Practical Assessment, Research, and Evaluation

Volume 14 *Volume 14, 2009* Article 19

November 2019

Differential Item Functioning Amplification and Cancellation in a Reading Test

Han Bao

C. Mitchell Dayton

Amy B. Hendrickson

Recommended Citation
Bao, Han; Dayton, C. Mitchell; and Hendrickson, Amy B. (2019) "Differential Item Functioning Amplification and Cancellation in a Reading Test," *Practical Assessment, Research, and Evaluation*: Vol. 14, Article 19.
DOI: https://doi.org/10.7275/6cmj-q724
Available at: https://scholarworks.umass.edu/pare/vol14/iss1/19

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 14, Number 19, October 2009 ISSN 1531-7714

Differential Item Functioning Amplification and Cancellation in a Reading Test

Han Bao
Shanghai Normal University, People's Republic of China

C. Mitchell Dayton
University of Maryland, College Park, U.S.A.

Amy B Hendrickson
College Board, U.S.A.

When testlet effects and item idiosyncratic features are both considered to be the reasons of DIF in educational tests using testlets (Wainer & Kiely, 1987) or item bundles (Rosenbaum, 1988), it is interesting to investigate the phenomena of DIF amplification and cancellation due to the interactive effects of these two factors. This research applies the 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.

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 that is required in item response theory. 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 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).

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). 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.

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. 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 detected 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. Study of DIF amplification and cancellation can be very useful for test construction purposes. Undetectable item DIF accumulates 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 cancels 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).

This study investigated the interactive effects of secondary testlet dimension and item attributes on the phenomena of DIF amplification and cancellation at both item and testlet levels in applications of the 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.*, 2003). The purpose of this study was to examine 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. The study was conducted using a real dataset.

THE MODEL

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 a 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.

Multiple-group Testlet Model

A 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.

As an extension and application of the random-effects approach using the testlet model, the main interest of this study lay in detecting whether and how testlets function differently for individuals with different group membership. That is, different manifest groups such as genders and ethnic groups may have different mental processes, levels of background knowledge, or learning experiences, which might 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:

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

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).

The addition of the γ parameter reflects the effect of this nuisance dimension. The value of $\gamma_{jd(i)g}$ is constant within a testlet for person j of group g, but the value of $\gamma_{jd(i)g}$ differs for each person of group g. 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.

A special case of Model (1) 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,

$$P(Y_{ijg}=1|\Omega_{ijg}) = \frac{e^{(a_{ig}(\theta_j - b_{ig} + \gamma_{jd(i)g}))}}{1+e^{(a_{ig}(\theta_j - b_{ig} + \gamma_{jd(i)g}))}}, \quad (2)$$

Glas *et al.* (2000) discussed three alternative ways of model formulation regarding to the testlet parameter: (1). as part of ability, assuming the item parameters were 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 through keeping the item parameters 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 research was aimed 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/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:

$$SIGNED-AREA=\int[P_R(\theta)-P_F(\theta)]d\theta, \quad (3)$$

$$UNSIGNED-AREA=\sqrt{\int[P_R(\theta)-P_F(\theta)]^2 d\theta} \quad (4)$$

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:

$$P(Y=1)=\frac{e^{(\psi)}}{1+e^{(\psi)}}, \qquad (5)$$

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, 1995).

Extending Zumbo's (1999) three steps of hypothesis testing, a five-step process is recommended to accommodate these 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 of 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 of 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: Finally, the interaction term based on the nuisance dimension is 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 to Model 3 with 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 was the difference in the R-squared values at each step of DIF modeling.

ΔR^2 is given as:

$$\Delta R^2 = R_2^2 - R_1^2 , \qquad (6)$$

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 (2001) presented guidelines for measurement of magnitude of overall DIF. Negligible DIF: Null hypothesis is retained or null hypothesis is rejected and ΔR^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.

ANALYSIS DESIGN

In order to investigate the phenomena of DIF amplification and cancellation using logistic regression procedure and signed/unsigned area indices, operational data from ACT (American College Testing) reading test made up of testlets was used.

The computer estimation programs WinBUGS1.4 and MATLAB7.2 were used in analyzing the set of real data. The data set was obtained from a released form of the American College Testing (ACT) in Reading (1995). The test chosen for analysis was 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 was consisted of four passages: Prose Fiction, Social Science, Humanities, and Natural Science. All four passages were given equal weight in scoring. There were 3078 females and 2875 males for the analysis of gender DIF and 1271 minority students and 3171 Caucasian students for the analysis of ethnic DIF. Cross-validation procedure was used by randomly partitioning the data into two subsets, one with 1528 females and1432 males and the other one with 1550 females and 1443 males for the gender example, and two samples each with 652 minority and 1550 Caucasians for the ethnicity example, such that the analysis were initially performed on the first subset, while the other subset were 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 model fit. The deviance information criterion (DIC)[1] was used as a model selection index. First, the 2-PL testlet model and 2-PL model were fit to the data to investigate whether there was local dependence due to testlet by comparing the model fit. Next, multiple-group 2-PLM testlet model or multiple- group 2-PLM model were fit to the data to detect whether there were DIF at the whole test level. Finally, a detailed examination of DIF at the item level and testlet level was constructed using ICC/TCC with signed/unsigned area indices, and logistic regression procedure. The cross-validation sample was then used for confirmatory analysis of those three steps.

RESULTS OF 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 existed in real data and its magnitude varied among different subjects of

[1] The deviance information criterion (DIC) is a hierarchical modeling generalization of the AIC (Akaike information criterion) and BIC (Bayesian information criterion, also known as the Schwarz criterion). It is particularly useful in Bayesian model selection problems where the posterior distributions of the models have been obtained by Markov chain Monte Carlo (MCMC) simulation. Like AIC and BIC it is an asymptotic approximation as the sample size becomes large. The deviance information criterion is calculated and provided by WinBUGS output.

content; second, within the same examination, magnitude of person-testlet interaction varied among different subgroups of examinees; third, the item characteristics interacted 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 fits better than the 2-PL model by taking the person–testlet interaction into consideration, it might indicate 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 the one-group 2-PL/2PL testlet models, it might suggest that the test functions differently between the reference group and focal group. Regarding the model identification issue, constraints were set to 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 are shown in Table 1 and Table 2. For both of the gender samples, the DIC's of 2-PLM testlet models were smaller than those of 2-PLM models and additionally, the DIC of the 2-group 2-PL testlet model was decreased by about 197 for sample 1 and by about 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 the 2-group 2-PL testlet model was 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 as well as 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 was 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 was 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

TABLE 1: 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 2: 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

characteristics. Similar results could be confirmed by the second ethnic sample.

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 to as reference group and females were referred to as focal group; for the race sample, Caucasians were referred to 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 additionally, 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 depended on the main dimension. The second sample of gender data and race data were used for this study.

Convergence of Models

Not surprisingly, the means and precisions (1/ variance) of the testlet distributions were 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 Brooks-Gelman-Rubin (BGR) diagnostic plots[2], it was shown that these means and precisions required a burn-in[3] of approximately 35,000 iterations. The one noteworthy indicator was that the BGR diagnostics had stabilized around one. 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 standard deviations to the MC-error ratios[4] were less than the recommended ratio of 0.05 (See Table 3: for gender sample and Table 4: for ethnic sample of testlet mean and precision parameters).

Magnitudes of Differences on Testlet Parameters

When people mention "the content of the ACT Reading Test", it might refer to two different things. The first type of content refers to the subject matter of the passage, which may be thought to be, the testlet effect. The second type refers to the sorts of questions asked about the passage, which may be thought to be, 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, are defined to be measured in ACT reading assessment and are assumed to be able to be acknowledged by every student through certain amount of training in class.

[2] A formal convergence ***diagnostic*** can be implemented using the option ***BGR*** diag in WinBUGS. The Gelman-Rubin statistic is based on the following procedure:1) estimate the model with a variety of different initial values and iterate for an n-iteration burn-in and an n-iteration monitored period; 2) take the n-monitored draws of m parameters and calculate the Gelman-Rubin statistic. Once convergence is reached, the Gelman-Rubin statistic should approximately equal one.

[3] Number of 'burn in' samples is the input required as part of the MCMC estimation procedure. The Metropolis-Hastings algorithm randomly samples from the posterior distribution. Typically, initial samples are not completely valid because the Markov Chain has not stabilized. The burn-in samples allow discarding these initial samples. The GIBBSIT procedure will estimate the necessary burn-in and sample size to collect from the posterior distribution.

[4] The index provided by WinBUGS to assess how well the estimation is doing by comparing the mean of the samples, and the true (estimated) posterior mean. The ratio is called the *MC error*. Rule of thumb indicated in WinBUGS manual is it is OK if this is less than5% of the true error.

TABLE 3: 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	10.4700
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	0.0055	0.0020	0.8535
taud2	**1.3140**	0.17190	0.008595	0.0064	0.0022	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 4: 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.

However, different subjects of passages might mean different things to minorities and majorities because of their different levels of familiarity with the cultures and also could mean different things to females and males because of certain different cognitive attributes between them, 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 of 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 are listed in Table 3 for gender sample and Table 4 for the ethnic sample.

For the first passage, males scored 0.9685 on average, which was less than females' scores since girls seemed to be more interested in and more familiar with the topics related to parties such as foods, dresses, *etc.*; not surprisingly, minorities scored lower on average than Caucasians by about 0.7 because of their 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 to 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 may be 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 four testlets. It could suggest that certain level of Natural Science background knowledge was highly required in order to understand the content of this passage.

In contrast with the R^2 based effect size indices of logistic regression procedure (in Appendix Table A-3 and Appendix Table A-7), for the gender sample in Appendix Table A-3, 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 Appendix Table A-7, 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.

Magnitudes of Difference on Item Characteristic Features

Appendix Table A-1 and Appendix Table A-2 show that items are 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 of the difference between the item difficulties and item discriminations in the two subgroups do not contain zero. Appendix Table A-3 lists the R^2 based effect sizes of logistic regression procedure of detecting DIF. Items that are bolded in the table are those for which the magnitudes of DIF were relatively larger than others. Appendix Table A-4 lists the regression coefficients. Again, the coefficients showing statistically significant different from zero are bolded in the Appendix Table A-4. The "+"sign of the values of regression coefficients indicates females group is favored and "-" sign indicates that the males group is favored.

For the item difficulty parameters, the largest DIF between two subgroups were 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 Appendix Table A-3, where $R_3^2 - R_2^2$ suggesting the magnitude of DIF was due to item difficulty after conditioning on the effects of testlet and main latent trait. However, for some reason, the large amount of

DOI: https://doi.org/10.7275/6cmj-q724

DIF on Item 7 was 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 Appendix Table A-4. 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 amounts of DIF around 0.4 ~ 0.5 were found on items 16, 23, 6 and 9. They were confirmed by the results in Appendix Table A-3 and Appendix Table A-4 except the DIF on item 9 could not be detected by R^2 based effect sizes. Finally, negligible amounts of DIF on item difficulty parameters were detected for Item 3, 5, 13, 14, 17, 18, 19, 22, 25, 26, 32 and 37.

As to the item discrimination parameter, a large amount of DIF was detected for Item 7 with the mean adif of 0.3199, indicating that the item discriminated more highly among males than females. The result was consistent with those of logistic regression procedures, where $R_4^2 - R_3^2$ and $R_5^2 - R_3^2$ indicated the interaction of item with the main dimension θ and with the secondary dimension γ separately. 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 interaction of subgroups and the main dimension θ, as τ_4 , and interaction of subgroups and the testlet dimension, as τ_5 , respectively). A small magnitude of DIF was detected on Items 12 and 26, and negligible magnitudes of DIF were detected on Items 10, 23 and 29.

Appendix Table A-5 and Appendix Table A-6 show 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. Appendix Table A-7 lists the R^2 based effect sizes of logistic regression procedure of detecting DIF. Appendix Table A-8 lists the regression coefficients.

Large magnitudes of DIF on item difficulty parameters were detected on Items 1, 4, and 34, where Caucasians were favored. Unfortunately, due to the reason mentioned above, DIF on item difficulty parameters of Items 1 and 4 was not detected by R^2 based effect size indices. Evidence of relatively moderate magnitudes of DIF was found on Items 8 and 9, where they seemed unusually easy for the minority group. A detailed study of these two items revealed that the simplest reading comprehension skill---identify specific details and facts directly in text was required but some Caucasian students seemed to have the incorrect answer because of their own empirical understandings about the content. For example, Item 9 asked the 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, many Caucasian students selected one of the distractions: "somewhat festive but socially insincere" while relatively more minority students gave correct answers to these items. Small or negligible amounts of DIF on item difficulty parameters were detected on Items 12 and 26. Moderate amount of DIF on item discrimination parameters were detected on Item 1; negligible amounts of DIF were detected on Items 12, 16, 25 and 28.

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

The evidence of DIF amplification and cancellation at item and testlet levels was investigated using signed/ 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. Appendix Table A-9 and Appendix Table A-10 list the results of the indices of the two examples.

DIF Amplification and Cancellation at the item level

Evidence of the phenomena of DIF amplification at the item level was found on one example of Item 4 nested within the first testlet of ethnic sample and the phenomena of DIF cancellation at the item level was 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 comparisons, Item 6 reflected the DIF attributed only to the testlet effect, where there were no significantly statistical differences on item difficulty parameters and item discrimination parameters between the two subgroups (see Figure 1). 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 2).

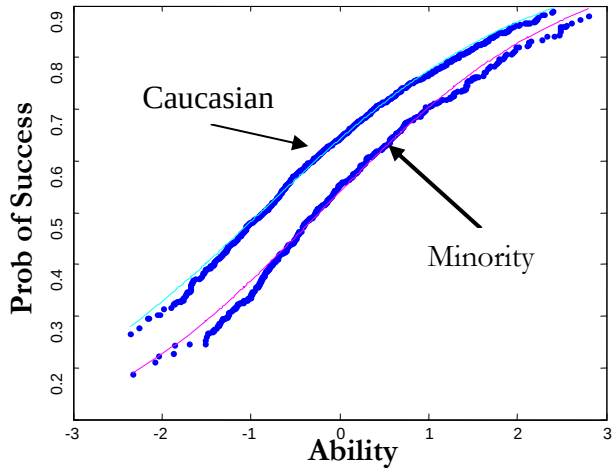

FIGURE 1: ICC of Item 6 of Ethnic Sample

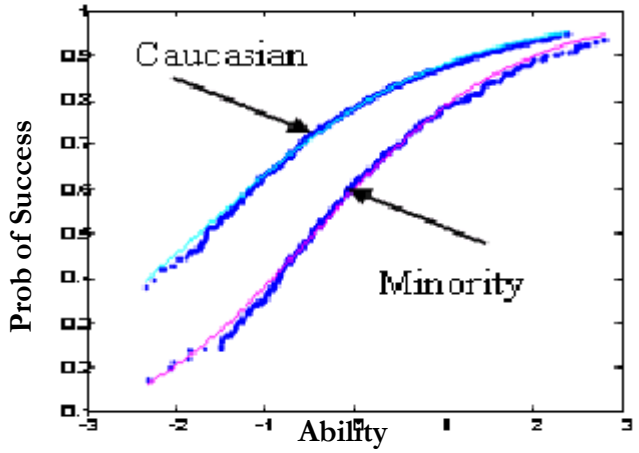

FIGURE 2: ICC of Item 4 of Ethnic Sample

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 that Caucasians had higher abilities on testlet dimension, the difference on the item difficulty parameters between the two subgroups was 0.4887 and the Minority group was favored (See Figure 3). The magnitudes of DIF of these three items measured by signed-area and unsigned-area indices are shown in Appendix Table A-10. The other kind of DIF cancellation at the item level because of crossing of ICCs was detected on Item 29 of the gender example (See Figure 4). 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 scored about 0.2 higher on the means of testlet distribution than males, the ICC of females group shifted slightly to the left and, thus, the 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 the signed-area index but was detected by the unsigned area index (See Appendix Table A-9).

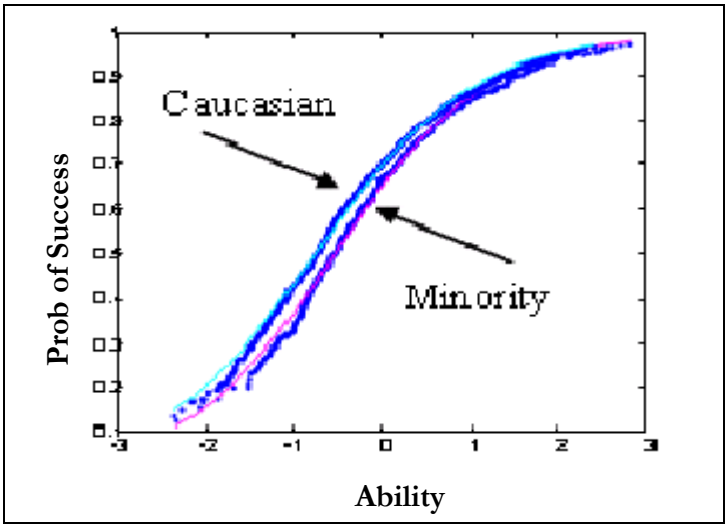

FIGURE 3: ICC of Item 8 of Ethnic Sample

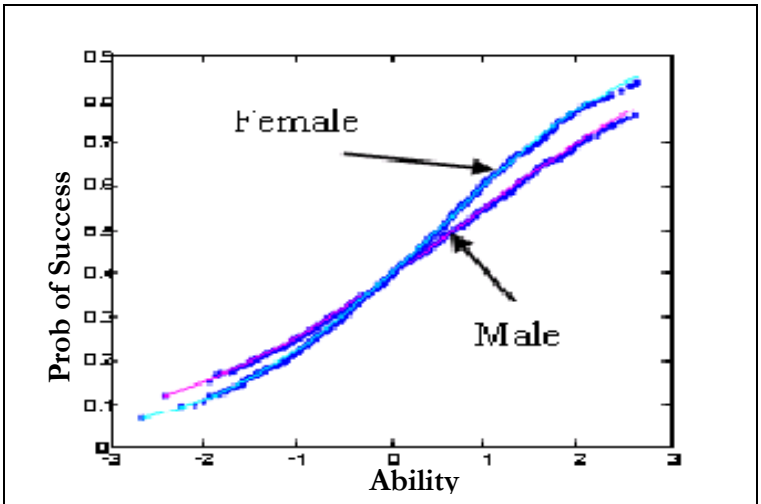

FIGURE 4: ICC of Item 29 of Gender Sample

DIF Amplification and Cancellation at the testlet level

Evidence of DIF amplification and cancellation at the testlet level was 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 5 as an example) and nearly half of them functioned the opposite way (See Figure 6 as an example), and, thus, the cumulative effect of DIF cancelled out at the testlet level (See Figure 7). See Appendix Table A-9 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 Appendix Table A-10 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 6 as an example).

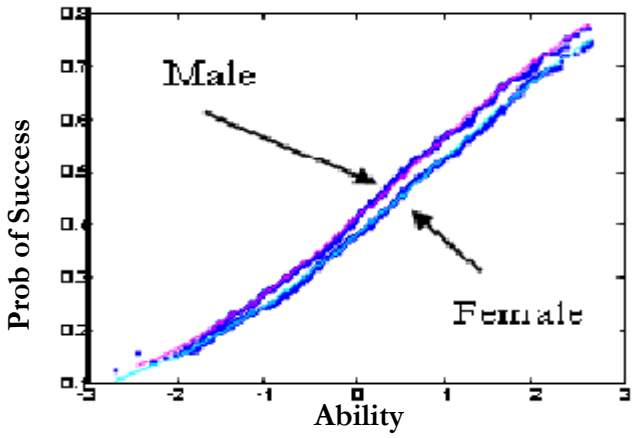

FIGURE 5: ICC of Item 39 of Gender Example

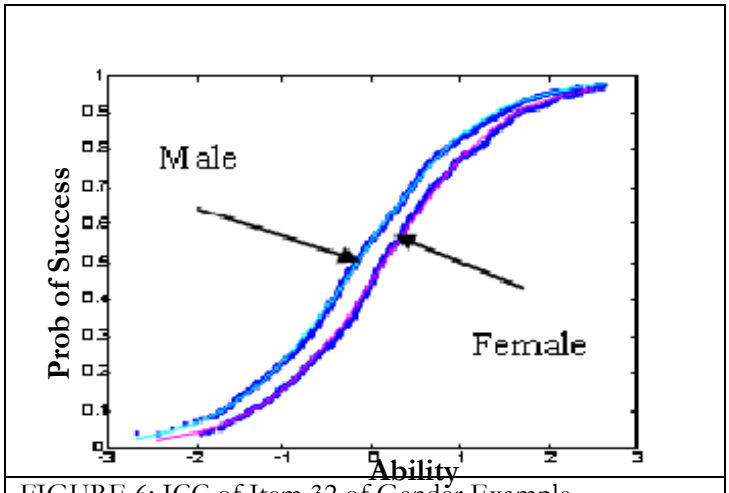

FIGURE 6: ICC of Item 32 of Gender Example

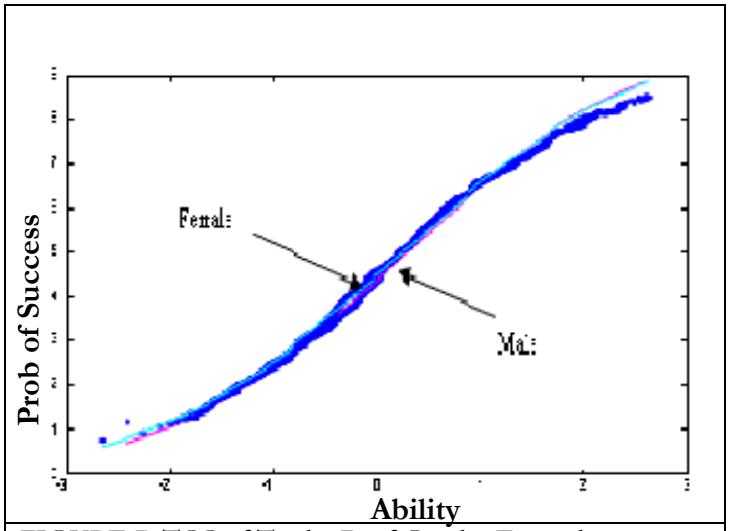

FIGURE 7: TCC of Testlet D of Gender Example

SUMMARY

In summary, in applications of logistic regression and signed/unsigned area indices, the analyses of the real data obtained from ACT reading test revealed that the person-testlet interaction effect existed and the phenomena of DIF amplification and cancellation could be attributed to comprehensive DIF effects of testlet distributions and idiosyncratic features of items within testlets. As indicated by the results of real data analysis, the magnitudes of person-testlet interaction effects, embodied in 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. Actually, larger magnitudes of difference on the testlet effect were found in ethnic samples than that in gender samples. The phenomena of

DIF amplification and cancellation were also detected in the real data analysis taking advantages of the statistical procedures applied in this study.

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 real data analysis, logistic regression procedure and signed/unsigned area indices based on item response theory demonstrated their effectiveness in assessing DIF at two levels. The signed/ unsigned area indices were useful for providing magnitude measure of DIF and logistic regression was useful for identifying items with DIF and, also, for explaining the sources of 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 took advantages of the multiple-group item response testlet model proposed by Li, *et al.* 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 analysis of real data led to the following conclusions:

First, the homogeneous functioning of testlet effect and item difficulty parameters between the two subgroups seemed to be 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 seemed to be the reason for DIF cancellation at the item level. More generally, the 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 might be due to the existence of testlet effects and the reason for DIF cancellation at the testlet level might be due to the heterogeneous functioning of individual items nested within the testlet.

Third, the person-testlet interaction effect was detected in real ACT test data. The magnitude of this effect seemed to vary 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 (1987) raised the issue that, because decisions were made at a level higher than the item, the study of DIF at the item level might only have limited importance. Since many current assessments, especially language tests, are made up of testlets, it is 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 an attractive solution to yield a set of DIF-balanced test construction units. 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 does 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, provides 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 is still important to assess whether the statistically significant amount of DIF present is of practical importance and also enough sample size is 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 can be found and it must also use the most efficient statistical model available to analyze data. One underlying assumption always exists albeit often overlooked is that IRT model that is assumed to underlie the individual item responses is 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 future study. Moreover, although manifest groups such as gender and racial groups can be easily identified for use in traditional DIF studies, there is the fact that manifest groups lack homogeneity and the possibility that these groups are

not really the groups affected by DIF. Thus, a latent class approach using latent grouping variables to allow for the assessment of DIF without tying that DIF to any specific set of variables is a possible future approach for making a more definitive investigation of the presence of DIF.

REFERENCES

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.

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.

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

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.

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.

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.

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129- 145). Hillsdale, NJ: Erlbaum.

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

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.

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

Li, Y., Bolt, D.M. & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement*, 30, 3-21.

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

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.

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.

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

Rosenbaum, P. R. (1988). Item bundles. *Psychometrika,* 53, 349-359.

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

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.

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.

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

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

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.

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

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

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.

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.

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.

APPENDIX

TABLE A-1: DIF Analysis of Item Difficulty Parameters of Gender Example

Node	Mean	SD	MC error	2.50%	Median	97.50%
Testlet A						
difb[1]	-0.10270	0.20660	0.004978	-0.50540	-0.10680	0.31760
difb[2]	-0.07048	0.17980	0.004094	-0.41650	-0.07379	0.29100
difb[3]	**-0.21630**	**0.12440**	**0.002524**	**-0.46050**	**-0.21610**	**0.02824**
difb[4]	0.16090	0.18090	0.003842	-0.18310	0.15740	0.52670
difb[5]	**-0.15610**	**0.10840**	**0.002581**	**-0.36800**	**-0.15620**	**0.05826**
difb[6]	**-0.34300**	**0.21360**	**0.004160**	**-0.74360**	**-0.35120**	**0.09494**
difb[7]	**0.88360**	**0.33190**	**0.011960**	**0.28000**	**0.86390**	**1.58400**
difb[8]	0.08549	0.11770	0.002637	-0.14440	0.08531	0.31640
difb[9]	**-0.40540**	**0.12650**	**0.002657**	**-0.65380**	**-0.40570**	**-0.15440**
difb[10]	0.16410	0.16700	0.002404	-0.14570	0.16000	0.50600
Testlet B						
difb[11]	0.01597	0.16680	0.003601	-0.32150	0.01744	0.33740
difb[12]	-0.04680	0.14830	0.003248	-0.35270	-0.04119	0.23120
difb[13]	**0.31000**	**0.08970**	**0.001404**	**0.13660**	**0.30830**	**0.48910**
difb[14]	**-0.30800**	**0.09294**	**0.001199**	**-0.49520**	**-0.30700**	**-0.13040**
difb[15]	0.13500	0.12880	0.002048	-0.12530	0.13710	0.38510
difb[16]	**0.40340**	**0.09899**	**0.001296**	**0.21090**	**0.40210**	**0.60010**
difb[17]	**0.25660**	**0.08075**	**0.001200**	**0.10070**	**0.25690**	**0.41620**
difb[18]	**-0.19280**	**0.08174**	**0.001069**	**-0.35260**	**-0.19280**	**-0.03259**
difb[19]	**0.22920**	**0.09441**	**0.001482**	**0.04450**	**0.22920**	**0.41400**
difb[20]	**-0.80270**	**0.16410**	**0.002941**	**-1.14300**	**-0.79610**	**-0.49920**
Testlet C						
difb[21]	-0.08023	0.09739	0.001630	-0.27200	-0.08039	0.11220
difb[22]	**0.23530**	**0.09491**	**0.001625**	**0.05234**	**0.23400**	**0.42730**
difb[23]	**0.58340**	**0.15050**	**0.003246**	**0.30310**	**0.57800**	**0.89270**
difb[24]	-0.05519	0.07604	9.42E-04	-0.20270	-0.05511	0.09284
difb[25]	**-0.21710**	**0.07815**	**9.01E-04**	**-0.37040**	**-0.21620**	**-0.06645**
difb[26]	**-0.32280**	**0.07504**	**0.001007**	**-0.46900**	**-0.32270**	**-0.17620**
difb[27]	-0.12410	0.06634	7.76E-04	-0.25570	-0.12410	0.00589
difb[28]	-0.16000	0.08541	0.00141	-0.32660	-0.16010	0.00774
difb[29]	-0.00080	0.13400	0.00231	-0.25550	-9.59E-04	0.27140
difb[30]	0.13990	0.09844	0.00124	-0.05149	0.13880	0.33680
Testlet D						
difb[31]	-0.08681	0.13550	0.00427	-0.36050	-0.08547	0.17710
difb[32]	**0.28450**	**0.12520**	**0.00421**	**0.03400**	**0.28510**	**0.53040**
difb[33]	-0.16760	0.13670	0.00470	-0.44040	-0.16550	0.09622
difb[34]	0.09170	0.13960	0.00473	-0.18770	0.09262	0.36390
difb[35]	0.01212	0.13890	0.00379	-0.26710	0.01253	0.28390
difb[36]	-0.12050	0.15710	0.00393	-0.43230	-0.11950	0.18770
difb[37]	**-0.33710**	**0.15930**	**0.00411**	**-0.65280**	**-0.33580**	**-0.02637**
difb[38]	0.07977	0.17350	0.00365	-0.26500	0.07946	0.42090
difb[39]	-0.39080	0.21130	0.00353	-0.81210	-0.38780	0.01833
difb[40]	0.63480	0.83860	0.03417	-0.96900	0.60770	2.37400

TABLE A-2: 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	-0.14940	0.07881	0.30970
difa[2]	-0.01535	0.14530	0.003642	-0.30090	-0.01630	0.27070
difa[3]	**0.21740**	**0.14540**	**0.002714**	**-0.06702**	**0.21750**	**0.50360**
difa[4]	-0.06170	0.10640	0.002250	-0.27130	-0.06188	0.14730
difa[5]	-0.13920	0.13307	0.002244	-0.40060	-0.13810	0.12220
difa[6]	0.12450	0.08405	0.001538	-0.04106	0.12470	0.29130
difa[7]	**0.31990**	**0.09535**	**0.002690**	**0.13110**	**0.32090**	**0.50300**
difa[8]	-0.09050	0.11020	0.001511	-0.30590	-0.09091	0.12570
difa[9]	-0.07703	0.09534	0.001304	-0.26500	-0.07645	0.10820
difa[10]	**0.20740**	**0.12170**	**0.002171**	**-0.03176**	**0.20720**	**0.44610**
Testlet B						
difa[11]	-0.08951	0.11040	0.002319	-0.30430	-0.09008	0.13000
difa[12]	**-0.30290**	**0.16100**	**0.003433**	**-0.62100**	**-0.30110**	**0.01032**
difa[13]	0.06761	0.15360	0.002659	-0.22960	0.06654	0.37040
difa[14]	-0.00111	0.12130	0.001851	-0.23920	-0.00173	0.23800
difa[15]	-0.15440	0.12020	0.001972	-0.39150	-0.15370	0.08113
difa[16]	-0.00907	0.10190	0.001300	-0.20790	-0.00938	0.19090
difa[17]	0.14470	0.12120	0.001637	-0.08955	0.14400	0.38640
difa[18]	0.19550	0.12009	0.001632	-0.03826	0.19510	0.43380
difa[19]	0.05060	0.11050	0.001496	-0.16420	0.05007	0.26930
difa[20]	0.08035	0.10140	0.001704	-0.11870	0.08043	0.27900
Testlet C						
difa[21]	-0.00249	0.14001	0.002385	-0.27820	-0.00107	0.27340
difa[22]	-0.01042	0.14170	0.002422	-0.28990	-0.00996	0.26810
difa[23]	0.03811	0.10690	0.002038	-0.16930	0.03726	0.24760
difa[24]	0.10450	0.13610	0.001879	-0.16000	0.10430	0.37220
difa[25]	-0.08945	0.12430	0.001860	-0.33330	-0.08915	0.15320
difa[26]	**0.34950**	**0.12970**	**0.001898**	**0.09668**	**0.34870**	**0.60820**
difa[27]	-0.05898	0.14190	0.002006	-0.33730	-0.05965	0.21880
difa[28]	-0.01824	0.12850	0.001975	-0.27140	-0.01882	0.22950
difa[29]	**-0.17120**	**0.08929**	**0.001328**	**-0.34510**	**-0.17110**	**0.00440**
difa[30]	0.08928	0.10840	0.001599	-0.12490	0.08962	0.29920
Testlet D						
difa[31]	0.01698	0.10940	0.002103	-0.19660	0.01607	0.23470
difa[32]	-0.05541	0.13210	0.002285	-0.31380	-0.05515	0.20210
difa[33]	-0.16700	0.12580	0.002305	-0.41130	-0.16800	0.08451
difa[34]	0.10000	0.11650	0.002019	-0.12620	0.09931	0.32960
difa[35]	-0.01602	0.10820	0.002039	-0.22790	-0.01531	0.19460
difa[36]	-0.13640	0.08587	0.001421	-0.30490	-0.13620	0.03172
difa[37]	-0.12680	0.08430	0.001301	-0.29390	-0.12690	0.03852
difa[38]	-0.03055	0.09003	0.001691	-0.20070	-0.03038	0.14500
difa[39]	0.03172	0.07344	0.001189	-0.11040	0.03122	0.17620
difa[40]	-0.06490	0.06345	0.002378	-0.18880	-0.06501	0.06116

TABLE A-3: 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						0.0155			
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						0.0156			
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						0.0167			
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						0.1503			
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 A-4: 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

TABLE A-5: DIF Analysis of Item Difficulty Parameters of Ethnic Example

Node	Mean	SD	MC error	2.50%	Median	97.50%
Testlet A						
difb[1]	**0.94980**	**0.3001**	**0.011230**	**0.41620**	**0.92890**	**1.59300**
difb[2]	-0.08906	0.1750	0.003675	-0.42870	-0.08808	0.25590
difb[3]	-0.03996	0.1413	0.002768	-0.31860	-0.04037	0.24130
difb[4]	**0.71400**	**0.2137**	**0.005366**	**0.32610**	**0.70430**	**1.15900**
difb[5]	-0.07201	0.1362	0.002701	-0.33850	-0.07184	0.19110
difb[6]	0.002077	0.1746	0.002986	-0.33500	-0.00136	0.35480
difb[7]	-0.32220	0.3198	0.009072	-0.96980	-0.31510	0.26850
difb[8]	**-0.48870**	**0.1317**	**0.002667**	**-0.75400**	**-0.48720**	**-0.23660**
difb[9]	**-0.48770**	**0.1549**	**0.003251**	**-0.79480**	**-0.48550**	**-0.18430**
difb[10]	-0.16630	0.1992	0.002514	-0.56550	-0.16520	0.21940
Testlet B						
difb[11]	-0.03778	0.19680	0.003747	-0.44170	-0.03154	0.33790
difb[12]	**-0.29210**	**0.21380**	**0.004772**	**-0.75020**	**-0.28020**	**0.09283**
difb[13]	-0.10860	0.12420	0.002150	-0.36040	-0.10630	0.12770
difb[14]	-0.10590	0.11170	0.001522	-0.33140	-0.10570	0.11250
difb[15]	0.09901	0.19330	0.003864	-0.29740	0.10340	0.46580
difb[16]	-0.00137	0.12270	0.001786	-0.23740	-0.00434	0.24270
difb[17]	0.02145	0.11960	0.001852	-0.20550	0.01868	0.26500
difb[18]	-0.06604	0.11010	0.001642	-0.27840	-0.06671	0.15180
difb[19]	0.12510	0.14560	0.002385	-0.14520	0.11860	0.42920
difb[20]	0.36620	0.27670	0.005379	-0.10590	0.34210	0.97820
Testlet C						
difb[21]	0.05679	0.12980	0.002525	-0.19790	0.05749	0.31070
difb[22]	-0.02023	0.11510	0.002046	-0.24680	-0.01959	0.20170
difb[23]	0.10200	0.17420	0.004429	-0.23670	0.10110	0.45260
difb[24]	-0.09023	0.09745	0.001268	-0.28000	-0.09094	0.10360
difb[25]	0.08529	0.10730	0.001242	-0.11950	0.08392	0.29990
difb[26]	**-0.18560**	**0.09121**	**0.001274**	**-0.36260**	**-0.18620**	**-0.00285**
difb[27]	-0.07009	0.09135	0.001334	-0.24630	-0.07095	0.11190
difb[28]	-0.05849	0.12440	0.002441	-0.28970	-0.06350	0.20140
difb[29]	0.20010	0.17390	0.003309	-0.11330	0.19010	0.57550
difb[30]	-0.01958	0.12910	0.001730	-0.25940	-0.02384	0.25060
Testlet D						
difb[31]	0.16350	0.16430	0.004512	-0.15460	0.16410	0.48640
difb[32]	0.17220	0.14840	0.004273	-0.12270	0.17220	0.46250
difb[33]	0.09963	0.16540	0.004896	-0.22650	0.10040	0.42310
difb[34]	**0.47430**	**0.17200**	**0.004935**	**0.13190**	**0.47410**	**0.81460**
difb[35]	0.23110	0.18850	0.004078	-0.12590	0.22690	0.61370
difb[36]	0.33140	0.17730	0.004319	-0.01078	0.32990	0.68310
difb[37]	0.02936	0.20500	0.004337	-0.36710	0.02721	0.43970
difb[38]	0.19740	0.22000	0.004296	-0.21120	0.19160	0.64470
difb[39]	-0.13540	0.23010	0.003916	-0.57260	-0.14060	0.33260
difb[40]	-1.56400	0.92510	0.035030	-3.35800	-1.56400	0.29540

TABLE A-6: DIF Analysis of Item Discrimination Parameters of Ethnic Example

Node	Mean	SD	MC error	2.50%	Median	97.50%
Testlet A						
difa[1]	**0.39900**	**0.14160**	**0.003524**	**0.13440**	**0.39490**	**0.6906**
difa[2]	0.11250	0.18750	0.004065	-0.24140	0.10680	0.49320
difa[3]	0.16670	0.19150	0.004114	-0.19870	0.16080	0.55190
difa[4]	0.13280	0.13000	0.002850	-0.11690	0.13060	0.39490
difa[5]	0.00419	0.15010	0.002514	-0.28440	8.64E-04	0.30470
difa[6]	0.03734	0.11470	0.001909	-0.18150	0.03619	0.26740
difa[7]	-0.02790	0.10770	0.002821	-0.23200	-0.03023	0.18430
difa[8]	0.02016	0.15450	0.002538	-0.27590	0.01740	0.32830
difa[9]	-0.11510	0.12210	0.002001	-0.35050	-0.11760	0.12540
difa[10]	0.00375	0.13740	0.002292	-0.25980	0.00118	0.27770
Testlet B						
difa[11]	-0.04498	0.13840	0.002763	-0.31140	-0.04596	0.23560
difa[12]	**-0.38330**	**0.17120**	**0.003960**	**-0.71700**	**-0.38510**	**-0.03942**
difa[13]	-0.06591	0.19230	0.003306	-0.43410	-0.07019	0.31870
difa[14]	-0.16020	0.15640	0.002466	-0.46010	-0.16310	0.15220
difa[15]	-0.11800	0.13760	0.002802	-0.38340	-0.11990	0.15670
difa[16]	**0.31420**	**0.13900**	**0.001874**	**0.05551**	**0.30970**	**0.59740**
difa[17]	-0.18960	0.14200	0.001916	-0.46420	-0.19180	0.09171
difa[18]	-0.18920	0.15240	0.002233	-0.47990	-0.19350	0.11740
difa[19]	-0.07618	0.12900	0.001852	-0.32510	-0.07699	0.18210
difa[20]	-0.19350	0.11830	0.002106	-0.42310	-0.19550	0.04458
Testlet C						
difa[21]	0.01375	0.16260	0.003117	-0.29980	0.01060	0.33690
difa[22]	-0.16690	0.18030	0.003254	-0.51190	-0.16990	0.19730
difa[23]	0.11930	0.13790	0.002958	-0.14410	0.11590	0.39580
difa[24]	0.04945	0.16760	0.002514	-0.27040	0.04534	0.38730
difa[25]	**-0.30740**	**0.15100**	**0.002287**	**-0.60190**	**-0.30910**	**-0.01030**
difa[26]	0.08698	0.18150	0.002829	-0.26350	0.08249	0.45380
difa[27]	-0.08755	0.17650	0.002740	-0.42520	-0.09159	0.26640
difa[28]	**-0.31840**	**0.15450**	**0.002867**	**-0.61360**	**-0.32180**	**-0.01224**
difa[29]	0.00213	0.11640	0.002054	-0.22050	7.70E-05	0.23640
difa[30]	0.00351	0.14570	0.002455	-0.27570	0.00243	0.29630
Testlet D						
difa[31]	-0.06947	0.13730	0.002603	-0.33270	-0.07260	0.20270
difa[32]	0.19430	0.17980	0.003076	-0.14280	0.18730	0.56570
difa[33]	-0.04290	0.14670	0.002605	-0.32180	-0.04696	0.25790
difa[34]	-0.13520	0.13490	0.002284	-0.39590	-0.13580	0.13360
difa[35]	-0.11070	0.12610	0.002297	-0.35050	-0.11210	0.13900
difa[36]	0.14190	0.12290	0.002250	-0.08962	0.13840	0.38940
difa[37]	-0.08644	0.10010	0.001574	-0.27810	-0.08710	0.11350
difa[38]	0.04689	0.11550	0.002294	-0.17110	0.04401	0.27910
difa[39]	0.18270	0.10680	0.001914	-0.01977	0.18060	0.39930
difa[40]	0.09759	0.07353	0.002411	-0.04192	0.09635	0.24480

TABLE A-7: 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	0.9320	0.9940	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 A-8: 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

TABLE A-9: 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 A-10: 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**

Citation

Bao, Han, Dayton, C. Mitchell, & Hendrickson, Amy B. (2009). Differential Item Functioning Amplification and Cancellation in a Reading Test. *Practical Assessment, Research & Evaluation*, 14(19). Available online: http://pareonline.net/getvn.asp?v=14&n=19.

Note

The first author was sponsored by the Shanghai Pujiang Program.

Authors' Correspondence:

Dr. Han Bao, Assistant Professor
100 Guilin Road, Shanghai, China,
College of Education, Shanghai Normal University,
People's Republic of China, 200234
E-mail: hbao2008 [at] shnu.edu.cn

Dr. C. Mitchell Dayton, Professor Emeritus
Department of Measurement, Statistics & Evaluation,
University of Maryland
College Park, MD 20742
E-mail: cdayton [at] umd.edu

Dr. Amy B. Hendrickson, Associate Psychometrician
The College Board, U.S.A.
1233 20th Street, NW Suite 600
Washington, DC 20036-2375
E-mail: ahendrickson [at] collegeboard.org