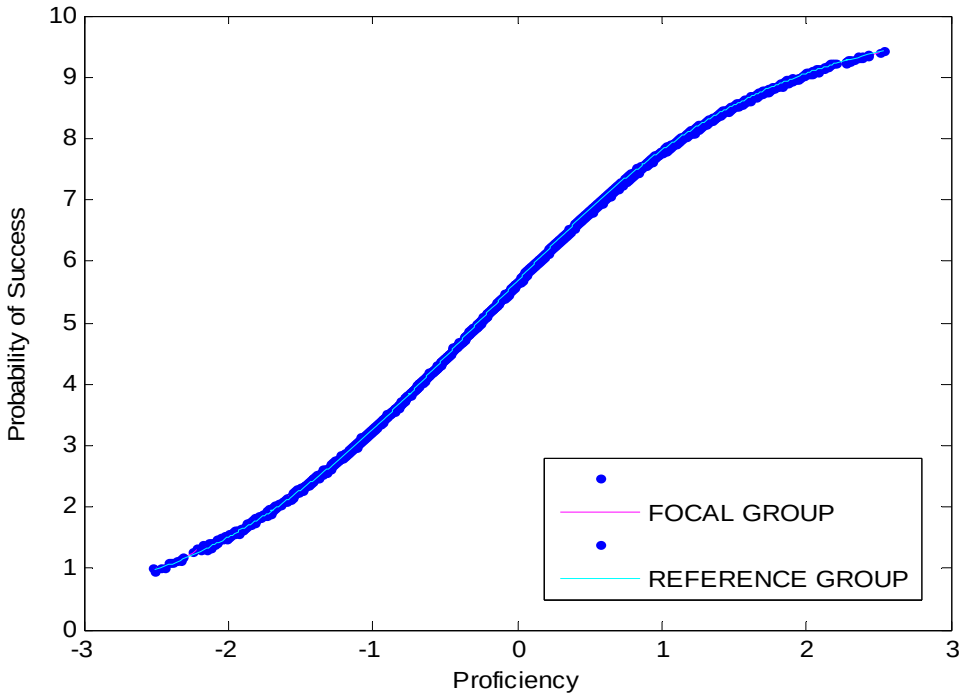


FIGURE 12: TCC of M2C1B1

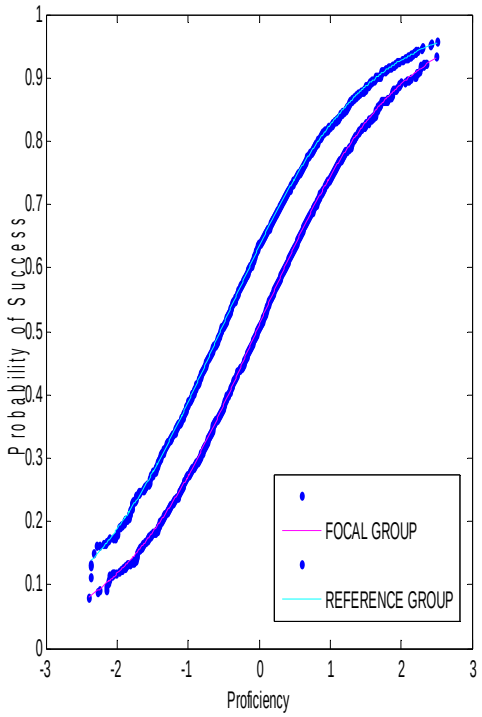


FIGURE 13: ICC of M2C2B1

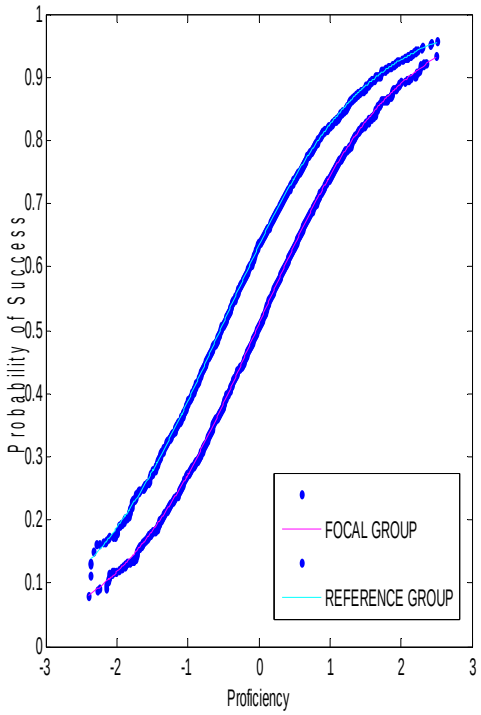


FIGURE 14: ICC of M2C2B1

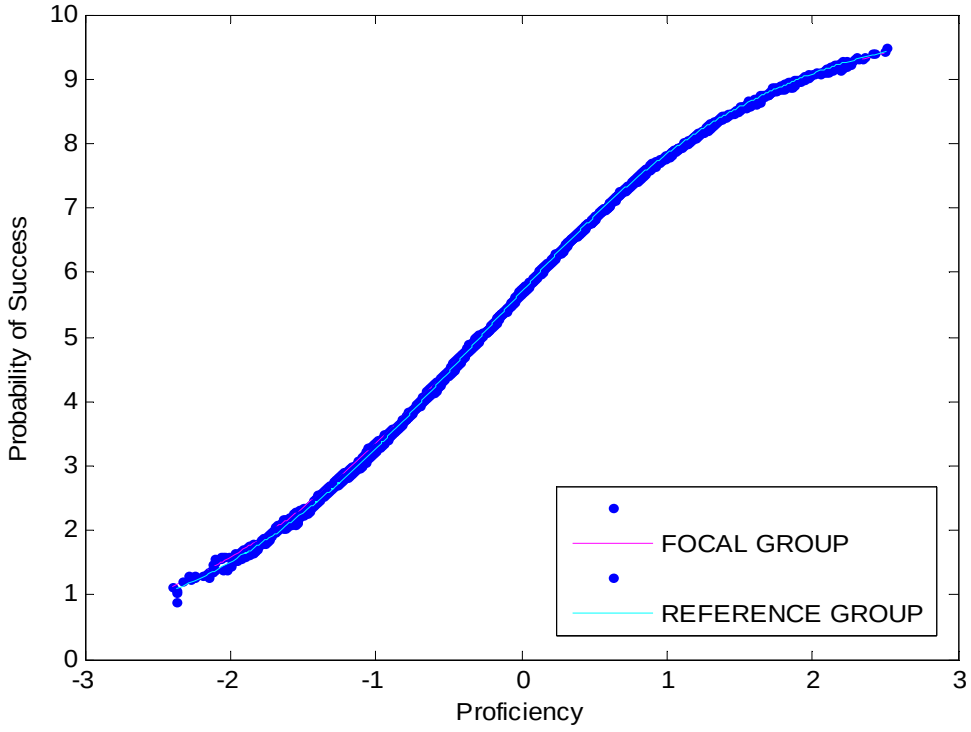


FIGURE 15: TCC of M2C2B1

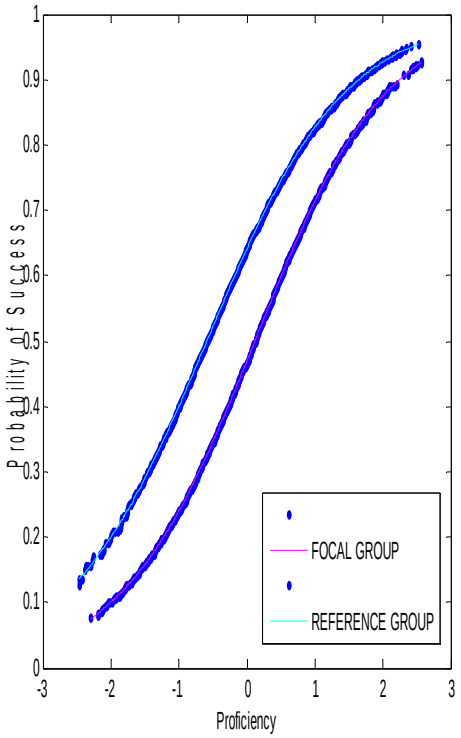


FIGURE 16: ICC of M3C1B1

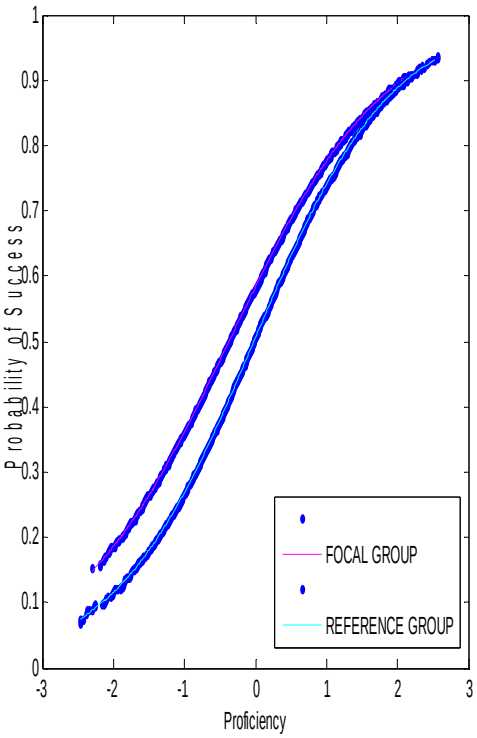


FIGURE 17: ICC of M3C1B1

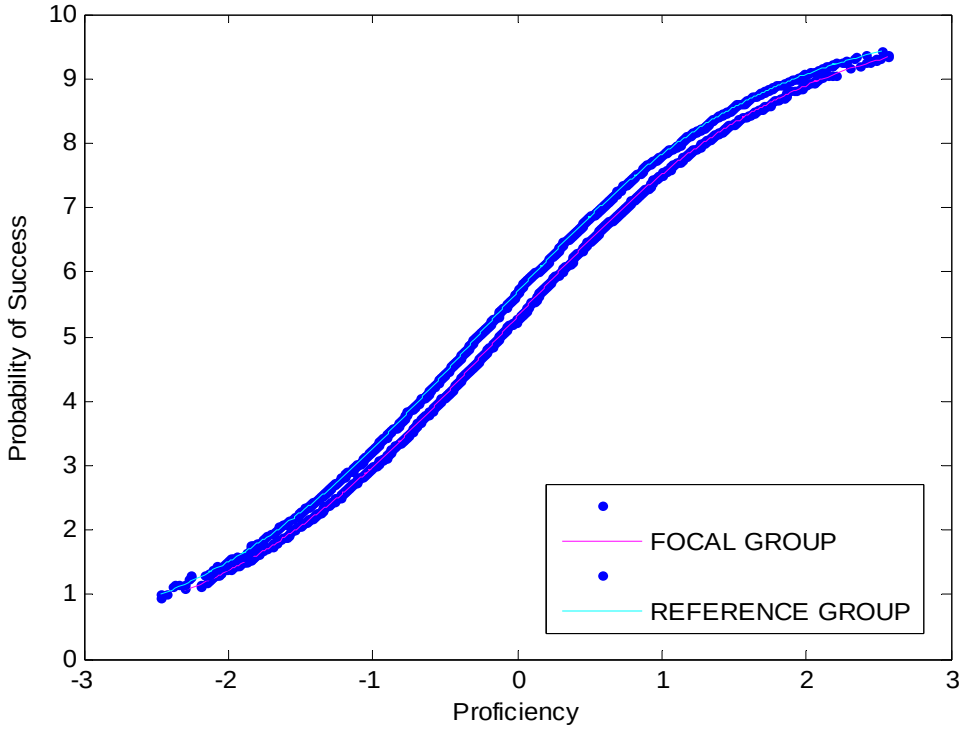


FIGURE 18: TCC of M3C1B1

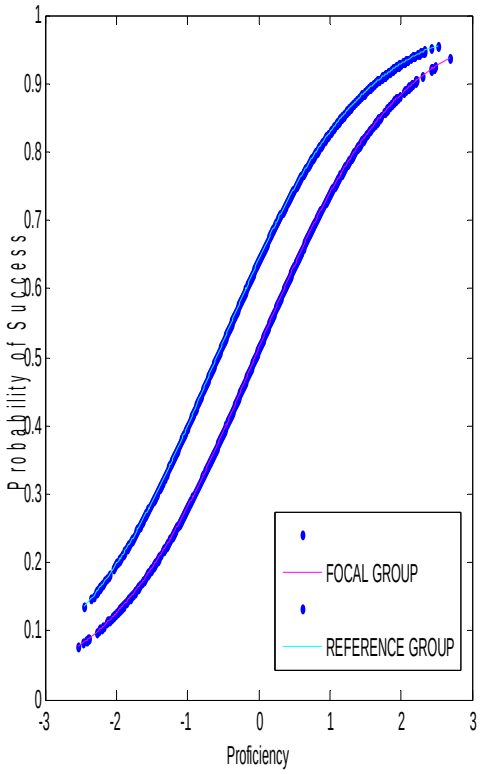


FIGURE 19: ICC of M1B2

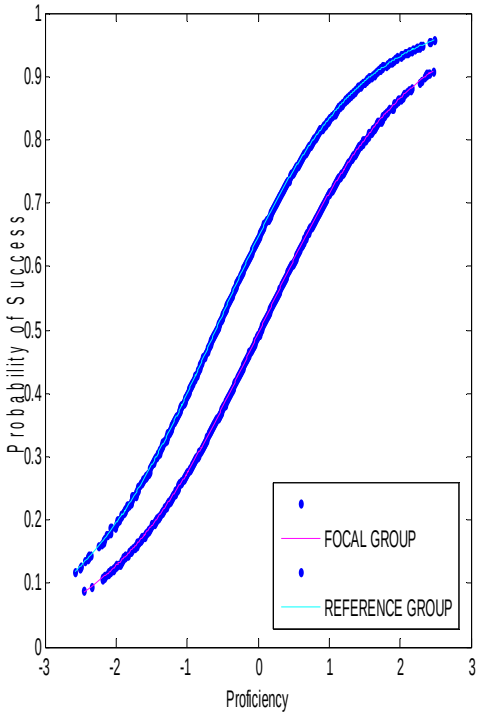


FIGURE 20: ICC of M2C1B2

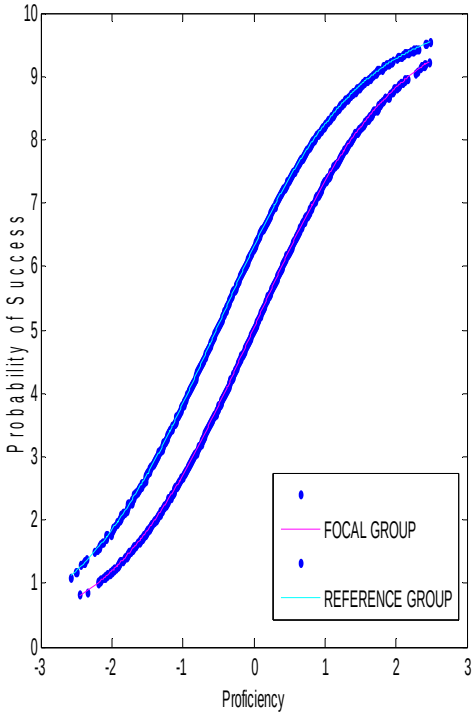


FIGURE 21: TCC of M2C1B2

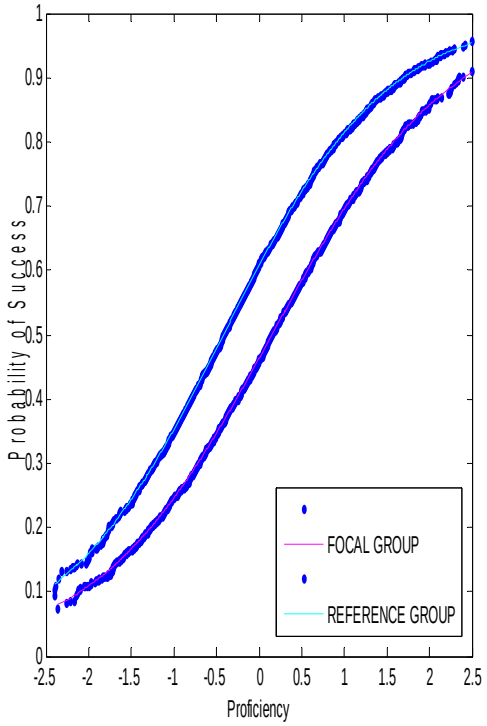


FIGURE 22: ICC of M2C2B2

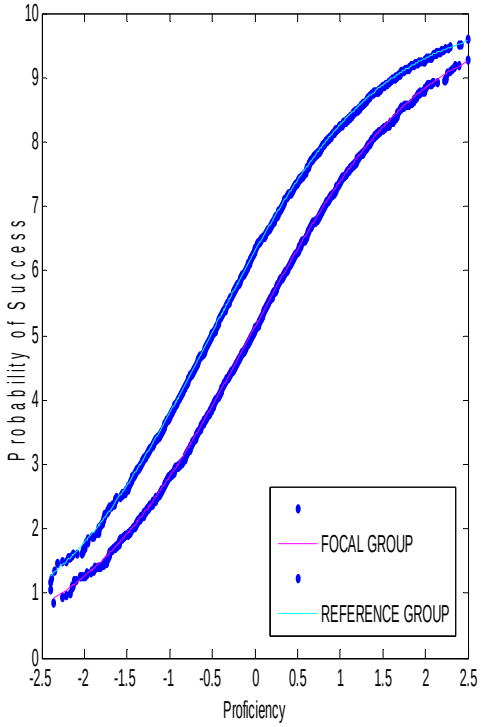


FIGURE 23: TCC of M2C2B2

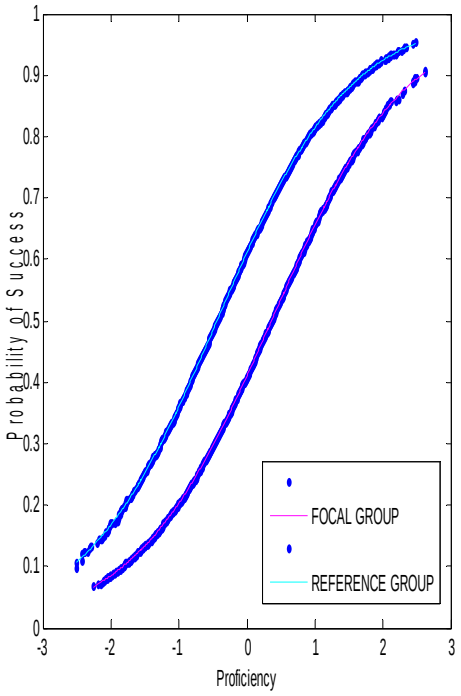


FIGURE 24: ICC of M3C1B2

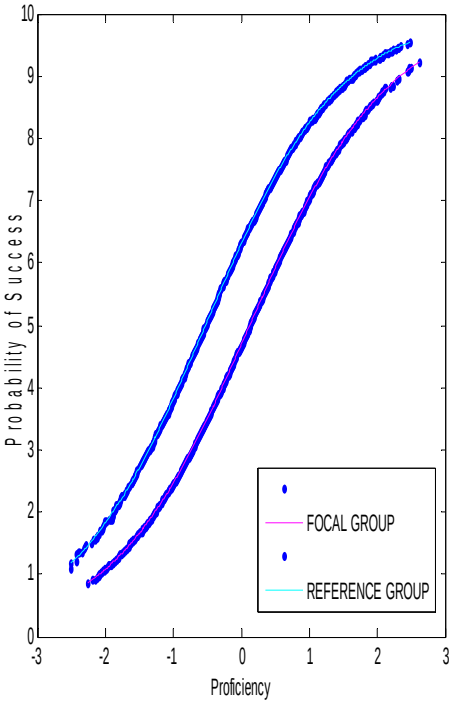


FIGURE 25: TCC of M3C1B2

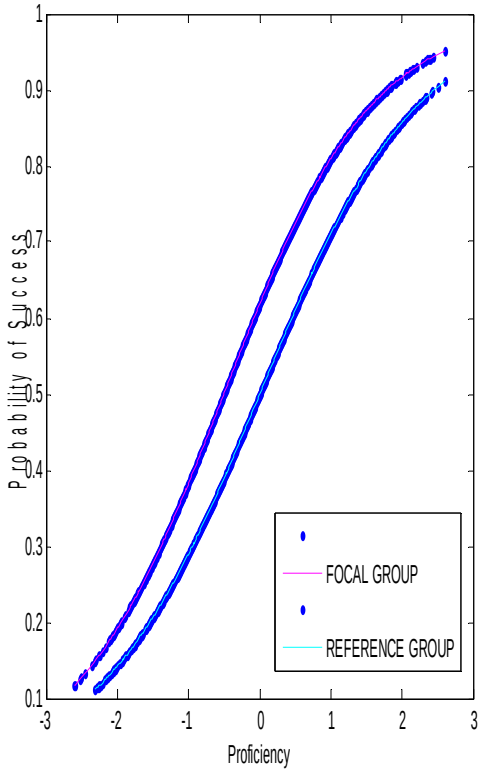


FIGURE 26: ICC of M1B3

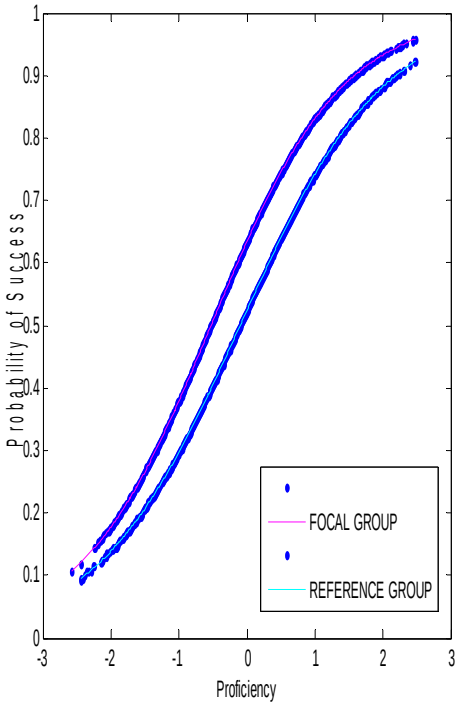


FIGURE 27: ICC of M2C1B3

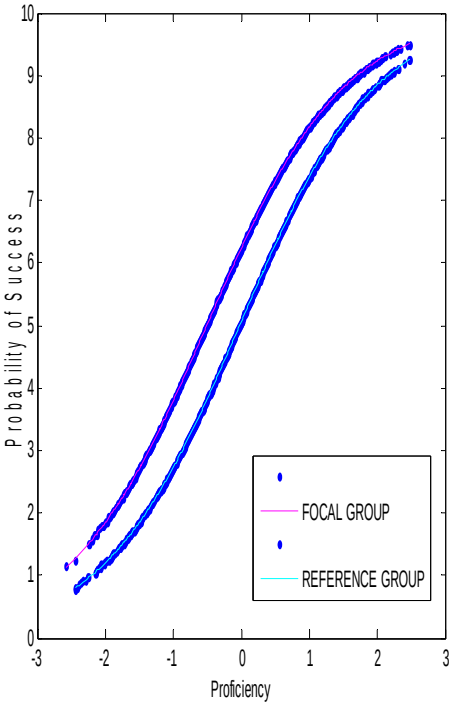


FIGURE 28: TCC of M2C1B3

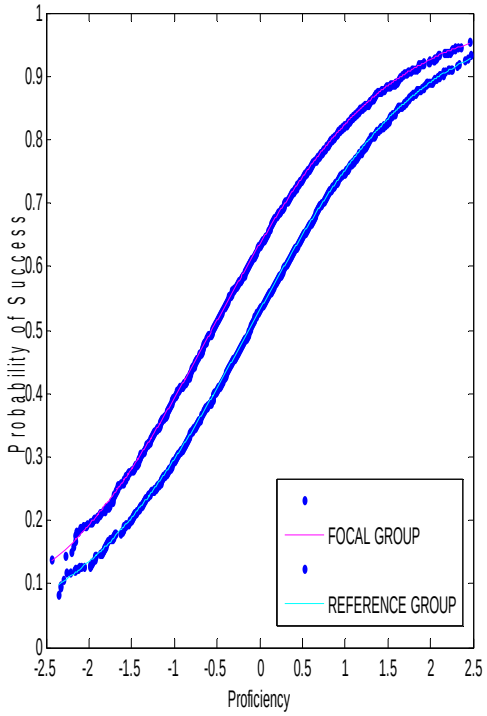


FIGURE 29: ICC of M2C2B3

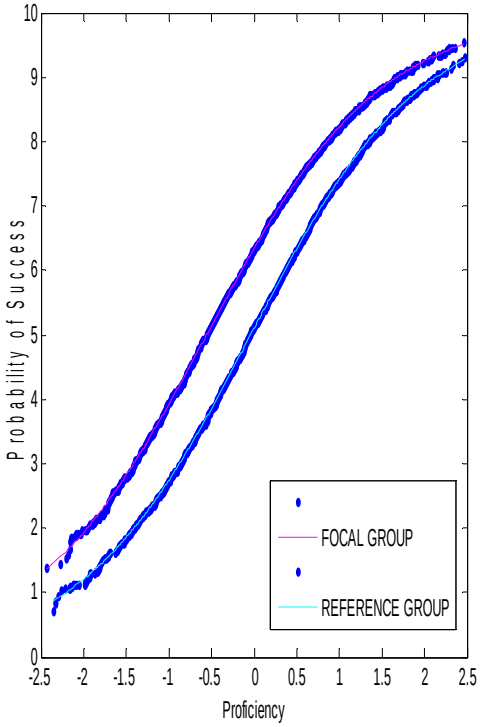


FIGURE 30: TCC of M2C2B3

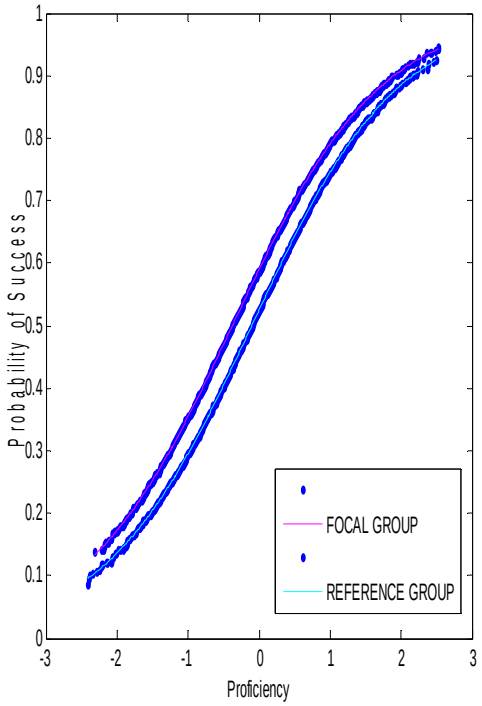


FIGURE 31: ICC of M3C1B3

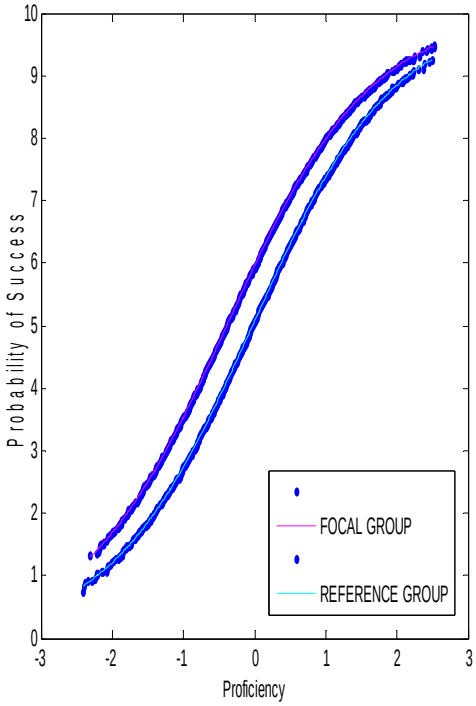


FIGURE 32: TCC of M3C1B3

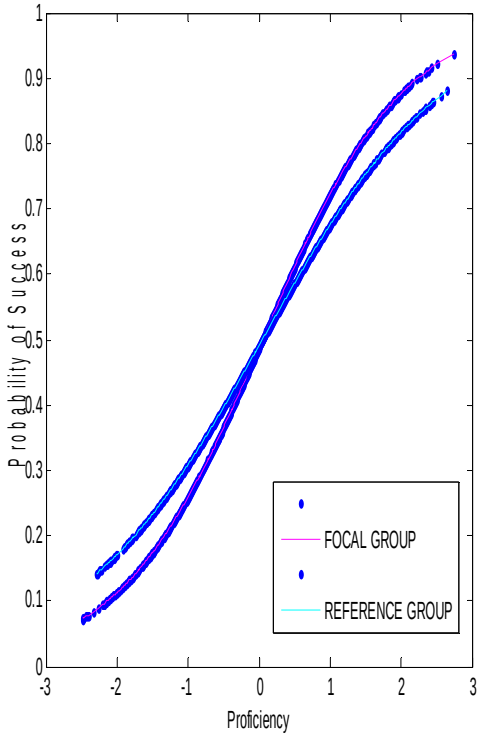


FIGURE 33: ICC of M1C1

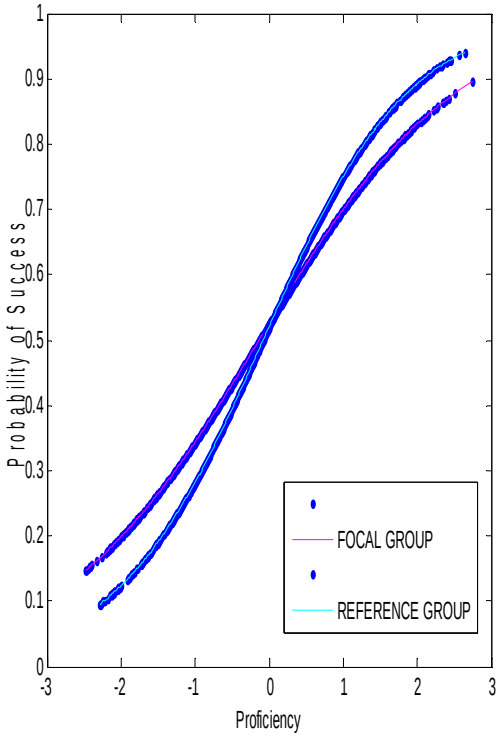


FIGURE 34: ICC of M1C1

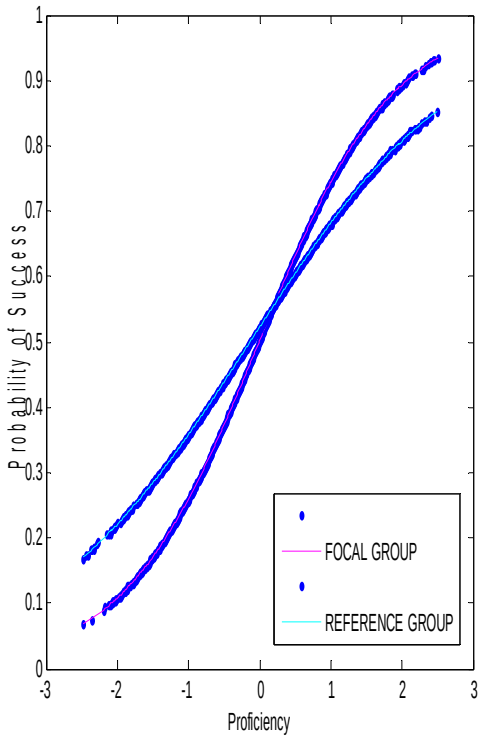


FIGURE 35: ICC of M2C1C1

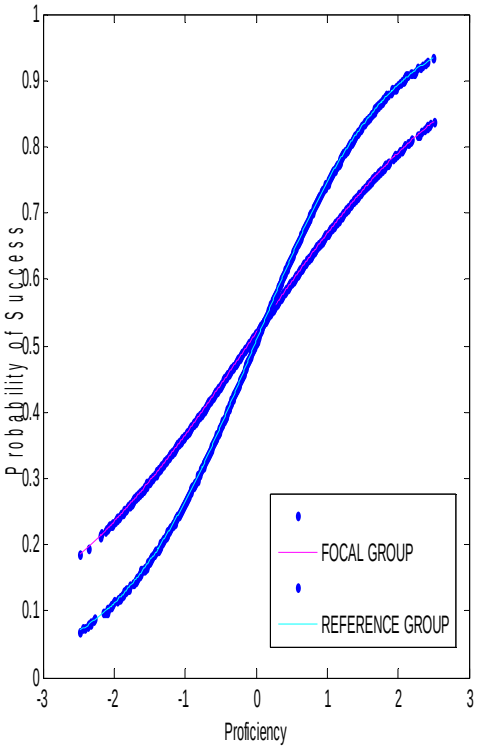


FIGURE 36: ICC of M2C1C1

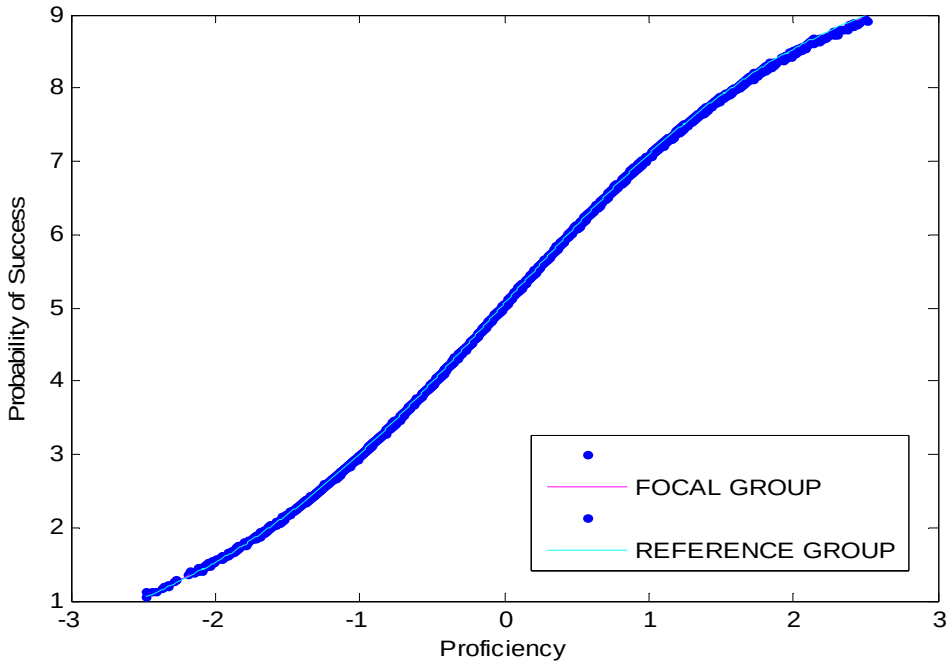


FIGURE 37: TCC of M2C1C1

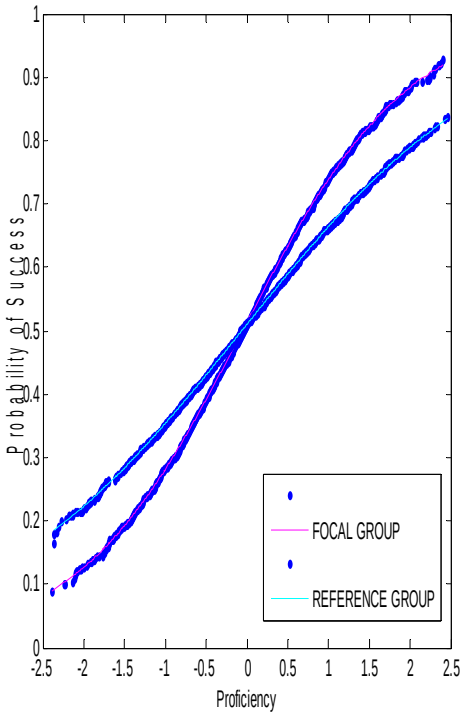


FIGURE 38: ICC of M2C2C1

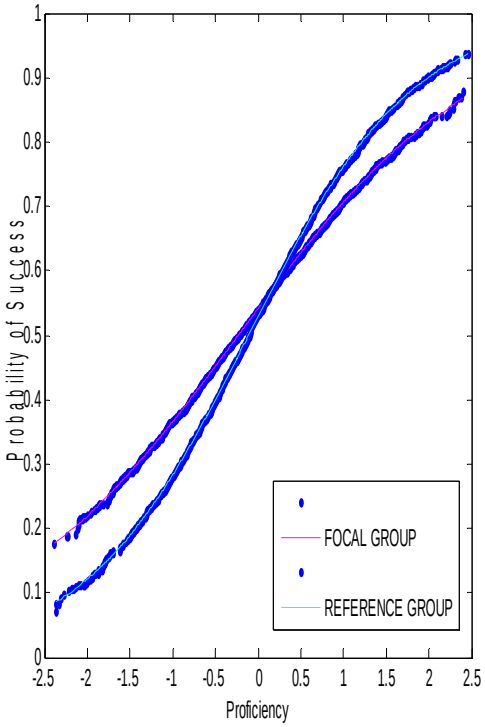


FIGURE 39: TCC of M2C2C1

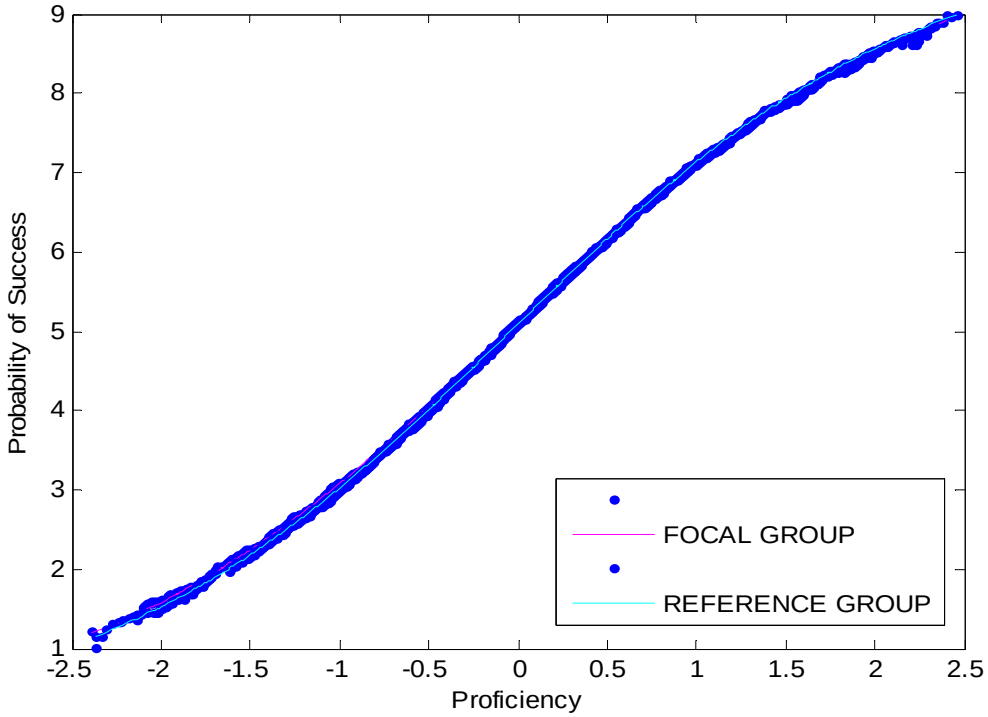


FIGURE 40: TCC of M2C2C1

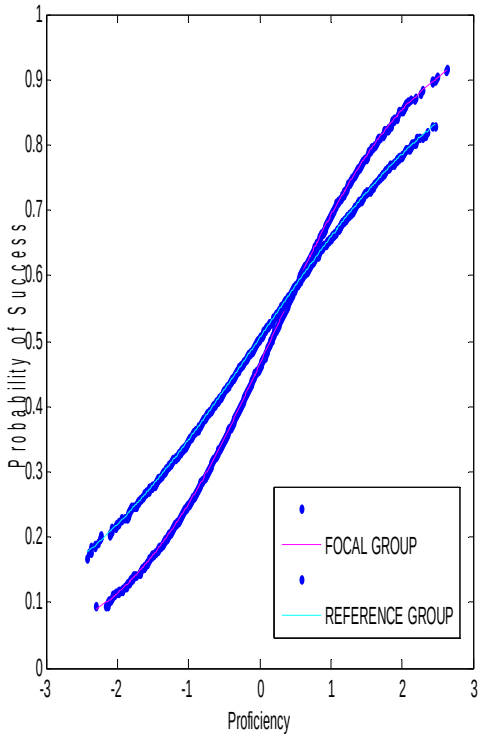


FIGURE 41: ICC of M3C1C1

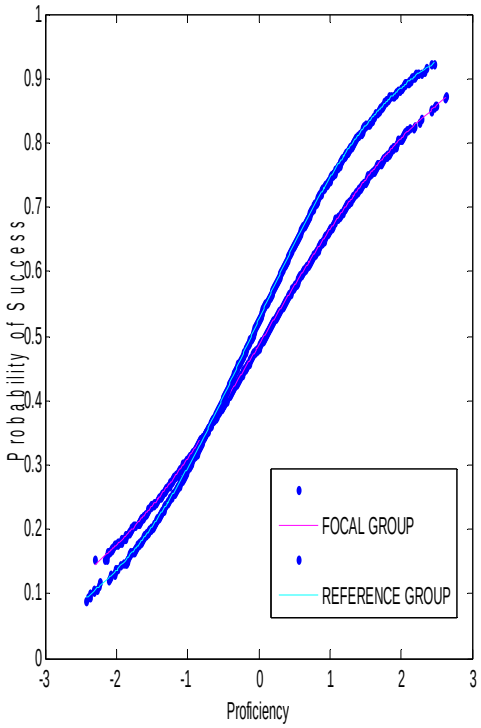


FIGURE 42: ICC of M3C1C1

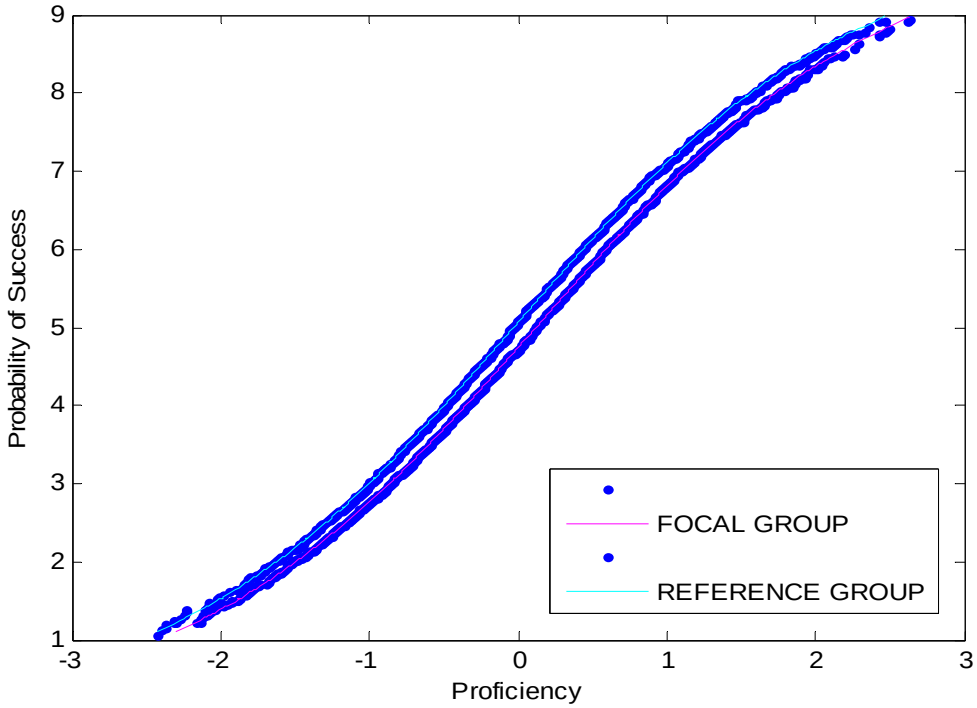


FIGURE 43: TCC of M3C1C1

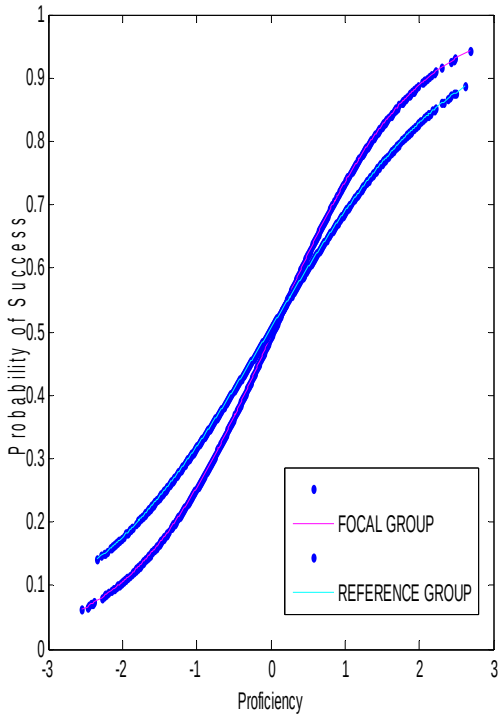


FIGURE 44: ICC of M1C2

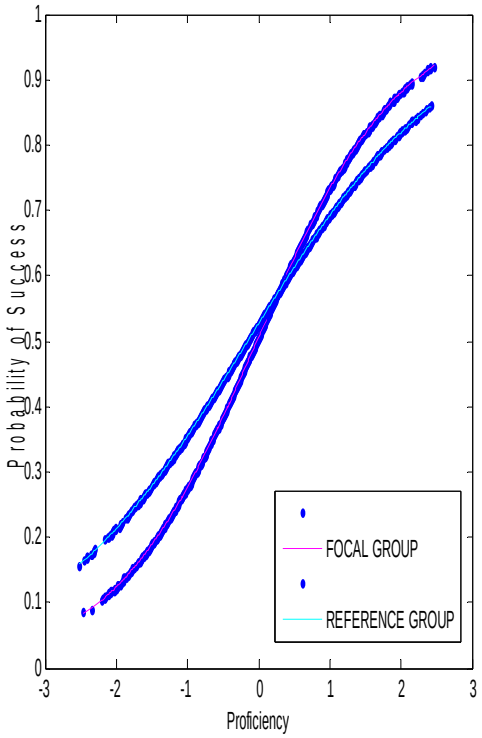


FIGURE 45: ICC of M2C1C2

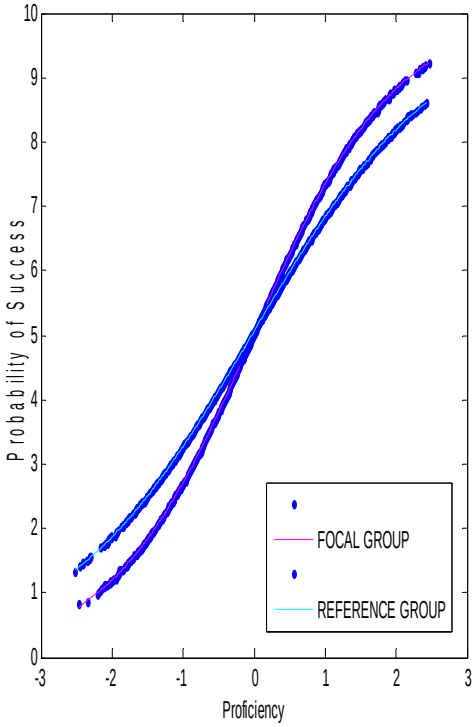


FIGURE 46: TCC of M2C1C2

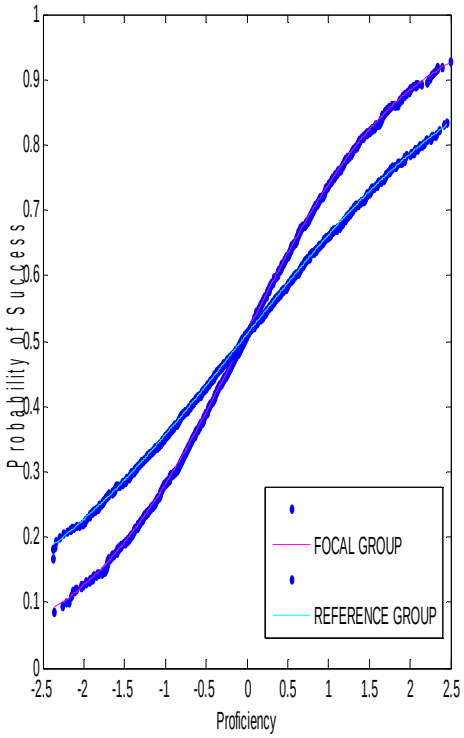


FIGURE 47: ICC of M2C2C2

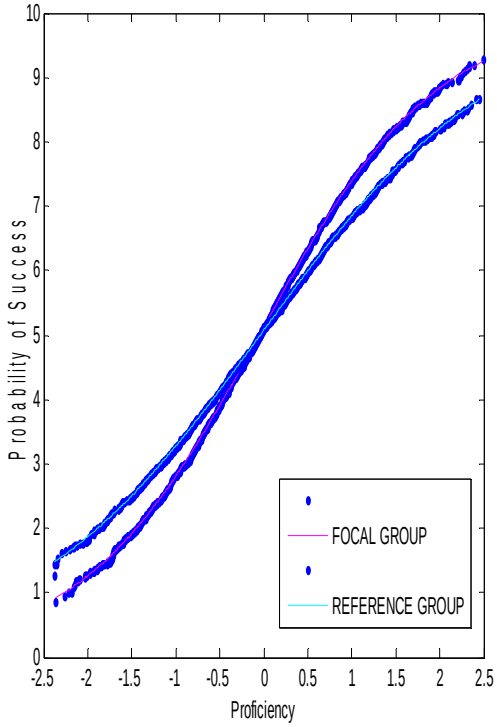


FIGURE 48: TCC of M2C2C2

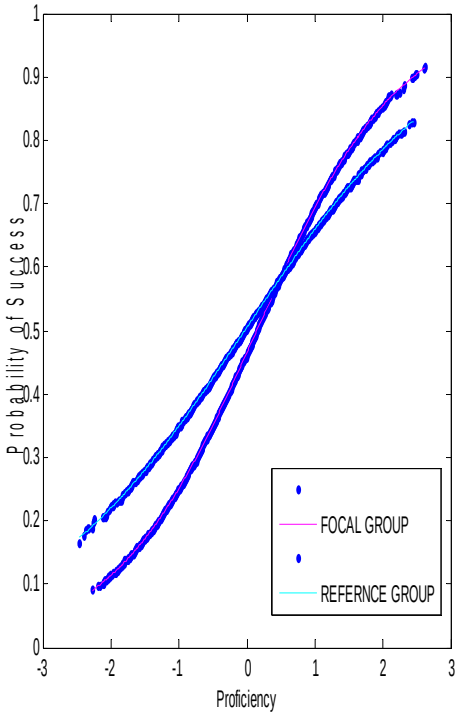


FIGURE 49: ICC of M3C1C2

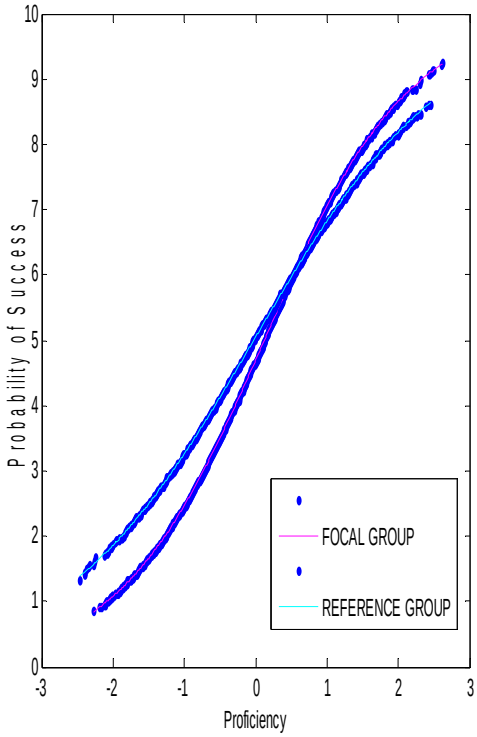


FIGURE 50: TCC of M3C1C2

Appendix C: Item Characteristic Curves of Representative Items in Simulation Study of Ethnic Example

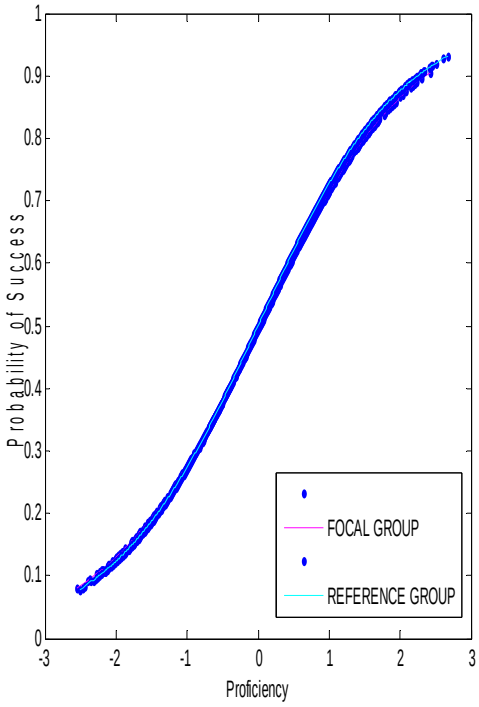


FIGURE 51: ICC of M1A

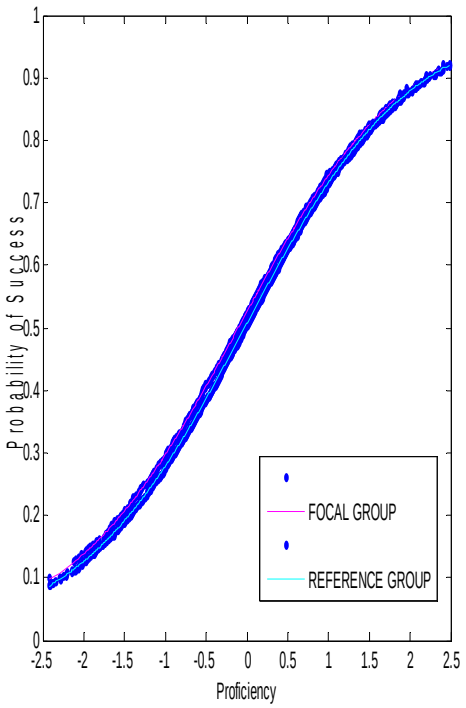


FIGURE 52: ICC of M2C1A

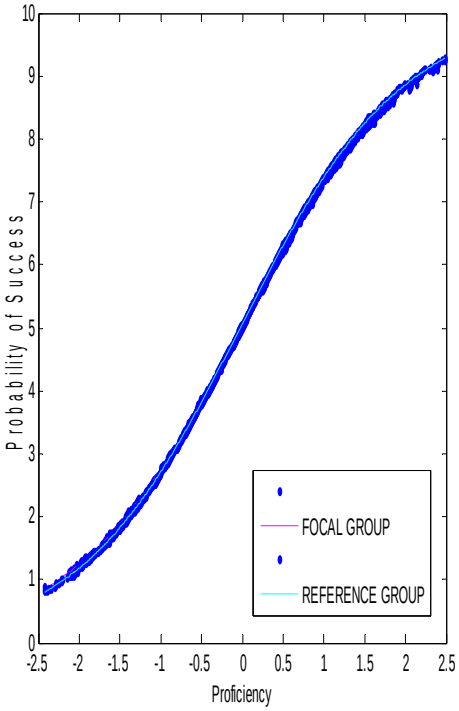


FIGURE 53: TCC of M2C1A

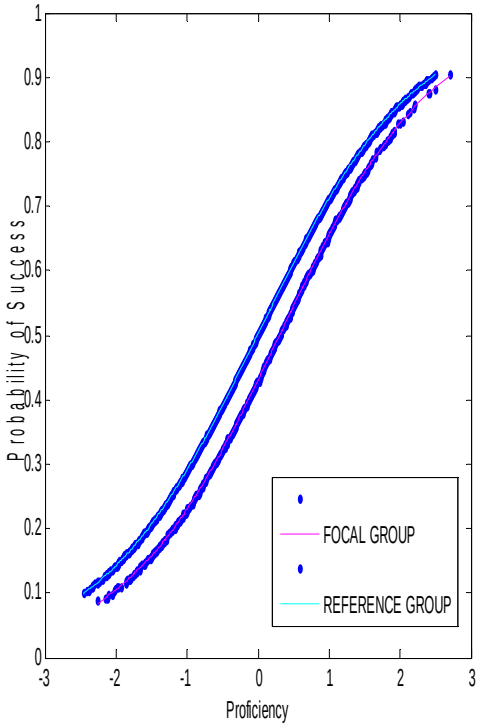


FIGURE 54: ICC of M3C1A

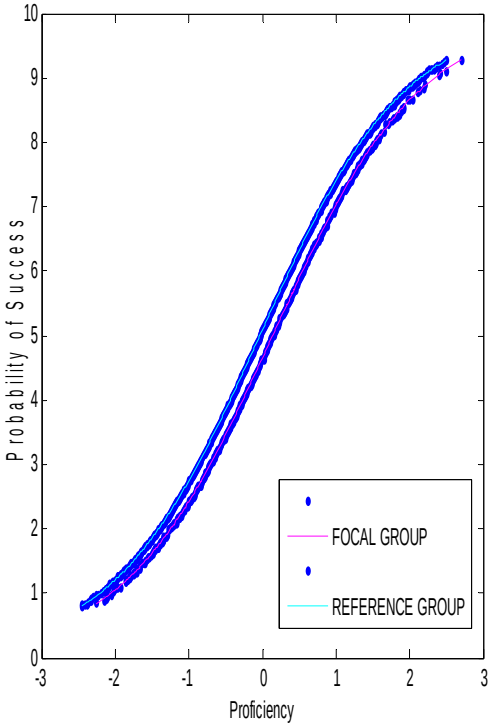


FIGURE 55: TCC of M3C1A

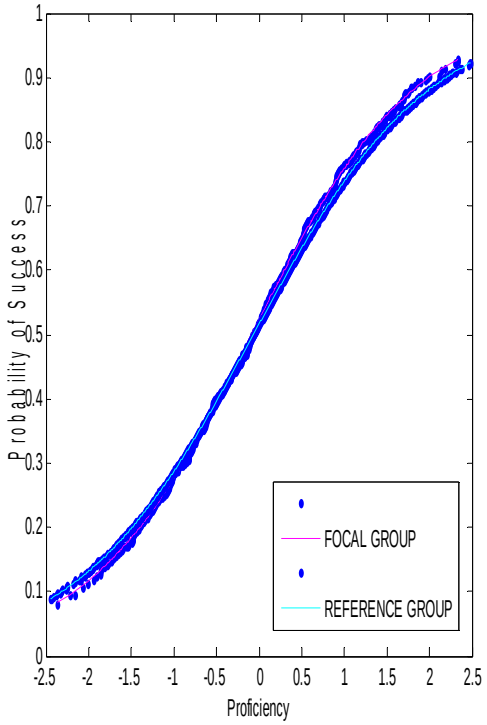


FIGURE 56: ICC of M3C2A

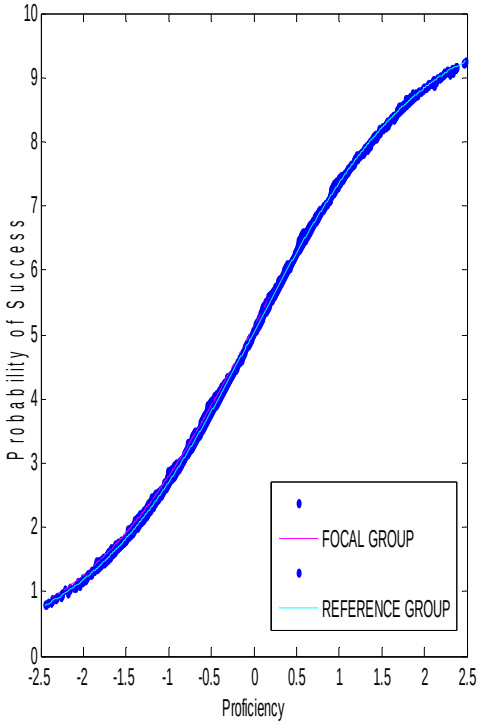


FIGURE 57: TCC of M3C2A

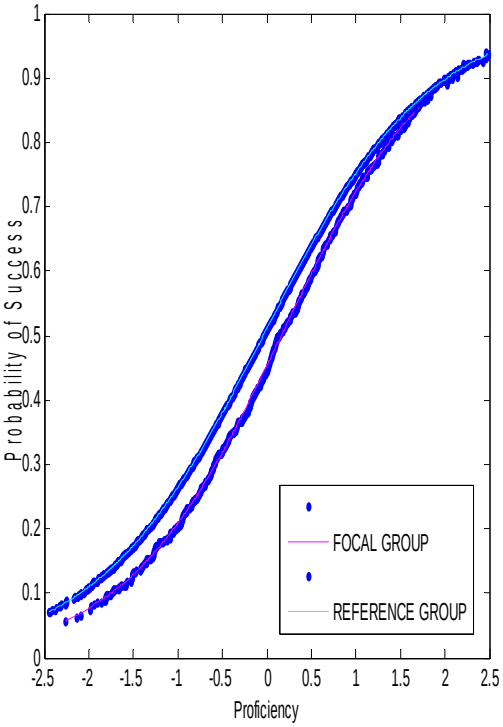


FIGURE 58: ICC of M3C3A

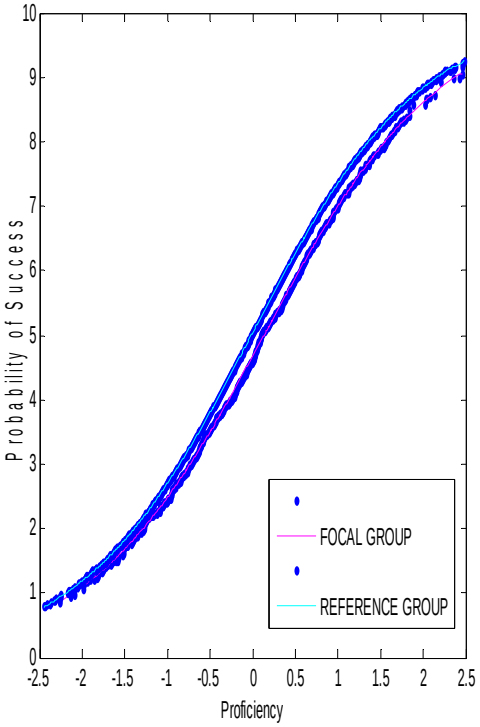


FIGURE 59: TCC of M3C3A

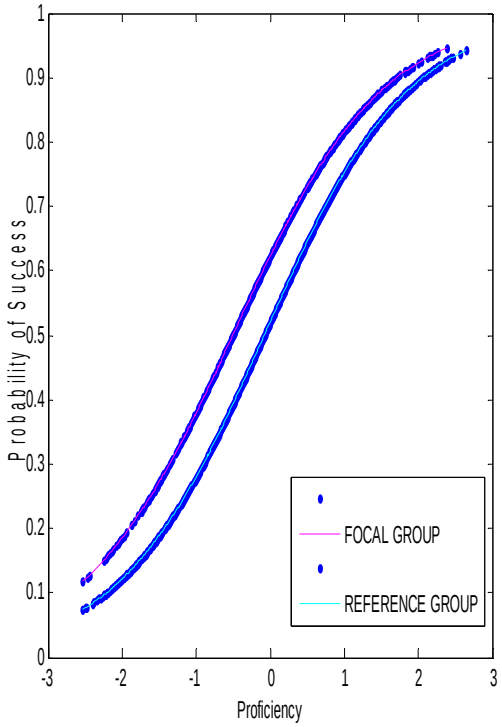


FIGURE 60: ICC of M1B1 (46)

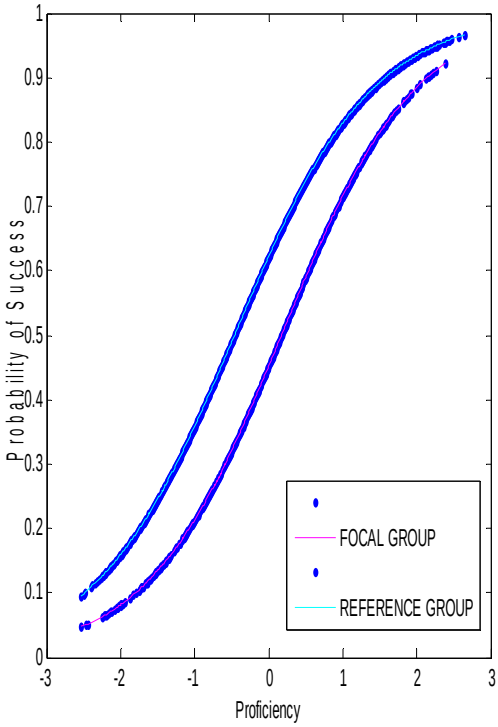


FIGURE 61: ICC of M1B1 (46)

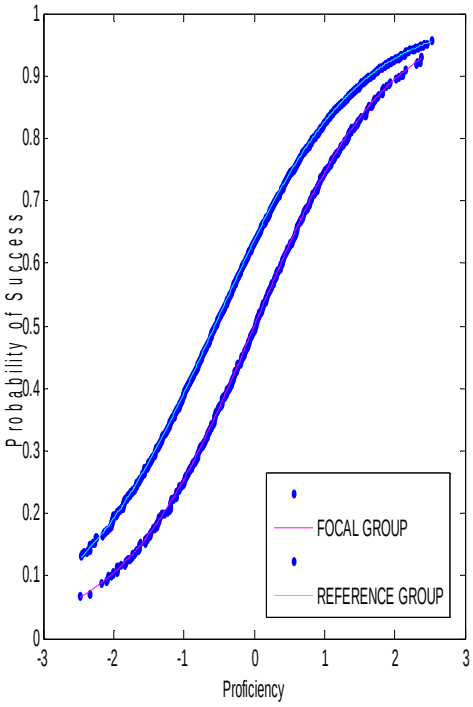


FIGURE 62: ICC of M2C1B1 (23)

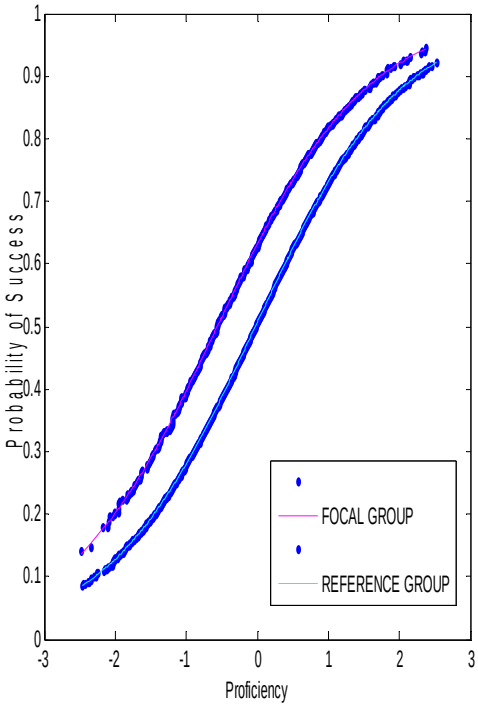


FIGURE 63: ICC of M2C1B1

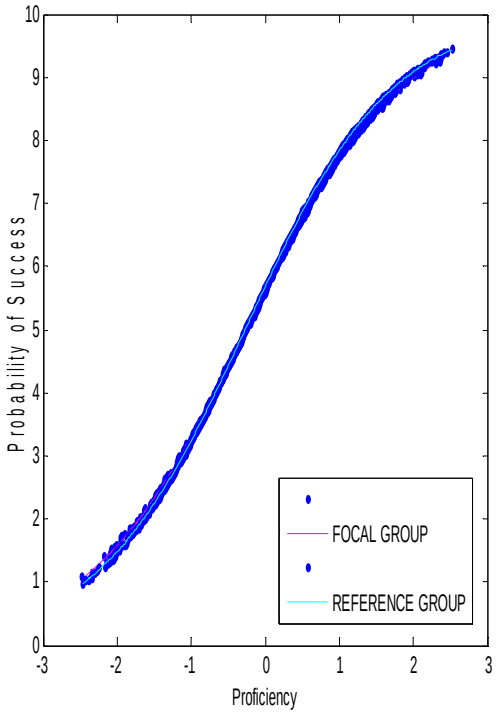


FIGURE 64: TCC of M2C1B1

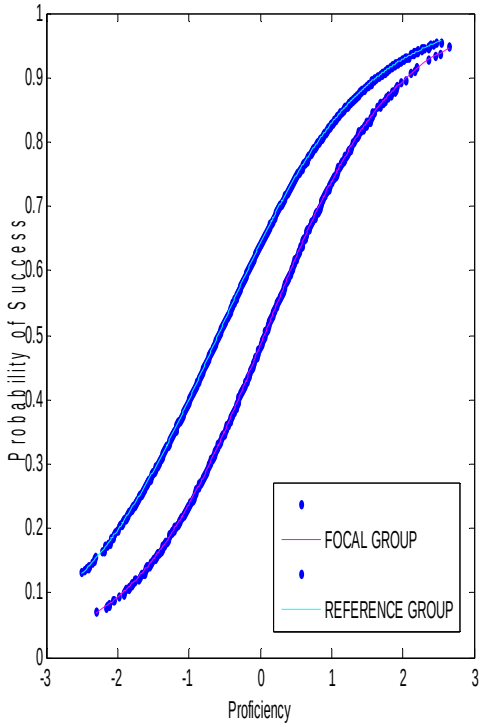


FIGURE 65: ICC of M3C1B2

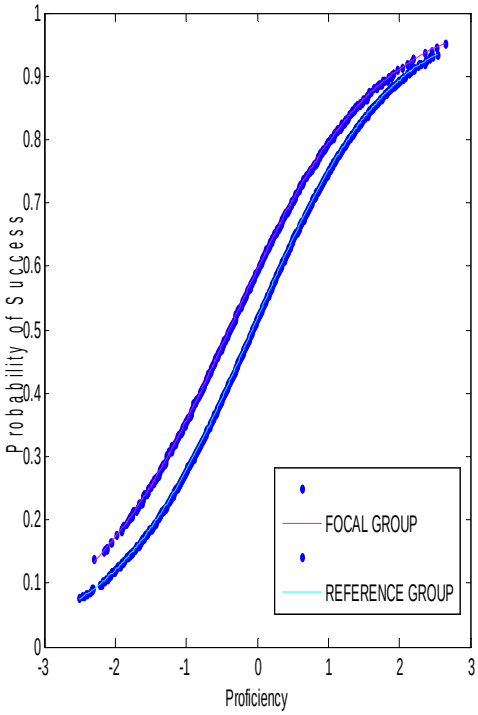


FIGURE 66: ICC of M3C1B2

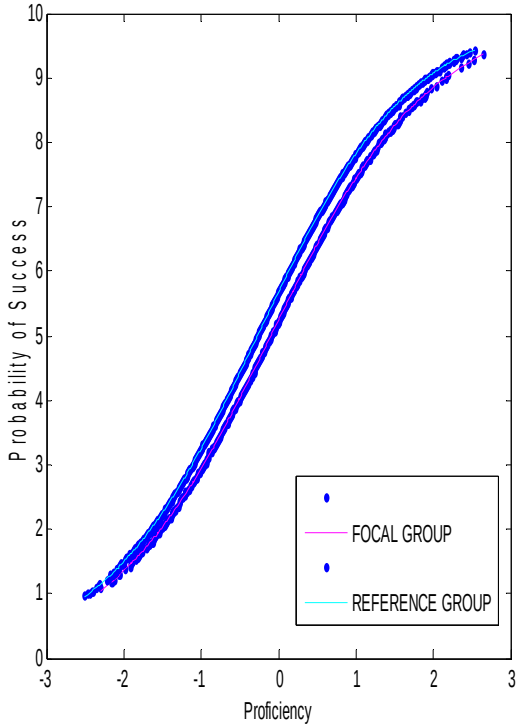


FIGURE 67: TCC of M3C1B2 (24)

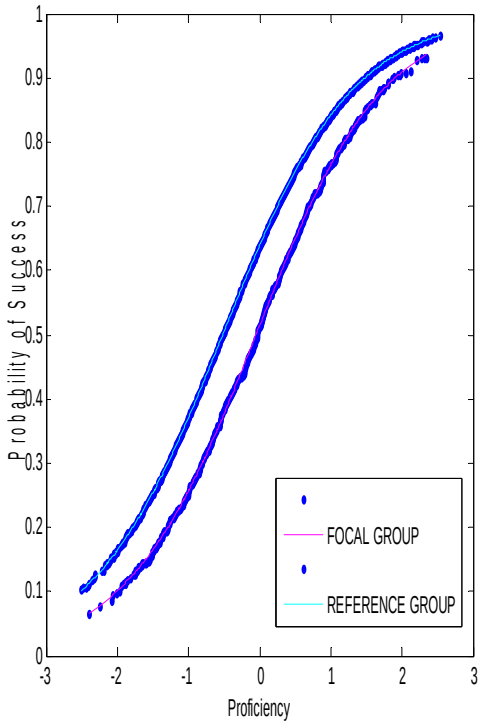


FIGURE 68: ICC of M3C2B1 (25)

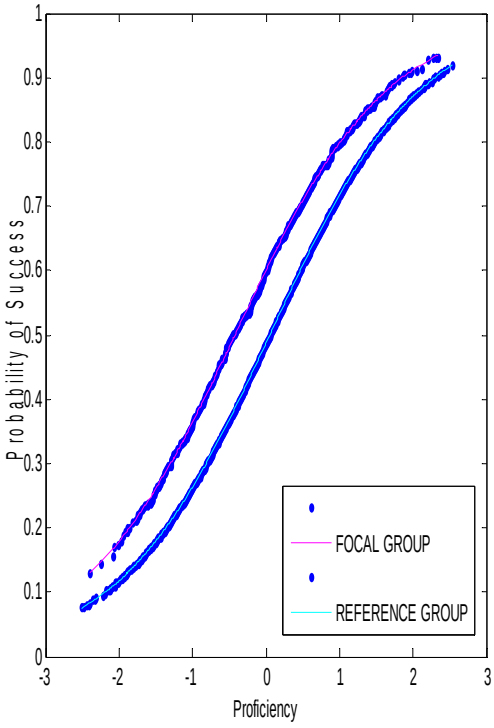


FIGURE 69: ICC of M3C2B1

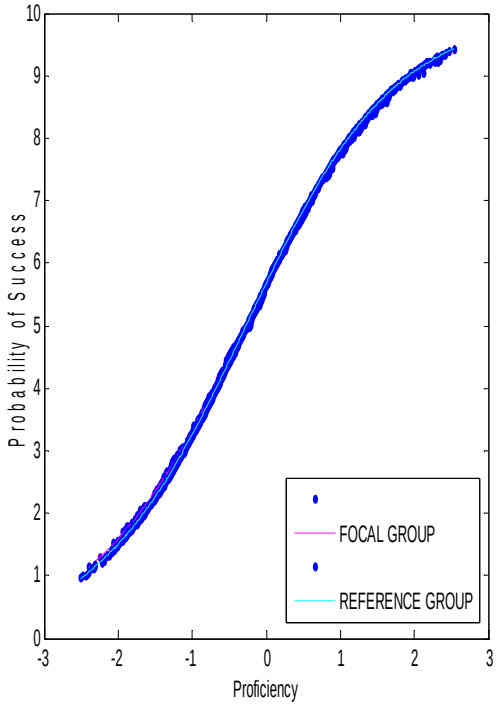


FIGURE 70: TCC of M3C2B1 (25)

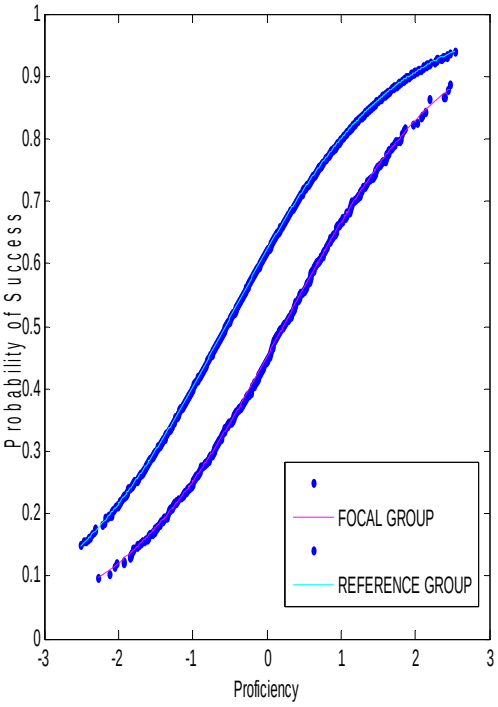


FIGURE 71: ICC of M3C3B1 (26)

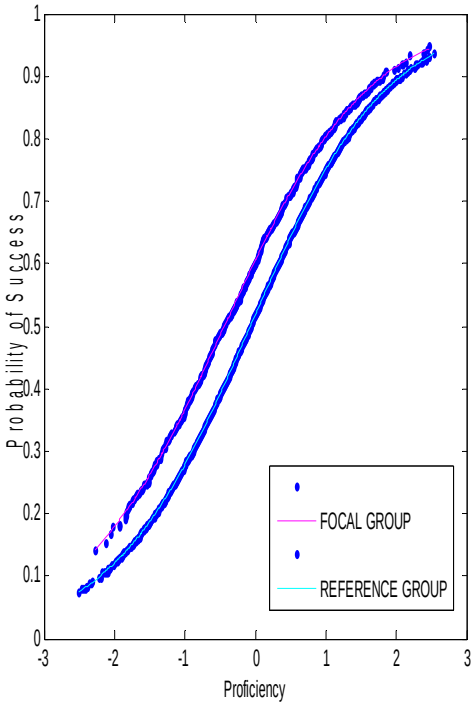


FIGURE 72: ICC of M3C3B1 (26)

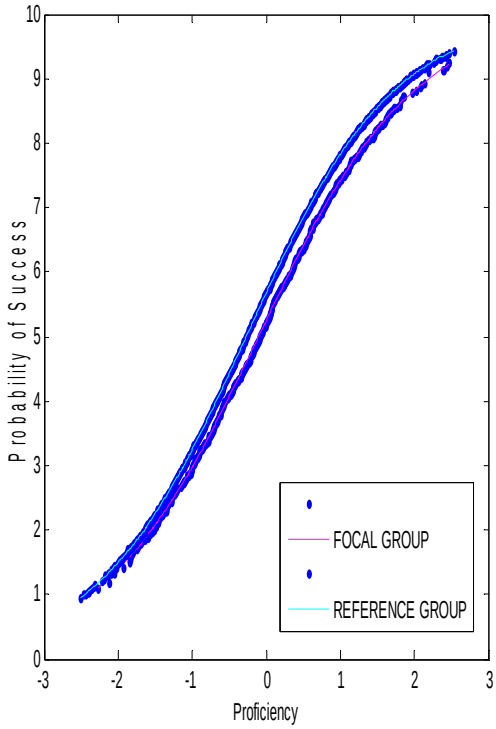


FIGURE 73: TCC of M3C3B1 (26)

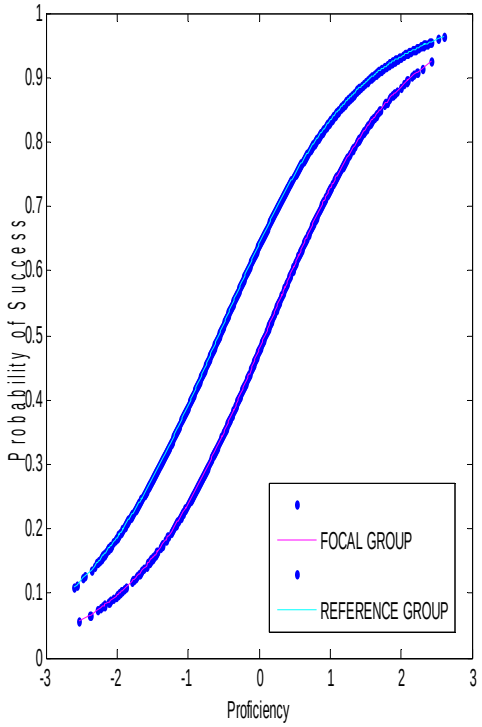


FIGURE 74: ICC of M1B2 (47)

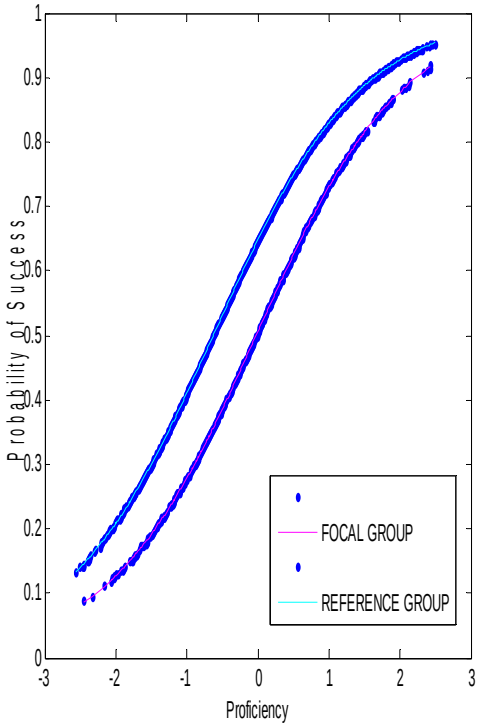


FIGURE 75: ICC of M2C1B2 (27)

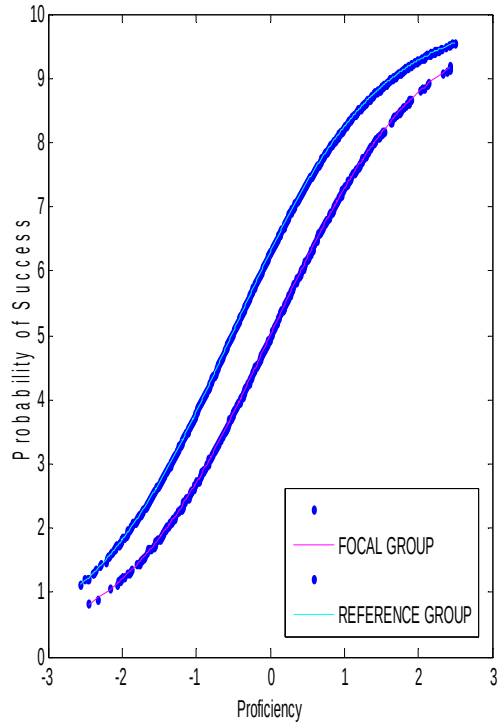


FIGURE 76: TCC of M2C1B2 (27)

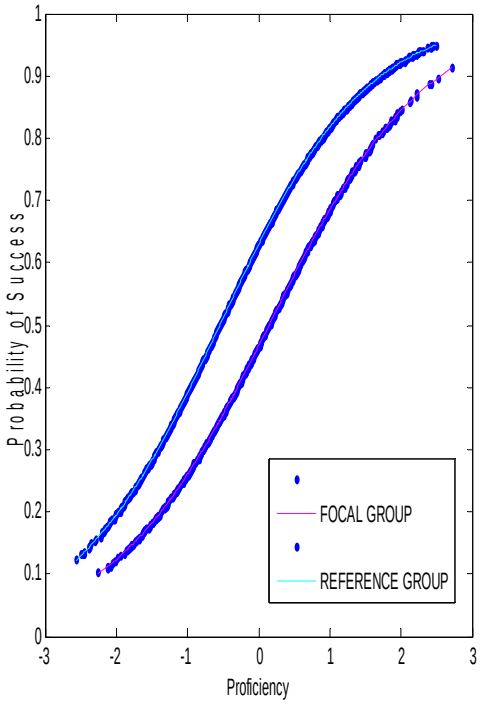


FIGURE 77: ICC of M3C1B2 (28)

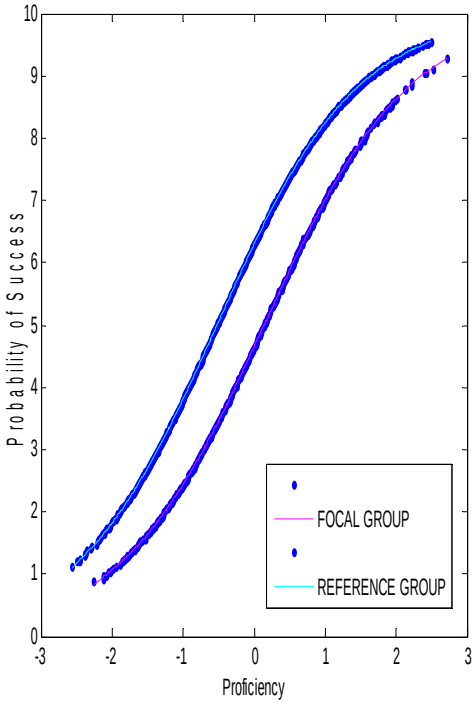


FIGURE 78: TCC of M3C1B2 (28)

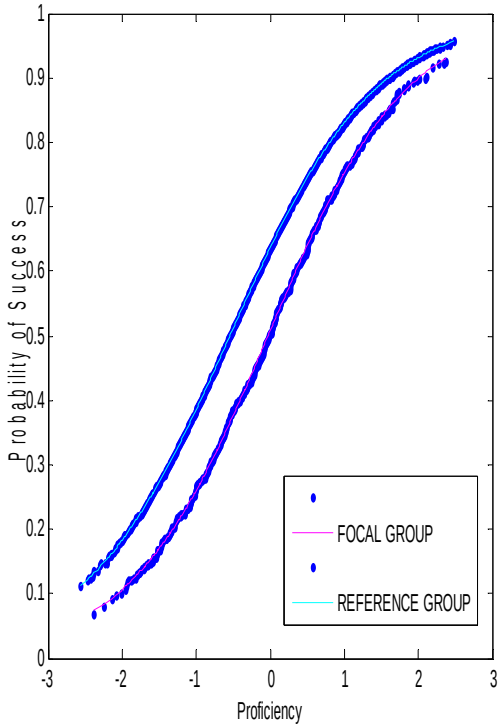


FIGURE 79: ICC of M3C2B2 (29)

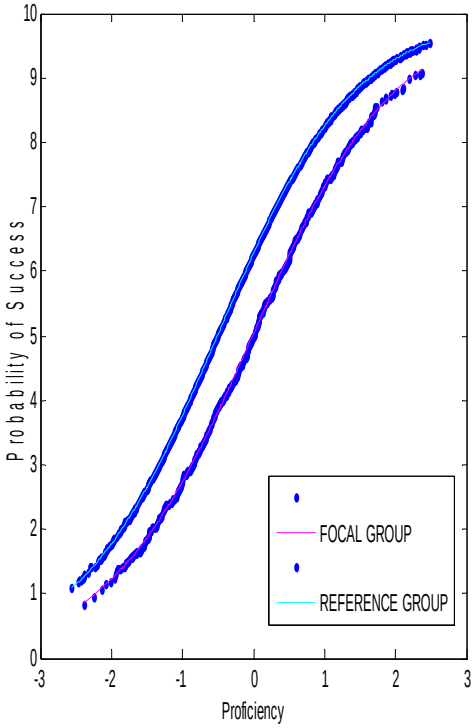


FIGURE 80: TCC of M3C2B2 (29)

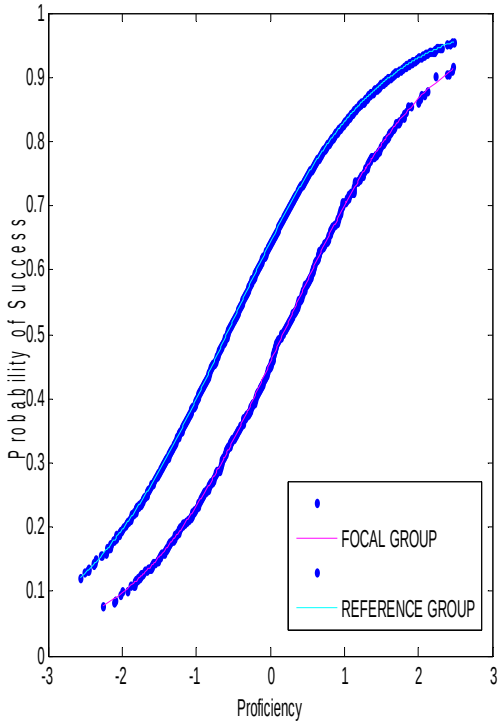


FIGURE 81: ICC OF M3C3B2 (30)

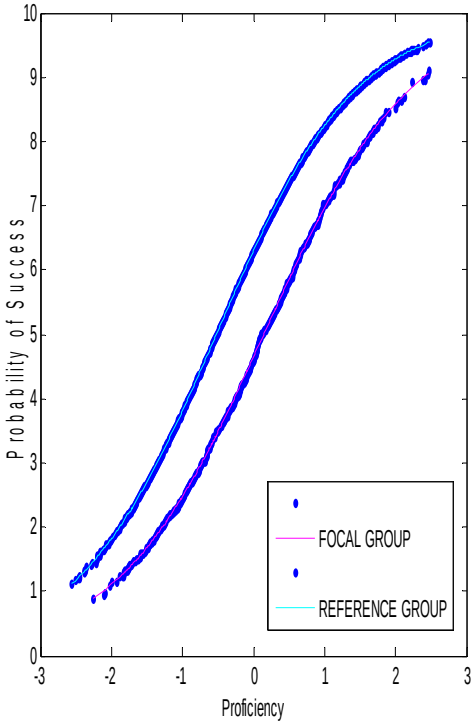


FIGURE 82: TCC of M3C3B2 (30)

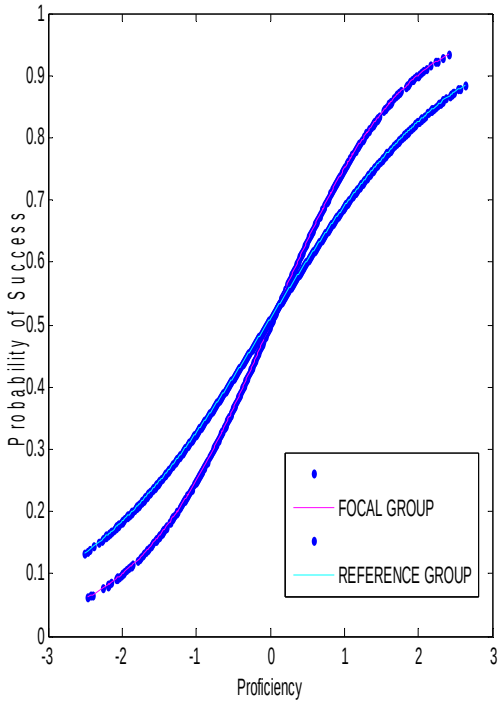


FIGURE 83: ICC of M1C1 (48)

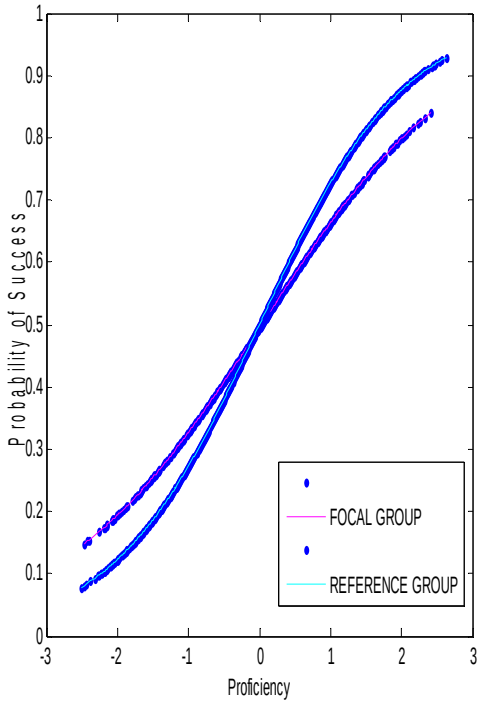


FIGURE 84: ICC of M1C1 (48)

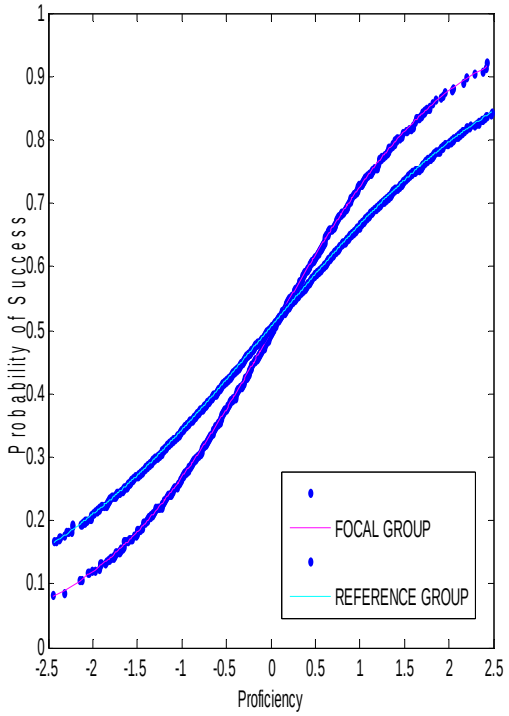


FIGURE 85: ICC of M2C1C1 (31)

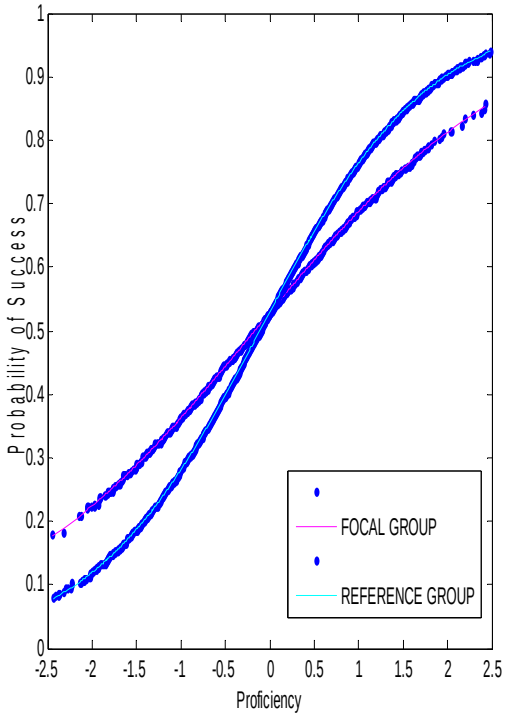


FIGURE 86: ICC of M2C1C1 (31)

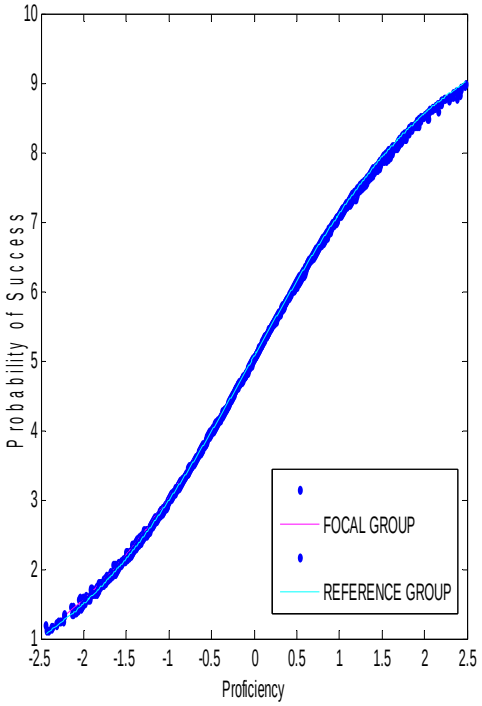


FIGURE 87: TCC of M2C1C1 (31)

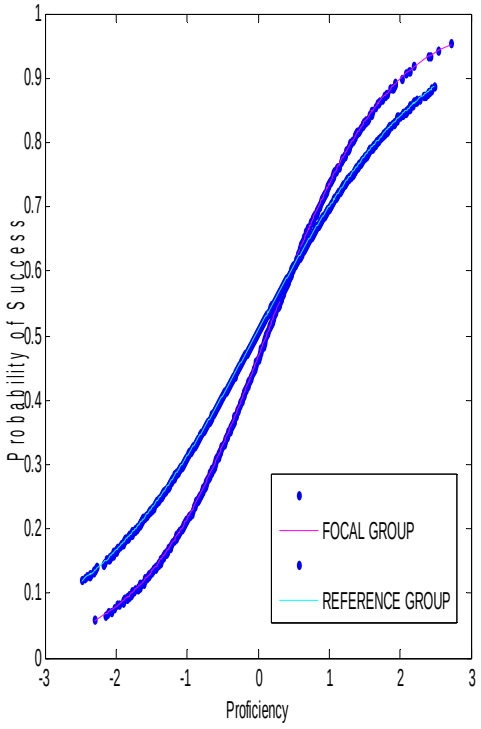


FIGURE 88: ICC of M3C1C1 (32)

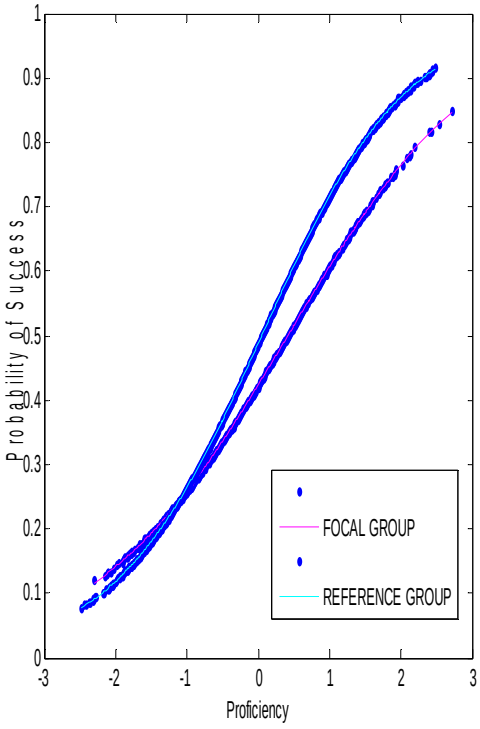


FIGURE 89: ICC of M3C1C1 (32)

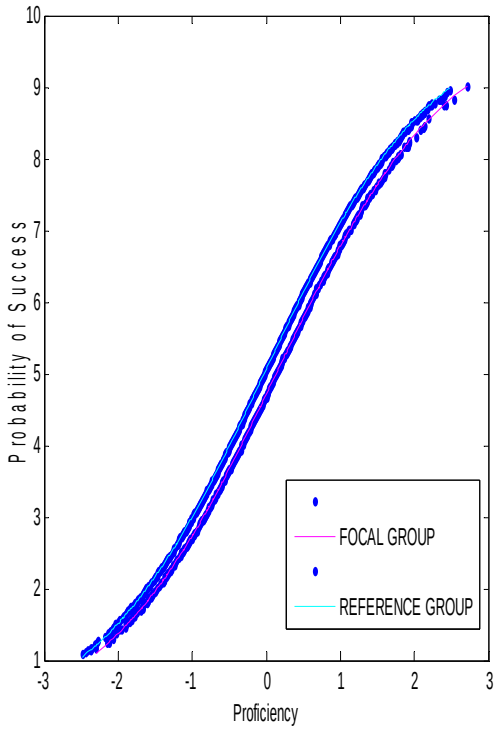


FIGURE 90: TCC of M3C1C1 (32)

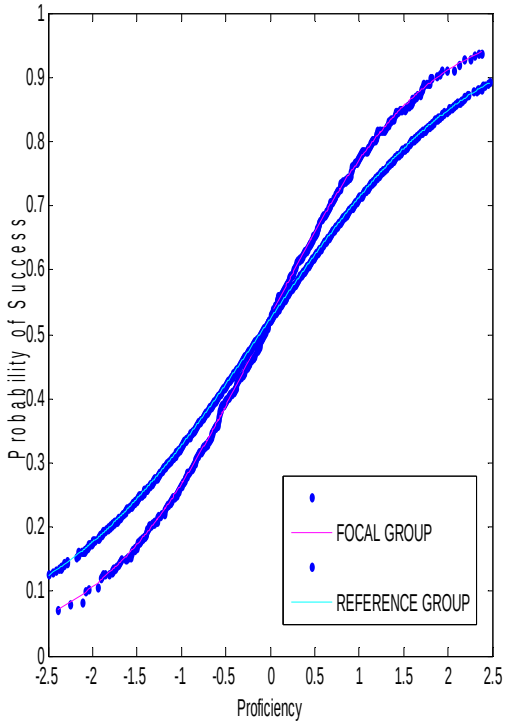


FIGURE 91: ICC of M3C2C1 (33)

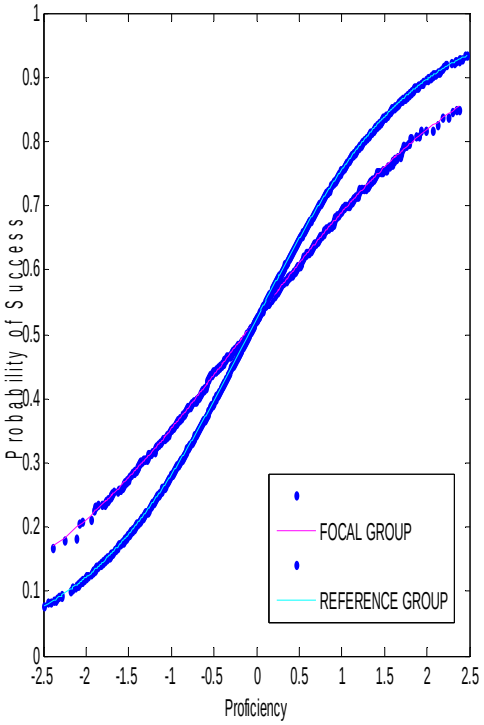


FIGURE 92: ICC of M3C2C1 (33)

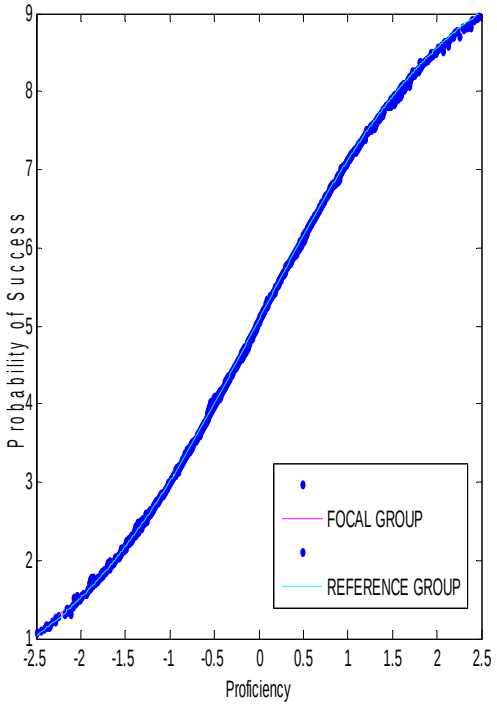


FIGURE 93: TCC of M3C2C1 (33)

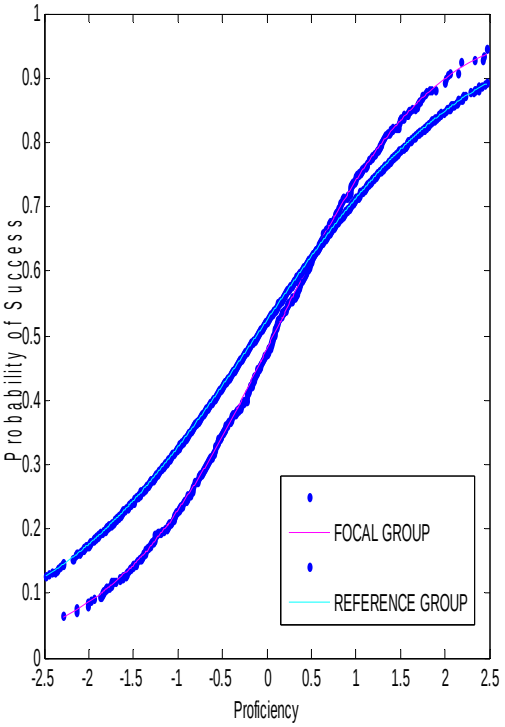


FIGURE 94: ICC of M3C3C1 (34)

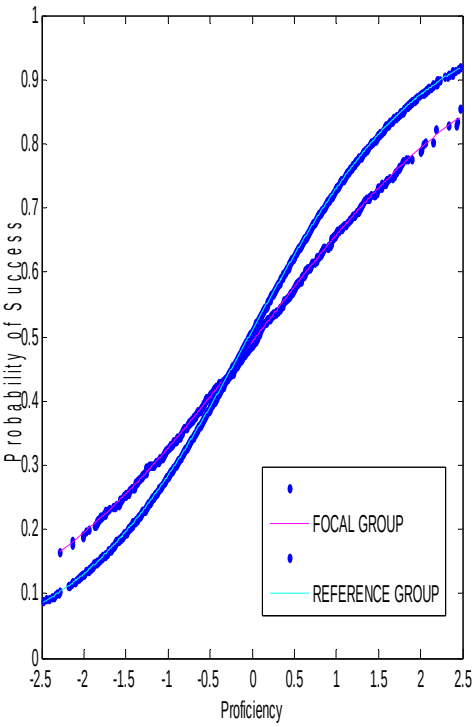


FIGURE 95: ICC of M3C2C1 (34)

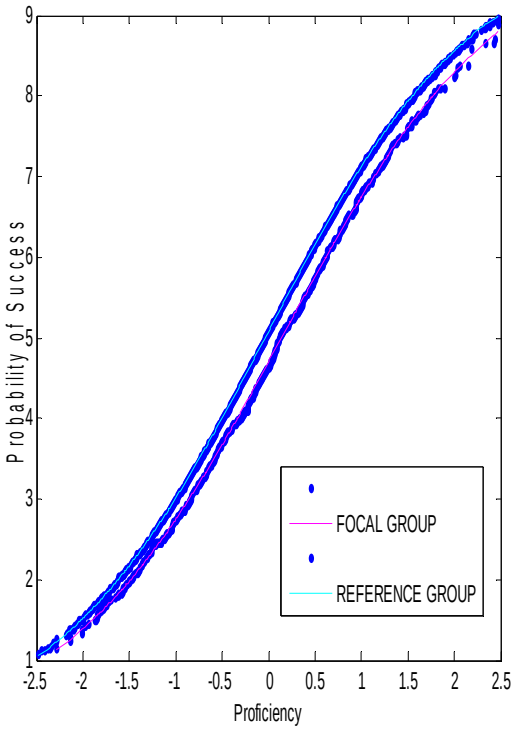


FIGURE 96: TCC of M3C3C1 (34)

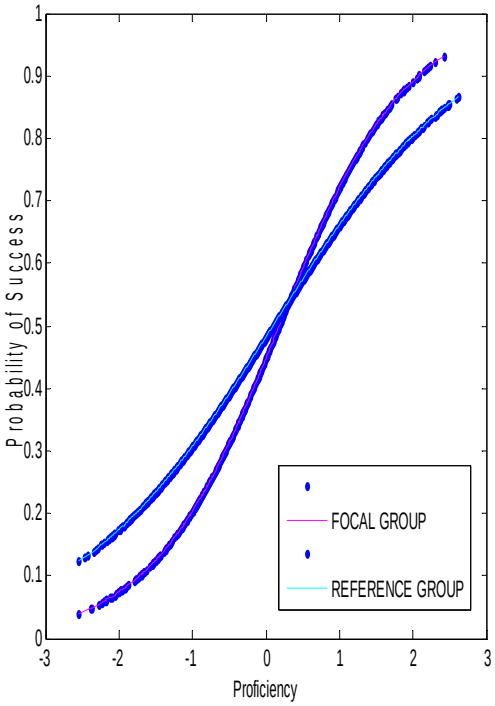


FIGURE 97: ICC of M1C2 (49)

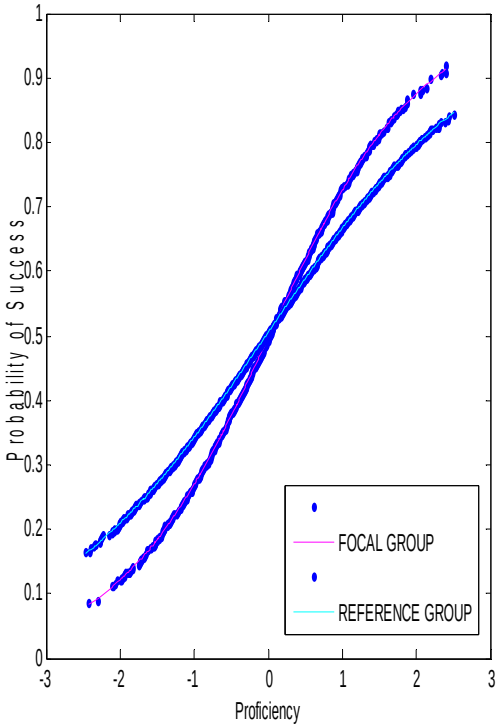


FIGURE 98: ICC of M2C1C2 (35)

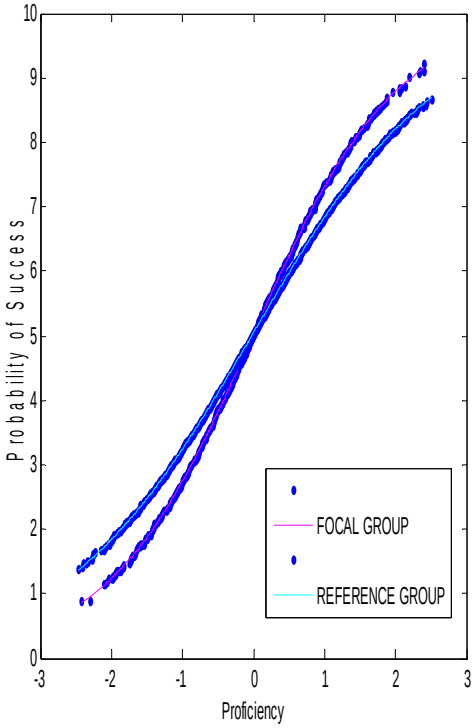


FIGURE 98: TCC of M2C1C2 (35)

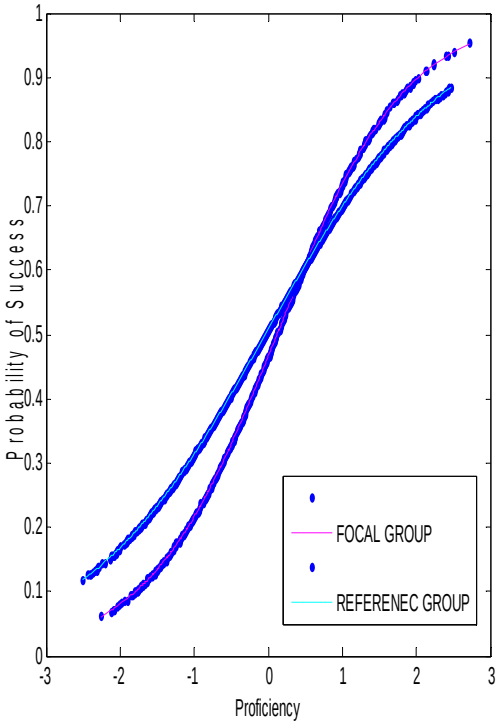


FIGURE 99: ICC of M3C1C2 (36)

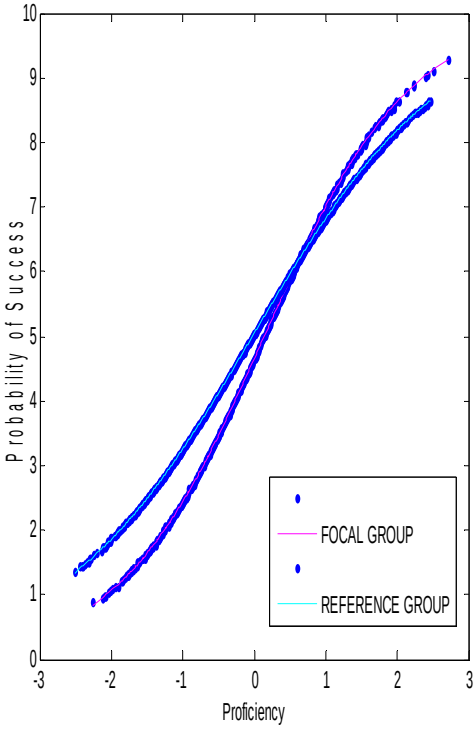


FIGURE 100: TCC of M3C1C2 (36)

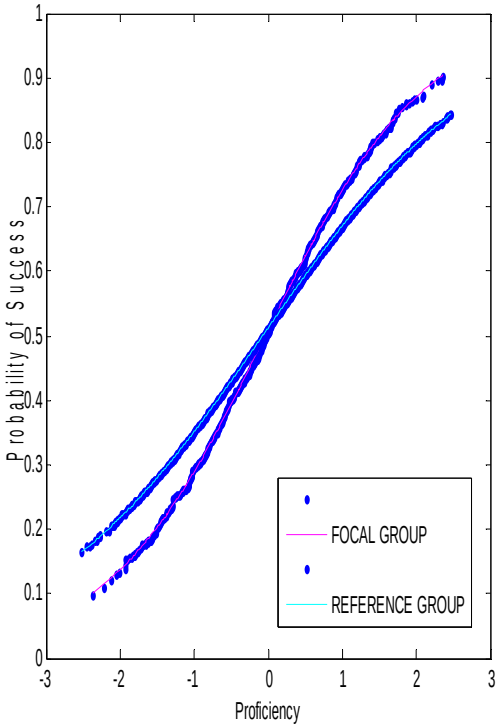


FIGURE 101: ICC of M3C2C2 (37)

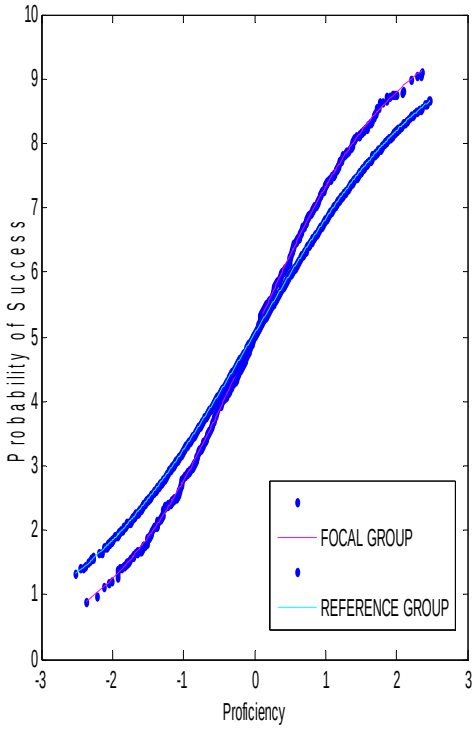


FIGURE 102: TCC of M3C2C2 (37)

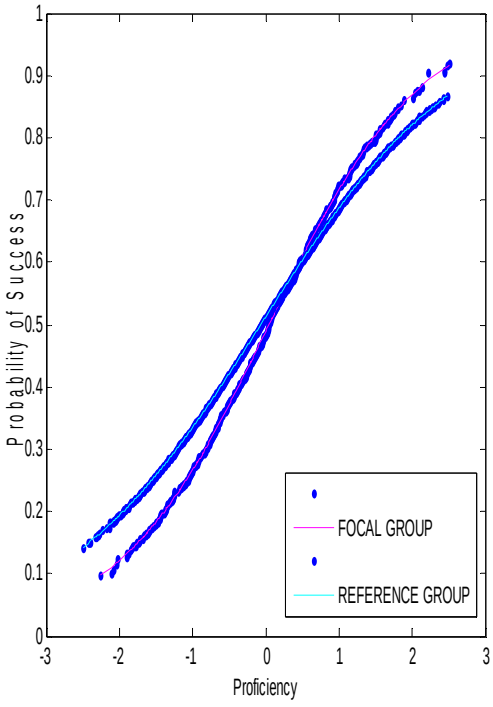


FIGURE 103: ICC of M3C3C2 (38)

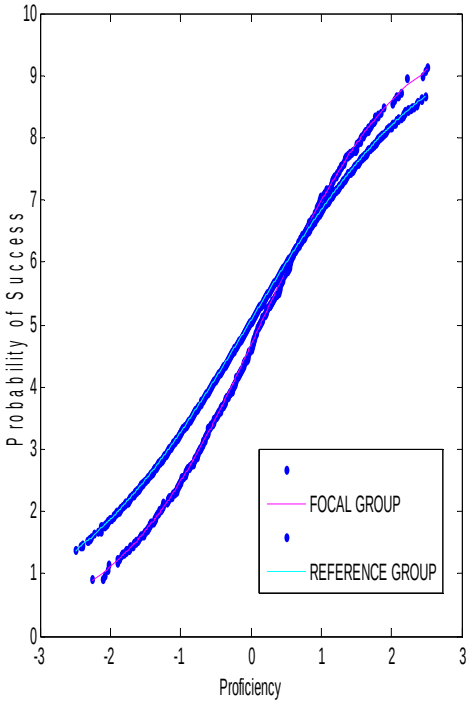


FIGURE 104: TCC of M3C3C2 (38)

Appendix D: Item Characteristic Curves and Testlet Characteristic Curves for Gender Example

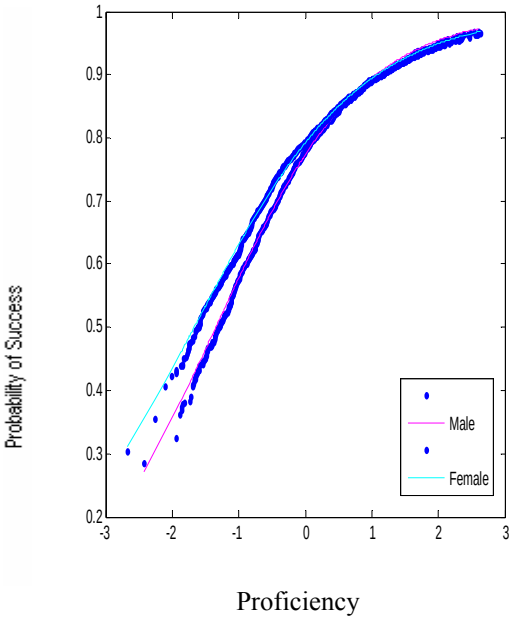


FIGURE 1: ICC of Item 1

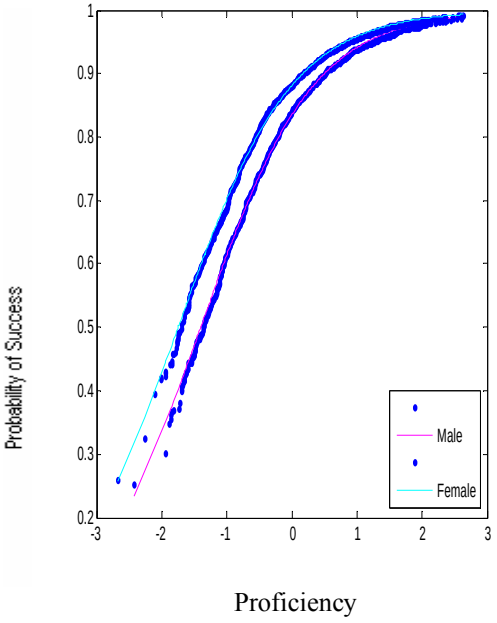


FIGURE 2: ICC of Item 2

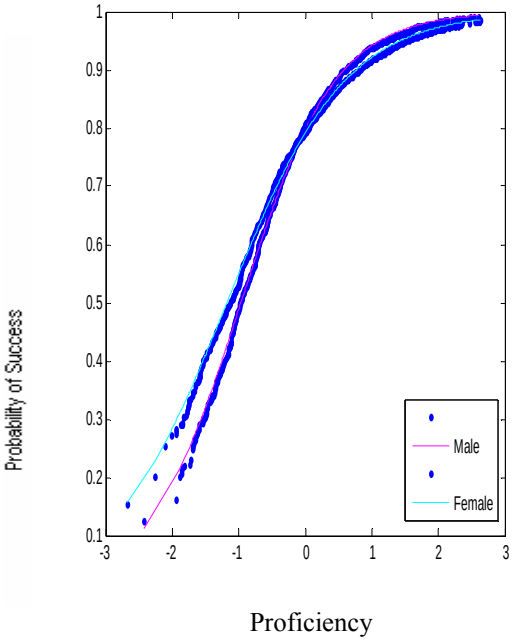


FIGURE 3: ICC of Item 3

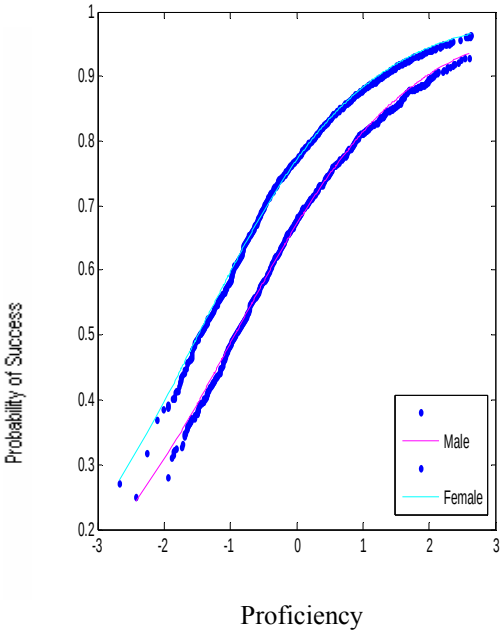


FIGURE 4: ICC of item 4

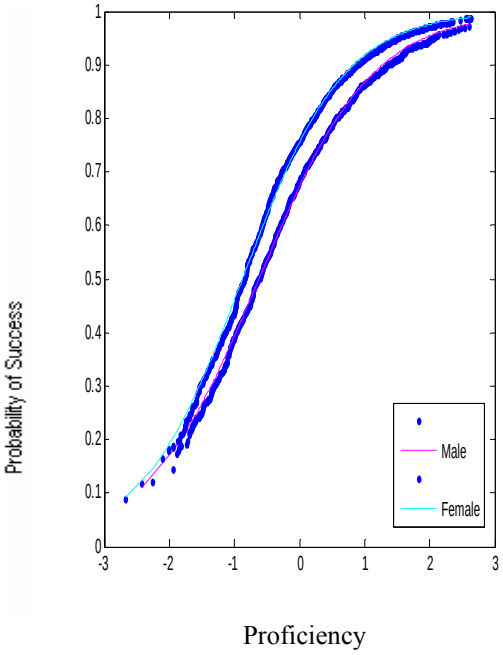


FIGURE 5: ICC of Item 5

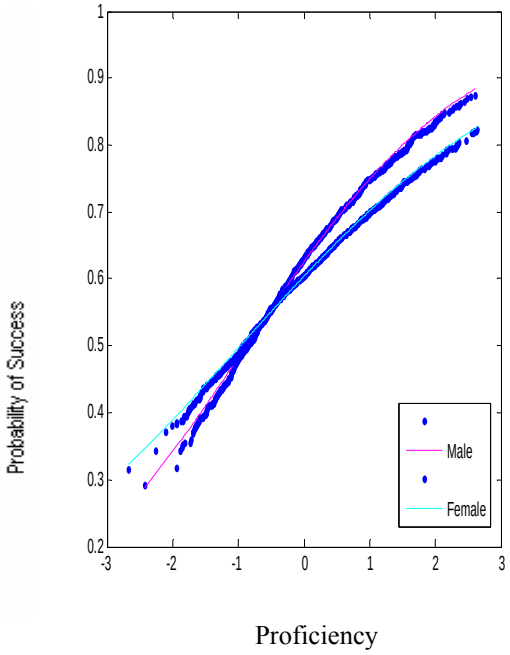


FIGURE 6: ICC of Item 6

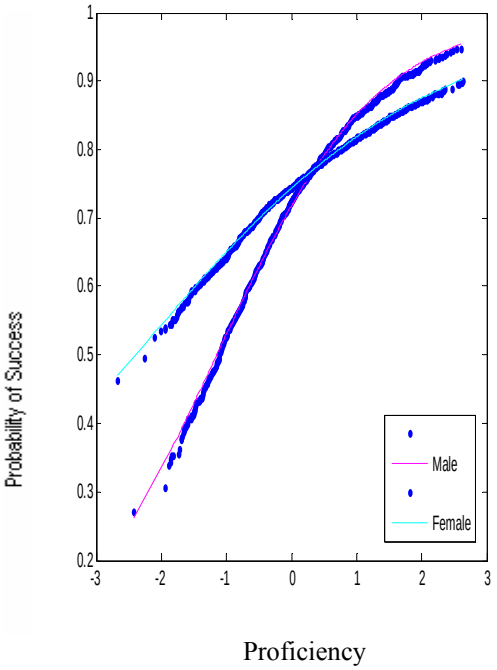


FIGURE 7: ICC of Item 7

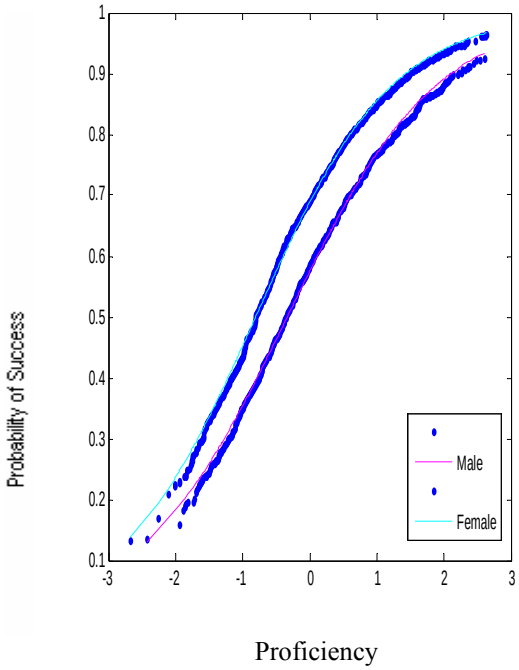


FIGURE 8: ICC of Item 8

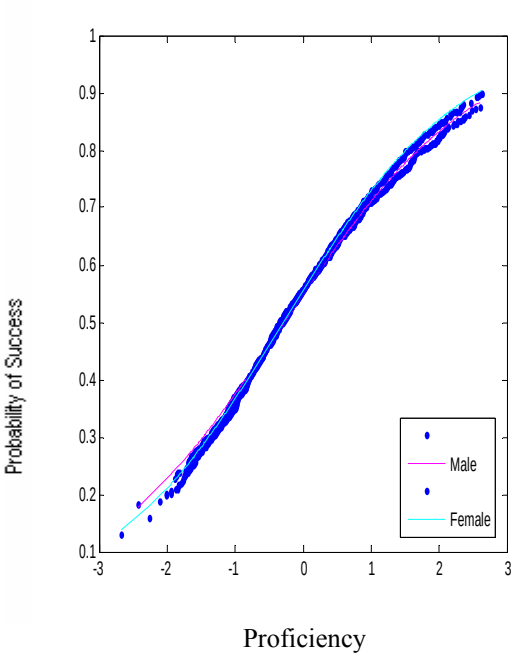


FIGURE 9: ICC of Item 9

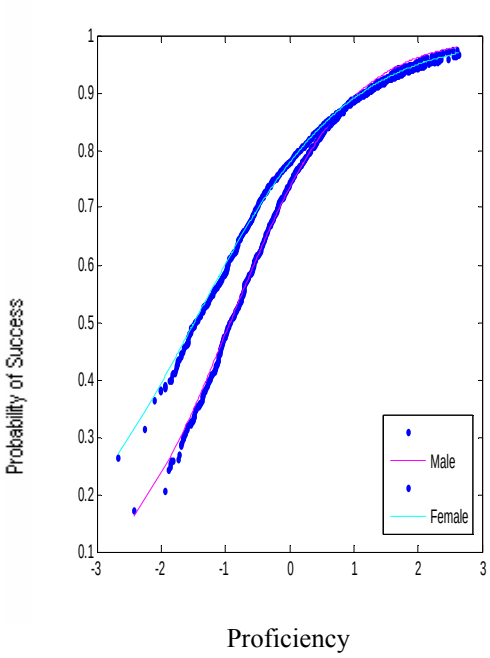


FIGURE 10: ICC of Item 10

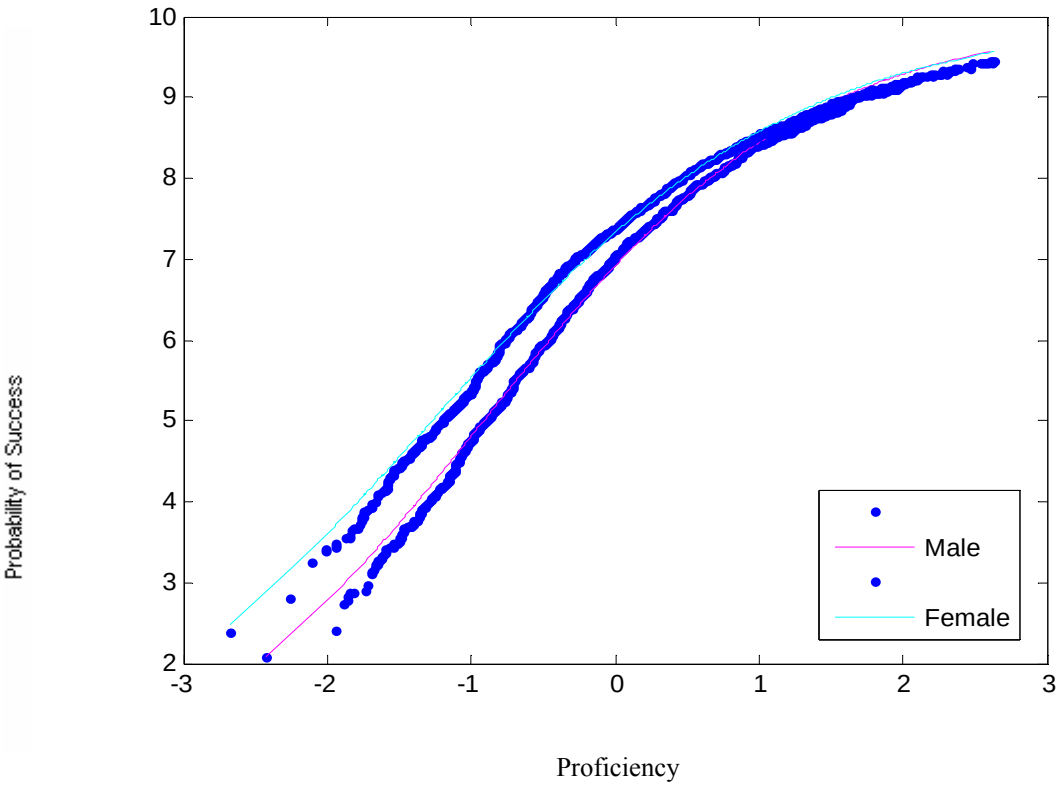


FIGURE 11: TCC of Testlet A

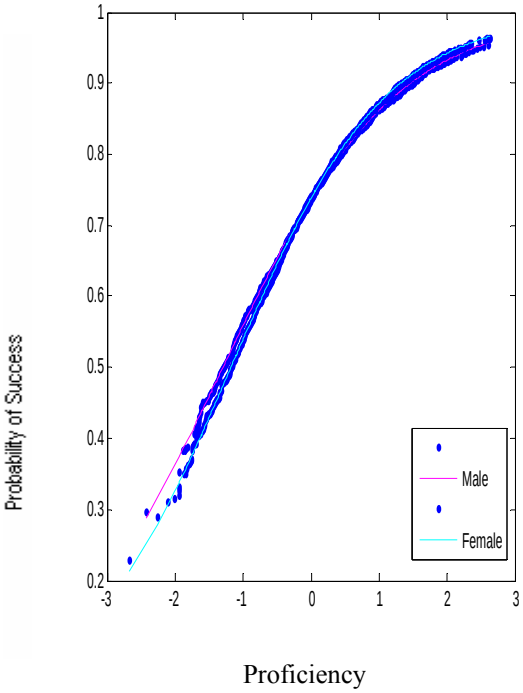


FIGURE 12: ICC of Item 11

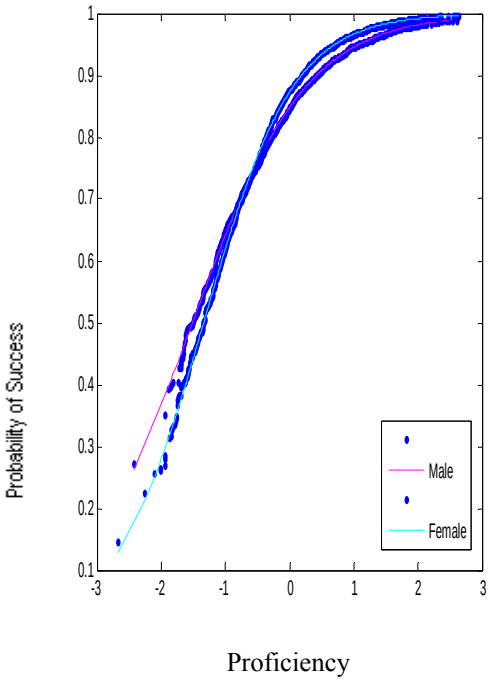


FIGURE 13: ICC of Item 12

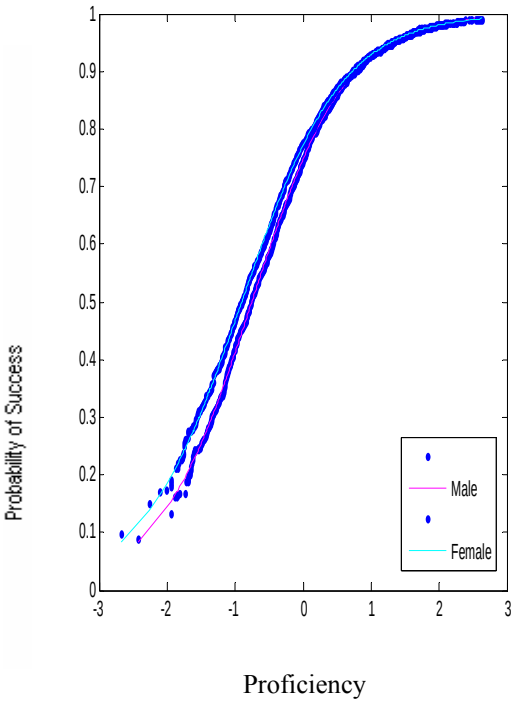


FIGURE 14: ICC of Item 13

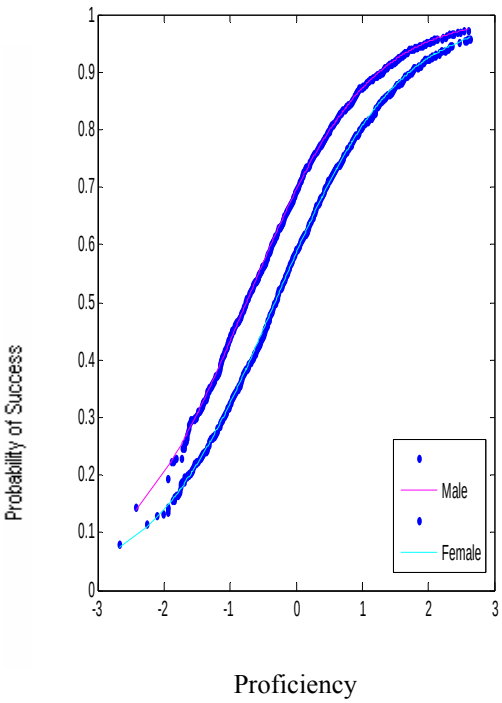


FIGURE 15: ICC of Item 14

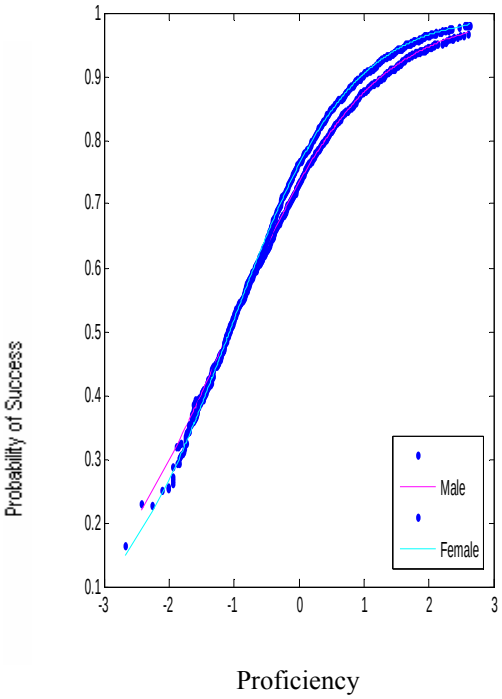


FIGURE 16: ICC of Item 15

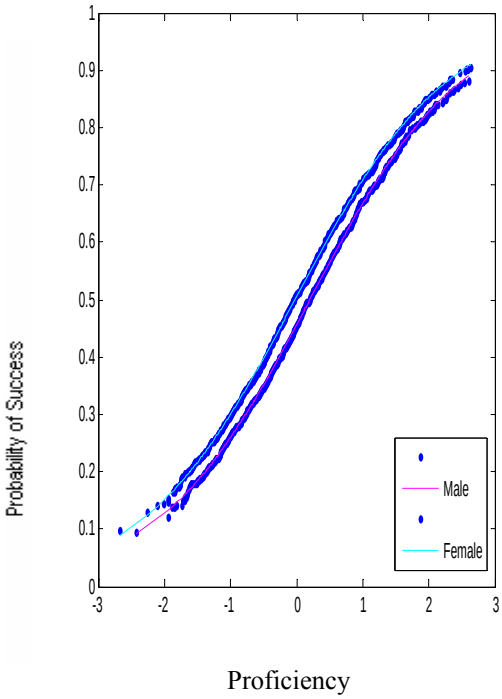


FIGURE 17: ICC of Item 16

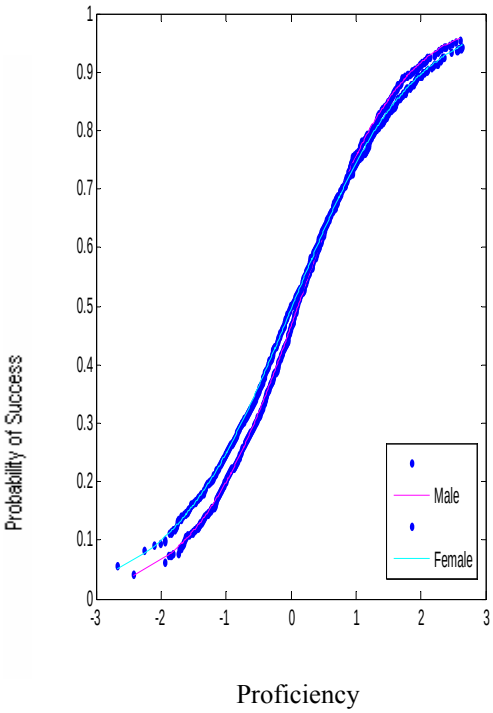


FIGURE 18: ICC of Item 17

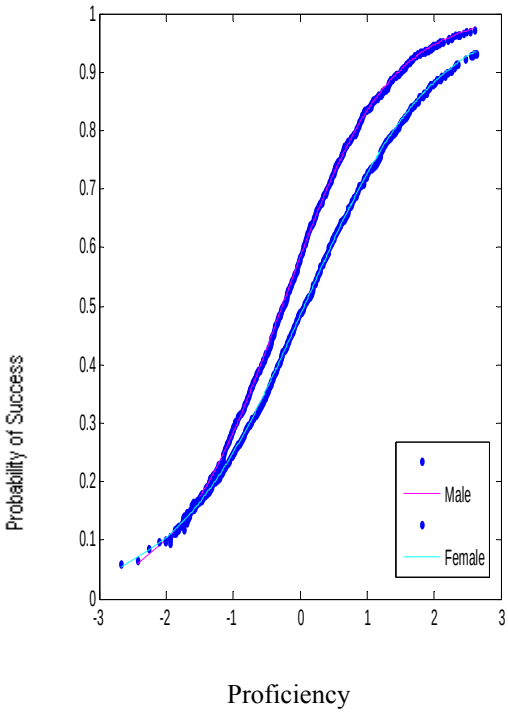


FIGURE 19: ICC of Item 18

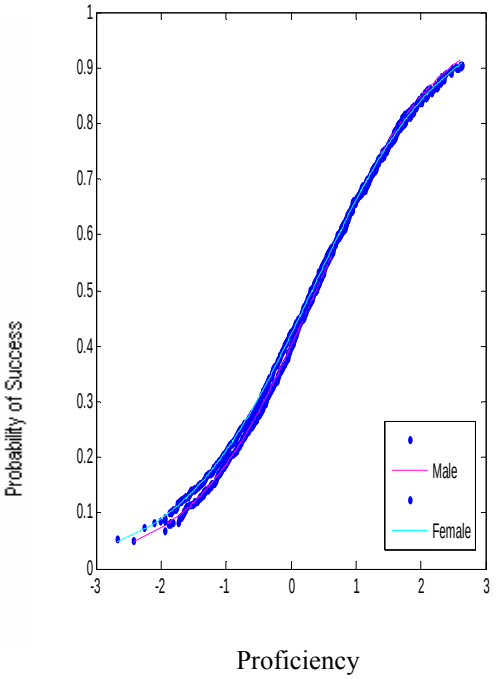


FIGURE 20: ICC of Item 19

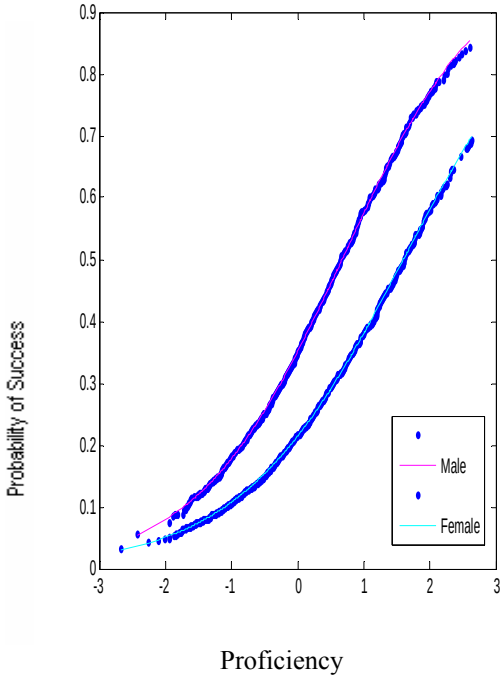


FIGURE 21: ICC of Item 20

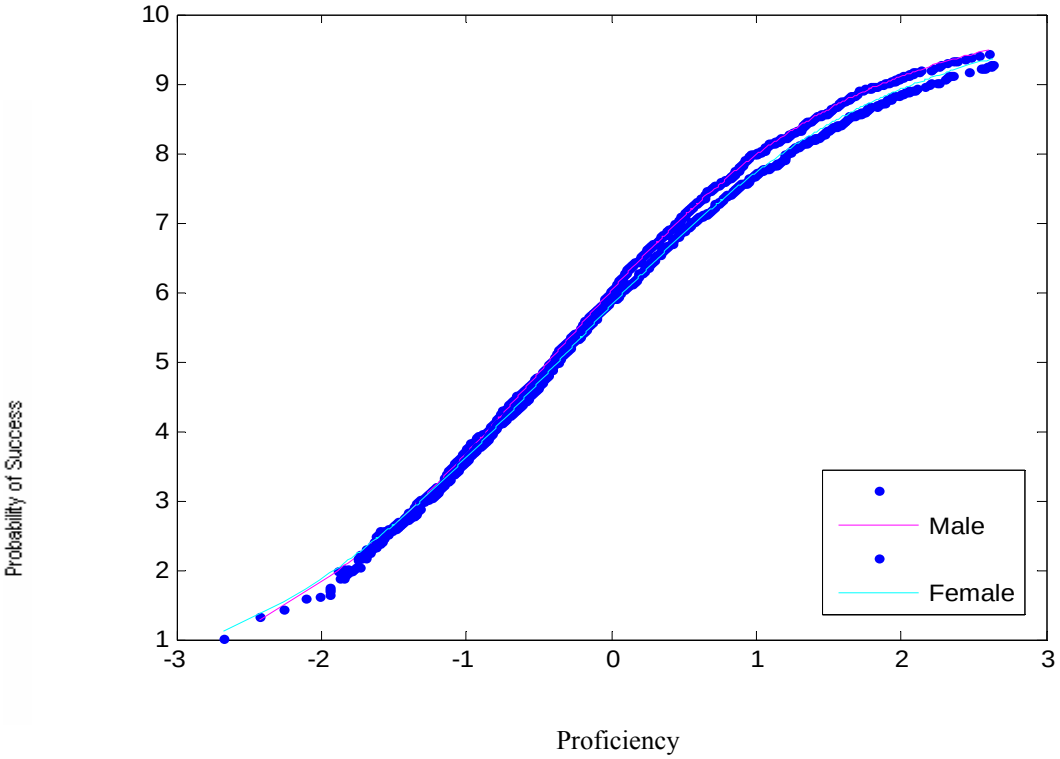


FIGURE 22: TCC of Testlet B

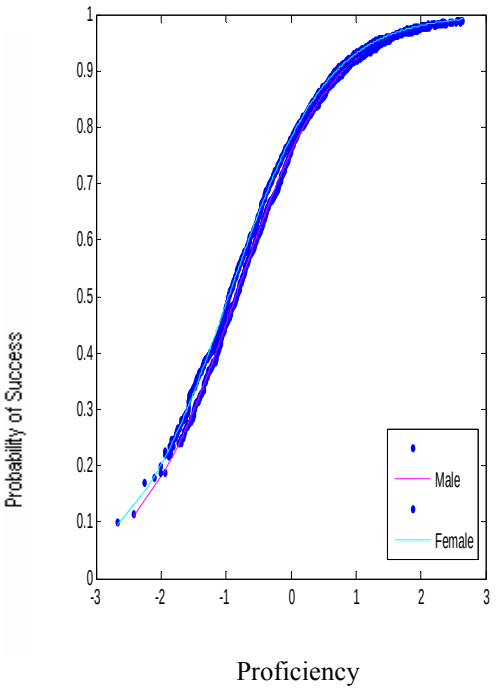


FIGURE 23: ICC of Item 21

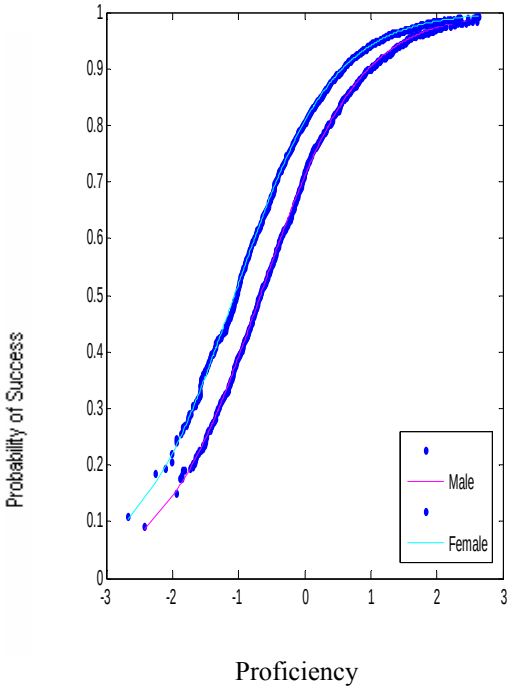


FIGURE 24: ICC of Item 22

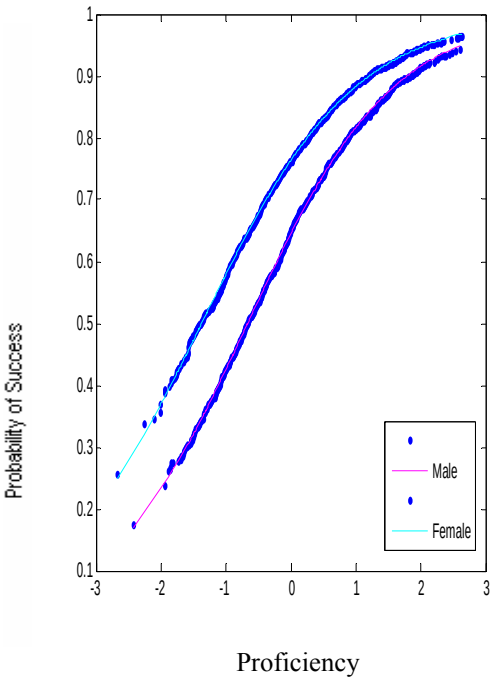


FIGURE 25: ICC of Item 23

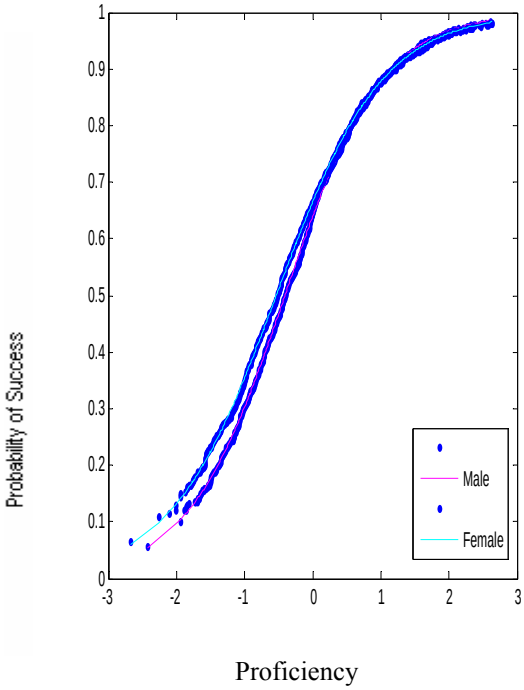


FIGURE 26: ICC of Item 24

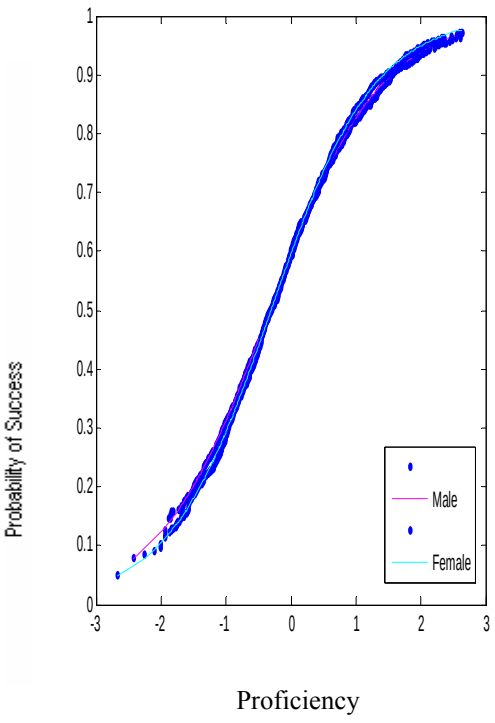


FIGURE 27: ICC of Item 25

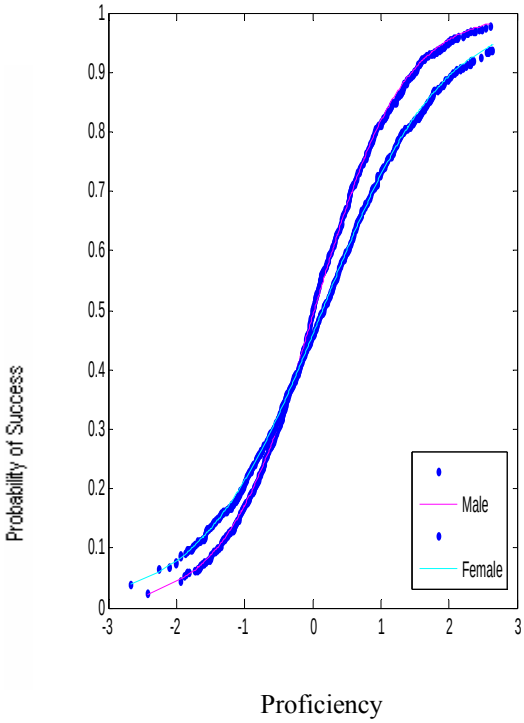


FIGURE 28: ICC of Item 26

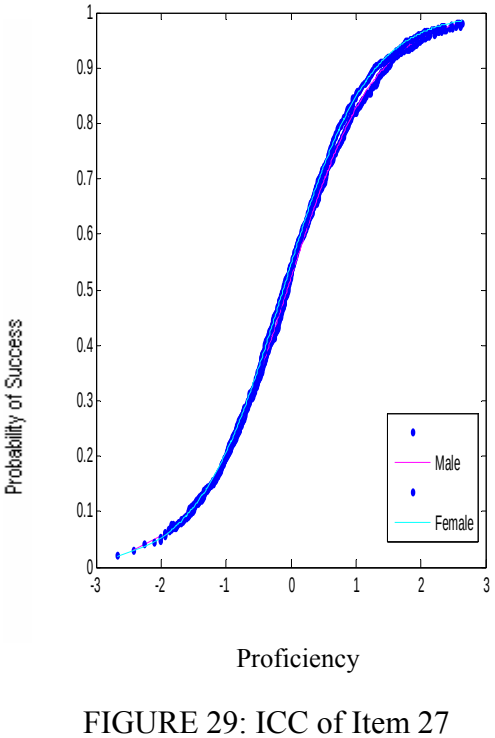


FIGURE 29: ICC of Item 27

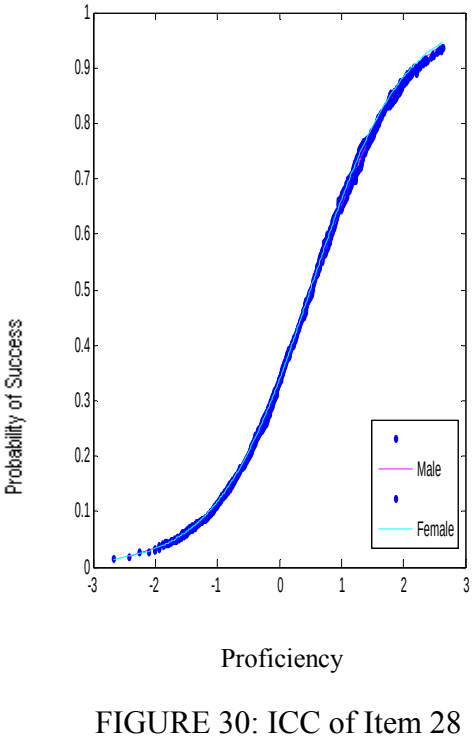


FIGURE 30: ICC of Item 28

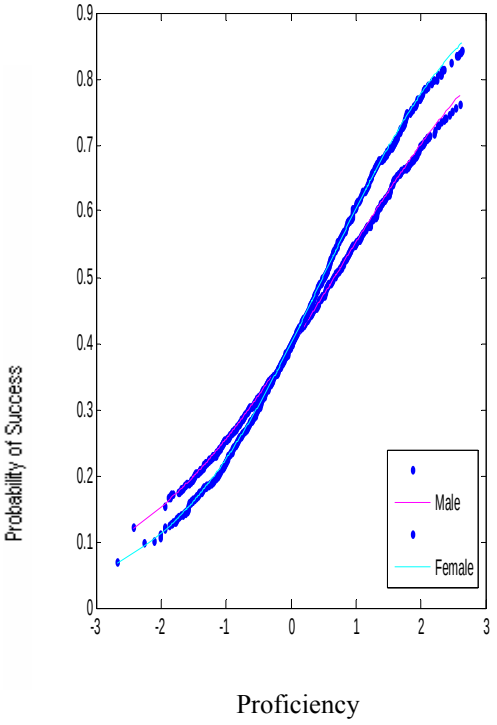


FIGURE 31: ICC of Item 29

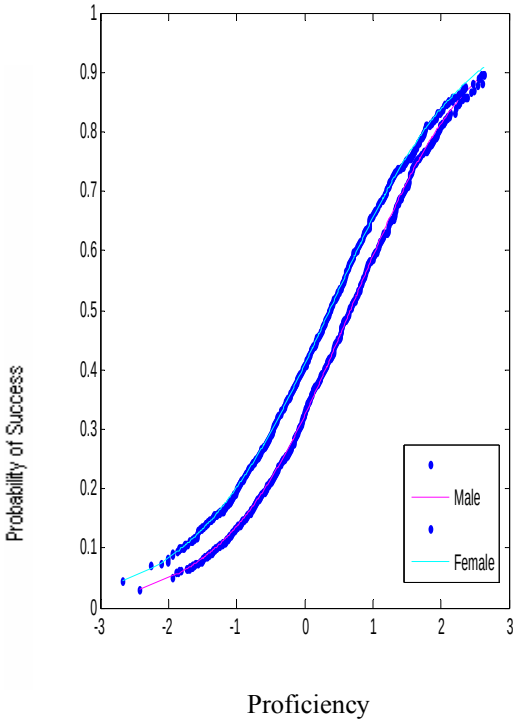


FIGURE 32: ICC of Item 30

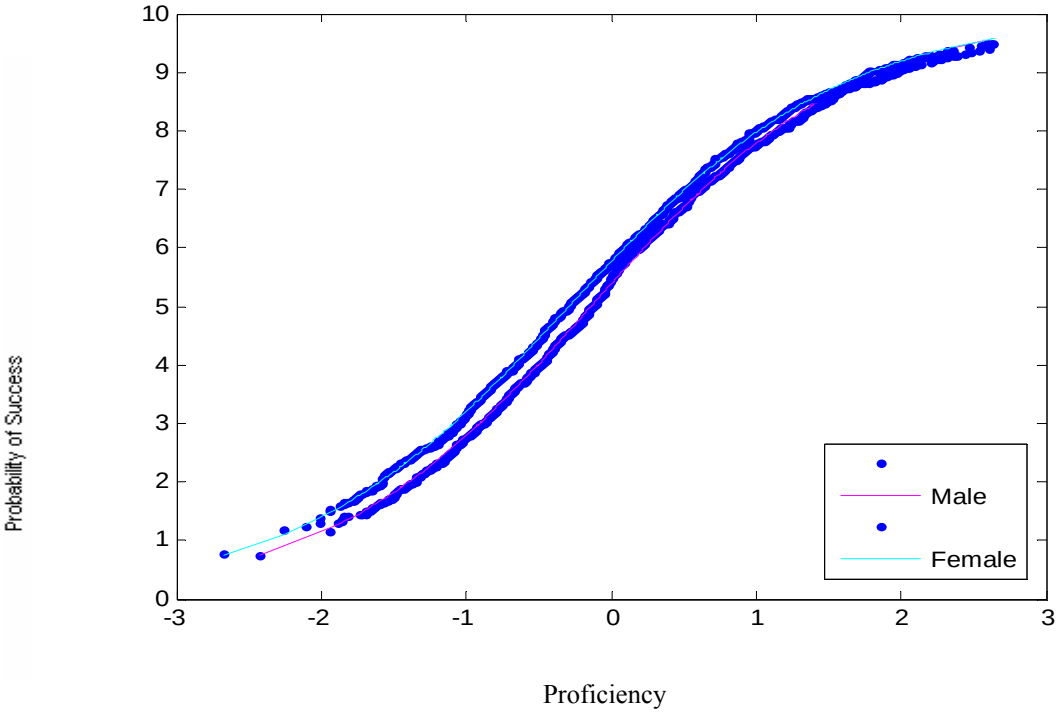


FIGURE 33: TCC of Testlet C

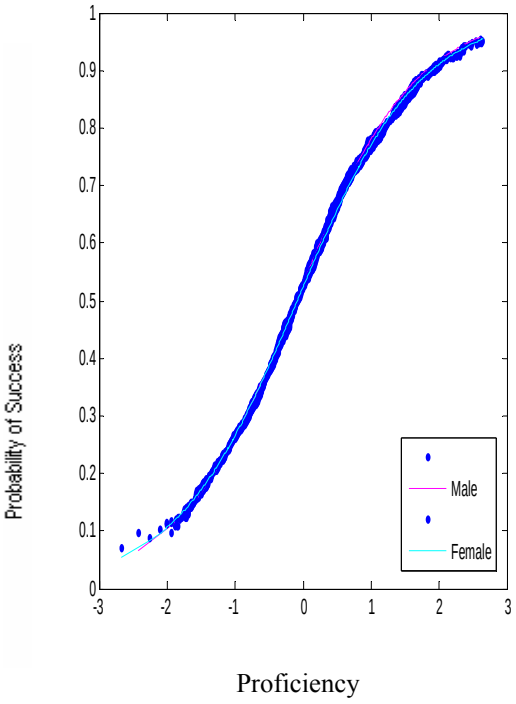


FIGURE 34: ICC of Item 31

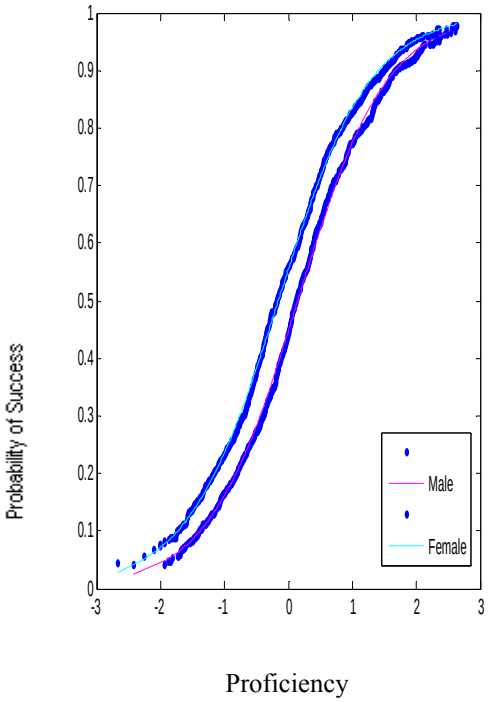


FIGURE 35: ICC of Item 32

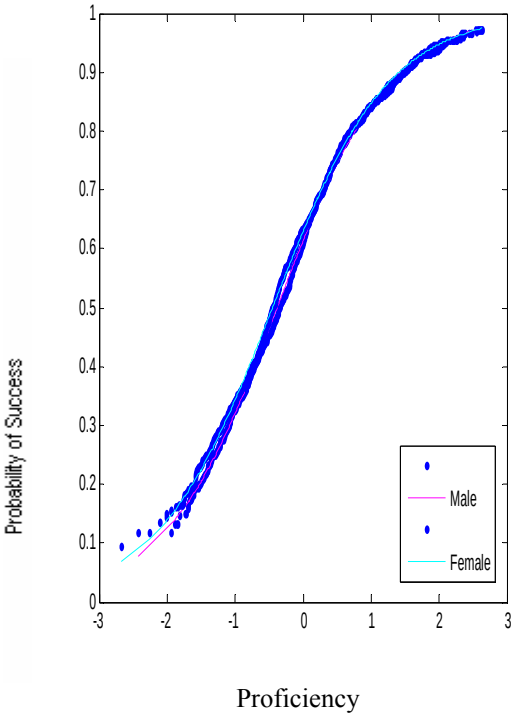


FIGURE 36: ICC of Item 33

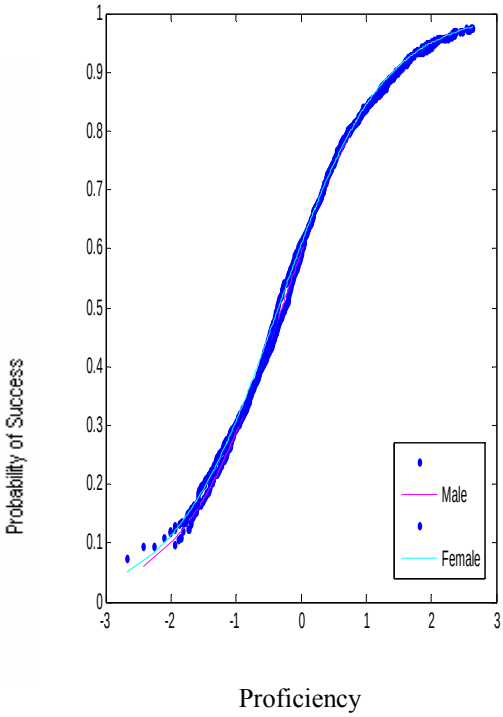


FIGURE 37: ICC of Item 34

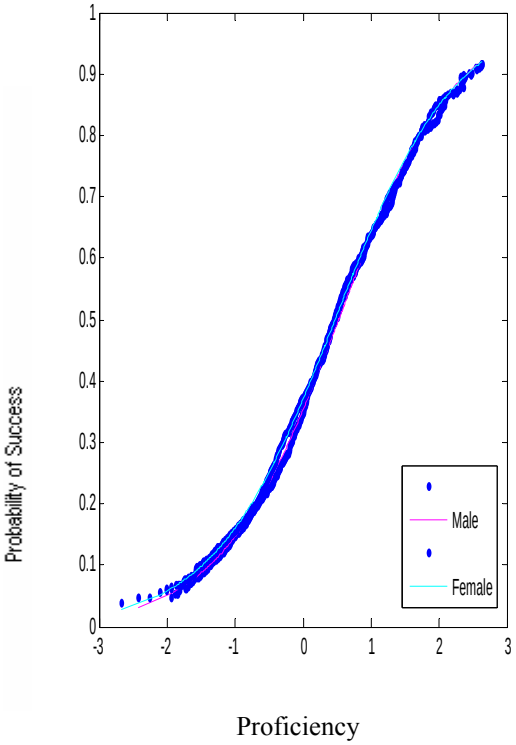


FIGURE 38: ICC of Item 35

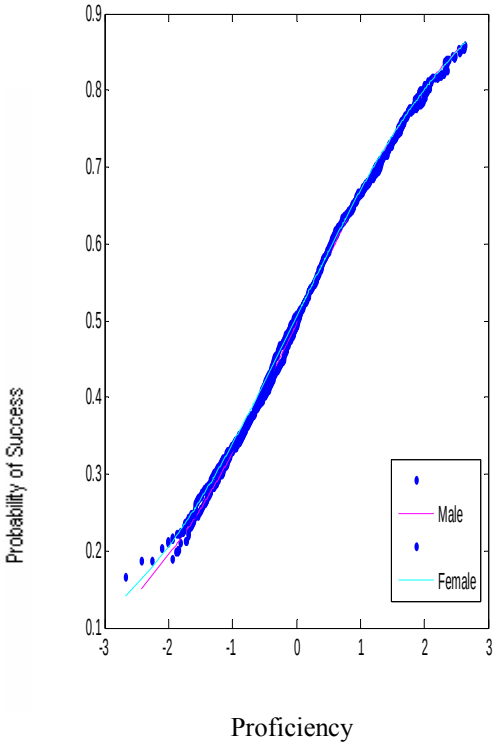


FIGURE 39: ICC of Item 36

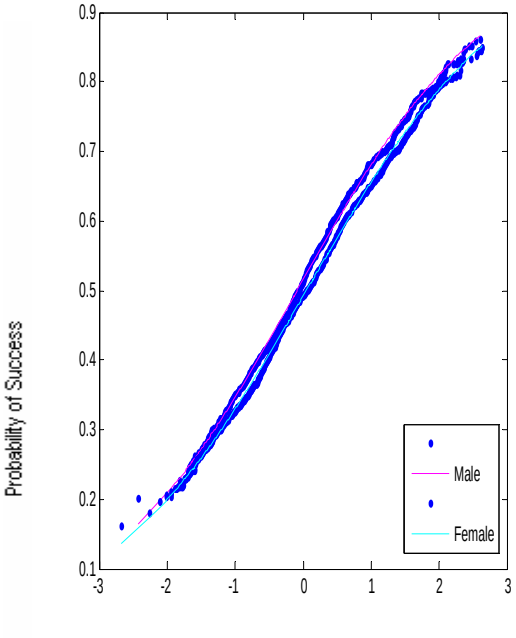


FIGURE 40: ICC of Item 37

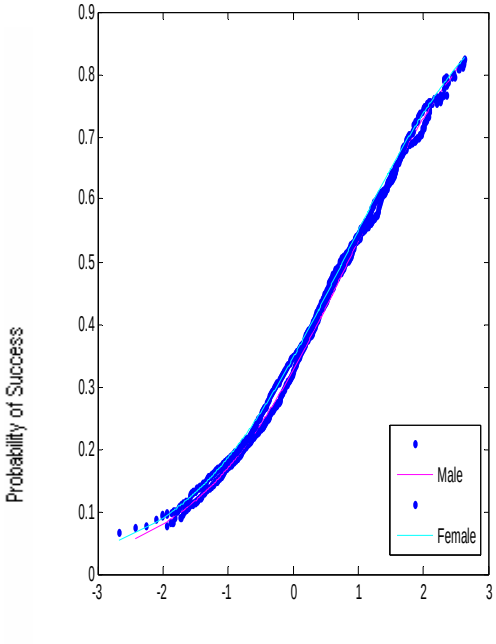


FIGURE 41: ICC of Item 38

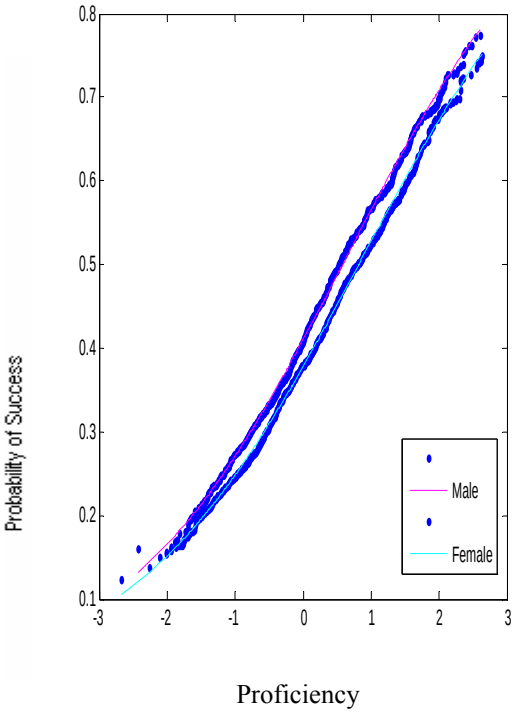


FIGURE 42: ICC of Item 39

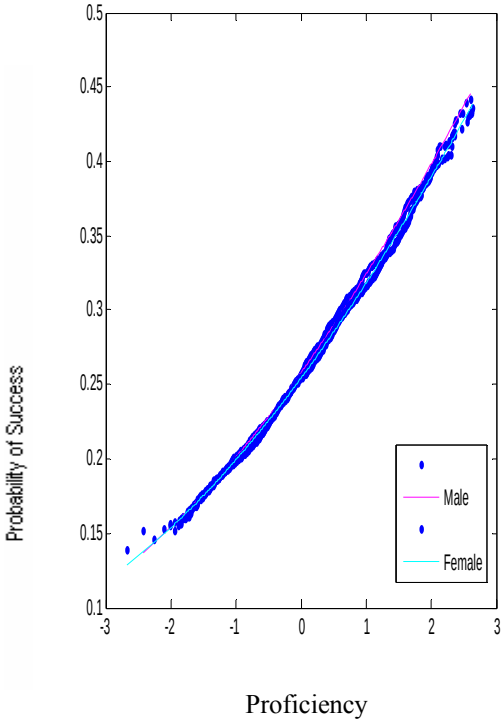


FIGURE 43: ICC of Item 40

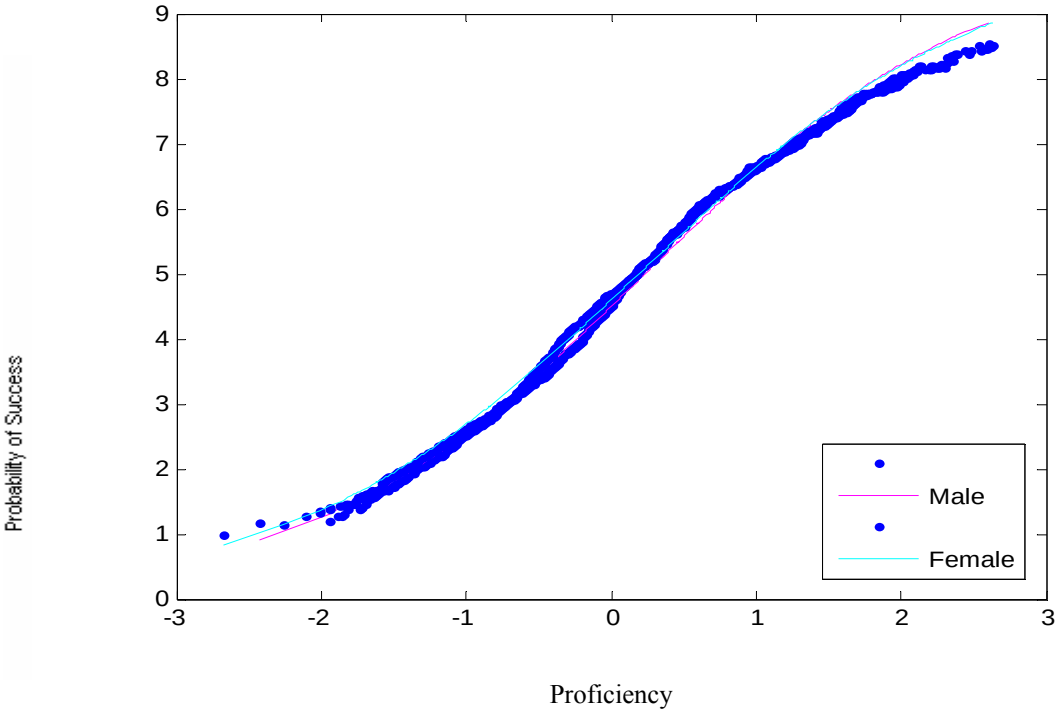


FIGURE 44: TCC of Testlet D

Appendix E: Item Characteristic Curves and Testlet Characteristic Curves for Ethnic Example

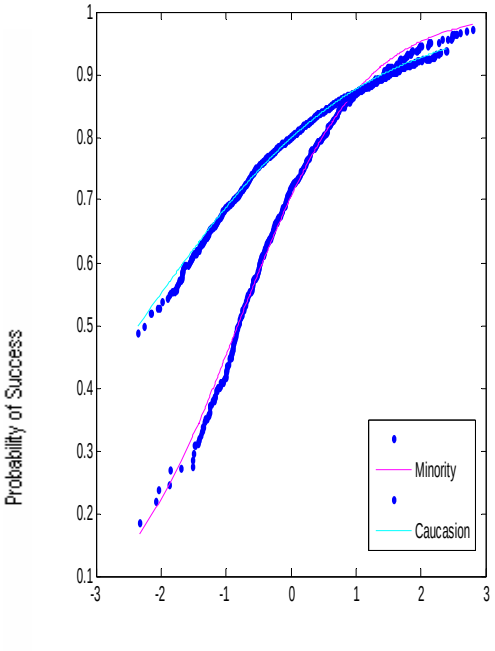


FIGURE 1: ICC of Item 1

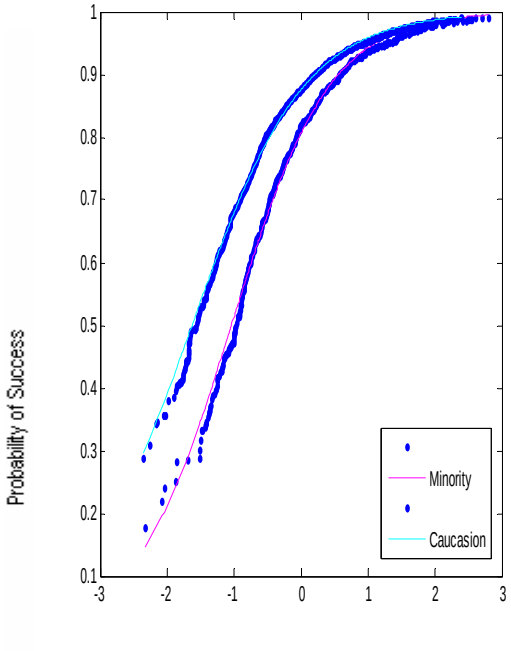


FIGURE 2: ICC of Item 2

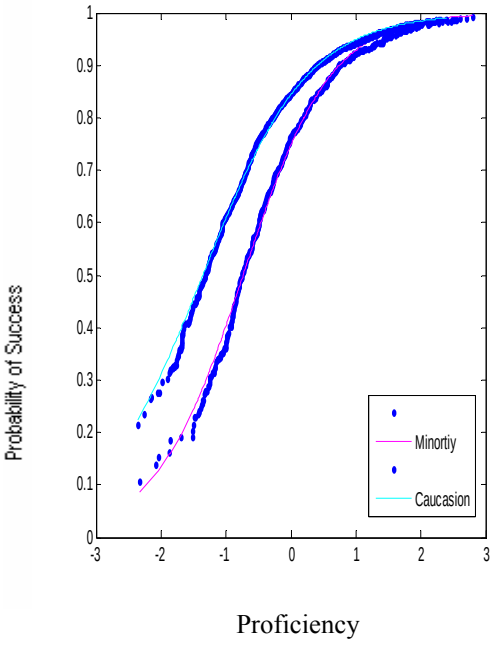


FIGURE 3: ICC of Item 3

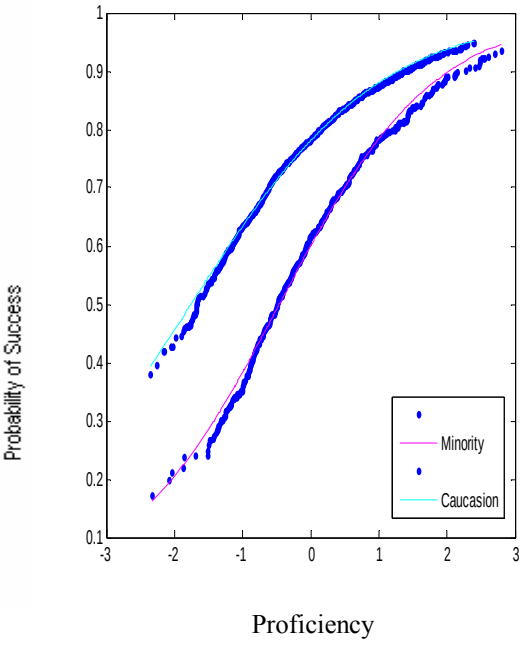


FIGURE 4: ICC of Item 4

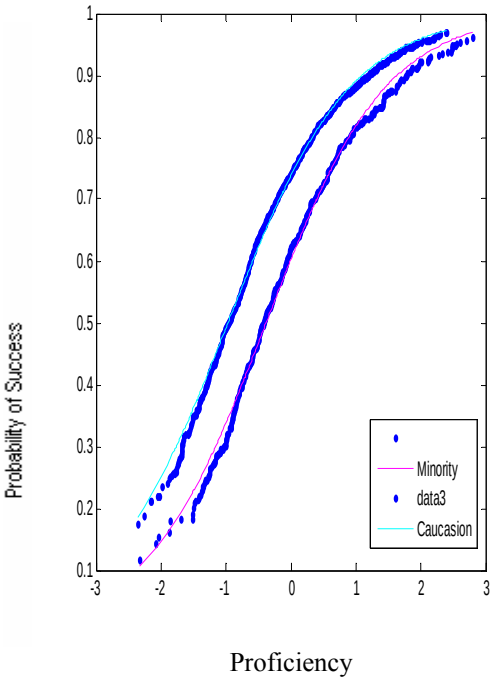


FIGURE 5: ICC of Item 5

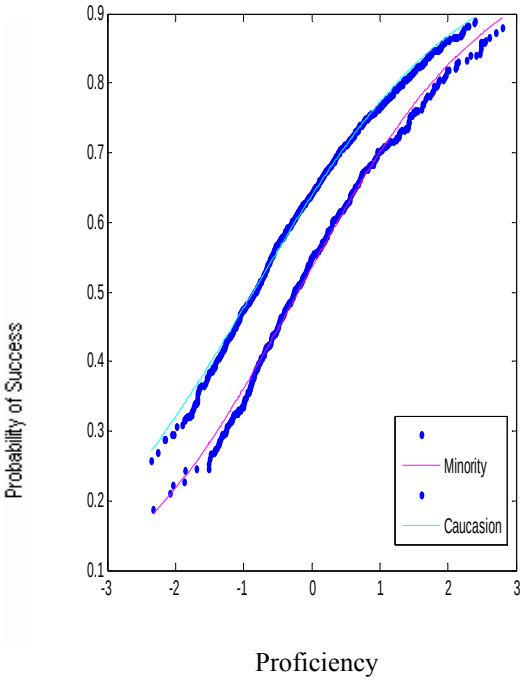
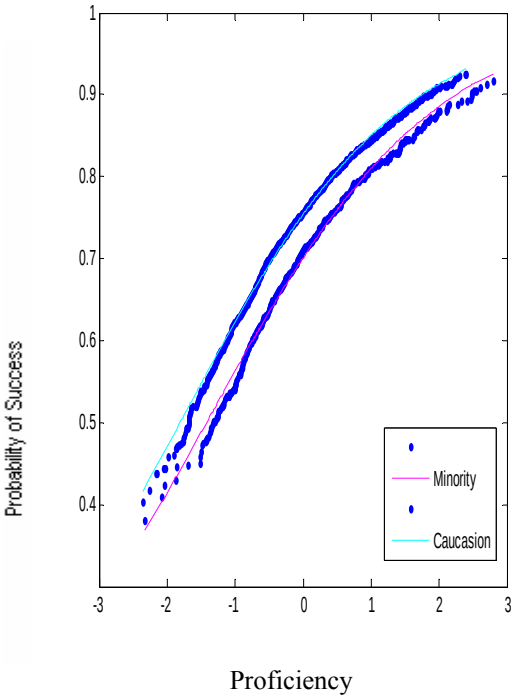


FIGURE 6: ICC of Item 6



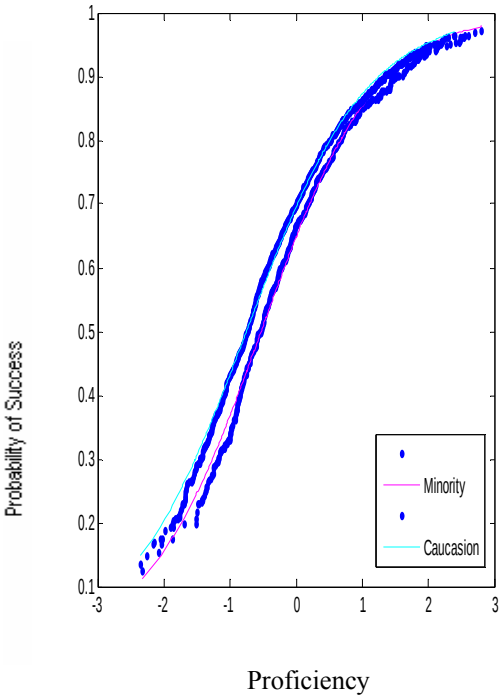


FIGURE 7: ICC of Item 7

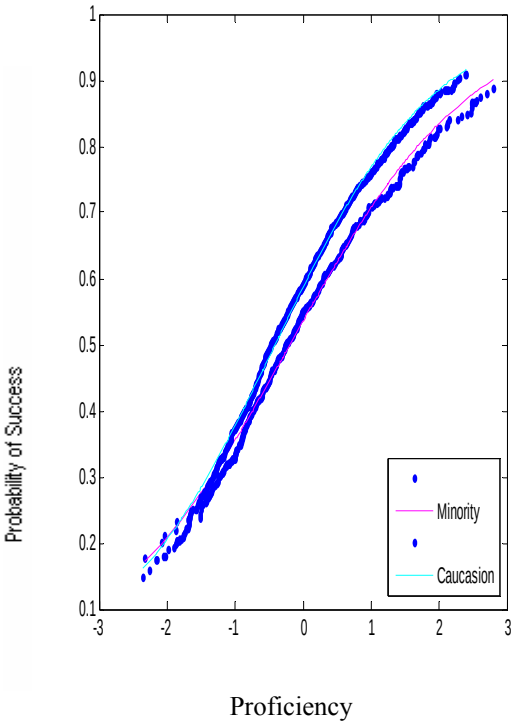


FIGURE 8: ICC of Item 8

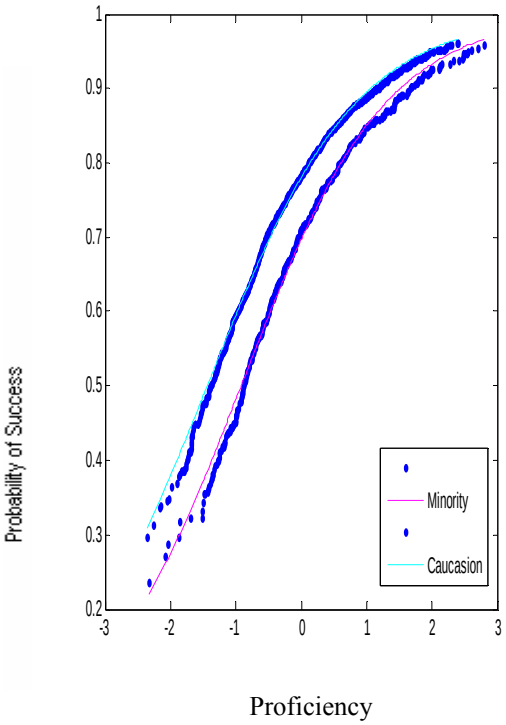


FIGURE 9: ICC of Item 9

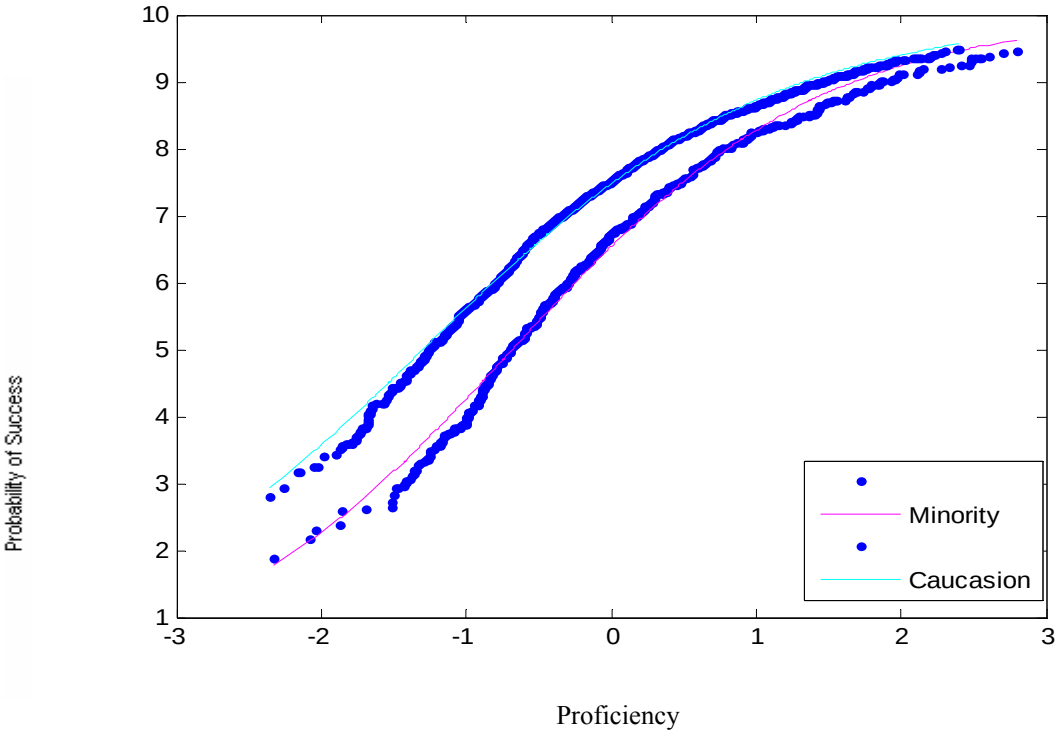


FIGURE 10: ICC of Item 10

FIGURE 11: TCC of Testlet A

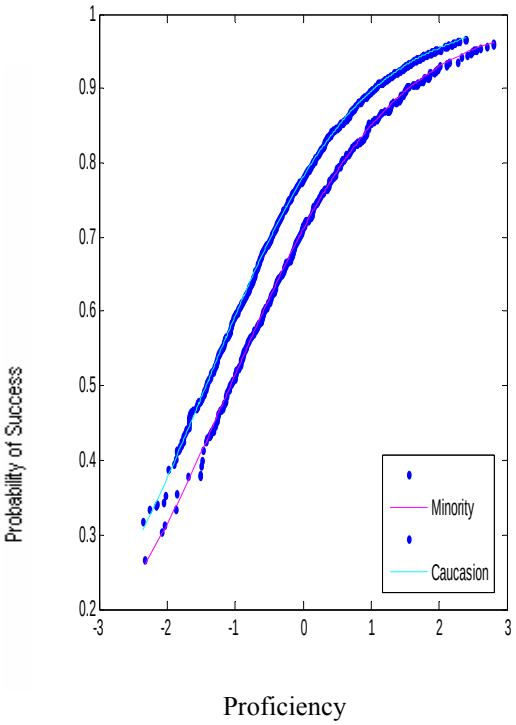


FIGURE 12: ICC of Item 11

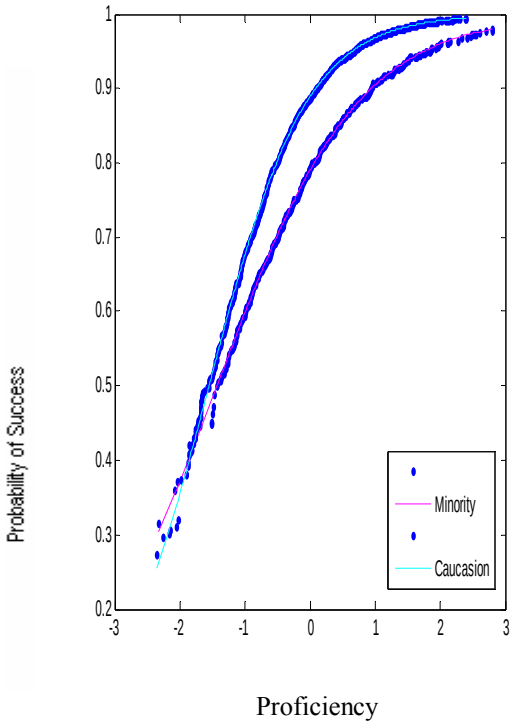


FIGURE 13: ICC of Item 12

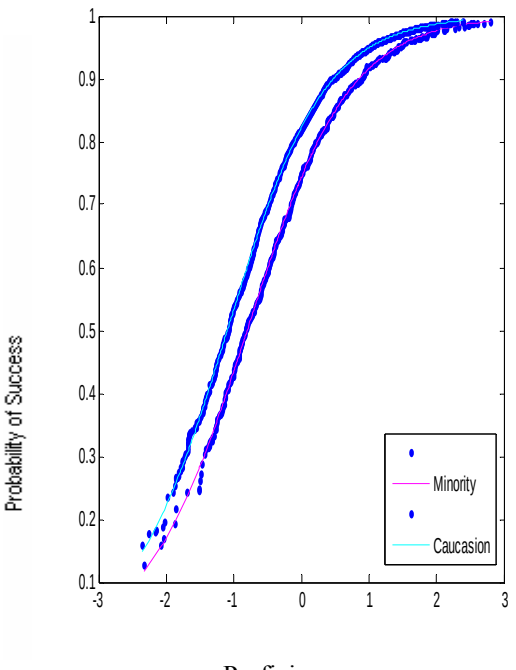


FIGURE 14: ICC of Item 13

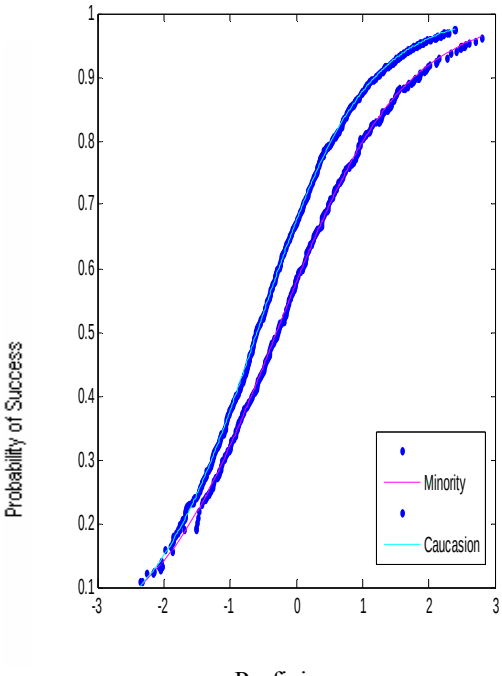


FIGURE 15: ICC of Item 14

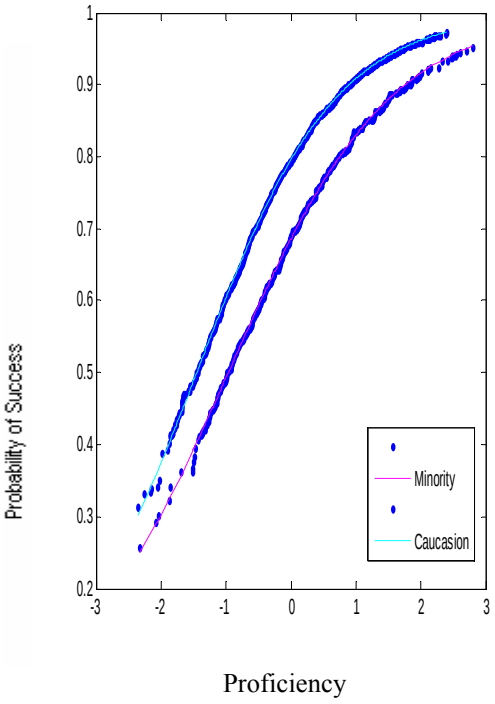


FIGURE 16: ICC of Item 15

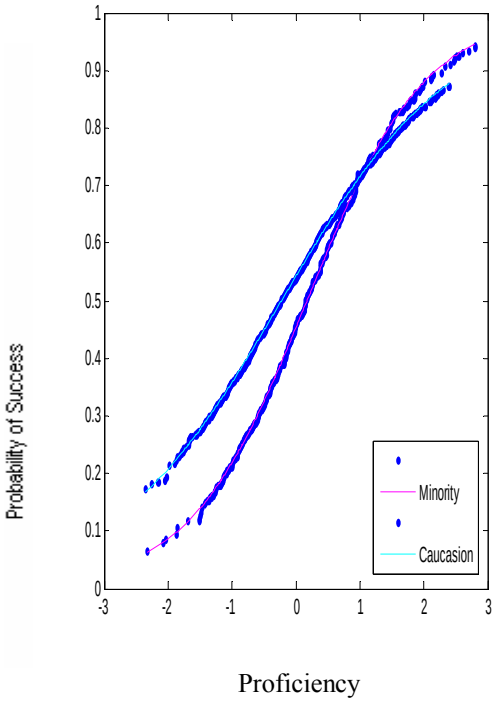


FIGURE 17: ICC of Item 16

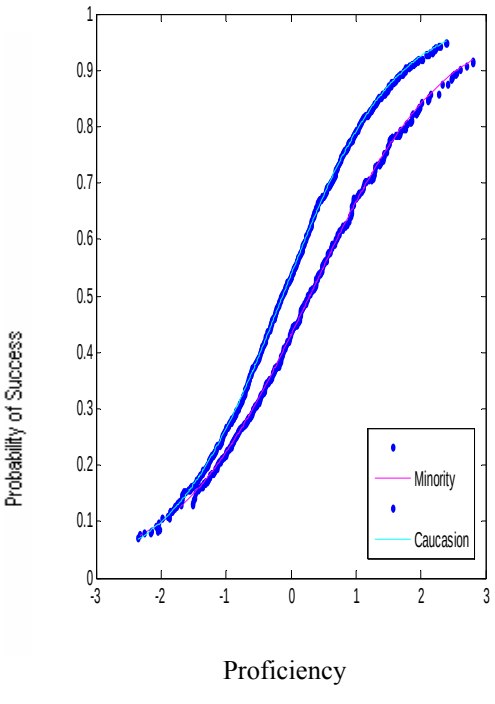


FIGURE 18: ICC of Item 17

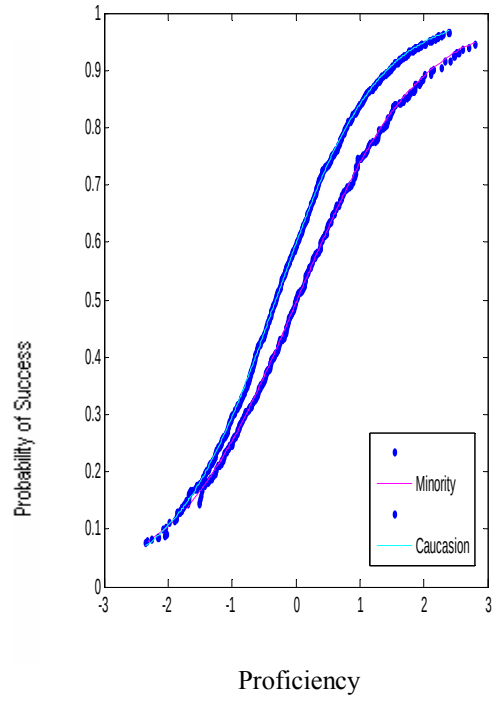


FIGURE 19: ICC of Item 18

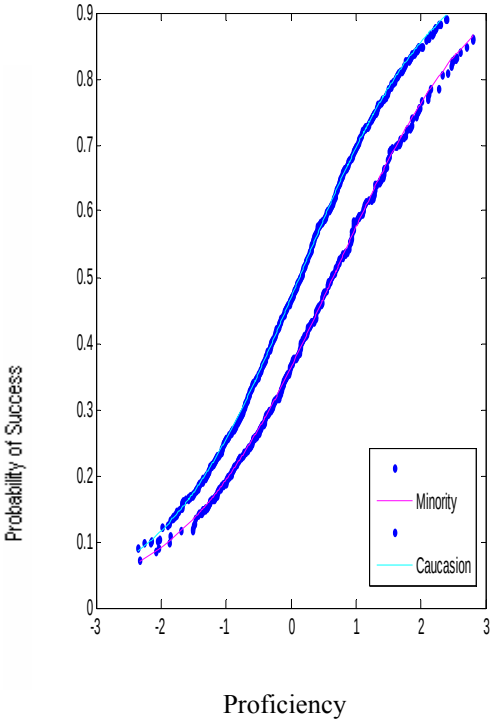


FIGURE 20: ICC of Item 19

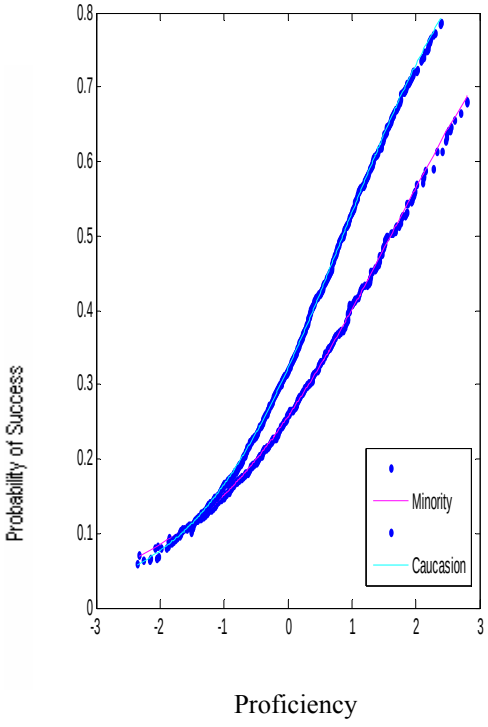


FIGURE 21: ICC of Item 20

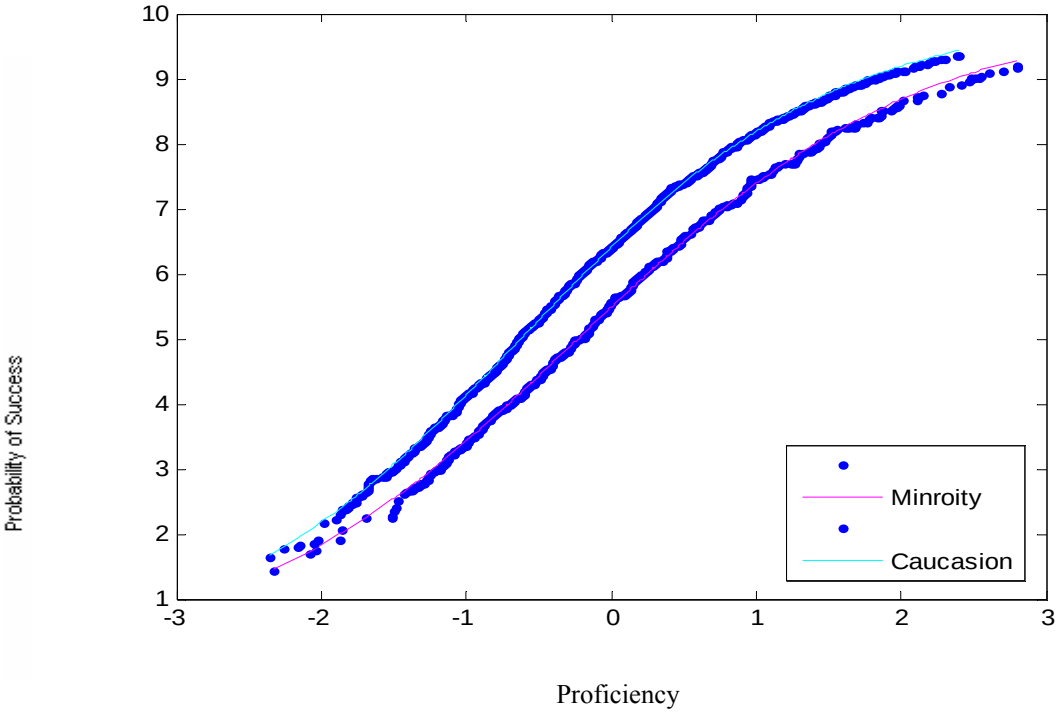


FIGURE 22: TCC of Testlet B

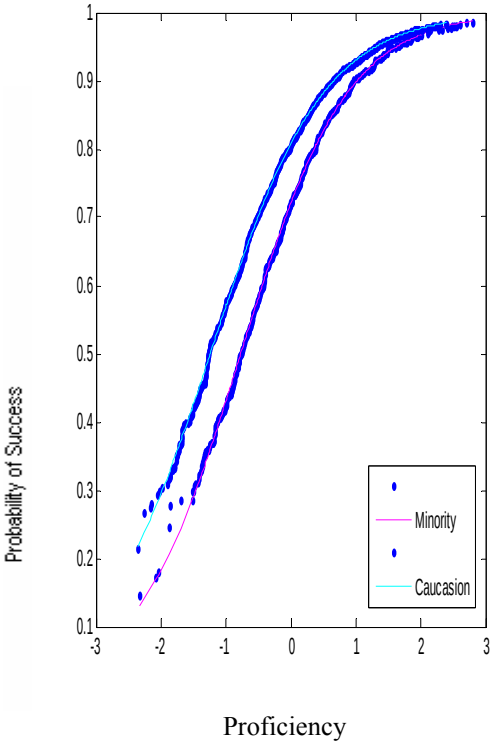


FIGURE 23: ICC of Item 21

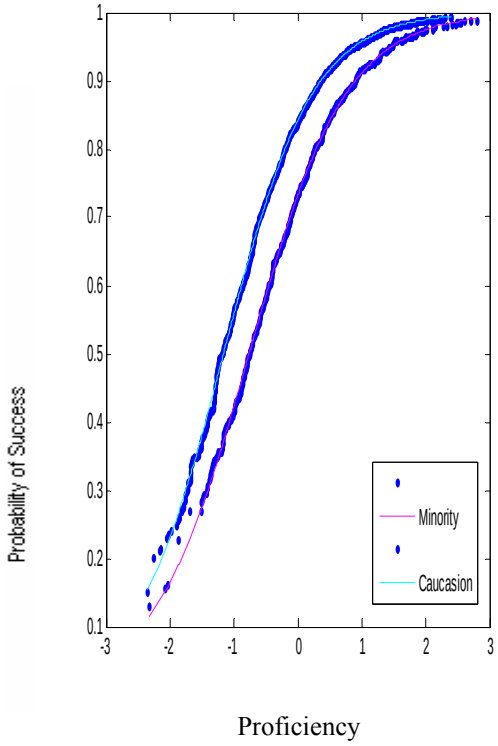


FIGURE 24: ICC of Item 22

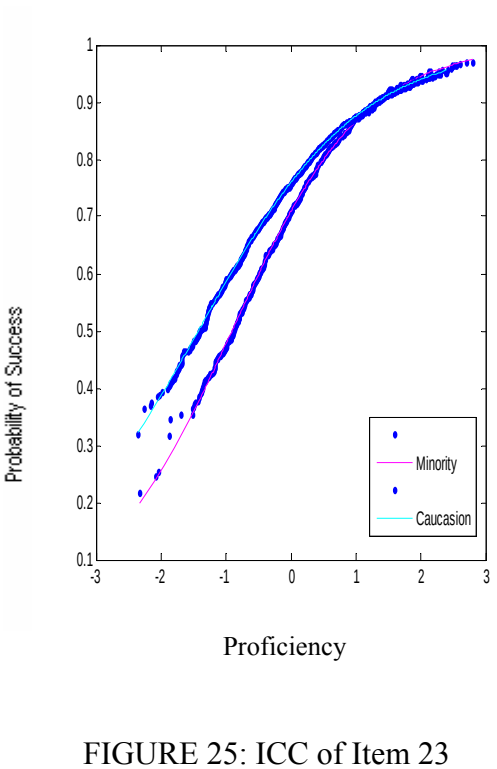


FIGURE 25: ICC of Item 23

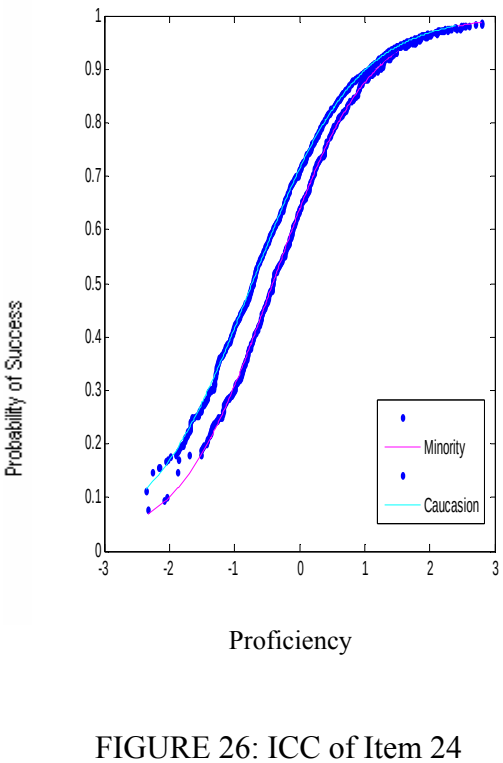


FIGURE 26: ICC of Item 24

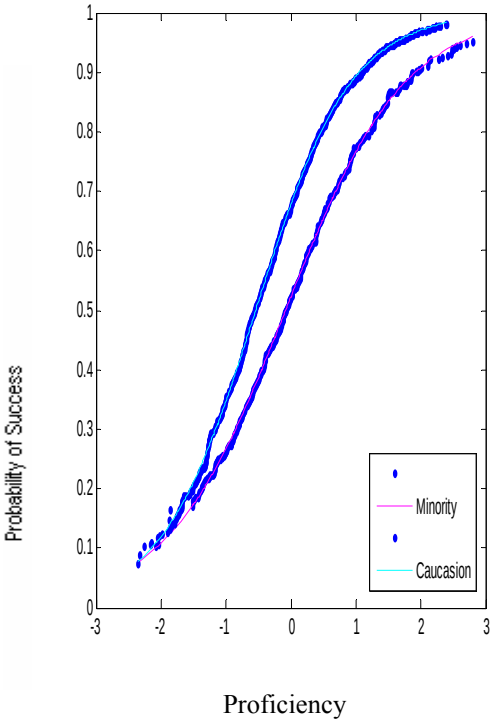


FIGURE 27: ICC of Item 25

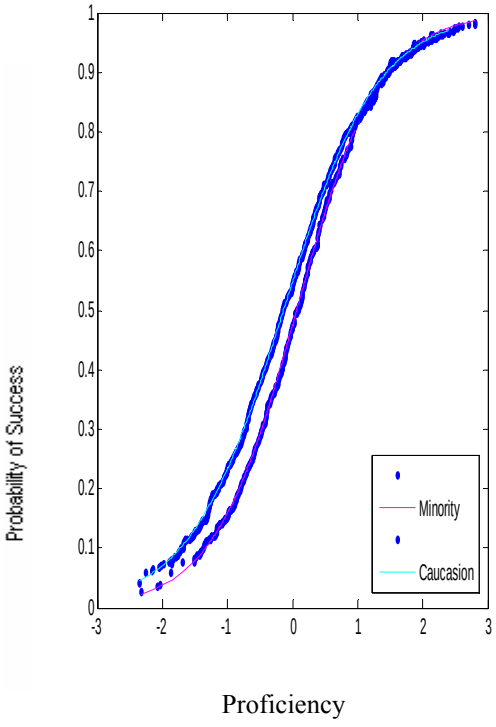


FIGURE 28: ICC of Item 26

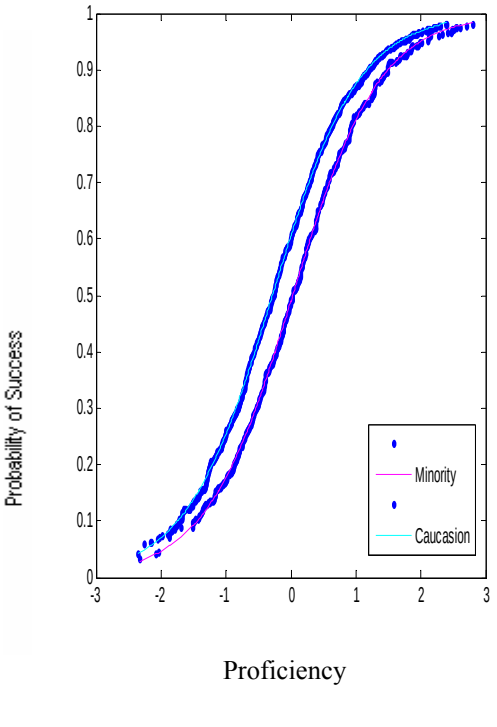


FIGURE 29: ICC of Item 27

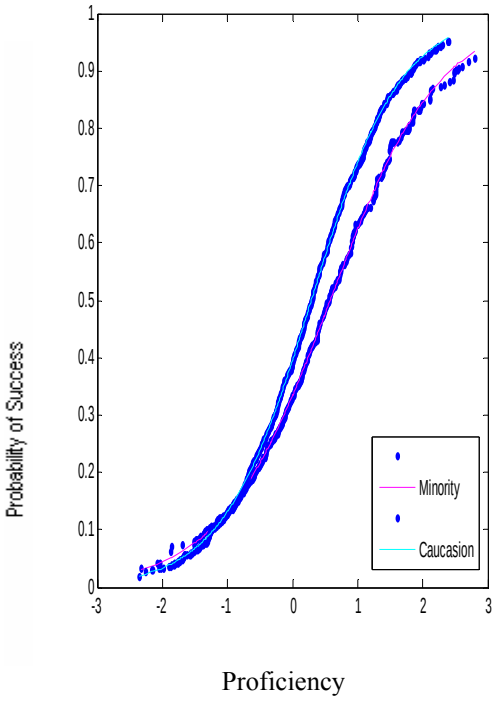


FIGURE 30: ICC of Item 28

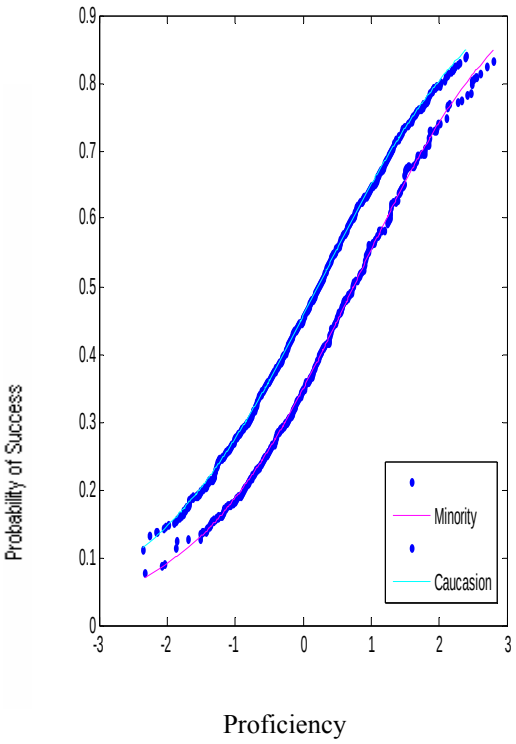


FIGURE 31: ICC of Item 29

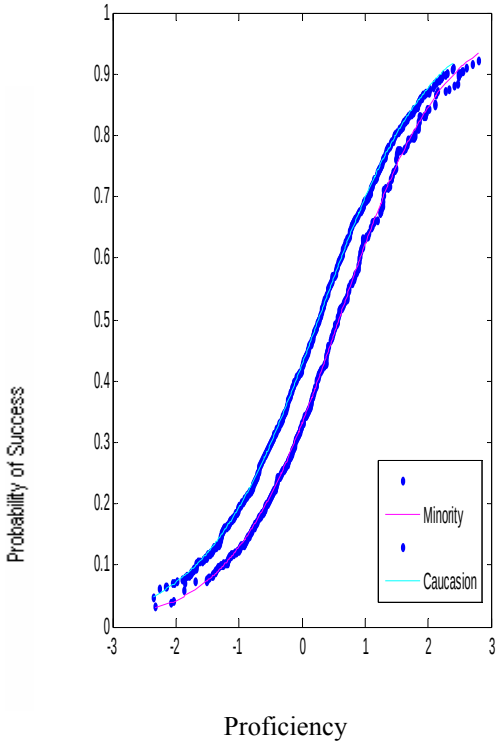


FIGURE 32: ICC of Item 30

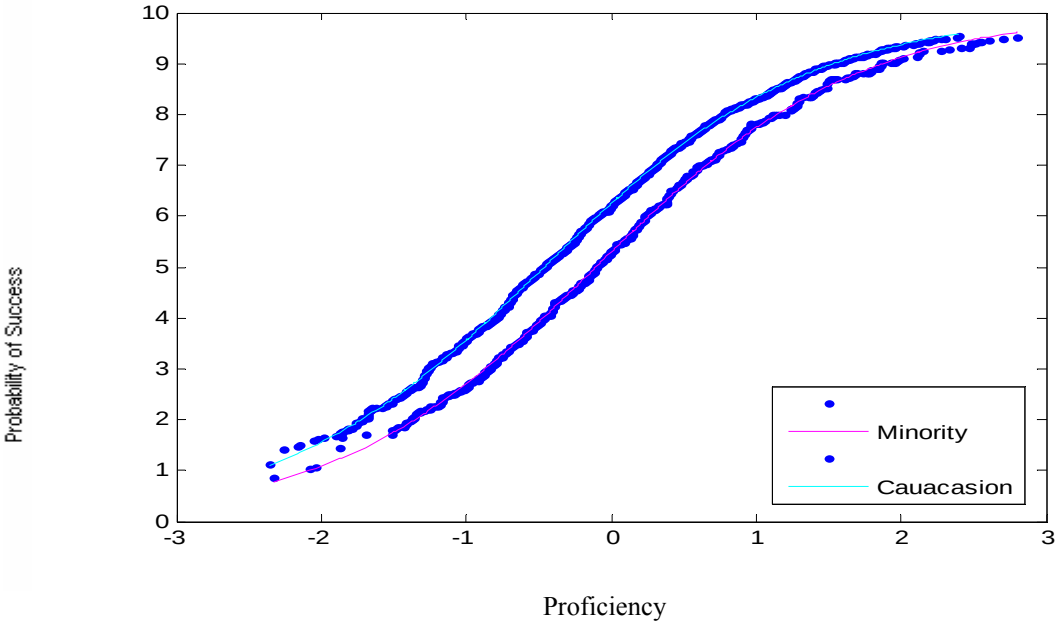


FIGURE 33: TCC of Testlet C

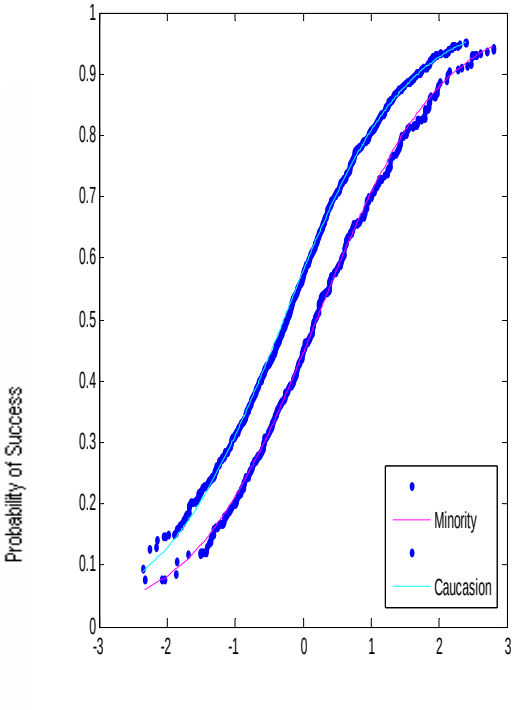


FIGURE 34: ICC of Item 31

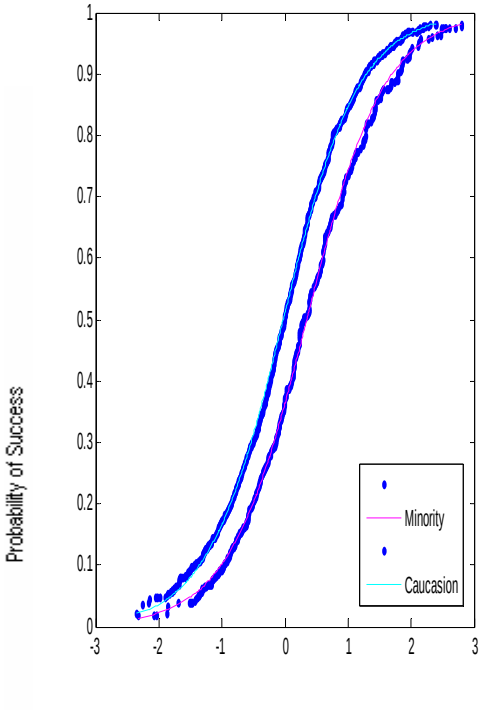


FIGURE 35: ICC of Item 32

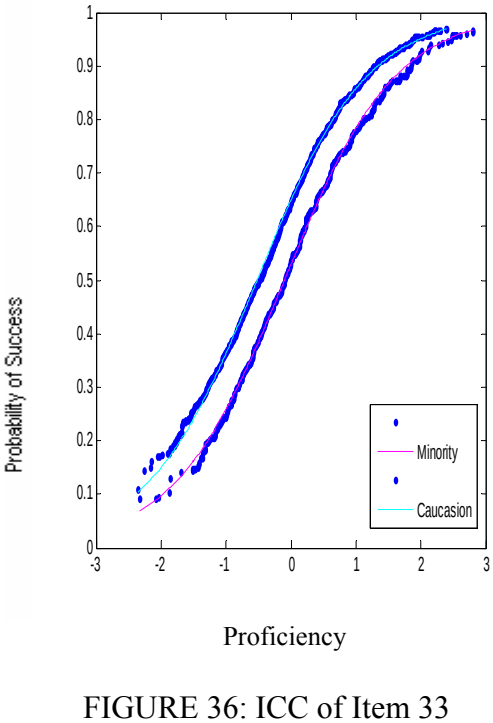


FIGURE 36: ICC of Item 33

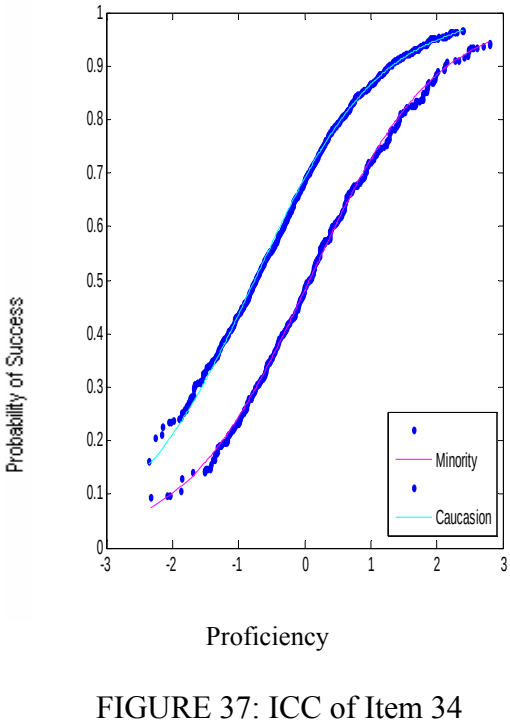


FIGURE 37: ICC of Item 34

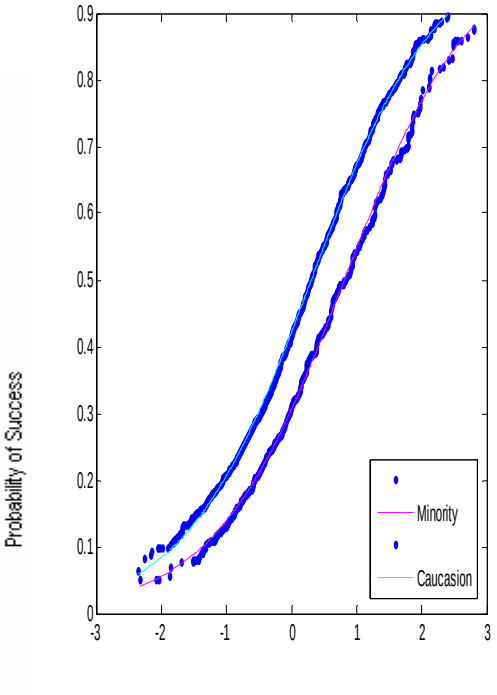


FIGURE 38: ICC of Item 35

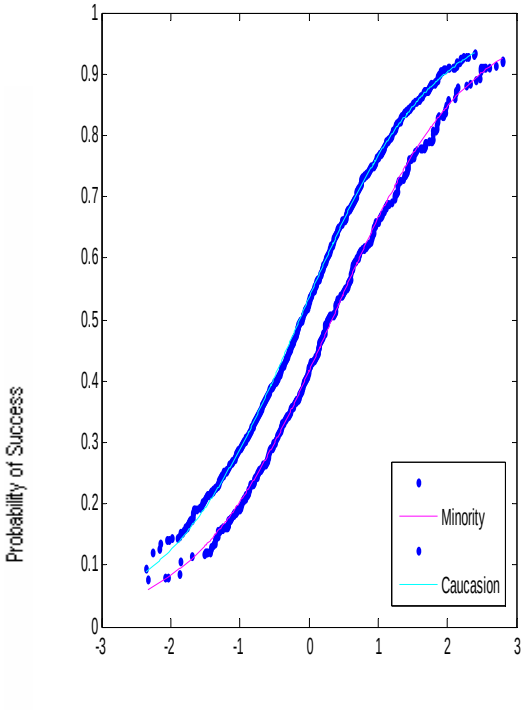


FIGURE 39: ICC of Item 36

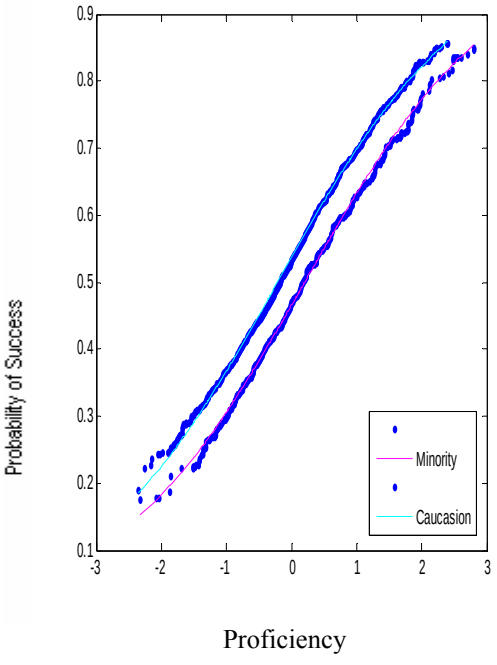


FIGURE 40: ICC of Item 37

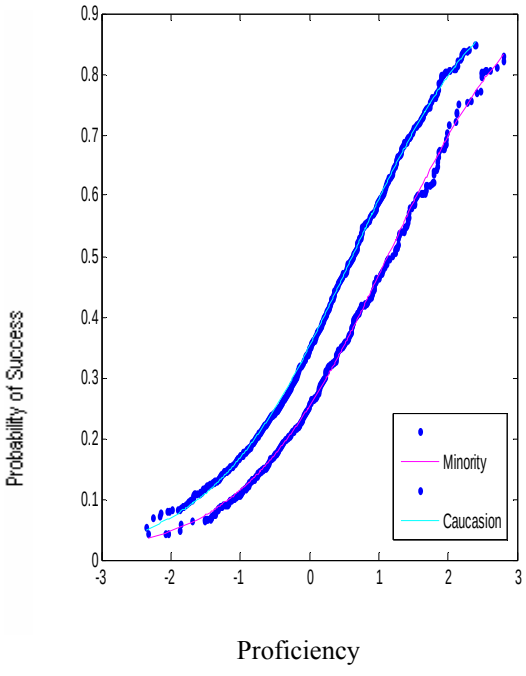


FIGURE 41: ICC of Item 38

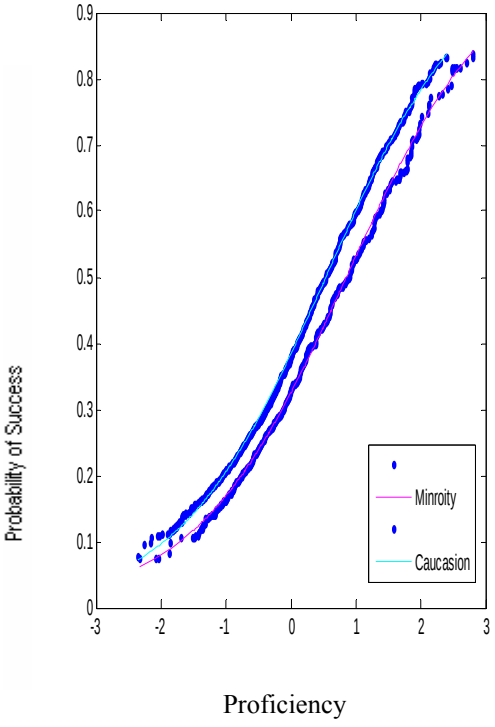


FIGURE 42: ICC of Item 39

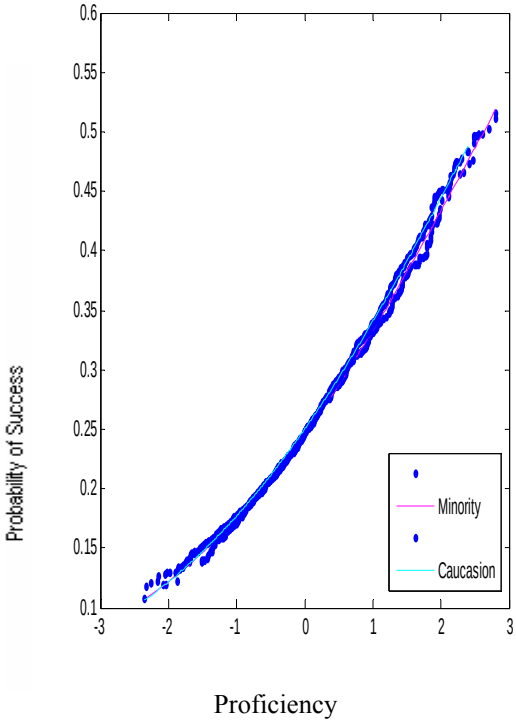


FIGURE 43: ICC of Item 40

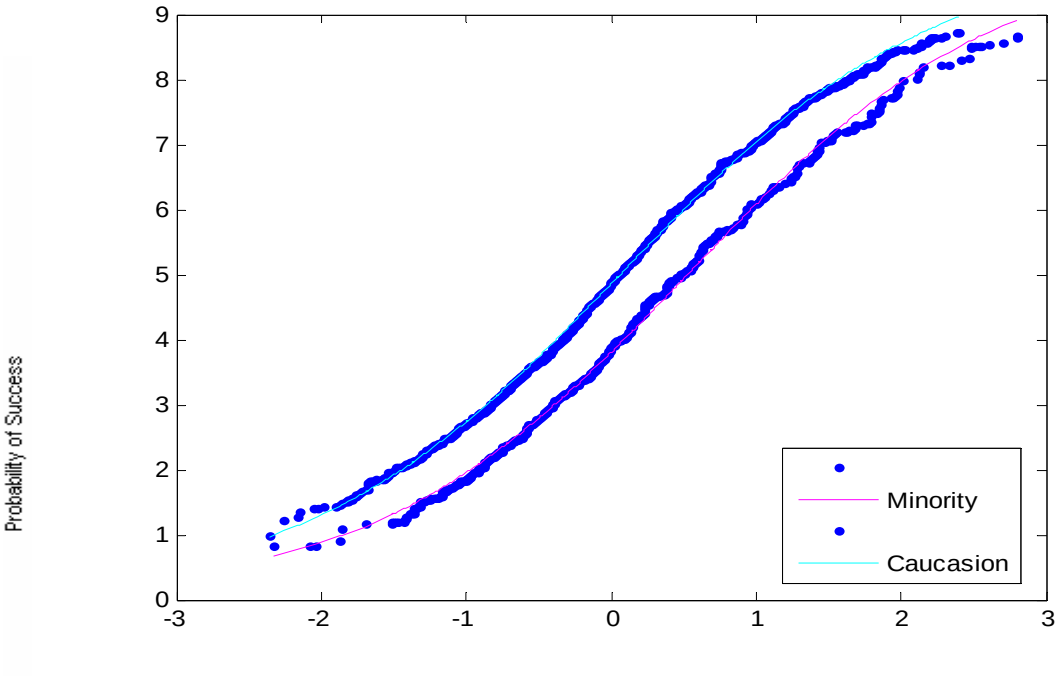


FIGURE 44: TCC of Testlet

Appendix F: WINBUGS Codes for MCMC Estimation

```
model  {
   muf ~ dnorm(0,10000);
   mur ~ dnorm(0,10000);
   tauf ~ dgamma(1,1);
   taur ~ dgamma(1,1);


for (k in 11:20) {
      a[k] ~ dlnorm(0, 4);
      b[k] ~ dnorm(0, 1);
  }

for (k in 21:25) {
      af[k] ~ dlnorm(0, 4);
      bf[k] ~ dnorm(0, 1);
      ar[k] ~ dlnorm(0, 4);
      br[k] ~ dnorm(0, 1);
  }

for (k in 26:30) {
      af[k] ~ dlnorm(0, 4);
      bf[k] ~ dnorm(0, 1);
      ar[k] ~ dlnorm(0, 4);
      br[k] ~ dnorm(0, 1);
  }

for (j in 1:500) {
     for (k in 1:10) {
       p[j,k] <- exp(a[k]*(theta[j]-b[k]))/(1+exp(a[k]*(theta[j]-b[k])));
       r[j,k] ~ dbern(p[j,k]);
      }

     for (k in 11:20) {
       p[j,k] <- exp(a[k]*(theta[j]-b[k]-eta[j]))/(1+exp(a[k]*(theta[j]-b[k]-eta[j])));
       r[j,k] ~ dbern(p[j,k]);
      }

    for (k in 21:I) {
       p[j,k] <- exp(af[k]*(theta[j]-bf[k]-etaf[j]))/(1+exp(af[k]*(theta[j]-bf[k]-etaf[j])));
       r[j,k] ~ dbern(p[j,k]);
      }
       eta[j] ~ dnorm(0,1);
       etaf[j] ~ dnorm(muf,tauf);
       theta[j] ~ dnorm(0,1);
```

```
  }

for (j in 501:N) {
     for (k in 1:10) {
       p[j,k] <- exp(a[k]*(theta[j]-b[k]))/(1+exp(a[k]*(theta[j]-b[k])));
       r[j,k] ~ dbern(p[j,k]);
      }

      for (k in 11:20) {
       p[j,k] <- exp(a[k]*(theta[j]-b[k]-eta[j]))/(1+exp(a[k]*(theta[j]-b[k]-eta[j])));
       r[j,k] ~ dbern(p[j,k]);
      }

     for (k in 21:I) {
        p[j,k] <- exp(ar[k]*(theta[j]-br[k]-etar[j]))/(1+exp(ar[k]*(theta[j]-br[k]-etar[j])));
        r[j,k] ~ dbern(p[j,k]);
      }
      eta[j] ~ dnorm(0,1);
      etar[j] ~ dnorm(mur,taur);
      theta[j] ~ dnorm(0,1);
  }
}

#initial vales
list(b(11:30) = c(5, 5, 5, 5, 5, 5, 5,5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5))
list(b(11:30) = c(-5, -5, -5, -5, -5, -5 -5, -5, -5, -5, -5, -5,-5, -5, -5, -5, -5, -5, -5, -5))

#data
list(N=1000,  I=30,
a=c(1.1, 1.0, 1.0, 1.3, 0.7, 1.4, 1.2, 1.4, 0.9, 0.6, NA, NA, NA, NA, NA, NA, NA, NA,
NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA),
b=c(-1.0, -0.9, 0.1, 1.1, 0.4, 0.1, 0.7, 0.6, 0.8, 1.0, NA, NA, NA, NA, NA, NA, NA, NA,
NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ),
 r=structure(.Data=c(
1,1,1,1,1,1,1,1,0,0,1,0,0,1,0,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
1,1,0,0,1,1,0,0,1,0,0,1,0,0,1,1,1,1,0,1,1,0,1,0,1,1,1,1,1,1,1,1,
.....
1,1,1,1,1,1,0,1,1,1,1,1,1,1,1,0,1,1,0,1,1,1,0,0,0,1,1,1,1,1,1,1,
1,0,1,0,0,1,0,1,1,1,1,1,1,1,0,1,0,0,0,0,0,1,1,0,1,1,1,1,1,0,1,0,
1,0,0,0,1,1,1,0,1,1,1,0,1,1,0,1,0,1,1,1,0,1,0,1,1,1,1,1,1,0,0,
1,1,1,1,0,1,1,1,0,0,1,1,1,0,1,1,1,1,0,1,0,1,1,1,1,1,1,1,0,1,1
), .Dim=c(1000, 30))
```

Appendix G: Explanation of Each Situation in the Simulation Study

Note: Example 1 (For Gender DIF 50 Males/50 Female**)**

Model 1, A: Assuming *local independence*; assuming item discrimination parameter function *homogeneously* between groups; assuming item difficulty parameter function *homogeneously* between groups; This is a baseline condition with ***no DIF***.

Model 1, B1: Assuming *local independence*; assuming item discrimination parameter function *homogeneously* between groups; for the *first five items* of interest, assuming the mean difference between the item difficulty parameters was 0.5 and the *reference group was favored,* for *the latter five items* of interest, assuming the mean difference between the item difficulty parameters was 0.5 and *the focal group was favored*; This is a condition of ***Uniform DIF*.**

Model 1, B2: Assuming *local independence*; assuming *small* conditional dependency within testlet; assuming item discrimination parameter function *homogeneously* between groups; for *all of the 10 items* of interest, assuming the mean difference between the item difficulty parameters was 0.5 and *the reference group was favored*; This is a condition of ***Uniform DIF***.

Model 1, B3: Assuming local independence; assuming item discrimination parameter function *homogeneously* between groups; for *all of the 10 items* of interest, assuming the

mean difference between the item difficulty parameters was 0.5 and *the focal group was favored*; This is a condition of ***Uniform DIF***.

Model 1, C1: Assuming local independence; assuming *small* conditional dependency within testlet; assuming item difficulty parameter function *homogeneously* between groups; for *the first 5 items* of interest, assuming the mean difference between the item discrimination parameters was 0.3 and the *reference group* was favored; *for the latter 5 items* of interest, assuming the mean difference between the *item discrimination parameters* was 0.5 and *the focal group was favored*; This is a condition of ***Non-Uniform DIF***. (In theory, ***DIF cancellation*** would happen at the ***item*** level because of Crossing of ICCs but still could be detected by unsigned-area index.)

Model 1, C2: Assuming local independence; assuming the mean difference between the item discrimination parameters was 0.3 and the *reference group* was favored; assuming item difficulty parameter function *homogeneously* between groups; This is a condition of ***Non-Uniform DIF***. (In theory, ***DIF cancellation*** would happen at the ***item*** because of Crossing of ICCs but still could be detected by unsigned-area index.)

Model 1, C3: Assuming local independence; assuming the mean difference between the item discrimination parameters was 0.3 and the *focal group* was favored; assuming item difficulty parameter function *homogeneously* between groups; This is a condition of ***Non-Uniform DIF***. (In theory, ***DIF cancellation*** would happen at the ***item*** because of Crossing of ICCs but still could be detected by unsigned-area index.)

Model 2, Condition I, A: Assuming local dependence and testlet effect function *homogeneously* between groups; assuming *small* conditional dependency within testlet; assuming item discrimination parameter function *homogeneously* between groups; assuming item difficulty parameter function *homogeneously* between groups; This is a baseline model with ***no DIF***.

Model 2, Condition I, B1: Assuming local dependence and testlet effect function *homogeneously* between groups; assuming *small* conditional dependency within testlet; assuming item discrimination parameter function *homogeneously* between groups; for the *first five items* within testlet, assuming the mean difference between the item difficulty parameters was 0.5 and the *reference group was favored,* for *the latter five items* within testlet, assuming the mean difference between the item difficulty parameters was 0.5 and *the focal group was favored* ; This is a condition of ***Uniform DIF*** when ***DIF cancellation*** would happen at the ***testlet*** level.

Model 2, Condition I, B2: Assuming local dependence and testlet effect function *homogeneously* between groups; assuming *small* conditional dependency within testlet; assuming item discrimination parameter function *homogeneously* between groups; for *all of the 10 items* within testlet, assuming the mean difference between the item difficulty parameters was 0.5 and *the reference group was favored*; This is a condition of ***Uniform DIF*** when ***DIF amplification*** would happen at the ***testlet*** level.

Model 2, Condition I, B3: Assuming local dependence and testlet effect function *homogeneously* between groups; assuming *small* conditional dependency within testlet; assuming item discrimination parameter function *homogeneously* between groups; for *all of the 10 items* within testlet, assuming the mean difference between the item difficulty parameters was 0.5 and *the focal group was favored*; This is a condition of ***Uniform DIF*** when ***DIF amplification*** would happen at the ***testlet*** level.

Model 2, Condition I, C1: Assuming local dependence and testlet effect function *homogeneously* between groups; assuming *small* conditional dependency within testlet; assuming item difficulty parameter function *homogeneously* between groups; for *the first 5 items* within testlet, assuming the mean difference between the item discrimination parameters was 0.3 and the *reference group* was favored; *for the latter 5 items* within testlet, assuming the mean difference between the *item discrimination parameters* was 0.5 and *the focal group was favored*; This is a condition of ***Non-Uniform DIF*** when ***DIF cancellation*** would happen at the ***testlet*** level. (In theory, ***DIF cancellation*** would happen at the ***item*** level because of Crossing of ICCs but still could be detected by unsigned-area index.)

Model 2, Condition I, C2: Assuming local dependence and testlet effect *function homogeneously* between groups; assuming *small* conditional dependency within testlet; assuming the mean difference between the item discrimination parameters was 0.3 and the *reference group* was favored; assuming item difficulty parameter function *homogeneously* between groups; This is a condition of ***Non-Uniform DIF*** when ***DIF***

amplification would happen at the ***testlet*** level. (In theory, ***DIF cancellation*** would happen at the ***item*** and ***testlet*** level because of Crossing of ICCs and TCCs but still could be detected by unsigned-area index.)

Model 2, Condition I, C3: Assuming local dependence and testlet effect *function homogeneously* between groups; assuming *small* conditional dependency within testlet; assuming the mean difference between the item discrimination parameters was 0.3 and the *focal group* was favored; assuming item difficulty parameter function *homogeneously* between groups; This is a condition of ***Non-Uniform DIF*** when ***DIF amplification*** would happen at the ***testlet*** level. (In theory, ***DIF cancellation*** would happen at the ***item*** and ***testlet*** level because of Crossing of ICCs and TCCs but still could be detected by unsigned-area index.)

Model 2, Condition II, A: Assuming local dependence and testlet effect function *homogeneously* between groups; assuming *large* conditional dependency within testlet; assuming item discrimination parameter function *homogeneously* between groups; assuming item difficulty parameter function *homogeneously* between groups; This is a baseline model with ***no DIF***.

Model 2, Condition II, B1: Assuming local dependence and testlet effect function *homogeneously* between groups; assuming *large* conditional dependency within testlet; assuming item discrimination parameter function *homogeneously* between groups; for the *first five items* within testlet, assuming the mean difference between the item difficulty parameters was 0.5 and the *reference group was favored,* for *the latter five items* within

testlet, assuming the mean difference between the item difficulty parameters was 0.5 and *the focal group was favored* ; This is a condition of ***Uniform DIF*** when ***DIF cancellation*** would happen at the ***testlet*** level.

Model 2, Condition II, B2: Assuming local dependence and testlet effect function *homogeneously* between groups; assuming *large* conditional dependency within testlet; assuming item discrimination parameter function *homogeneously* between groups; for *all of the 10 items* within testlet, assuming the mean difference between the item difficulty parameters was 0.5 and *the reference group was favored*; This is a condition of ***Uniform DIF*** when ***DIF amplification*** would happen at the ***testlet*** level.

Model 2, Condition II, B3: Assuming local dependence and testlet effect function *homogeneously* between groups; assuming *large* conditional dependency within testlet; assuming item discrimination parameter function *homogeneously* between groups; for *all of the 10 items* within testlet, assuming the mean difference between the item difficulty parameters was 0.5 and *the focal group was favored*; This is a condition of ***Uniform DIF*** when ***DIF amplification*** would happen at the ***testlet*** level.

Model 2, Condition II, C1: Assuming local dependence and testlet effect function *homogeneously* between groups; assuming *large* conditional dependency within testlet; assuming item difficulty parameter function *homogeneously* between groups; for *the first 5 items* within testlet, assuming the mean difference between the item discrimination parameters was 0.3 and the *reference group* was favored; *for the latter 5 items* within testlet, assuming the mean difference between the *item discrimination parameters* was

0.5 and *the focal group was favored*; This is a condition of ***Non-Uniform DIF*** when ***DIF cancellation*** would happen at the ***testlet*** level. (In theory, ***DIF cancellation*** would happen at the ***item*** and ***testlet*** levels because of Crossing of ICCs and TCCs but still could be detected by unsigned-area index.)

Model 2, Condition II, C2: Assuming local dependence and testlet effect *function homogeneously* between groups; assuming *large* conditional dependency within testlet; assuming the mean difference between the item discrimination parameters was 0.3 and the *reference group* was favored; assuming item difficulty parameter function *homogeneously* between groups; This is a condition of ***Non-Uniform DIF*** when ***DIF amplification*** would happen at the ***testlet*** level. (In theory, ***DIF cancellation*** would happen at the ***item*** and ***testlet*** level because of Crossing of ICCs and TCCs but still could be detected by unsigned-area index.)

Model 2, Condition II, C3: Assuming local dependence and testlet effect *function homogeneously* between groups; assuming *large* conditional dependency within testlet; assuming the mean difference between the item discrimination parameters was 0.3 and the *focal group* was favored; assuming item difficulty parameter function *homogeneously* between groups; This is a condition of ***Non-Uniform DIF*** when ***DIF amplification*** would happen at the ***testlet*** level. (In theory, ***DIF cancellation*** would happen at the ***item*** and ***testlet*** level because of Crossing of ICCs and TCCs but still could be detected by unsigned-area index.)

Model 3, Condition I, A: Assuming local dependence and testlet effect function *heterogeneously* between groups, the mean difference between the groups was 1.2247 and the *reference group was favored*; assuming *moderate* conditional dependency within testlet; assuming item discrimination parameter function *homogeneously* between groups; assuming item difficulty parameter function *homogeneously* between groups; This is a condition ***Uniform DIF*** when only testlet parameter function differently between groups and ***DIF amplification*** would happen at the ***testlet*** level.

Model 3, Condition I, B1: Assuming local dependence and *testlet effect function heterogeneously* between groups, the mean difference between the groups was 1.2247 and the *reference group* was favored; assuming *moderate* conditional dependency within testlet; assuming item discrimination parameter function *homogeneously* between groups; for the *first five items* within testlet, assuming the mean difference between the item difficulty parameters was 0.5 and the *reference group was favored*, for *the latter five* items within testlet, assuming the mean difference between the item difficulty parameters was 0.5 and the *focal group was favored*; This is a condition of ***Uniform DIF*** when ***DIF amplification*** would happen for each of the first five items, ***DIF cancellation*** would happen for each of the latter five items and additionally, ***DIF amplification*** might happen at the ***testlet*** level.

Model 3, Condition I, B2: Assuming local dependence and testlet effect function *heterogeneously* between groups, the mean difference between the groups was 1.2247 and the *reference group was favored*; assuming *moderate* conditional dependency within

testlet; assuming item discrimination parameter function *homogeneously* between groups; for *all of the 10 items* within testlet, assuming the mean difference between the item difficulty parameters was 0.5 and *the reference group was favored*; This is a condition of ***Uniform DIF*** when ***DIF amplification*** would happen at both of ***item and testlet*** level.

Model 3, Condition I, B3: Assuming local dependence and testlet effect function *heterogeneously* between groups, the mean difference between the groups was 1.2247 and the *reference group was favored*; assuming *moderate* conditional dependency within testlet; assuming item discrimination parameter function *homogeneously* between groups; for *all of the 10 items* within testlet, assuming the mean difference between the item difficulty parameters was 0.5 and *the focal group was favored*; This is a condition of ***Uniform DIF*** when ***DIF cancellation*** might happen at the ***item*** level and ***DIF amplification*** might happen at the ***testlet*** level.

Model 3, Condition I, C1: Assuming local dependence and testlet effect function *heterogeneously* between groups, the mean difference between the groups was 1.2247 and the *reference group was favored*; assuming *moderate* conditional dependency within testlet; for *the first five items* within testlet, assuming the mean difference between the item discrimination parameters was 0.3 and the *reference group was favored*; for *the latter five items* within testlet, assuming the mean difference between the *item discrimination* parameters was 0.3 and the *focal group was favored* ; assuming item difficulty parameter function *homogeneously* between groups; This is a condition of ***Non-Uniform DIF*** and ***DIF cancellation*** would happen at the ***testlet*** level. (In theory, ***DIF***

cancellation would happen at the ***item*** level because of Crossing of ICCs but still could be detected by unsigned-area index.)

Model 3, Condition I, C2: Assuming local dependence and testlet effect function *heterogeneously* between groups, the mean difference between the groups was 1.2247 and the *reference group was favored*; assuming *moderate* conditional dependency within testlet; for *all of the 10 items* within testlet, assuming the mean difference between the item discrimination parameters was 0.3 and the *reference group was favored*; assuming item difficulty parameter function *homogeneously* between groups ; This is a condition of ***Crossing DIF*** when ***DIF amplification*** would happen at the ***testlet*** level. (In theory, ***DIF cancellation*** would happen at the ***item*** and ***testlet*** levels because of Crossing of ICCs and TCCs but still could be detected by unsigned-area index.)

Model 3, Condition I, C3: Assuming local dependence and testlet effect function *heterogeneously* between groups, the mean difference between the groups was 1.2247 and the *reference group was favored*; assuming *moderate* conditional dependency within testlet; for *all of the 10 items* within testlet, assuming the mean difference between the item discrimination parameters was 0.3 and the *focal group was favored*; assuming item difficulty parameter function *homogeneously* between groups ; This is a condition of ***Crossing DIF*** when ***DIF amplification*** would happen at the ***testlet*** level. (In theory, ***DIF cancellation*** would happen at the ***item*** and ***testlet*** levels because of Crossing of ICCs and TCCs but still could be detected by unsigned-area index.)

Example 2 (For Ethnic DIF 80 Majority/20 Minority)**:**

Model 1, A;

Model 1, B1; Model 1, B2; Model 1, B3;

Model 1, C1; Model 1, C2; Model 1, C3;

(Same as Example 1)

Model 2, Condition I, A: Assuming local dependence and testlet effect function *homogeneously* between groups; assuming *small* conditional dependency within testlet; assuming item discrimination parameter function *homogeneously* between groups; assuming item difficulty parameter function *homogeneously* between groups; This is a baseline model with ***no DIF***.

Model 2, Condition I, B1: Assuming local dependence and testlet effect function *homogeneously* between groups; assuming *small* conditional dependency within testlet; assuming item discrimination parameter function *homogeneously* between groups; for the *first five items* within testlet, assuming the mean difference between the item difficulty parameters was 0.5 and the *reference group was favored,* for *the latter five items* within testlet, assuming the mean difference between the item difficulty parameters was 0.5 and *the focal group was favored* ; This is a condition of ***Uniform DIF*** when ***DIF cancellation*** would happen at the ***testlet*** level.

Model 2, Condition I, B2: Assuming local dependence and testlet effect function *homogeneously* between groups; assuming *small* conditional dependency within testlet;

assuming item discrimination parameter function *homogeneously* between groups; for *all of the 10 items* within testlet, assuming the mean difference between the item difficulty parameters was 0.5 and *the reference group was favored*; This is a condition of ***Uniform DIF*** when ***DIF amplification*** would happen at the ***testlet*** level.

Model 2, Condition I, C1: Assuming local dependence and testlet effect function *homogeneously* between groups; assuming *small* conditional dependency within testlet; assuming item difficulty parameter function *homogeneously* between groups; for *the first 5 items* within testlet, assuming the mean difference between the item discrimination parameters was 0.3 and the *reference group* was favored; *for the latter 5 items* within testlet, assuming the mean difference between the *item discrimination parameters* was 0.5 and *the focal group was favored*; This is a condition of ***Non-Uniform DIF*** when ***DIF cancellation*** would happen at the ***testlet*** level. (In theory, ***DIF cancellation*** would happen at the ***item*** level because of Crossing of ICCs but still could be detected by unsigned-area index.)

Model 2, Condition I, C2: Assuming local dependence and testlet effect *function homogeneously* between groups; assuming *small* conditional dependency within testlet; assuming the mean difference between the item discrimination parameters was 0.3 and the *reference group* was favored; assuming item difficulty parameter function *homogeneously* between groups; This is a condition of ***Non-Uniform DIF*** when ***DIF amplification*** would happen at the ***testlet*** level. (In theory, ***DIF cancellation*** would

happen at the ***item*** and ***testlet*** level because of Crossing of ICCs and TCCs but still could be detected by unsigned-area index.)

Model 3, Condition I, A: Assuming local dependence and testlet effect function *heterogeneously* between groups, the mean difference between the groups was 1.2247 and the *reference group was favored*; assuming *small* conditional dependency within testlet; assuming item discrimination parameter function *homogeneously* between groups; assuming item difficulty parameter function *homogeneously* between groups; This is a condition ***Uniform DIF*** when only testlet parameter function differently between groups and ***DIF amplification*** would happen at the ***testlet*** level.

Model 3, Condition I, B1: Assuming local dependence and *testlet effect function heterogeneously* between groups, the mean difference between the groups was 1.2247 and the *reference group* was favored; assuming *small* conditional dependency within testlet; assuming item discrimination parameter function *homogeneously* between groups; for the *first five items* within testlet, assuming the mean difference between the item difficulty parameters was 0.5 and the *reference group was favored*, for *the latter five* items within testlet, assuming the mean difference between the item difficulty parameters was 0.5 and the *focal group was favored*; This is a condition of ***Uniform DIF*** when ***DIF amplification*** would happen for each of the first five items, ***DIF cancellation*** would happen for each of the latter five items and additionally, ***DIF amplification*** might happen at the ***testlet*** level.

Model 3, Condition I, B2: Assuming local dependence and testlet effect function *heterogeneously* between groups, the mean difference between the groups was 1.2247 and the *reference group was favored*; assuming *small* conditional dependency within testlet; assuming item discrimination parameter function *homogeneously* between groups; for *all of the 10 items* within testlet, assuming the mean difference between the item difficulty parameters was 0.5 and *the reference group was favored*; This is a condition of ***Uniform DIF*** when ***DIF amplification*** would happen at both of ***item and testlet*** level.

Model 3, Condition I, C1: Assuming local dependence and testlet effect function *heterogeneously* between groups, the mean difference between the groups was 1.2247 and the *reference group was favored*; assuming *small* conditional dependency within testlet; for *the first five items* within testlet, assuming the mean difference between the item discrimination parameters was 0.3 and the *reference group was favored*; for *the latter five items* within testlet, assuming the mean difference between the *item discrimination* parameters was 0.3 and the *focal group was favored* ; assuming item difficulty parameter function *homogeneously* between groups; This is a condition of ***Non-Uniform DIF*** and ***DIF cancellation*** would happen at the ***testlet*** level. (In theory, ***DIF cancellation*** would happen at the ***item*** level because of Crossing of ICCs but still could be detected by unsigned-area index.)

Model 3, Condition I, C2: Assuming local dependence and testlet effect function *heterogeneously* between groups, the mean difference between the groups was 1.2247 and the *reference group was favored*; assuming *small* conditional dependency within

testlet; for *all of the 10 items* within testlet, assuming the mean difference between the item discrimination parameters was 0.3 and the *reference group was favored*; assuming item difficulty parameter function *homogeneously* between groups ; This is a condition of ***Crossing DIF*** when ***DIF amplification*** would happen at the ***testlet*** level. (In theory, ***DIF cancellation*** would happen at the ***item*** and ***testlet*** levels because of Crossing of ICCs and TCCs but still could be detected by unsigned-area index.)

Model 3, Condition II, A: Assuming local dependence and testlet effect function *heterogeneously* between groups, the means of distributions of testlet parameter are same for focal group and reference group, and the variance of distribution of testlet parameter of focal group is *larger* than that of reference group, reflecting *larger* conditional dependency within testlet for focal group than that of reference group; assuming item discrimination parameter function *homogeneously* between groups; assuming item difficulty parameter function *homogeneously* between groups; This is a condition when only testlet parameter function differently between groups and ***if there was DIF***, ***DIF amplification*** would happen at the ***testlet*** level.

Model 3, Condition II, B1: Assuming local dependence and testlet effect function *heterogeneously* between groups, the means of distributions of testlet parameter are same for focal group and reference group, and the variance of distribution of testlet parameter of focal group is *larger* than that of reference group, reflecting *larger* conditional dependency within testlet for focal group than that of reference group; assuming item discrimination parameter function *homogeneously* between groups; for the *first five items*

within testlet, assuming the mean difference between the item difficulty parameters was 0.5 and the *reference group was favored*, for *the latter five* items within testlet, assuming the mean difference between the item difficulty parameters was 0.5 and the *focal group was favored*; This is a condition of ***Uniform DIF*** when ***DIF cancellation*** might happen at the ***testlet*** level.

Model 3, Condition II, B2: Assuming local dependence and testlet effect function *heterogeneously* between groups, the means of distributions of testlet parameter are same for focal group and reference group, and the variance of distribution of testlet parameter of focal group is *larger* than that of reference group, reflecting *larger* conditional dependency within testlet for focal group than that of reference group; assuming item discrimination parameter function *homogeneously* between groups; for *all of the 10 items* within testlet, assuming the mean difference between the item difficulty parameters was 0.5 and *the reference group was favored*; This is a condition of ***Uniform DIF*** when ***DIF amplification*** would happen at both of ***testlet*** level.

Model 3, Condition II, C1: Assuming local dependence and testlet effect function *heterogeneously* between groups, the means of distributions of testlet parameter are same for focal group and reference group, and the variance of distribution of testlet parameter of focal group is *larger* than that of reference group, reflecting *larger* conditional dependency within testlet for focal group than that of reference group; for *the first five items* within testlet, assuming the mean difference between the item discrimination parameters was 0.3 and the *reference group was favored*; for *the latter five items* within

testlet, assuming the mean difference between the *item discrimination* parameters was 0.3 and the *focal group was favored* ; assuming item difficulty parameter function *homogeneously* between groups; This is a condition of ***Non-Uniform DIF*** and ***DIF cancellation*** would happen at the ***testlet*** level. (In theory, ***DIF cancellation*** would happen at the ***item*** level because of Crossing of ICCs but still could be detected by unsigned-area index.)

Model 3, Condition II, C2: Assuming local dependence and testlet effect function *heterogeneously* between groups, the means of distributions of testlet parameter are same for focal group and reference group, and the variance of distribution of testlet parameter of focal group is *larger* than that of reference group, reflecting *larger* conditional dependency within testlet for focal group than that of reference group; for *all of the 10 items* within testlet, assuming the mean difference between the item discrimination parameters was 0.3 and the *reference group was favored*; assuming item difficulty parameter function *homogeneously* between groups ; This is a condition of ***Crossing DIF*** when ***DIF amplification*** would happen at the ***testlet*** level. (In theory, ***DIF cancellation*** would happen at the ***item*** and ***testlet*** levels because of Crossing of ICCs and TCCs but still could be detected by unsigned-area index.)

Model 3, Condition III, A: Assuming local dependence and testlet effect function *heterogeneously* between groups, the mean difference of testlet parameters between the groups was 1.2247 and the *reference group was favored*, and the variance of distribution of testlet parameter of focal group is *larger* than that of reference group, reflecting *larger*

conditional dependency within testlet for focal group than that of reference group; assuming item discrimination parameter function *homogeneously* between groups; assuming item difficulty parameter function *homogeneously* between groups; This is a condition ***Uniform DIF*** when only testlet parameter function differently between groups and ***DIF amplification*** would happen at the ***testlet*** level.

Model 3, Condition III, B1: Assuming local dependence and testlet effect function *heterogeneously* between groups, the mean difference of testlet parameters between the groups was 1.2247 and the *reference group was favored*, and the variance of distribution of testlet parameter of focal group is *larger* than that of reference group, reflecting *larger* conditional dependency within testlet for focal group than that of reference group; assuming item discrimination parameter function *homogeneously* between groups; for the *first five items* within testlet, assuming the mean difference between the item difficulty parameters was 0.5 and the *reference group was favored*, for *the latter five* items within testlet, assuming the mean difference between the item difficulty parameters was 0.5 and the *focal group was favored*; This is a condition of ***Uniform DIF*** when ***DIF amplification*** would happen for each of the first five items, ***DIF cancellation*** would happen for each of the latter five items and additionally, ***DIF amplification*** might happen at the ***testlet*** level.

Model 3, Condition III, B2: Assuming local dependence and testlet effect function *heterogeneously* between groups, the mean difference of testlet parameters between the groups was 1.2247 and the *reference group was favored*, and the variance of distribution

of testlet parameter of focal group is *larger* than that of reference group, reflecting *larger* conditional dependency within testlet for focal group than that of reference group; assuming item discrimination parameter function *homogeneously* between groups; for *all of the 10 items* within testlet, assuming the mean difference between the item difficulty parameters was 0.5 and *the reference group was favored*; This is a condition of ***Uniform DIF*** when ***DIF amplification*** would happen at both of ***item and testlet*** level.

Model 3, Condition III, C1: Assuming local dependence and testlet effect function *heterogeneously* between groups, the mean difference of testlet parameters between the groups was 1.2247 and the *reference group was favored*, and the variance of distribution of testlet parameter of focal group is *larger* than that of reference group, reflecting *larger* conditional dependency within testlet for focal group than that of reference group; for *the first five items* within testlet, assuming the mean difference between the item discrimination parameters was 0.3 and the *reference group was favored*; for *the latter five items* within testlet, assuming the mean difference between the *item discrimination* parameters was 0.3 and the *focal group was favored* ; assuming item difficulty parameter function *homogeneously* between groups; This is a condition of ***Non-Uniform DIF*** and ***DIF cancellation*** would happen at the ***testlet*** level. (In theory, ***DIF cancellation*** would happen at the ***item*** level because of Crossing of ICCs but still could be detected by unsigned-area index.)

Model 3, Condition III, C2: Assuming local dependence and testlet effect function *heterogeneously* between groups, the mean difference of testlet parameters between the

groups was 1.2247 and the *reference group was favored*, and the variance of distribution of testlet parameter of focal group is *larger* than that of reference group, reflecting *larger* conditional dependency within testlet for focal group than that of reference group; for *all of the 10 items* within testlet, assuming the mean difference between the item discrimination parameters was 0.3 and the *reference group was favored*; assuming item difficulty parameter function *homogeneously* between groups ; This is a condition of ***Crossing DIF*** when ***DIF amplification*** would happen at the ***testlet*** level. (In theory, ***DIF cancellation*** would happen at the ***item*** and ***testlet*** levels because of Crossing of ICCs and TCCs but still could be detected by unsigned-area index.)

References

Ackerman, T. A., Gierl, M. J., & Walker C. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice, 22,* 37-53.

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

Alderson, J. C. (2000). *Assessing reading*. Cambridge, UK: Cambridge University Press.

Allalouf, A., Hambleton, R., & Sireci, S. (1999). Identifying the causes of translation DIF on verbal items. *Journal of Educational Measurement, 36,* 185-198.

Angoff, W. (1993). Perspective on differential item functioning methodology. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-24). Hillsdale, NJ: Lawrence Erlbaum.

Baker, F.B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques*. New York: Dekker.

Benson, J., & Tippets, E. (1990). Confirmatory factor analysis of the Test Anxiety Inventory. In C. D. Spielberger & R. Diaz-Guerrero (Eds.), *Cross-cultural anxiety* (Vol. 4, 149-156). New York: Hemisphere/Taylor-Francis.

Birbaum, A. (1958a). *On the estimation of mental ability*. Series Report No. 15. Project No. 77755-23, USAF School of Aviation Medicine, Randolph Air Base. Texas.

Birbaum, A. (1958b). *Further considerations of efficiency in tests of a mental ability*. Technical Report No. 17, Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Base. Texas.

Birbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M. R. Novick. *Statistical theories of mental test scores* (Chapters 17-20). Reading, MA: Addison-Wesley.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores*. *Reading,* Mass.: Addison-Wesley Publishing.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more latent categories. *Psychometrika, 37*, 29-51.

Bock, R.D., & Aitkin, M. (1980). Marginal maximum likelihood estimation of item

parameters: An application of the EM algorithm. *Psychometrika, 46*, 443-449.

Bock, R. D., & Jones, L. V. (1968). *The measurement and prediction of judgment and choice*. San Francisco: Holden-Day.

Bolt, D., Froelich, A., Habing, B., Hartz, S., Roussos, L., & Stout, W. (1999, September). An *applied and foundational research project addressing DIF, impact, and equity with applications for ETS test development* ETS Research Report (RR-03-06). Princeton, NJ: Educational Testing Service.

Bond, L. (1993). Comments on the O'Neill and McPeek paper. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 277-279). Hillsdale: NJ: Lawrence Erlbaum.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64,* 153-168.

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items.* Newbury Park: Sage.

Chang, H.H., Mazzeo, J. & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33*, 333-353.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify Differentially functioning test items. *Educational Measurement: Issues and Practice, 17,* 31-44.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel And Standardization. In P. W. Holland & H. Wainer (Eds.) *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Erlbaum.

Douglas, J. Kim, H. R., Roussos, L., Stout, W., & Zhang, J. (1999). *LSAT dimensionality Analysis for the December 1991, June 1992, and October, 1992, administrations.* (Statistical Report No. 95-05). Newton, PA: Law School Admission Council.Multidimensional Framework 22.

Douglas, J., Roussos, L., and Stout, W., (1996). Item bundle DIF hypothesis testing: Identifying suspect bundles and assessing their DIF. *Journal of Educational Measurement, 33,* 465-484.

Downing, S. M., & Haladyna, T. M. (2004). Validity threats: overcoming interference with proposed interpretations of assessment data. *Medical Education*, 38 (3), 327–333.

Drasgow, F., & Levine, M.V. (1986). Optimal detection of certain forms of inappropriate test scores. Applied Psychological Measurement, 10, 59-67

Dresher, A. R. (2002). *The examination of local item dependency of NAEP assessments using the testlet model*. Unpublished doctoral dissertation, University of Pittsburgh.

Dresher, A. R. (2004). *An Empirical Investigation of LID using the Testlet Model: A Further Look*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Du, Z. (1998). *Modeling conditional item dependencies with a three-parameter logistic testlet model*. Unpublished doctoral dissertation, Columbia University.

Englehard, G,. Hansche, L., & Rutledge, K. E. (1990). Accuracy of bias review judges in identifying differential item functioning on teacher certification tests. *Applied Measurement in Education, 3,* 347-360.

Everson, H. T., Millsap, R. E., & Rodriguez, C. M. (1991). Isolating gender differences in test anxiety: A confirmatory factor analysis of the Test Anxiety Inventory. *Educational and Psychological Measurement, 51,* 243-251.

Froelich, A. G. (2000). *Assessing the unidimensionality of test items and some asympototics of parametric item response theory.* Unpublished Doctoral Dissertation. University of Illinois at Urbana-Champaign, Department of Statistics.

Froelich, A. G., & Habing, B. (2001). *Refinements of the DIMTEST methodology for Testing unidimensionality and local independence.* Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Gallagher, A. M., De Lisi, R., Holst, P. C., McGillicuddy-De Lisi, A. V., Morely, M., & Cahalan, C. (2000). Gender differences in advanced mathematical problem solving. *Journal of Experimental Child Psychology, 75,* 165-190.

Geisser, S., & Eddy, W. F. (1979). A Predictive approach to model selection. *Journal of American Statistical Association*, 74, 153-160.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of American Statistical Association*, 85, 398-409.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *Institute of Electrical and Electronics Engineers, Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.

Gierl, M. J., Bisanz, J., Bisanz, G., & Boughton, K. (in press). Identifying content and Cognitive skills that produce gender differences in mathematics: A demonstration of the DIF analysis framework. *Journal of Educational Measurement*.

Gierl, M. J., Bisanz, J., Bisanz, G., Boughton, K., & Khaliq, S. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice, 20,* 26-36.

Gierl, M. J., & Bolt, D. (2003, April*). Implications of the multidimensionality-based DIF Analysis framework for selecting a matching and studied subtest.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests. *Journal of Educational Measurement, 38,* 164-187.

Gierl, M. J., & Rogers, W. T. (1996). A confirmatory factor analysis of the Test Anxiety Inventory using Canadian high school students. *Educational and Psychological Measurement, 56,*315-324.Multidimensional Framework 23

Gierl, M. J., Rogers, W. T., & Klinger, D. (1999, April). *Consistency between statistical procedure and content reviews for identifying translation DIF.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montréal, Quebec, Canada.

Gill, J. (2002). *Bayesian Methods for the Social and Behavioral Sciences*. Chapman & Hall.

Glas, C. A. W., Wainer, H., & Bradlow, E. T. (2000). MML and EAP estimation in testlet-based adaptive testing. In W.J. van der Linden & C. A. W. Glas (Eds), Computerized adaptive testing: Theory and practice (pp. 271-287). Dordrecht, Netherlands: Kluwer.

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

Hamilton, L., Nussbaum, M., Kupermintz, H., Kerkhoven, J., & Snow, R. (1995). Enhancing the validity and usefulness of large-scale educational assessments: NELS:88 science achievement. *American Educational Research Journal, 32,* 555-581.

Holland & H. Wainer (Eds.) *Differential item functioning* (pp. 337-347). Hillsdale, NJ: Erlbaum.

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..

Hoskens, M. & Boeck, D. (1997). A parametric model for local item dependences among test items. *Psychological Methods*, 2, 261-277.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluation 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.

Kim, S.-H., & Cohen, A. S. (1992). Effects of linking methods for estimation of DIF. *Journal of Educational Measurement*, 29, 51-66.

Kupermintz, H., Ennis, M., Hamilton, L., Talbert, J., & Snow, R. (1995). Enhancing the validity and usefulness of large-scale educational assessments: NELS:88 mathematics achievement. *American Educational Research Journal, 32,* 524-554.

Lee, G., Dunbar, S. B., & Frisbie, D. A. (1999). *Measurement models for a testlet-based tests.* Paper presented at the annual meeting of the National Conference of Measurement in Education.

Lee, G., & Frisbie, D. A. (1999). Estimating reliability under a generalizability theory model for test scores composed of testlets. *Applied Measurement in Education, 12*, 237-255.

Li, Y., Bolt, D. M. J., & Fu, J. (2004). A comparison of alternative models for testlet. *Applied Psychological Measurement*, Vol. 30, No.1, 3-21.

Li, Y., & Cohen, A. S. (2003). *Equating tests composed of testlets: A comparison of a testlet response model and four polytomous response models*. Paper presented at the annual meeting of the National Council on Measurement in Education. Chicago, IL.

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

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

Mazor, K. M., Kanjee, A., & Clauser, B. E. (1995). Using logistic regression and the Mantel-Haenszel with multiple ability estimates to detect differential item functioning. *Journal of Educational Measurement, 32,* 131-144.

The Mathworks Inc. (2004). *MATLAB version 7.0*.4. Natick, Massachusetts:The MathWorks Inc.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for Assessing measurement bias. *Applied Psychological Measurement, 17,* 297-334.

Muraki, E., & Bock, R. D. (1998). *PARSCALE: Parameter scaling of rating data*. Chicago, IL: Scientific Software, Inc.

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. Journal of Educational Measurement, 30, 293-311.

Neyman, J., & Scott, B. (1948). Consistent Estimation from Partially Consistent Observations, *Econometrica*, Vol 16, 1-32.

O'Neill, K. A., & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 255-276). Hillsdale, NJ: Lawrence Erlbaum.

Oshima, T. C., Raju, N. S., Flowers, C. P., & Slinde, J. A. (1998). Differential bundle functioning using the DFIT framework: Procedures for identifying possible sources of differential functioning. *Applied Measurement in Education, 11,* 353-369.

Patz, R.J., & Junker, B.W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*, 146-178.

Plake, B. S. (1980). A comparison of a statistical and subjective procedure to ascertain Item validity: One step in the validation process. *Educational and Psychological Measurement, 40,* 397-404.

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomous scored items: A framework for classification and evaluation. *Applied psychological Measurement*, *1*, 23-37.

Ramsey, P. A. (1993). Sensitivity review: The ETS experience as a case study. In P. W. Holland & H. Wainer (Eds.) *Differential item functioning* (pp. 367-388). Hillsdale, NJ: Erlbaum.

Raju, N. S. (1988). *MHDIP*. Chicago:Department of Psychology, Illinois Institute of Technology.

Reckase, M. D., & Mckinley, R. L. (1983, April). *The definition of difficulty and discrimination for multidimensional item response theory models*. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal.

Reith,J., & Roznowski, M. (1991, April). *Predictive relations of tests containing differentially functioning items: Do biased items result in biased test?* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Roznowski, M. (1987). Use of tests manifesting sex differences as measures of intelligence: Implications for measurement bias. *Journal of Applied Psychology,* 72, 480-483.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20,* 355-371.

Rudner, L. M. (1977, April). *An approach to biased item identification using latent trait measurement theory*. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Rudner, L.M., Getson, P.R., & Knight, D.L. (1980). Biased item detection techniques. *Journal of Educational Statistics*, *5*, 213-233.

Rudner, L. M. (1977, April). *An approach to biased item identification using latent trait measurement theory*. Paper presented at the Annual Meeting of American Educaitonal Research Association, NewYork, NY.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980). Biased item detection techniques. *Journal of Educational Statistics, 5*, 213-233.

Sahu, S. K. (2002). Bayesian Estimation and Model Choice in Item Response Models. *Journal of Statistical Computation and Simulation*, *72*, 217-232.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, No. 17.

SAS Institute. (2002). *SAS/STAT, Version 8.2.* Cary, NC: Author.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*, 159-194.

Smith, A. F. M. (1991). Discussion of 'posterior Bayes factors' by M. Aitken. *Journal of the Royal Statistical Society B*, 53, 132-133.

Spielberger, C. D., Gonzalez, H. P., Taylor, C. J., Anton, W. D., Algaze, B., Ross, G. R., & Westberry, L. G. (1980). *Preliminary Profession Manual for the Test Anxiety Inventory.*

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Gilks, W. (1996). *BUGS 0.5* Bayesian Inference Using Gibbs Sampling Manual* (Version II). [Computer Program.], MRC Biostatistics Unit, Cambridge.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. (2003). *WinBUGS 1.4* User Manual*. [Computer Program.], MRC Biostatistics Unit, Cambridge.

Ip, E, H. (2002). Locally dependent latent trait model and the Dutch identity revisited. *Psychometrika, 67,* 367-386.

Stout, W. (2002). Psychometrics: From practice to theory and back. *Psychometrika, 67,* 485-518.

Stout, W. & Roussos, L. (1995). *SIBTEST manual.* University of Illinois: Department of Statistics, Statistical Laboratory for Educational and Psychological Measurement.

Sudweeks, R. R., & Tolman, R. R. (1993). Empirical versus subjective procedures for identifying gender differences in science test items. *Journal of Research in Science Teaching, 30,* 3-19.

Swaminathan. H., & Gifford, J.A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Measurement*, 28.237-247

Swaminathan, H., & Gifford, J.A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika*. 50, 349-364.

Swaminathan, H., & Gifford, J.A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika*, 51, 589-601.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*, 361-370.

Thissen, D. (2001*). IRTLRDIF v.2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning*. University of North Carolina at Chapel Hill.

Thissen, D., Steinberg, L., & Mooney, J. (1989). Trace lines for testlets: A use of multiple-categorical response models. *Journal of Educational Measurement*, *26*, 247-260.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response thoery in the study of group differences in trace lines. In H. Wainer & H.I. Braun (Eds.), *Test Validity* (pp. 147-169). Hillsdale, NJ: Lawrence Erlbaum.

Thissen, D. (1991). *MULTILOG user's guide: Multiple, categorical item analysis and test scoring using item response theory* [computer program]. Chicago: Scientific Software.

Tiemey, Luke (1994). Markov chain for exploring posterior distributions. *Annals of*

statistics, 22, 1701-1728.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Low School Admission Test as an example. *Applied Measurement in Education*, *8*(2), 157-187.

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3PL useful in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245-270). Boston, MA: Kluwer-Nijhoff.

Wainer, H., & Lewis, C. (1990). Towards a psychometrics for testlets. *Journal of Educational Measurement, 27*, 1-14.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15*, 22-29.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement*, *28*, 197-219.

Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case of testlets. *Journal of Educational Measurement*, *24*, 185-202.

Wang, W.-C., & Wilson, M. R. (2005A). The Rasch testlet model. *Applied Psychological Measurement*, 29, 126-149.

Wang, W.-C., & Wilson, M.R. (2005b). Assessment of differential item functioning in testlet-based items using the Rasch testlet model. *Educational and Psychological Measurement*, 65, 549-576.

Wollack, J. A., & Cohen, A. S. (2004). *A Model for Simulating Speeded Test Data*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Wollack, J. A., Bolt, D. M., Cohen, A. S., & Lee, Y.-S. (2002). Recovery of item parameters in the nominal response model: A comparison of marginal likelihood estimation and Markov Chain Monte Carlo estimation. *Applied Psychological Measurement, 26(3)*. 339-352.

Ware, W. B., Galassi, J. P., & Dew, K. M. H. (1990). The Test Anxiety Inventory: A confirmatory factor analysis. *Anxiety Research, 3,* 205-212.

Wilson, M., & Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics, 26,* 283-306.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30,* 187-213.

Zhang, J. & Stout, W. (1999). The theoretical detect index of dimensionality and its application to approximate simple structure. *Psychometrika, 64,* 231-249.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 337-347). Hillsdale, NJ: Lawrence Erlbaum.

Zellner, A. (1971). An introduction to Bayesian inference in Econometrics. New York: John Wiley.

Zimmerman, D. W., & Zumbo, B. D. (1990). The effects of outliers on the relative power parametric and nonparametric statistical tests. *Perceptual and Motor Skills, 71*, 339-349.

Zumbo, B. D. (1999). *A handbook on the theory and methods for differential item functioning: Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores.* Ottawa, ON: Directorate of Human Resources Research and Evaluation Department of National Defense.