The Pennsylvania State University

The Graduate School

College of Education

DEVELMOPMENT OF A DIFFERENTIAL ITEM FUNCTIONING (DIF) PROCEDURE USING THE HIERARCHICAL GENERALIZED LINER MODEL: A COMPARISON STUDY WITH LOGISTIC REGRESSION PROCEDURE

A Thesis in

Educational Psychology

by

Wonsuk Kim

© 2003 Wonsuk Kim

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2003

We approve the thesis of Wonsuk Kim.

Date of Signature

Hoi K. Suen
Professor of Educational Psychology
Thesis Advisor
Chair of Committee

Jaeyong Lee
Assistant Professor of Statistics

Michael J. Rovine
Associate Professor of Human Development
& Family Studies

Rayne A. Sperling
Assistant Professor of Education

Robert J. Stevens
Associate Professor of Educational
Psychology
Professor in Charge of Graduate Programs
in Educational Psychology

ABSTRACT

Based on Kamata's item analysis model (2001), an extension for differential item functioning procedure was developed and the applicability was examined. The Kamata's item analysis model is a type of hierarchical generalized linear model for item analysis, in which items are considered as nested in examinees. Kamata showed that the item difficulty estimates and person ability estimates are mathematically equivalent to the Rasch model. Extending the Kamata model by adding matching variable and group membership variable and their interaction terms in the level-2 model, the approach to DIF, which is named the Hierarchical Generalized Linear Modeling DIF, was constructed. The main purpose of this study was to examine the equivalency of this procedure with a traditional DIF method, the logistic regression DIF method. An English test that consisted of 60 items with 4 sub-scales for freshmen's diagnostic purpose was analyzed to answer the research questions in this study. For the extensive investigation of the two DIF methods, four different group membership variables were administrated; Male vs. Female, White vs. Asian, White vs. Black, Asian vs. Black. The comparisons were made with the coefficients for the group membership variable and interaction term. In addition to that, *p*-values of the regression coefficients were compared. The results of the empirical comparison show that the DIF outputs are highly similar to each other in both uniform DIF and non-uniform DIF. Beyond the similarity with the logistic regression DIF method, features of the HGLM DIF method and its strengths and weaknesses were discussed. Furthermore non-parametric examination of the DIF item by using the TestGraf program was considered in order to find the prominent difference in pattern of DIF identified by specific procedures.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

The purpose of this study is to evaluate a statistical procedure for detecting items on a test that are differentially more difficult for one group of examinees than another after controlling for possible group differences in the attribute being measured by the test. This type of analysis is one of many procedures used to evaluate bias in a test and its items.

Test bias is essentially a deficiency in the validity of inferences made from test scores. The dilemma of possible gender and/or ethnic biases in the use of many tests has drawn concerns among test developers as well as the public in general. Court decisions have been made in some instances to prohibit the use of certain tests for admission decisions because of evidence that these tests are biased against female and/or minority examinees (Linn & Drasgow, 1987).

Given that a test is never perfect, scores on many tests do in fact reflect, to some degree, variables other than the intended attribute or construct. This phenomenon implies a threat to the validity of inferences and uses of test scores. In a general sense, the inferences and uses of scores on such tests can be considered biased against individuals with certain characteristics on the unintended variable. However, the term test bias is usually restricted to evidence of bias that is systematic against a particular group. Specifically, when the unintended variables are correlated with characteristics such as gender, ethnicity, and socio-economic background, the use of these test scores for certain inferences or decisions may give unfair advantages to some demographic groups of

individuals. When this association occurs, the test is said to be biased in a more restricted sense. Therefore, a more appropriately restricted definition of a biased test is one the score on which reflects not only the quantity of an unintended construct, but some other socio-demographic variables as well (Ackerman, 1992).

Before proceeding in detail, it is necessary to clarify some related terms such as test bias, item impact, differential item functioning (DIF), and item bias. First, *test bias* is one in which there are systematic differences in the meaning of test scores associated with group membership. It is a characteristic of the test as a whole, and is analyzed at the level of total test score. It is not simply a compound concept because the characteristics of the sum of items can be quite different from the characteristics of the items. For example, the absence of biased items is not a sufficient safeguard against the biased uses of, or inferences from, total scores. Furthermore, if the number of items that favors a group is approximately equal to the number of items that disadvantage the group, the biases may cancel out in the summed, total score so that it has little or no predictive bias (Hong & Roznowski, 2001).

Second, a distinction can be made between *item impact* and DIF (Dorans & Holland, 1993). Item impact refers to a group difference in measured performance on a test or items. Whenever groups differ on the attributes measured by tests, impact is unavoidable - mean total score differences will lead to different passing rates by groups for items. In contrast, DIF refers to differences in the functioning of an item among groups that are *conditioned* on the attribute measured by the test. Specifically item impact is based on the proportion passing the item regardless of total score, whereas DIF indexes compare the proportion passing *conditional* on total score or some estimate of the

latent attribute. While DIF procedures assume a calculation after matching on the underlying construct that item is intended to measure, the item impact occurs when examinees from different groups have differing probabilities of responding correctly to an item because there are true differences between the groups in the underlying ability being measured by the item (Camilli & Shepard, 1994).

Camilli and Shepard recommend that items that show the statistical discrepancy known as DIF undergo a judgmental procedure to uncover the source of the unintended group differences in item difficulty. If it is found that the source of the group difference is due to an unintended construct that is irrelevant to the attribute that the test was designed to measure, then the item is considered biased. Thus, DIF is required, but not sufficient for item bias. Furthermore, DIF is a statistical term and item bias is a judgmental term.

It is useful to regard group differences in item (and test) performance as consisting of two components: the "true" difference between the groups and an "artificial" difference brought about by the use of inappropriate and irrelevant (biased) items. These two effects, in sum, have been referred to as *impact.* The use of the total score criterion or latent trait variable as a matching variable serves to separate out these two effects.

Brief historical overview of DIF methods

Although it is necessary to construct items carefully to avoid possible items with DIF, there is no guarantee that a carefully constructed item will be free of bias. A

number of statistical methods have been developed to assess the possibility of item bias and DIF based on test data.

Concerns about DIF, the statistical investigation of item bias, have a long history. In the first stage of the development of DIF, researchers used traditional statistical techniques (Angoff, 1993) such as the analysis of variance procedure to test the interaction of item difficulty with group (Camilli & Shepard, 1987), the delta-plot or *transformed* item difficulty, and the Golden rule procedure, which gives preferences in test construction to items with group item difficulty differences of 0.15 or less (Faggen, 1987; Linn & Drasgow, 1987). These methods involve the calculation of item difficulty values for each group, which can be quite inaccurate under common circumstances (Pennock-Roman, 1986). When groups differ substantially in average ability score, items that have good discrimination between low and high scoring examinees will have substantially different item difficulty values as compared to items with more moderate discrimination parameters. Under these conditions, items with high discrimination that are not biased will appear to be biased, and biased items with low discrimination will be overlooked (Camilli & Shepard, 1994). The other methods using *Chi-square* tests have good accuracy, but they are less satisfactory on other grounds. For example, Scheunemann's *Chi-square* lacks both an inferential test and an adequate descriptive effect size index, and the full chi-square method also lacks an effect size index.

Early methods for assessing measurement bias generally have been replaced by more sophisticated statistical techniques, such as the Mantel-Haenszel procedure (Holland & Thayer, 1988), the standardization method (Dorans & Kulick, 1986), logistic regression procedure (Swaminathan & Rogers, 1990), and a number of alternatives based

on item response theory (Tissen, Steinberg, & Wainer, 1988). These sophisticated statistical techniques are known for having desirable characteristics as DIF procedures. The most important characteristic is the accuracy of the DIF detection, which can be expressed as statistical power and Type I error rate in statistical terminology. Power is the probability of identifying the DIF item as DIF, while the Type I error rate is the probability of identifying an item as DIF when no DIF is present.

In addition, based upon these DIF procedures, which have focused on dichotomously scored items, generalized DIF procedures for polytomously scored items have been developed. Increasing interests in DIF analysis techniques for polytomous item scores are due to the increased use of performance and constructed-response items in achievement tests. Moreover, there have been interests in investigating DIF in personality, attitude, and other affective tests, which generally have polytomous items. There have been concerns that these affective items are subject to DIF due to cultural or language differences as well as the more traditional concerns of gender or ethnic difference. In terms of polytomous item scores, preliminary evidence suggests that the following methods look promising in terms of high accuracy and ease of application: some variations of Mantel-Haenszel method, the IRT based graded response model, the logistic discriminant function, and the standardization method. These methods will be described in Chapter 2 in detail.

As demonstrated above, a variety of approaches in DIF analysis have been constantly proposed, based upon various theoretical backgrounds and/or diverse practical purposes. It is necessary to sort them out in a consistent and elaborative way. Millsap

and Everson (1993) and Potenza and Dorans (1995) provided a comprehensive review of DIF methods.

Potenza and Dorans (1995) provide a good framework for classification of various DIF methods. They classified various DIF methods into four groups with two dimensions. One dimension is that of observed score approach vs. latent score approach; the other dimension is that of parametric approach vs. nonparametric approach.

The first dimension can be thought of as the type of matching variable. DIF indicates a difference in item performance between two comparable groups of examinees; that is, groups that are matched with respect to the construct being measured by the test. The comparison of matched or comparable groups is critical because explicit difference between two groups is compounded with true group difference and item specific difference. The matching variable is also called the conditioning variable; it is the variable representing the construct to be measured. The matching variable must be controlled for in the DIF method in order to identify DIF, as opposed to item impact. Both observed score approaches and latent variable approaches assume that the items measure the same dimension as the matching variable (i.e., they presume unidimensionality). The difference is that the former uses the observed score as the matching variable, but the latter uses an estimate of latent trait level, which is a function of observed data. This distinction has implications for how DIF is defined and measured.

The second dimension put forth by Potenza and Dorans (1995) is related to the existence of specific models or function for DIF analysis. Parametric approaches require the specification of a particular function to describe the relationship between item score and underlying construct or total score on the test. For nonparametric approaches, no

such function is specified. Parametric approaches have the risk of misspecification of the model between item performance and the matching variable. They also tend to have very large sampling covariation among parameter estimates. On the other hand, nonparametric procedures require sufficient data to directly estimate the regressions of item score on test score.

A DIF approach using multilevel item analysis

The statistical procedure that is developed in this research can be classified in Potenza and Dorans' classification framework as a parametric, latent variable approach. This procedure is extended from the Kamata's multilevel item analysis model (Kamata, 2001). Since Kamata's multilevel item analysis model is based on the multilevel or hierarchical model framework, the DIF method that adopted the model, also can be treated as being in the category of multilevel analysis model.

The multilevel or hierarchical linear model (HLM) framework has been widely used in social science research fields such as institutional research, meta analysis, and longitudinal research in developmental studies (Bryk & Raudenbush, 1992). The HLM procedures are designed to handle the analysis of data sets that contain clustering which may cause a lack of statistical independence among study subjects as indicated by a non-trivial intra-class correlation. When data demonstrate clustering, the use of traditional statistical procedures leads to underestimates of standard errors and ultimately to overly liberal null hypothesis tests.

In contrast to traditional regression, HLM or multilevel linear regression models allow each cluster to have its own intercept and regression coefficients for some or all variables. These regression coefficients are often considered normally distributed with mean equal to the effects of cluster characteristics as specified in the higher level, between-cluster model. The errors from the prediction are referred to as random effects, which are assumed to be normally distributed with a mean of zero and a variance/covariance matrix. The resulting variances from each level give information about its respective amount of unexplained variation. Such models for continuous outcomes have been well developed by researchers using different estimation methods. For example, Bryk and Raudenbush (1992) used the EM algorithm and Goldstein (1995) used iterative generalized least squares as estimation methods.

The use of HLM has opened up many possibilities for researchers: One is the improved estimation of effects within individual units (e.g., developing an improved estimate of a regression model for an individual school by borrowing strength from the fact that similar estimates exist for other schools). A second is the formulation and testing of hypotheses about cross-level effects (e.g., how varying school size might affect the relationship between social class and academic achievement within schools). A third is the partitioning of variance and covariance components among levels (e.g., decomposing the correlation among sets of student-level variables into within- and between-school components). Goldstein (1995), Bryk and Raudenbush (1992) and Longford (1995) have given detailed accounts of applications and methodology of these models in social and educational contexts.

Although it has not been unusual to conceive of a continuous and normal distribution for human characteristics, certain types of educational data cannot be normally distributed. Since the outcomes of dichotomous variables, which is usual in educational context, are either right or wrong, the usual linear model that assumes a normal random error fails. Instead, logistic regression, one of the generalized linear models, is used to model such outcomes. The logistic model uses the logit (the natural log of the odds ratio) of the dependent variable as the outcome variable.

For analyzing nested non-normal data such as binary data and count data, the multilevel generalized linear model with random effects is a natural outgrowth of both generalized linear models and hierarchical linear models. It incorporates generalized linear models into the framework of the hierarchical linear model. The logic of a random coefficients model changes little when moving from the linear framework of continuous outcomes to the logistic regression framework of a dichotomous outcome (Reise, 2000).

From the perspective of HLM, it is possible to regard the IRT as an example of hierarchical generalized linear modeling (HGLM). That is, one can say that the multilevel formulation of IRT is inherited from the HLM model itself. This perspective is characterized by the treatment of person ability parameters as random parameters in an IRT model, a treatment originally intended to facilitate the marginal maximum likelihood estimation (MMLE) item parameter estimation (Bock & Aitkin, 1981).

More recently, Adams and Wilson (1996) proposed a more general model, the random coefficient multinomial logit model (RCMLM), which subsequently has been further generalized to its multidimensional form (Adams, Wilson, & Wang, 1997). The RCMLM is formulated so that person parameters are represented as random variables.

Consequently, these models can include person-characteristic variables as predictors. The RCMLM is general enough to subsume a wide range of Rasch models, including both dichotomous and polytomous models.

In the similar stream of perspective for item analysis model, Kamata (2001) provided that an alternative formulation of a multilevel item response model as a two-level model, which can be used for the framework of a DIF technique. She showed that the two-level item analysis model was algebraically equivalent to the Rasch model.

Kamata's two-level item analysis model has some potential advantages over traditional item analysis procedures. First, it can provide empirical Bayes estimates of a person's ability parameter. Since the empirical Bayes (EB) estimates use other person's data as well as the data of the person of interest, it is known to be a more stable estimate. Second, the HLM procedure provides an estimate of the reliability of the estimation of the empirical Bayes ability parameters. In HLM, reliability refers to the percentage of the total variance around each parameter that is parameter variance. The total variance of each parameter consists of both parameter variance and sampling variance. Parameter variance can be explained by the between-unit models. There is also sampling variance around the parameters from sampling error within the level-2 units, however, and this cannot be explained by the between-unit model because it is essentially error. Reliability thus indicates how much of the total variance can be explained by the between-unit models. So, this is conceptually the same as the classical definition of the reliability. Third, and most importantly, the approach allows explanatory variables to be entered into the model so that regression analyses can be administrated in the item analysis. In such regression model, the dependent variable of regression is not just the observed score of

measurement but the construct which is estimated from the item analysis model. In item analyses using HLM scheme, additional explanatory variables can be included into the regression model easily.

Based upon these advantages of Kamata's model, possibility of extending it into DIF analyses has been investigated in this study. After considering various possible approaches in the application of Kamata's models in terms of DIF procedures, one particular model approach has been finalized. This alternative extension model for DIF from Kamata's model will be called the hierarchical generalized linear DIF model (HGLM DIF model).

This study is intended to examine the compatibility of the HGLM DIF model with traditional DIF methods, in particular, the logistic regression procedure for DIF. The reason to choose logistic regression DIF method as the counter part for comparison is that it has quite similar components and structure to the HGLM DIF model. In addition, it is a well-established DIF method which has shown acceptable results in some simulation studies (Jodoin, & Gierl, 2001; French & Timothy, 1996; Mazor, Kanjee, & Clauser, 1995).

In Chapter 2, the main components and issues of this DIF analysis will be explained along with the descriptions of some traditional DIF methods. To review the components and issues of DIF analysis methods is an essential step prior to formulating and evaluating an alternative model for DIF analysis. This review will help to enhance detailed discussions about the HGLM DIF model later in the chapter.

Chapter 2

REVIEW OF DIF METHODS

This chapter consists of two parts, one for the review of the current DIF analysis procedures, and the other for the introduction of the hierarchical generalized linear model for DIF analysis (HGLM DIF) extended from the two-level hierarchical item analysis model by Kamata. In the first part, a range of conventional DIF methods will be organized according to the classification framework introduced in the previous chapter. Along with the classification of current DIF methods, the critical components of DIF analysis will be addressed. These components will be specified in the attempt to identify the characteristics of the HGLM DIF method.

In the second part of the chapter, the statistical model of HGLM DIF will be explained. First, it will be demonstrated that Kamata's two-level item analysis serves as the basis for the HGLM DIF, and then the need for hierarchical modeling and the need to extend it further to HGLM will be addressed. Finally, the HGLM DIF will be shown to be equivalent to more conventional DIF methods, like logistic regression DIF method.

Review of traditional DIF methods

Before describing the components of DIF along Potenza and Dorans (1995) classification framework, it is necessary to begin with a definition of DIF. Though every DIF method is slightly different according to diverse theoretical backgrounds, a general

definition can be given that applies to the various methods, including logistic regression and HLGM DIF, which are of primary concern in this study.

Denote the observed scores provided by an item as a random variable *Y*. Ordinarily, *Y* is discretely measured, but the number of possible values may be large. Examinees are divided into two or more populations on the basis of group variables denoted as *G*, which can be multivariate. The variables in *G* are usually demographic information, such as ethnicity or gender. It is assumed that these are known and measured without error. Finally, define *X* as a latent variable for which *Y* is the intended observed indicator.

Conditional independence holds for *Y* in relation to *X* and *G* if $P(Y|X=x, G=g) = P(Y|X=x)$ for all values of *X* and *G*. Here, $P(Y|X=x)$ is the conditional probability function for *Y* given that *X* assumes the value *x*. If this equation holds for *Y*, the relationship between *Y* and the *X* is independent of group membership *G*. Among individuals with common values on *X*, the distribution of *Y* is the same across populations defined by *G*. *Y* is unbiased as a measure of *X* if the conditional independence in the above equation holds. Conversely, differential item functioning of *Y* in relation to *G* occurs if the conditional independence is violated. The idea of conditioning on *X* in defining bias is important for distinguishing item bias from ordinary group differences, or impact.

As shown by the definition, three components, group variable (*G*), item score (*Y*), and the matching variable (*X*) are the basic components in DIF analysis. The group variable (*G*) consists of coding the examinees based on membership in some group of interest. Often group membership is based on gender, ethnicity, primary language

spoken, or cultural background. It should be noted that it is currently more popular to look at DIF applications examining cross-cultural factor structure, beyond gender, ethnicity, and language difference (Hambleton & Kenjee, 1995; Whittney & Schmitt, 1997; Terreci, 2001).

The second component, the item score (*Y*), is the target of the DIF analysis. Conventionally, the type of item score is dichotomous, that is, coded as either a correct response or an incorrect response. However, polytomous coding such as partial credit or Likert scale ratings are widely used so many polytomous DIF methods have been developed. For the purpose of DIF analysis with polytomous item scores, some additional statistical or asymptotic considerations must be made. However, most DIF methods that have been used with dichotomous item scores can be extended to polytomous item score cases, although these extensions can lead to various consequences based on the method being used.

The matching variable (*X*) is the third basic component of DIF analyses and often is called the conditional variable. All the methods available to identify DIF are designed to match the groups, either directly or indirectly, on the proficiency measured by the items. This component is the most extensively debated component in the field of DIF research. Without this component, making a distinction between impact and bias would be impossible. If this were the case, DIF would have little significance in the measurement field. The matching variable is what the test is supposed to measure. If the item score followed the matching variable perfectly, there would be no DIF. However, an item has DIF if, given the matching variable (*X*), there is variation due to group membership based on the grouping variable (*G*). All the methods that have been

developed to identify DIF assume that the items are homogeneous and unidimensional. This assumption is explicitly made for the IRT models and at least implicitly made for the Mantel-Haenszel (M-H) method, the standardization method, and logistic regression methods.

The type of matching variables used results in one of the dimensions for the classification of the DIF methods by Potenza and Dorans. When explicit matching is carried out on an observed score, as it is for the chi-square, Mental-Haenszel, standardization, logistic regression and logistic discrimination methods, the matching variable should be stratified as finely as possible, consistent with the amount of data on hand. In addition to the type of matching variable, the number of items that included in matching variable, whether the studied item should be included in the matching variable or not, the purification issues, and the justification of using external matching variable are the important issues in the matching variable. Part of the above issues will be mentioned later in this chapter in detail.

Based on the three basic components (*Y*, *G*, and *X*), we can identify some criteria for the classification of diverse DIF methods. From the item score variable (*Y*), the distinction between dichotomous vs. polytomous type of item score can be drawn. From the matching variable (*X*), the basic differentiation is about the type of the matching variable, that is, either observed score or latent trait score. In addition to these two criterions, whether the specific relation is assumed among the three variables is another aspect to be considered in the classification of DIF. This consideration results into parametric vs. non-parametric model criteria for classification.

Following the classification made by Potenza and Dorans (1995), we can make the classification of the current DIF methods shown in Table **2-1**.

Table **2-1**: Classification framework of DIF methods

Type of procedure / *matching variable*	*Parametric*	*Nonparametric*
	Dichotomous DIF	
Observed Score	Logistic Regression Logistic Discriminant Function Analysis	MH or STND
Latent Variable	General IRT likelihood ratio Limited Information IRT LR Loglinear IRT LR IRT-D2 Lord's chi-square	SIBTEST
	Polytomous DIF	
Observed Score	Polytomous logistic regression Logistic Discriminant Function Analysis	Polytomous MH
Latent variable	General IRT likelihood ratio Partial Credit Model	Polytomous SIBTEST

Based upon this classification, some DIF methods such as M-H, IRT DIF, logistic regression, logistic discriminant function analysis, and SIBTEST will be explained. Among these five DIF methods, four methods except for the logistic discriminant function analysis are representative DIF methods in each category for the dichotomous

item score DIF. Since the polytomous DIF methods are basically based on the dichotomous DIF, reviews of the DIF methods in this chapter will be confined to the dichotomous cases.

The logistic discriminant function analysis belongs to the same category as logistic regression DIF. The reason that it is reviewed here is that it can be used as an excellent example to reveal the characteristics that are required for polytomous DIF analyses. Actually, the rest of the methods are selected in order to show specific issues around the DIF methods such as the necessity of the effect size index, the purification process of the matching variable, the need to deal with the non-uniform pattern DIF, the application to the polytomous item score, and the dimensionality issue in DIF analysis. These issues are necessary to assess or appreciate any DIF method to be developed. Even when a new approach like the HGLM DIF in this research, is introduced, these issues needed to be discussed.

These five issues are critical in DIF research and it has been recognized that corresponding methods of each issue seem to give a breakthrough for corresponding issues. That's why each of these 5 methods is combined with specific issue in this chapter.

Mantel-Haenszel DIF method and the effect size index

The M-H DIF method (Mantel & Haenszel, 1959), which was extended by Holland and Thayer (1988) for use in detecting DIF, falls into the category of the observed score matching variable and the non-parametric method in the above

classification framework. It is an observed score matching variable approach since it uses the observed total score intervals as the matching variable. It is non-parametric since it does not assume a specific function between the item score and the observed total score.

In order to calculate the M-H DIF index for an item, it is necessary to make a contingency table that describes the frequencies of correct (1) and incorrect (0) performance of the focal (*f*) and referent (*r*) groups, both of whom score in the same total score interval on the test. The focal group is the group that is being considered as the subject of DIF analysis, while the reference group is the remainder of the population not in the focus group. In this operational sense, the focal and reference groups are matched on the ability most nearly relevant to the ability measured by the item.

The M-H index is calculated first at each total score interval (*i*) of the test of which the item is a part. It can be written as:

$$\alpha_i = \frac{\frac{p_{r_i}}{q_{r_i}}}{\frac{p_{f_i}}{q_{f_i}}} \qquad (2.1)$$

where the α_i is the ratio of the odds that the reference group students have answered the item correctly to the odds that the focal group students have answered the item correctly. If there is no difference in the performance of the two groups on this item within this total score interval, then α_i will be equal to 1. If, however, the two groups function differently (e.g., the total score interval the focal group performances better on the item than the

reference group) then α_i would be less than 1. If, on the other hand, the reference group performs better than the focal group, α_i would be greater than 1.

The M-H DIF procedure estimates a common odds ratio across all matched categories. The form of its index is given as follows:

$$a_{MH} = \frac{\sum_i \frac{p_{r_i} q_{f_i} N_{r_i} N_{f_i}}{N_i}}{\sum_i \frac{q_{r_i} p_{f_i} N_{r_i} N_{f_i}}{N_i}} \qquad (2.2)$$

where N_{ri}, and N_{fi} indicate the number of examinees in the reference group and the focal group in the total score interval respectively.

The index is weighted by the number of cases in the interval; also, that the interval in which the numbers of cases in the two groups are more nearly equal receives the heavier weight.

For the sake of convenience, α_{MH} is transformed to another scale, yielding an index that is referred to as M-H D-DIF, by means of the conversion, M-H D-DIF = -2.35*$ln(\alpha_{MH})$. This transformation centers the index about the value 0, which corresponds to the absence of differential item functioning, and reverses the index so that positive values of M-H D-DIF indicate that the item favors the focal group; negative values indicate that the item favors reference group and disfavors the focal group.

Another DIF assessment method, which is a highly related procedure with the M-H DIF, is the standardization approach developed by Dorans and Kulick (1986; Dorans & Holland, 1993). While the M-H procedure is a statistically powerful technique for

detecting measurement bias at the item level, the standardization method is a more easily understood procedure for describing and explaining the nature of the measurement bias. The results of the standardization method are usually in close agreement with those of the M-H procedure.

The effect size index is used to compare, independent of sample size, across items and studies. The M-H DIF makes use of the effect size, which is the common odds ratio α_{MH}. The use of the effect size provides the same scale across different data sets or items. The statistical test of M-H is the α divided by its standard error. Holland and Thayer (1988) suggested that the M-H procedure provides both a significance test and a measure of the effect size.

IRT DIF method and the purification of matching variable

The IRT DIF method belongs to the category of the latent variable and parametric approach in the classification framework as HGLM DIF. The IRT DIF method has strong theoretical justification and sophisticated analysis procedure, such that it is a quite reliable method for DIF analysis.

The reasons for the strong endorsement of IRT DIF include the use of the item response curve and the purification procedure of matching variable. IRT provides a class of models describing the relationship between individual item responses and the construct measured by the test. These models are called item response curves or item characteristic curves. In practical applications of IRT, the item response curve takes some specified functional form. The value of the item response curve at each level of proficiency

parameter is the conditional probability of a correct response given that level of ability or proficiency. If we are considering the possibility that an item may function differently (i.e., exhibit DIF) for some focal group relative to some other referent group, then in the context of IRT we are considering whether the item response curves differ for the two groups. If the item response curves are the same, there is no DIF; if the item response curves differ, there is DIF.

Because the item response curve for an item is determined by the item parameters, the question of DIF detection could be approached by computing estimates of the item parameters within each group. Like other DIF methods, the aim is to compare the item parameters between two groups by matching the proficiency parameters between the two groups. Technically, this means that the proficiency parameters need to be obtained from the data of two groups. This seems to be obvious, seeing that if the two proficiency parameters and item parameters are estimated from separated two group data, the comparison of item parameter for DIF analysis become meaningless because the two proficiency parameter estimates are not on the same scale. On the one hand, we need to get two separated item parameters from two groups in order to do DIF analysis, but on the other hand, we need to get the proficiency parameter estimate from both groups to get the equal scale. This is why we need the purification process in the IRT DIF method. Even though several purification processes have been developed, the basic idea is the same. The principal aim of the purification process is to select a set of items that have no DIF. In order to do that, it is necessary to split the items into studied item(s) and anchored items, which will be used as matching variable. For the studied item (for DIF), two sets of item parameters and proficiency parameter are estimated from two groups.

But, for the anchored items, the item parameters and proficiency parameters are estimated from all group data, assuming these items are clean. If the studied item does not show any DIF signal, it becomes one of the anchored items. If it shows DIF, it is not included in the anchored item. This process needs to be done iteratively.

Which linking function to use is an important issue in IRT based DIF methods. When using internal criterion, it is recommended to use iterative methods to purify criterion after items with DIF are found, eliminating DIF items from the calculation of proficiency parameter estimates. Evidence from simulation studies show that removing biased items iteratively provides better detection of DIF items than a single iteration method (Millsap & Everson, 1993). Though several linking methods are proposed (Lord, 1980; Park & Lautenshlager, 1990; Candell & Drasgow, 1988), the anchor test method is usually recommended (Camilli & Shepard, 1994).

After a suitable adjustment to correct for possible differences in the distribution of proficiency parameter within the two groups, a statistical test of whether the item parameters differed significantly between the two groups provided evidence of DIF. Lord (1980) proposed two tests for evaluating the significance of DIF; the simpler of the two compared the difficulty parameters for the two groups

$$d_i = \frac{\hat{b}_{Fi} - \hat{b}_{Ri}}{\sqrt{Var\,\hat{b}_{Fi} + Var\,\hat{b}_{Ri}}} \qquad (2.3)$$

in which b_{gi} is the maximum likelihood estimate of the item difficulty parameter b_i in group g and $Var\ b_{gi}$ is the corresponding estimate of the sampling variance of b_{gi}. Lord further proposed a more general test of the joint difference between $[a_i, b_i]$ for the two groups, $D_i^2=V_i'\Sigma'_iV_i$ where $\mathbf{V}_i$ is $[b_{fi}-b_{ri}, a_{fi}-a_{ri}]$, Σ is the estimate of the sampling variance-covariance matrix of the difference between the item parameter estimates, and D_i^2 is distributed as chi-square on 2 degrees of freedom for large samples. A problem arising in the practical application of Lord's procedure for DIF detection is that the estimate of the standard errors of the item parameters obtained from the so-called joint maximum likelihood algorithm are not entirely accurate.

There have been substantial advances in the statistical technology involved in model-fitting and hypothesis-testing in the context of item response models in the past decade. Thissen, Steinberg, and Wainer (1993) extensively explains the four types of IRT based DIF methods including the general IRT likelihood ratio, the log-linear IRT likelihood ratio, the limited-information IRT likelihood ratio, and the IRT-D square methods. They are different in terms of the choice of IRT model and the method of estimation. However, only one DIF detection procedure is considered, in a sense that procedure is to fit an item response model to the data for the reference and focal groups, and test the significance of the difference between the item parameter estimates for the two groups; if the difference is significant, the item is said to exhibit DIF.

One of the most used IRT DIF methods among them is the model comparison method, that is, the generalized IRT likelihood ratio test method. The weight of evidence so far supports the likelihood ratio test over other methods (Millsap & Everson, 1993). This method focuses on the log likelihood ratio of fit of two models; simpler (compact)

and more complex (augmented) models. Variations in models are the number of parameters in the augmented model. First of all, it is needed to calculate log-likelihoods for compact and augmented model to get overall measure, then do individual tests on *a*, *b*, *c* item parameters or whichever ones varied.

The likelihood ratio test is currently thought of as the best measure of statistical significance but not a good effect size index, that is, it depends on sample size and units are not equal across studies. As for the effect size of the IRT DIF method, there is so-called area method that focuses on calculating the area between two item response curve (Raju, 1988, 1990), or the difficulty parameter difference index that is focus on the difference between two difficulty parameters. The difficulty parameter difference index is widely used due to its simplicity, and it is a good effect size index because it is on the same scale across studies if it has mean of zero and standard deviation of 1. However, it is not desirable when there is non-uniform DIF caused by differences in item discrimination parameter or guessing parameter. It is recommended to do graphs to inspect differences when log likelihood ratios are significantly different.

Logistic regression DIF and the non-uniform DIF

Swaminathan and Rogers (1990) introduced a DIF method based on the logistic regression model, which is able to deal with both uniform and non-uniform DIF. Uniform DIF can show that the difference in the probability of success between groups is consistent across all levels of ability, for example, when the item favors all focal group members regardless of ability. On the contrary, non-uniform DIF can show that the

difference of the probability of success between groups is not constant across ability levels. In other words, there is an interaction effect between group membership and ability. It is based on statistical modeling of the probability of responding correctly to an item by group membership and a matching variable. This matching variable is usually the scale or subscale total score but sometimes a different measure of the same variable. This method belongs to the category of observed matching variable and parametric approach.

The logistic regression procedure uses the item response as the dependent variable, with group membership variable, total scale score for each subject and a group membership by total scale score interaction as independent variables. The method provides a test of DIF conditionally on the relationship between the item response and the total score, testing the effects of group membership for uniform DIF, and the interaction of group membership with total score to assess non-uniform DIF.

The logistic regression equation is

$$Y = \ln\left[\frac{p_i}{(1-p_i)}\right] = b_0 + b_1 X + b_2 G + b_3 (XG) \qquad (\ 2.4\ )$$

where Y is a natural log of the odds ratio, p is the proportion of individuals that endorse the item in the direction of the latent variable. X and G indicate the observed total score and group variable, respectively.

The regression parameters in the above equation can be estimated using maximum likelihood and can be tested for significance. If the item is unbiased, only b_0

and b_1 should be nonzero. A model that includes b_0, b_1 and b_2 corresponds to an item that shows uniform bias. If the interaction parameter b_3 is nonzero, non-uniform bias is present. Swaminathan and Rogers (1990) argue that the M-H procedure can be thought of as being based on a logistic regression model where the ability variable is discrete and no interaction term between the group membership variable and ability is permitted.

Testing for the statistical significance of DIF follows naturally from its definition. That is, DIF modeling has a natural hierarchy of entering variables into the model. That is, (1) one first enters the conditioning variable, (2) then, the group variable is entered, and finally (3) the interaction term is entered into the equation.

With this information and the Chi-square test for logistic regression one can compute the statistical tests for DIF. That is, one obtains the Chi-squared value for step (3) and subtracts from it the Chi-square value for step (1). The resultant Chi-square value can then be compared to its distribution function with 2 degrees of freedom. The resulting two-degree of freedom Chi-square test is a simultaneous test of uniform and non-uniform DIF (Swaminathan & Rogers, 1990).

Logistic discriminant function analysis and the polytomous DIF

Polytomously scored measures are becoming more important due to increased emphasis on performance-based measurement and constructed-response items. Ability or achievement test items that are awarded partial credit are polytomous scored, as are many attitude and personality scales. Testlets are another source of polytomous measures. In the case of IRT for polytomous data, more parameters are needed than in the

dichotomous case. Furthermore, the increase in response options means that the number of possible response patterns across measures is greatly increased, complicating the assessment of fit. In some cases, goodness-of-fit tests will have little power except in very large samples (Millsap & Everson, 1993).

One of promising method for the polytomous DIF is the logistic discriminant function analysis (LDFA). Miller and Spray (1993) developed an extension of logistic regression procedures that holds promise for identifying DIF in polytomously scored items. They argued that it is reasonable, employing a logistic form of the posterior probability used in discriminant analyses to estimate the probability of group variable to be group member given the matching variable and item response, *P(G|X,Y)*, even though the group variable, *G*, is fixed and item response, *Y,* is random. Thus, the logistic discriminant function analysis for DIF detection with polytomously scored responses is written as

$$P(G \mid X,Y) = \frac{\exp(1-G)(-\alpha_0 - \alpha_1 X - \alpha_2 Y - \alpha_3 XY)}{[1+\exp(-\alpha_0 - \alpha_1 X - \alpha_2 Y - \alpha_3 XY)]} \qquad (2.5)$$

where group membership is denoted by *G*=0 for the focal group and *G*=1 for the reference group. The response variable *Y* needs not to be restricted to dichotomously scored categories, but can assume polytomously scored values.

Simultaneous item bias test and dimensionality

An important issue in any item bias investigation is the dimensionality of the matching variable. If the bias investigation is to be meaningful, the matching variable must be limited to include only those dimensions for which the item response is intended to be an indicator. There may exist additional latent variables that influence the item response in unanticipated ways. Given the intended purpose for the item response, the additional latent variable may be considered a nuisance variable (Ackerman, 1992; Shealy & Stout, 1993).

In these conceptions, test item performance is determined by a latent "target" trait that the test is intended to measure and one or more additional latent "nuisance" trait. These nuisance traits may influence performance, but are irrelevant to the purpose of the test. Test-wiseness or contextual clues might be examples for such nuisance traits. However, the existence of these nuisance traits is not sufficient to show bias. Another requirement is that the groups being compared must differ in their distributions on these nuisance traits. This multidimensional perspective on bias is valuable in providing a coherent theoretical account of why bias may appear when unidimensional latent variable models are used.

Shealy and Stout (1993) proposed a bias detection procedure that builds on the multidimensional conception of bias. Stout's Simultaneous Item Bias Test (SIBTEST) detects bias by comparing the responses of examinees in the referent and focal groups that have been allocated by using their scores on a "matching subtest" (Roussos & Stout, 1996). The matching subtest is a subset of items that, ideally, are known to be unbiased.

In this framework, DIF is conceptualized as a difference in the probability of endorsing a keyed item response when individuals in groups have the same levels of the latent attribute of interest. The DIF item is considered to possess different amounts of nuisance abilities that influence responding.

In most practical applications, however, the user does not have accurate a priori knowledge regarding bias. Therefore, if neither visual inspection nor classical statistics are useful for identifying a valid matching subtest, one can conduct an "automatic DIF analysis". In that case, SIBTEST will be run successively for *n* items, where on a given trial, the *i*th item is the object of study and the remaining *n* - 1 items constitute the matching subtest. In this way, the procedure resembles the IRT DIF purification process without imposing a formal measurement model. It can be extended to evaluate a group of items and evaluate for DIF simultaneously. In that sense, the SIBTEST can be classified into the category with the latent matching variable and the non-parametric approach.

To calculate the DIF index in the SIBTEST procedure, one calculates the difference in item difficulty values between reference and focal group for the studied item using only those people having a particular score level on the anchor test. Then weighted item difficulty parameter differences by the proportion of cases in the focal group at each score level are summed. This index (*B*) indicates the average weighted item difficulty difference controlling for the matching variable. When having one item as the studied item, the item difficulty value difference has the same metric across studies, but this is not true when several items are tested simultaneously. The statistical test of the SIBTEST is the ratio of *B* to its standard error, which would be normally distributed.

Summary of five DIF methods

Those DIF procedures above have been compared in terms of theoretical and practical perspectives. The M-H procedure is used widely but it cannot cover every issue in DIF research. Unlike the IRT method, the M-H and standardization methods provide indices of differential item difficulty, but not of differential item discrimination. It is possible for an item with the same difficulty parameter in the two groups, but with different slope parameters, to yield a DIF index of zero when analyzed by all but the IRT method. It would be therefore advisable to plot graphs to see the pattern of DIF.

Roussos and Stout (1996) compared the M-H DIF methods and the SIBTEST. They found that when differences in distribution of measured ability across the examinee group were present, both procedures displayed an inflated Type I error for certain items, though MH displayed greater inflation. However, when no latent distributional differences were present, both procedures performed satisfactorily under all conditions. On the other hand, Narayanan and Swaminathan (1996) compared the logistic regression DIF method and SIBTEST procedures for detecting non-uniform DIF. Their results showed that overall there was high power and agreement between both methods in detecting non-uniform DIF, but both procedures had an inflated Type I error rate.

Research comparing the logistic regression procedure to the M-H procedure (Swaminathan & Rogers, 1990) suggests that the logistic regression model is as powerful for detecting uniform DIF as the MH procedure, and more powerful for detecting non-uniform DIF. However, the logistic regression DIF method would produce poor results, if the total score is a poor proxy for the latent trait when the item responses follow multi-

parameter IRT models. Also, it is difficult to apply to polytomous DIF; polytomous data must be recoded into a number of dichotomous sets, each of which is suitable for a separate regression analysis. French and Timothy (1996) introduced three different coding schemes called continuation ratio logits, cumulative logits, and adjacent categories.

Logistic regression models, although useful for detecting uniform and non-uniform DIF, become computationally cumbersome when applied to polytomous items. The logistic discriminant function analysis approach, on the other hand, may provide an elegant solution by treating the item response as an independent variable and requiring only one regression model per item. The LDFA DIF method goes in to the same category with the logistic regression DIF method, that is, observed matching variable and parametric approach.

Among the above five DIF methods, the logistic regression DIF method is most similar to the HGLM DIF method in this study. The logistic regression method allows for the inclusion of curvilinear terms and other factors such as examinee characteristics like test anxiety or instructional opportunity that may be relevant factors for exploring possible causes of DIF. The logistic regression procedure is flexible because it can be extended to multiple examinee groups (Agresti, 1990; Miller & Spray, 1993). The flexibility of the logistic regression DIF method is also expected to apply to the HGLM DIF method.

HGLM DIF method using the hierarchical generalized linear model

The extension of Kamata model for DIF using multilevel model (HGLM DIF) is based on the two-level item analysis model by Kamata (2001). First, it will be shown that the framework of the two-level item analysis model by Kamata results from the extension of the standard hierarchical linear modeling to the hierarchical generalized linear modeling. The model for HGLM DIF based on the Kamata's two-level item analysis model will be represented after that.

The two-level item analysis model by Kamata

Kamata model that provides the base for the extension to DIF procedure is an explicit formulation of item analysis model with multilevel model framework. Some multilevel formulation of item analysis models have been presented by several researchers (Adams, Wilson & Wu, 1997; Hedcker & Gibbons, 1993). As mentioned earlier, the marginal maximum likelihood estimation (MMLE) technique which was developed to surmount the limitation of the simultaneous estimation of item and person parameters for item response model is regarded as an example of this formulation. In this framework, item parameters are fixed and person abilities are random. The person parameter may be decomposed into a linear combination of fixed and random effects. In short, the essence of multilevel formulation of item analysis model is that it is able to consider person ability parameter as random effect, therefore making possible to include person characteristic variable such as group membership variable as one-step analysis in the model. Even though various approaches to the multilevel item analysis model have

been proposed, the Kamata model was selected in this study because it provide the most recent multilevel item analysis model explicitly and it is possible to take advantages of simplicity of the model enough to extend to the DIF situation.

Kamata model combines the logistic regression model with the multilevel model by using sampling model and linking function, which will be elaborated later in this section. In case of the item analysis model in the HGLM framework, the items are supposed to be nested into the examinees. Consequently, the items are the level-1 unit, and the examinees consist of the level-2 unit. This means that every level-2 unit, examinee has their own equation with coefficients for level-1 unit, item. Now, the transformed predicted value, in this case the probability that item *i* being corrected by person *j*, is related to the predictors of the model, in this case the dummy variables indicating item, through the level-1 structural model. The level-1 structural model of the multilevel item analysis model can be represented like this:

$$\begin{aligned} \log\left(\frac{P_{ij}}{1-P_{ij}}\right) &= \eta_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \ldots + \beta_{(k-1)j}X_{(k-1)ij} \\ &= \beta_{0j} + \sum_{q=1}^{k-1}\beta_{qj}X_{qij} \\ &= \beta_{0j} + \beta_{qj} \end{aligned} \qquad (2.6)$$

where, P_{ij} is the probability that person *j* gets item *i* correct, and η_{ij} is the level-1 structural model. X_{qij} is the *q*th dummy variable for person *j*, with values 1 when $q = i$ and 0 when $q \neq i$ for item *i*. β_{0j} is intercept term, and β_{qj} is coefficient associated with X_{qij}, where $q=1,\ldots,k$. This is the specific effect of the *q*th dummy variable, consequently

the effect of the *i*th item. As shown in this equation, every person *j* has their own linear equation with two type of coefficient for items. Among them, the intercept term can be any item in a test, referring to as the "reference" item. The other individual item effect is defined as the difference of effect from the intercept term. The specific meaning of these terms in the level-1 model will be clear in the level-2 model.

This is the level-2 model.

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + u_{0j} \\ \beta_{qj} &= \gamma_{q0} \end{aligned} \qquad (2.7)$$

First of all, the intercept values across all examinees can be decomposed to the grand mean, γ_{00}, and deviations from the mean, u_{0j}. It is a random component of β_{0j} and normally distributed with the mean of 0 and variance. When considered together with the level-1 model, person proficiencies vary across persons and it is fixed across items. Here, u_{0j} is considered to be the proficiency of person *j*. Second, it is shown that item effect, β_{qj}, is assumed to be constant across persons, *j,* because of the lack of random effect. This means that item parameters are fixed across persons and vary across items. The random coefficient values of each term (item) in the level-1 model is summarized into its fixed mean, γ_{q0}, which imply the estimate of the difficulty for the corresponding item *q*.

The two-level item analysis model like Kamata model, which is called the hierarchical generalized linear model (HGLM) for item analysis, is originally formulated by combining the generalized linear modeling (GLM) with the HLM perspective. According to the GLM framework, a sampling distribution of item responses, its

expectation and variance, a link function, and a linear predictor model that are expressed mainly in the level-1 model above have to be specified. Then, following the HLM framework, level-2 models are formulated. The multilevel linear model can be viewed as a special case of the HGLM where the sampling model is normal and the link function is the identity link.

In order to make the multilevel model for item analysis with a level-1 item unit and level-2 person unit, it is necessary to set up the appropriate distribution for the level-1 model. The level-1 model for item analysis with a binary outcome in the HGLM may consist of three parts: a sampling model, a link function, and structural model. While standard HLM uses a normal sampling model and identity link function, the binary outcome model uses a binomial sampling model and a logit link function. Only the level-1 model differs from the linear case.

As for the sampling model, suppose that Y_{ij} represents the number of successful n_{ij} trials. Then the binomial distribution can be written as $Y_{ij}\left|p_{ij}\right. \sim Bin(n_{ij}, p_{ij})$ to denote that Y_{ij} has a binomial distribution with n_{ij} trials and probability of success p_{ij}. According to the binomial distribution, the expected value and variance of Y_{ij} are then $E(Y_{ij}\left|p_{ij}\right.) = n_{ij}p_{ij}$, $Var(Y_{ij}\left|p_{ij}\right.) = n_{ij}p_{ij}(1-p_{ij})$, respectively.

When n_{ij}=1, Y_{ij} may take on values of either zero or unity. This is a special case of binomial distribution known as the Bernoulli distribution. Essentially, the Bernoulli distribution is the mathematical abstraction of coin tossing. It is usually stated in terms of a sequence of generic trials that satisfy the following assumptions: (1) Each trial has two possible outcomes, generically called *success* and *failure*, (2) The trials are independent,

and (3) On each trial, the probability of success is $\boldsymbol{p}$ and the probability of failure is $1 - \boldsymbol{p}$. For the Bernoulli case, the predicted value of the binary Y_{ij} is equal to the probability of a success, p_{ij}. The Bernoulli trials process is characterized by a single parameter p_{ij}.

As for the link function, HGLM uses the logit link function $\eta_{ij} = \ln(\frac{p_{ij}}{1-p_{ij}})$, when the level-1 sampling model is binomial. Specifically, η_{ij} is the log of the odds of success. Thus if the probability of success, p_{ij}, is 0.5, the odds of success is 1.0 and the log-odds or "logit" is zero. When the probability of success is less than 0.5, the odds are less than one and the logit is negative, when the probability is greater than 0.5, the odds are greater than unity and the logit is positive. Thus, while p_{ij} is constrained to be in the interval (0,1), η_{ij} can take on any real value.

Using an analogy from institutional research would provide clearer explanation. For example, suppose that there is a model in which students are nested in school. A dependent variable in the student level, for example student SAT score, is predicted by other student level variables, for example IQ, SES etc., in each school unit in the form of the structural level-1 model, and the consequent random coefficients in the level-1 model are predicted by level-2 unit variable, school variable, in the level-2 model. In the item analysis case, all level-2 units have the same number as the level-1 unit, the number of items. The probability that *j* person gets item *i* correct is predicted by the item dummy variables, which are used to specify the corresponding item. If there are 60 items for each examinee, the level-1 model will have an intercept β_0, and 59 β_{qi} coefficients for corresponding 59 dummy variables for one person, a level-2 unit. Then these β_0 and β_{qi} coefficients are predicted by the person level variable, which is not included for the two-

level item analysis model, because the purpose of the model is to estimate the person ability and item parameter.

Integration of the above two models (Equation 2.6 and 2.7) will show the meaning of the parameters clearly.

$$\log(\frac{P_{ij}}{1-P_{ij}}) = \eta_{ij} = \gamma_{00} + u_{0j} + \gamma_{q0}$$

$$P_{ij} = \frac{1}{1+\exp\{-[u_{0j} - (-\gamma_{q0} - \gamma_{00})]\}} \qquad (2.8)$$

$$P_{ij} = \frac{1}{1+\exp\{-[\theta_j - \delta_i]\}}$$

Combining the level-1 with the level-2 equation, the linear predictor model (2.6) becomes $\eta_{ij} = \gamma_{00} + u_{0j} + \gamma_{q0}$ for a specific person *j* and a specific item *i* for *i*=*q*. Then, the probability that person j answers a specific item I correctly is expressed as in the second equation in (2.8). This equation is algebraically equivalent to the Rasch model, which represented in the third equation. As mentioned earlier, u_{oj} can be considered as person ability parameter, θ_j, and the γ_{q0} and γ_{00} is considered as representing item difficulty parameter, δ_i. In HLM program, the log odds of a correct response is the sum of the item difficulty and the person random effect, while in conventional IRT approaches, the log odds of a correct response is the difference of the item difficulty and person ability. For this reason, the coefficients in the fixed effects table corresponded to (-1) items the item difficulty one would get from conventional IRT analyses (Luppescu, 2002). As shown, the two-level item analysis model formulated consequently is equivalent to the Rasch

model algebraically. Note that only the Rasch model can be equivalent to this model, because this model considers only item difficulty.

HGLM DIF model

The basic idea of using the HGLM in the DIF study is due to the capability of the HGLM to include both matching variable and demographic variables in the level-2 model. By adding group variable in the level-2 equation, specifically in the equation which represents the item difficulty parameter, it is possible to examine the different item difficulties among the groups on the condition that the matching variable is constant. This capability can be extended in the situation that more than two grouping variables are included simultaneously. In that case, the extended model can show an alternative way of exploring the reason of the DIF of an item, which is somewhat difficult in the traditional DIF research, since so many other explanations exist.

Because the Kamata's two-level item analysis model is equivalent to the Rasch model and can deal with only the item difficulty parameter, the extension into the DIF method cannot deal with the discrimination parameter nor the guessing parameter as in the cases of 2-parameter or 3-parameter IRT DIF method. The DIF method based on the Rasch model defines a DIF item as the difference in item difficulty conditional on a person's ability. This is also true of the HGLM DIF method.

As shown in the previous part of this chapter, the level-1 model for the HGLM DIF method is equal to the level 1 model of the two-level item analysis model. The sampling model is the binomial distribution, and the linking function is the logit

transformation. The difference occurs in the level-2 model. While the level-2 model has no explanatory variables in the two-level item analysis model, the level-2 model for the DIF analysis includes the necessary variables. Consequently, in the level-2 model, the β_{qj} coefficients have been explained by matching variable, group variable, and interaction between matching variable and group variable. These variables are the same as the logistic regression model for DIF analysis in terms of their role in the DIF analysis. Here is the level-2 model for HGLM DIF.

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + u_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11} \cdot X_{1j} + \gamma_{2} \cdot G_{1j} + \gamma_{31}(X \cdot G)_{1j} \\ &\vdots \\ \beta_{qj} &= \gamma_{q0} + \gamma_{q1} \cdot X_{qj} + \gamma_{q2} \cdot G_{qj} + \gamma_{q3} \cdot (X \cdot G)_{qj} \end{aligned} \qquad (\ 2.9\ )$$

where the variables *X* and *G* indicate the matching variable and group variable, respectively. The person ability estimates from the first run with the Kamata's model are used as matching variable, *X*, in subsequent runs. The γ_{q1} is the coefficient of the matching variable for the q^{th} item, γ_{q2} is the coefficient of the group variable for the q^{th} item, and γ_{q3} is the coefficient of the interaction between mating variable and group variable of q^{th} item. Note that if there are *N* items, *(N-1)* equations will have the above coefficients for *(N-1)* items.

Like the case of the logistic regression, the HGLM DIF also has group effect regression weight γ_{q2} and interaction term coefficient γ_{q3} as effect sizes of DIF. The unit of γ is free if researchers agree to code group such that the focal group code is 1, and reference group is coded 0. By testing the coefficient γ_{q2} and γ_{q3}, we can see if the item *q*

is DIF or not. If the γ_{q2} is not zero, the corresponding item will be flagged as having uniform DIF. And if the γ_{q3} is not zero, the matching item will be flagged as having non-uniform DIF. If both coefficients are not zero, the item will be treated as having no DIF.

Since it is reasonable to believe that there was no significant variation among individuals in their slopes (β_{qj}) in the level-1 equation, it is possible to drop the error term in the level-2 equations representing each item. Consequently, if there is only one γ term left in the level-2 equations except for the first equation, it is equivalent to treat them as fixed, as opposed to random, coefficients. In the above equation, the error term was dropped from either equation and the additional explanatory variables are included. This would be equivalent to specifying the slope as fixed but varying – its variation across subjects is completely accounted for by explanatory variables such as matching variable, group variable, and interaction variable. On the contrary, the β_0 term is treated as having random effect, since the error term u_{0j} is included as shown in the first equation in the level-2 model. Treating the intercept term β_0 as having random effect was significant when the two-level item analysis model was seeking ability estimate from the u_{0j}. However, since the u_{0j} term is essentially a residual term, it takes on different meaning in the HGLM DIF model.

Combining the level-1 with the level-2 equation, Equation 2-10 shows how the item difficulty parameter is divided.

$$\log\left(\frac{P_{ij}}{1-P_{ij}}\right) = \eta_{ij} = \gamma_{00} + u_{0j} + \gamma_{q0} + \gamma_{q1}X_{qj} + \gamma_{q2}G_{qj} + \gamma_{q3}(XG)_{qj}$$
$$P_{ij} = \frac{1}{1+\exp\{-[u_{0j} - (-\gamma_{q0} - \gamma_{q1}X_{qj} - \gamma_{q2}G_{qj} - \gamma_{q3}(XG)_{qj} - \gamma_{00})]\}} \quad (2.10)$$

Note that even though the signs are still the same, u_{0j} in the HGLM DIF model has different meaning from the two-level item analysis case. It was the estimate of the person's ability parameter in the original Kamata's model, but it is now the residual after all explanatory variables have been taken into account. Although person's ability estimates does not hold its values anymore in the HGLM DIF model, it is likely to have potential advantages in DIF analysis. This will be discussed in a later chapter.

Chapter 3

METHOD

Research questions

In Chapter 2, the five important aspects of DIF research were listed as following: effect size of DIF, purification procedure, non-uniform DIF, polytomously scored items, and dimensionality. Even though it is necessary to explore which of these the HGLM DIF method addresses, it is more imperative to have confidence that the HGLM DIF produces results as reasonable as any traditional DIF method. Therefore, the research questions of this study focus on the empirical investigation of the equivalence of the HGLM DIF method to the logistic regression DIF method as a representative method of traditional DIF method. However, since the HGLM DIF model itself based on the Kamata's two-level item analysis model and used the ability estimation of it as the matching variable in the HGLM DIF method, it is required to endorse that Kamata's model is legitimate model which can be replaced with the Rasch model.

The first research relates to the validation of Kamata's proposal that her model is equivalent to the Rasch model. The validation itself is important not only because the HGLM DIF is based on her model, but also because the HGLM DIF method borrows the ability estimates of that model. If it is possible to answer the questions positively, the HGLM DIF are subject to further investigation as a DIF method. It is anticipated that the results will be quite similar because of the structural equivalence between the two

models, presented in Chapter 2. However, since there are obvious differences between two methods, the results will not be perfectly same. The comprehensive discussion of the similarities and discrepancies between two models will be elucidated in Chapter 5. The second group of research questions, and the main questions of this study, relate to the empirical investigation of HGLM DIF. It focuses on whether the HGLM DIF method results in a similar output as the logistic regression DIF method.

The purpose of this study is to answer the following questions.

1. Is Kamata's two-level item analysis model equivalent to the Rasch model?

    A. Are ability estimations from Kamata's model equivalent to those from the Rasch model?

    B. Are difficulty estimations from Kamata's model equivalent to those from the Rasch model?

2. When the HGLM DIF procedure is used as a DIF detection method, how similar are the results one obtains as compared to the logistic regression DIF method?

    A. Are the uniform DIF items detected by the HGLM DIF procedure the same items identified by the logistic regression procedure?

    B. Are the non-uniform DIF items detected by the HGLM DIF procedure the same as those identified by the logistic regression procedure?

Sample Characteristics

The subjects for this study were a sample of more than 10,000 students who had taken the prerequisite tests called the First-year Testing, Counseling and Advising Program (FTCAP) before their enrollment as freshmen at PSU in the 2000 fall semester. Within the sample, the male and female group had almost the same proportion. The ethnic composition of the sample was 80 % *White*, 5% *Black*, 5% *Asian*, and 10% with missing race identification. Those identifying themselves as Hispanic were an extremely small proportion of the sample and were thus not included in the study.

It was not necessary to use the entire sample of students in the DIF analyses done in this study. This was also not feasible because the *White* group was so much larger thank the *Black* and *Asian* groups. Therefore, a random sample was taken from the *White* group to create a smaller "*White*" group roughly equal to the *Black* and *Asian* groups. This was appropriate because the purpose of this study was not to make an inference from a sample to a more general population, and a random sampling procedure taken in this study should not impact the DIF analysis. The composition of the actual sample used in the study (N = 1654) is shown below in Table **3-1**.

Table **3-1**: Sample Size of Three Ethnic Groups and Gender Groups

		GENDER GROUP		*Total (%)*
		Male	Female	
ETHNIC GROUP	Black	224	320	544 (32.9)
	White	285	227	512 (31.0)
	Asian	319	279	598 (36.2)
Total (%)		828 (50.1)	826 (49.9)	1654 (100)

A total of four pairs of group comparisons were made in the following manner: *Male-Female*, *White-Black*, *White-Asian*, and *Asian-Black* for the purposes of examining item performance. Each pair of groups was analyzed with both the logistic regression DIF procedure and the HGLM DIF model.

Data Collection Procedure

The data used for this study consisted of 60 items in the FTCAP English test. This English test is one part of a series of tests called the FTCAP, which is designed to provide information to the students' adviser and family as well as the students themselves for the better academic progress in the university.

The FTCAP program is the first stage of academic advising, and introduces students to the academic structure and degree programs of the University. The purpose of this program includes determining starting level in English and mathematics, checking basic skills and providing academic advising. This program consists of two parts: the

testing component and the educational planning and academic advising component. The testing component of FTCAP is used for educational planning and academic advising purposes, to determine a student's appropriate starting levels in English, mathematics, and chemistry course sequences, and for basic skills purposes. This study used only the items from the English test. The purpose of the English test is to measure competence in the following areas: *Spelling*, *Vocabulary*, *Punctuation*, and *Grammar*.

As for the unidimesionality issue of the DIF procedure, it is necessary to mention that the FTCAP English test may not suitable perfectly in examining the validity of DIF procedure because it has four sub categories of English literature testing. These sub categories may be used separately in DIF analysis by using four separate sub-total scores for each separated DIF procedure. However, the FTCAP English test was analyzed as a whole test assuming that it measures a global English literature ability of students, so as to focus on the applicability of the proposed DIF procedure. The violation of the unidimensionality in using in this study will be addressed in the context of describing the limitation of this study in the discussion chapter.

The test data coded 0 for incorrect and 1 for correct was provided from the University Testing Service (UTS). Student identification numbers were deleted from the data set for the purpose of maintaining confidentiality. The FTCAP data were combined with some student demographic information such as gender and ethnicity, also provided by the university. UTS merged the two sets of data for use in this research. The un-reached items were treated as incorrect items.

Item means and variances, point-biserial correlation, and classical reliability analysis were obtained for the instrument using ITEMAN and SPSS WINDOWS

version10.4 (SPSS, 1997). The results of the item analyses are presented in the Appendix A.

Mean scores of different groups are shown in Table **3-2** as a preliminary investigation. Note that there is a substantial difference between the *Black* group and the other two ethnic groups in means. These mean differences were significant at the level of 0.05, as shown in Table **3-3**. Two groups can differ on mean levels in a construct even the items measure the construct in the same way for two groups, that is, there is no DIF. Hence the mean difference tells us nothing about the presence or absence of DIF in the items. However, one needs to be aware, that when the difference between two groups and the item discrimination are high, the item can be flagged as DIF even it has no DIF if one is using DIF methods based on proportion correct or point-biserial statistics or their transformations.

Table **3-2**: Mean Difference in Gender and Ethnicity Groups

		Mean	*N*	*Std. Deviation*
Gender	*Male*	23.17	828	9.596
	Female	23.63	826	10.018
	Total	23.40	1654	9.809
Ethnicity	*Black*	19.15	544	8.464
	White	25.30	512	10.042
	Asian	25.65	598	9.502
	Total	23.40	1654	9.809

Table **3-3**: T-Test for Group Mean Difference

	t	*df*	*p-value*	*mean difference*	*standard error*
Male vs. Female	0.955	1652	0.34	0.46	0.482
White vs. Black	-10.79	1054	< 0.00	-6.154	0.5703
White vs. Asian	-0.584	1108	0.56	-0.343	0.5874
Asian vs. Black	-12.153	1140	< 0.00	-6.497	0.534

Data Analysis Procedure

Investigation of Equivalence of the Kamata Model with the Rasch Model

Because the HGLM DIF is an extension of Kamata's two-level item analysis model, it is necessary to examine the equivalence of Kamata's model with the Rasch model in advance. Before running the HGLM DIF model, Kamata's model with 60 English items was embedded in the HLM version 5 program (Raudenbush, Bryk, Cheong, & Congdon, 2001). The residual output was saved as a file to get the students' ability estimations. These residuals, represented as u_{0j} in Chapter 2, will be used as the matching variable in the HGLM DIF method. In separate analyses, the Rasch model, 2-parameter logistic IRT model and 3-parameter logistic IRT model were applied to the same data so that 3 different sets of ability estimates were obtained by using the BILOG-MG program (Mislevy & Bock, 1984). Even though the 2-parameter logistic IRT model and 3-parameter logistic IRT model are not the main focus of this study, the ability and

item difficulty estimations are obtained in order to see how much different they are in terms of the distribution of ability estimation and item difficulty range. For the 3-parameter logistic IRT model, 0.2 guessing initial value was set up because the items have 5 options. The correlations among four different ability estimations from four models were obtained, and the mean and standard deviations of ability estimates from each model were compared.

In addition to the ability estimation, item difficulty parameters were estimated and compared with each other. Only item difficulty parameters were a major concern in this study. The correlation between item difficulty parameters will show whether Kamata's model is equivalent to the traditional one-parameter IRT model.

HGLM DIF Analysis and Logistic Regression DIF Analysis

The HGLM DIF model described in Chapter 2 was analyzed with the HLM program (Bryk, Raudenbush, & Congdon, 1996). For the HGLM DIF analysis, level-1 and level-2 data sets were made. The level-1 data set included the identification column with student ID, the whole item vector variable nested in the students and 59 dummy coding variables. These 59 dummy codes are for the 60 items. For the purpose of the comparison, one item should be saved and coded as 0. The criterion item can be any item, but this study used the last item (Item #60). All item responses of each student were aggregated into one column so that it produces one long one-item vector variable. The level-2 data set consists of the identification column with student ID, ability estimates from the previous Kamata's model analysis, group membership variable, and

interaction between ability estimates and group membership variable. An example of the data set used in the HGLM DIF procedure is shown in Appendix B. Figure **3-1** Figure 3-1 provides one example of a system of equations implemented in the HLM program for one analysis.

```
      Level-1 Model
      Prob(Y=1|B) = P

      log[P/(1-P)] = B0 + B1*(ITEM1) + B2*(ITEM2) + B3*(ITEM3) +
B4*(ITEM4) + B5*(ITEM5) + B6*(ITEM6) + B7*(ITEM7) + B8*(ITEM8) +
B9*(ITEM9) + … + B58*(ITEM58) + B59*(ITEMT5Q59)

      Level-2 Model
      B0 = G00 + U0
      B1 = G10 + G11*(ETHNICAB) + G12*(ZKAMATA) + G13*(ABZKAMAT)
      B2 = G20 + G21*(ETHNICAB) + G22*(ZKAMATA) + G23*(ABZKAMAT)
      B3 = G30 + G31*(ETHNICAB) + G32*(ZKAMATA) + G33*(ABZKAMAT)
      B4 = G40 + G41*(ETHNICAB) + G42*(ZKAMATA) + G43*(ABZKAMAT)
      B5 = G50 + G51*(ETHNICAB) + G52*(ZKAMATA) + G53*(ABZKAMAT)
      B6 = G60 + G61*(ETHNICAB) + G62*(ZKAMATA) + G63*(ABZKAMAT)
      B7 = G70 + G71*(ETHNICAB) + G72*(ZKAMATA) + G73*(ABZKAMAT)
      B8 = G80 + G81*(ETHNICAB) + G82*(ZKAMATA) + G83*(ABZKAMAT)
      B9 = G90 + G91*(ETHNICAB) + G92*(ZKAMATA) + G93*(ABZKAMAT)
      …
      B58 = G580+G581*(ETHNICAB)+G582*(ZKAMATA)+G583*(ABZKAMAT)
      B59 = G590+G591*(ETHNICAB)+G592*(ZKAMATA)+G593*(ABZKAMAT)
```

Figure **3-1**: Model Specification of HGLM DIF Procedure in HLM program

The level-1 model can be expressed as $\ln\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{i=1}^{59} \beta_i Q_i$ where Q_i is the dummy code for the *i*th item. The coefficient of each dummy code variable is explained by the three-predictor variables in the level 2: the matching variable (*ZKAMATA*), the group variable (*ETHNICAB*), and the interaction variable (*ABZKAMATA*).

The Kamata model was used to produce the estimates of the ability parameters that were included in the level-2 data set. These ability parameters were used as the matching variable in the HGLM DIF model, similar to the use of the total scale score in logistic regression DIF. These estimates were standardized into a scale with a mean of 0 and standard deviation of 1, in order to be commensurate with the standardized total observed score of the logistic regression DIF analysis in this study. This standardization is not necessary when either method is used for detecting DIF items in a certain test. But the standardization process was used in this study to enable the comparison of coefficients directly between the two approaches.

After specifying the equations in the HLM program, the optional specifications for a nonlinear model were chosen. A Bernoulli model (0 or 1) and the penalized quasi-likelihood (PQL) parameter estimation method were chosen, which is the default method for parameter estimation in the non-linear model.

Since extending HLM to HGLM requires a doubly-iterative algorithm, it significantly increased computational time. The first iteration, called a "micro iteration," can be represented with computing a linearized dependent variable as in the generalized linear model of McCullagh & Nelder (1989). Basically, this iteration involves the use of a standard HLM model with the introduction of a special weighting at level-1. After this standard HLM analysis has converged, the linearized dependent variable and weights must be recomputed in the so called "macro iteration" step. Then, the standard HLM analysis is re-computed. The double iterative process continues until estimates converge. The standard HLM iterations are called *micro iterations* and the re-computations of the linearized dependent variable and the weights are called the *macro iterations*. Fifty

macro iterations and 0.01 stopping criterion for macro iteration and 200 micro iterations and 0.001 stopping criterion were set as an option.

The HGLM procedure produces three sets of results: those for the normal linear model with an identity link function, those for the unit-specific model with a logit link function, and those for the population-average model with a logit link function. The linear model with identity link is estimated simply to obtain starting values for the estimation of the model with a logit link and is of no interest in and of itself. Only the unit-specific model with a logit link function is relevant for drawing conclusions in this research context.

For the logistic regression DIF analysis, SPSS for Windows version 10.1 was used. Since the logistic regression DIF is based on the independent item level, 60 analyses were needed for a group pair. Each item was treated as a dependent variable, and standardized total score variable, group dummy coding variable and the interaction term of total score and group dummy coding were included as predictor variables. For the purpose of strict comparison between the two models, total observed scores were transformed into standard total score having distribution of 0 mean and 1 standard deviation.

Four pairs of comparison groups were used in four separate analyses, for which the referent group was coded as "0" and the focal group was coded as "1." The referent group was the *Male* group in the gender group comparison, the *White* group in the *White-Black* and *White-Asian* analyses, and the *Asian* group in the *Asian-Black* group analyses.

Given that the size of the regression coefficients is affected by the scale of the independent variables, it should be noted that efforts were made to place them on

commensurate scales. That is, when the ability estimation from Kamata model was used as a matching variable in the HGLM DIF process, the values were standardized as having mean of zero and variance of one. Comparatively, the observed total score used in the logistic regression DIF process was also standardized for the purpose of comparison study.

Comparisons of the Coefficients and *p*-values in Results from Two Methods

Since the main purpose of this study is to compare the findings of the two models for DIF, all regression coefficients with *p*-values of 0.05 and below are investigated. A correlation between the magnitude of the HGLM DIF regression weight for the group variable and the corresponding logistic regression weight for the group variable was calculated across items and 4 pairs of references and focal groups. Another correlation between the size of the regression weights for the cross-product term of HGLM DIF with the corresponding weights in the logistic regression model was also calculated. The scatter gram plots were made to show the relation of coefficients. There were 236 pairs of observations (4 groups X 59 items, matched on item and group comparison) in each of the two correlations and in each of the two contingency tables because the HGLM DIF excludes one item that is used as an anchor in the estimation process.

For the *p*-values comparison, the degree of agreement at various α levels was examined. Even though the *p*-values are not in the interval scale and the interpretation of it should be the answer for either/or question about the hypothesis, the degree of agreement at various levels is analyzed for the further investigation purpose. Since the

sample size and the matching variable and group variable have the same scale, the *p*-values are comparable.

Apart from the comparison of the coefficients and their *p*-values between the two results, detecting DIF for the FTCAP English test followed a different procedure. There are two alternatives in logistic regression DIF methods: the simultaneous hypothesis testing procedure by Swaminathan and Rogers (1990), and the evaluation of R-square changes in a three step procedure recommended by Zumbo (1999). However, in the HGLM DIF method, it is difficult to get the variance/covariance matrix of coefficients for each item for the simultaneous hypothesis testing procedure and the R-square change values for the three steps procedure. So the two methods above for detecting DIF in the logistic regression DIF method cannot be used in this study. Instead of using the two methods above, in terms of evaluating the English test for DIF, the *p*-values for either group variable or the interaction variable is considered. Specifically, a much smaller *p*-value was used (0.012) as the criterion identifying the DIF item based on the Bonferroni multiple comparison procedure.

Chapter 4

RESULTS

This chapter is organized in two sections. The first section shows the verification of the equivalency of the HGLM analysis using the Kamata model and the Rasch model. The second section shows the results of the comparison of the DIF analyses for the logistic regression and the HGLM analyses, which is the main focus of this research.

Equivalence of Kamata Model with Rasch Model

Ability Parameter Estimations from Kamata or Rasch model

In order to confirm the equivalency of the Kamata model with the Rasch model, ability estimates for the same data were generated from these two models, as well as from the two-parameter logistic item response model (2PL) and the three parameter logistic item response model (3PL) using the BilogMG program. The descriptive statistics for these estimates are listed in Table **4-1**. It can be seen that the means for the Kamata and Rasch model are identical up to the second decimal place, but that the variance of Kamata estimation was smaller than Rasch model. The other estimates also showed statistics similar to those of Kamata's model, but to a lesser degree. This "shrunken" variation in the estimations from Kamata's model can be attributed to the empirical Bayesian approach in the HLM program. The empirical Bayesian estimation is known to provide more stable estimates than the OLS method. However, it is necessary to note that this

procedure yields estimates with smaller variance than ones from other estimation processes.

Table **4-1**: Descriptive Statistics of Ability Parameter Estimates from Four Different Item Response Models

	N	*Mean*	*S. D.*	*Skewness* (*Std. Error*)	*Kurtosis* (*Std.Error*)
Kamata model	1654	0.0000	0.7771	0.180 (0.060)	-0.008 (0.120)
Rasch model	1654	-0.0067	0.9500	0.173 (0.060)	-0.004 (0.120)
2P IRT model	1654	-0.0058	0.9555	0.256 (0.060)	0.095 (0.120)
3P IRT model	1654	-0.1465	0.9892	0.144 (0.060)	-0.205 (0.120)

As shown in Table **4-2**, the correlation between the Kamata and Rasch estimates is almost perfect. While the similarity between the Kamata estimates and the 2PL or 3PL estimates was slightly lower, the correlations were still close to perfect. The greater divergence in estimates occurs because the 2PL and 3PL models have different calibration and scoring processes, whereas the Rasch and Kamata model proceed from the assumptions of equal discrimination among items and no guessing parameter.

Previous descriptive item statistics showed that the item-total correlation ranged widely from 0.11 to 0.49, which indicates that models allowing for variation in the discrimination parameter would provide a more realistic fit. In addition, the items in FTCAP English test have 5 options, for which a model with a pseudo-guessing parameter would be more appropriate. Nevertheless, the high correlations between the Kamata ability estimates and the three IRT ability estimates demonstrates that it provides a matching variable that is very close to that provided by more traditional IRT methods.

Table **4-2**: Correlation of Ability Parameter Estimates among Four Different Item Response Models

Pearson Correlation of ability estimations	*Kamata*	*Rasch*	*2PL*	*3PL*
Kamata model	1.000			
Rasch model	.998(**)	1.000		
2PL model	.991(**)	.993(**)	1.000	
3PL model	.985(**)	.986(**)	.993(**)	1.000

Place Table Here

(** *Correlation is significant at the 0.01 level (2-tailed)*

In addition, it is useful to point out that the Kamata's ability estimates have a different metric from the other item response models. This is caused by arbitrary choices in the definitions of scale units. That is, the Kamata model imposes no scale restrictions on the residuals used as ability estimates other than a mean of zero, so that the scale does

not have fixed values for the standard deviation of the ability estimates. In contrast, by tradition, the Rasch model typically is expressed in the logistic metric which has a standard deviation fixed at 0.588. Hence in comparing this model to the Rasch or other models with a different metric, care must be taken to equate the θ scales before comparing any consequent estimation among them.

Item Difficulty Parameters Estimations from Kamata or Rasch Model

The next step was to compare the estimates of item difficulty parameters. Only item difficulty estimation was compared because there are no discrimination and guessing parameters in the Rasch and Kamata models.

As shown in Table **4-3**, the item difficulty estimates from the Kamata model had the lowest mean (0.5805) and the smallest standard deviation (0.9692) among the four different models. However, it is shown that the rank orders of item difficulties between Kamata and Rasch model were exactly the same (r=1.00), as shown in the Table **4-4**. As with the ability estimation, the results from the 2PL and/or 3PL depart more from the results of the Kamata model.

Table **4-3**: Descriptive Statistics for Item Difficulty Parameter Estimates from Four Different Item Response Models

	N	*Minimum*	*Maximum*	*Mean*	*S. D.*
Kamata model	60	-1.54	2.95	0.5805	0.9692
Rasch model	60	-1.88	3.59	0.7097	1.1813
2 PL IRT model	60	-1.88	5.53	0.9326	1.4686
3 PL IRT model	60	-1.21	3.73	0.8102	1.2382

Table **4-4**: Correlation of Item Difficulty Parameter Estimations among Four Different Item Response Models

Correlation of item difficulty estimations	*Kamata modle*	*Rasch model*	*2 PL IRT model*	*3 PL IRT model*
Kamata model	1.000			
Rasch model	1.000(**)	1.000		
2 PL IRT model	.971(**)	.971(**)	1.000	
3 PL IRT model	.950(**)	.950(**)	.948(**)	1.000

The difference of the mean of item difficulty parameter among the models seems to result from the different scale or metric used in the ability estimation. What is important here is that the Kamata model can provide the item difficulty parameter estimate, which has the same rank order to Rasch model. In conclusion, the assertion that Kamata's two-level hierarchical item analysis model is equivalent to the Rasch model is confirmed by this analysis, although the theta scale (and also the item difficulty scale) is in different units.

Comparison of the HGLM DIF and the Logistic Regression DIF results

DIF items in the FTCAP English test

For the purpose of the main comparison study, the regression coefficients and their *p*-values are analyzed with correlation and contingency table analyses, respectively. Before the presentation of correlation and contingency table analyses, the number of items identified as having DIF are listed in Table 4-5 according to the four subscales in the test and four studying groups.

For the purposes of considering item content, a stricter *p*-value was used to define an item having DIF owing to the high probability of "false positives" when so many comparisons are done. Specifically, when an item showed a *p*-value of less than 0.002 in either the group term or the interaction term, it was flagged as having DIF. This criterion was applied to both methods. The criterion of less than 0.002 *p*-value came from the Bonferroni multiple comparison procedure, so that the overall α level for all 60 items

becomes 0. 21 (i.e., 0.002 x 60 x 2 comparisons) per pair of groups. Even though there are two type of DIF items, uniform DIF or non-uniform DIF, identified in the FTCAP English test, they are not separated in the Table **4-5**.

Table **4-5**: Number of DIF item in four sub-tests in FTCAP English test

Sub scales	*Male vs.Female*		*White vs. Black*		*White vs. Asian*		*Asian vs. Black*	
(No. of items, % DIF)	*HGLM*	*LR*	*HGLM*	*LR*	*HGLM*	*LR*	*HGLM*	*LR*
Spelling (10, 18%)			4	4	3	3		
Vocabulary (25, 19%)	2	2	7	7	7	7	4	3
Punctuation (12, 4%)							2	2
Grammar (13, 3%)			0	1	1	1		
Total (60)	2	2	11	12	11	11	6	5

As shown in the *Table* 4-5, the *Vocabulary* sub-test had the most DIF items across four comparison groups. In addition, *White vs. Black* and *White vs. Asian* comparison groups resulted in more DIF items than other comparison groups. There were only two disagreements in the flagging of items as DIF by the HGLM DIF method and the logistic regression DIF methods, which are the one in the *Grammar* sub-test for the *White vs. Black* groups, and the other in the *Vocabulary* sub-test for *Asian vs. Black* groups.

Comparison of the Group and Interaction Coefficient from Two Methods

It is useful to demonstrate an example output of the HGLM DIF analysis before presenting the comparison results of the coefficients from two methods. Table **4-6** shows a section of the fixed effects output table for one HGLM DIF analysis. This is for the analysis with *Black vs. White* groups. The coefficient for *INTRCPT2* of *G00* is the grand mean, and the coefficients for *INTRCPT2* for *Gq0*, for example *G10*, are the intercepts. The *Gq2*, gamma 01 in the *Equation* 2-9, is the coefficient of the matching variable, in this case the standardized Kamata's ability estimation (*ZKAMATA*). It is expected to show the significant *p-values*, in general. The coefficients of *Gq1* and *Gq3* in this output are the DIF index for the uniform DIF and the non-uniform DIF, respectively. For example, the regression coefficients of the matching variables (ZKAMATA) for ITEM 2 and ITEM 3 are 0.8487 and 1.2556, respectively; they are significant at the level of 0.001. However, the regression coefficients of group membership variable (ETHNICBW) are different. ITEM 3 has the regression coefficient of matching variable different from zero significantly, but ITEM 2 does not. However, neither regression coefficients for interaction term (BWZKAMATA) are significant, so they show no non-uniform DIF in either ITEM 2 or ITEM 3.

Table **4-6**: Section of Output for fixed effect in the HGLM DIF analysis

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. d.f.	P-value
For INTRCPT1, B0					
INTRCPT2, G00	-1.709419	0.082210	-20.793	1141	<0.001
For ITEM1 slope, B1					
INTRCPT2, G10	2.338455	0.125642	18.612	68283	<0.001
ETHNICBW, G11	0.368548	0.157833	2.335	68283	0.020
ZKAMATA, G12	1.083938	0.124597	8.700	68283	<0.001
BWZKAMAT, G13	0.194223	0.183950	1.056	68283	0.292
For ITEM2 slope, B2					
INTRCPT2, G20	-0.060418	0.154611	-0.391	68283	0.696
ETHNICBW, G21	**0.203329**	0.180793	1.125	68283	0.261
ZKAMATA, G22	**0.848750**	0.119790	7.085	68283	<0.001
BWZKAMAT, G23	**-0.086053**	0.186522	-0.461	68283	0.644
For ITEM3 slope, B3					
INTRCPT2, G30	2.226735	0.126292	17.632	68283	<0.001
ETHNICBW, G31	**0.614472**	0.160819	3.821	68283	<0.001
ZKAMATA, G32	**1.255568**	0.131369	9.558	68283	<0.001
BWZKAMAT, G33	**-0.053497**	0.187555	-0.285	68283	0.775
For ITEM4 slope, B4					
INTRCPT2, G40	1.664214	0.121389	13.710	68283	<0.001
ETHNICBW, G41	0.030684	0.135121	0.227	68283	0.821
ZKAMATA, G42	0.855086	0.106481	8.030	68283	<0.001
BWZKAMAT, G43	0.001872	0.156332	0.012	68283	0.991
For ITEM5 slope, B5					
INTRCPT2, G50	1.974859	0.123241	16.024	68283	<0.001
ETHNICBW, G51	0.235799	0.145728	1.618	68283	0.105
ZKAMATA, G52	1.079177	0.119235	9.051	68283	<0.001
BWZKAMAT, G53	0.179715	0.177310	1.014	68283	0.311
…					
For ITEM56 slope, B56					
INTRCPT2, G560	1.251695	0.124106	10.086	68283	<0.001
ETHNICBW, G561	-0.447824	0.147764	-3.031	68283	0.003
ZKAMATA, G562	0.875725	0.106225	8.244	68283	<0.001
BWZKAMAT, G563	0.562668	0.188598	2.983	68283	0.003
For ITEM57 slope, B57					
INTRCPT2, G570	-1.428223	0.246289	-5.799	68283	<0.001
ETHNICBW, G571	0.597016	0.291761	2.046	68283	0.040
ZKAMATA, G572	1.070708	0.171544	6.242	68283	<0.001
BWZKAMAT, G573	-0.464123	0.259773	-1.787	68283	0.074
For ITEM58 slope, B58					
INTRCPT2, G580	0.644308	0.133195	4.837	68283	<0.001
ETHNICBW, G581	-0.397535	0.159925	-2.486	68283	0.013
ZKAMATA, G582	0.835908	0.107933	7.745	68283	<0.001
BWZKAMAT, G583	-0.086349	0.175644	-0.492	68283	0.623
For ITEM59 slope, B59					
INTRCPT2, G590	0.812085	0.131495	6.176	68283	<0.001
ETHNICBW, G591	-0.333780	0.154622	-2.159	68283	0.031
ZKAMATA, G592	0.995960	0.112629	8.843	68283	<0.001
BWZKAMAT, G593	-0.047855	0.178509	-0.268	68283	0.789

Note that the coefficients in the HGLM DIF and the logistic regression DIF output are represented as logit scale. There is a direct relationship between the coefficients produced by logit and the odds ratios. A logit is defined as the log base *e* (log) of the odds, that is, *log(p/q)*. Logistic regression is using the logit as the response variable, $log(p/q) = a + bX$. This means that the coefficients in logistic regression are in terms of the log odds. This logit can be expressed in odds by getting rid of the log. This is done by taking *e* to the power for both sides of the equation, that is, $p/q = e^{a + bX}$. The odds can be computed by raising *e* to the power of the regression coefficient. For example, the regression coefficient 0.6145 of the group membership variable in ITEM 3 implies that one unit change in ethnic group from White to Black group results in a 0.6145 units change in the log of the odds. The odds of a correct answer is 1.8487 ($e^b = e^{0.6145} = 1.8487$). On the other hand, the odds of an incorrect answer is 0.5409 (=1/1.8487). Therefore the odds ratio, the ratio of odds of correct answer to the odds of incorrect answer, is 3.4178 (=1.8487/0.5409). The interpretation of this odds ratio would be that the odds of a Black examinee correct answer are 3.4 times greater than for a White examinee.

Table **4-7** summarizes the descriptive statistics of both group and interaction coefficients from both HGLM DIF and LR DIF methods. All 60 items in the test for four comparison groups sum up 240 items to be studied. Note that the number of items in HGLM DIF has 4 less items than ones in LR DIF, because the HGLM DIF excluded one item in each of the comparisons. That item was used as the default for the dummy coding scheme.

Table **4-7**: Descriptive Statistic of Coefficients of Group and Interaction terms from Two Methods

Coefficients	*N*	*Minimum*	*Maximum*	*Mean*	*S. D.*
Group in HGLM	236	-1.06	1.26	0.0098	0.3317
Group in LR	240	-1.07	1.24	0.0020	0.3312
Interaction in HGLM	236	-0.78	0.61	-0.0216	0.2245
Interaction in LR	240	-0.82	0.61	-0.0164	0.2214

As shown in Figure **4-1**, the correlation of group coefficients from the two methods is almost perfect (r=0.99). The correlation of the coefficients of interaction terms show a slightly lower correlation (r=0.97), but the 2 coefficients are still high. (See the Figure **4-2**).

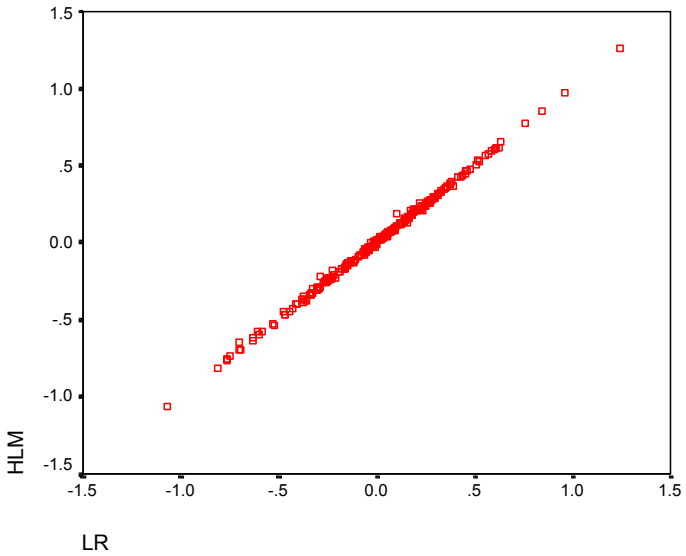


Figure **4-1**: Scatterplot of group coefficient from two methods

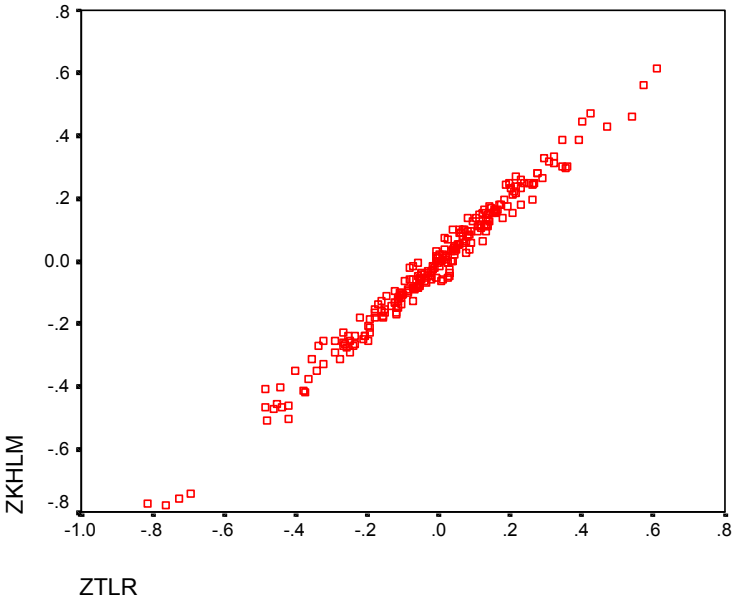


Figure **4-2**: Scatterplot of interaction coefficients from two methods

Comparison of *p*-values between two DIF methods

The contingency table analysis of *p*-values of the group coefficients shown in the Table **4-8** also indicates high degree of relationship. The *p*-values of the coefficients of group variable and interaction term of both models are classified into 5 categories. Agreement in the same significance category is 95.8%. And, when collapsing categories from 1 to 4 into one category, the agreement as to significant/non-significant is 100%.

Table **4-8**: Contingency Table of *p*-values of the Group Coefficient between Two DIF Methods

	From HGLM-DIF					*Total Row*
From LR-DIF	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*	*Count (%)*
(1) p < .0001	20 (8.5)	0	0	0	0	20 (8.5)
(2).0001 < p < .001	2 (0.8)	5 (2.1)	2 (0.8)	0	0	9 (3.8)
(3).001 < p < .01	0	0	9 (3.8)	5 (2.1)	0	14 (5.9)
(4).01 < p < .05	0	0	1 (0.4)	38 (16.1)	0	39 (16.5)
(5) p > .05	0	0	0	0	154 (65.3	154 (65.3)
Total Column Count (%)	22 (9.3)	5 (2.1)	12 (5.1)	43 (18.2)	154 (65.3)	236 (100.0)
Total Diagonal Cell Count (%)			226	(95.8)		

As shown in the Table **4-9**, the contingency table of *p*-values of the interaction coefficients displays similar picture to the case of the group coefficients. Note that there five items above the diagonal that were identified by HGLM DIF more sensitively, while

the logistic regression DIF method was more sensitive to detecting DIF for the five items below the diagonal. Agreement in exactly in the same significance category is 95.8%. And, when collapsing categories from 1 to 4 into one category, the agreement as to significant/non-significant is 98.3%.

Table **4-9**: Contingency Table of *p* values of the Interaction Coefficient between Two DIF Methods

	From HGLM-DIF					*Total Row*
From LR-DIF	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*	*Count (%)*
(1) p < .0001	2 (0.8)	0	0	0	0	2 (0.8)
(2).0001 < p < .001	1 (0.4)	0	0	0	0	1 (0.4)
(3).001 < p < .01	0	1 (0.4)	5 (2.1)	2 (0.8)	0	8 (3.4)
(4).01 < p < .05	0	0	2 (0.8)	13 (5.5)	3 (1.3)	18 (7.6)
(5) p > .05	0	0	0	1 (0.4)	206 (89.3)	207 (87.7)
Total Column Count (%)	3 (1.3)	1 (0.4)	7 (3.0)	16 (6.8)	209 (88.6)	236 (100.0)
Total Diagonal Cell Count (%)			226 (95.8)			

In conclusion, the comparison study of the HGLM DIF method and the logistic regression DIF method produced strong evidence of similarity between the two methods. However, a little difference shown in the interaction term may call for further investigation.

Chapter 5

DISCUSSIONS

In this chapter, I will discuss the similarities and discrepancies between the HGLM DIF method and the logistic regression DIF method. This discussion will answer any questions as to why the results of my analyses were so close, yet not identical. Based on the discussion of similarities and discrepancies, I will explain some features of the HGLM DIF method as a DIF method. Finally, the application of the HGLM DIF method as well as future research will be discussed.

The empirical investigation in this study that compare the HGLM DIF method to the logistic regression DIF method has provided evidence for the assertion that the HGLM DIF procedure is equivalent to the logistic regression DIF method. The equivalence between these two methods has been identified in various aspects of the results, including correlation of group and interaction coefficients, and their probabilities. The correlation of group coefficients from the two methods is almost perfect. The contingency table analysis of the probability values of the group coefficients indicates almost identical results. The correlation of the coefficients of interaction terms shows that they are also quite close to each other. This is also true for the contingency table analysis of the probability values of the interaction coefficients.

Graphical Examination of DIF Items

Now that the equivalence of the two methods has been demonstrated, it should be noted that neither method will provide diagnostic information as to why an item is identified as DIF. One potentially useful method to accompany the HGLM DIF method is to supplement the information with graphs so that test developers can be better informed as to potential areas of concern in a DIF item.

In addition to providing diagnostic information, graphical examinations of the DIF items were used to determine the specific features of the identified items as DIF by the two different DIF procedures used in this study; i.e., logistic regression and HGLM DIF. If there is a discrepancy in identifying a particular DIF item between the two methods, the graphical examination may give information to determine which method is correct. For the purpose of this type of investigation, the graphs based on a parametric model such as the logistic regression or HGLM model are not suitable, because any parametric model based graph should use the parameter estimates of its own in drawing graphs. Moreover, the parametric model based graph will make only graphs resulting from the parametric assumptions, for example, monotonously increasing graphs. Graphs based on a non-parametric approach were thus used. A non-parametric model based graph can show unusual features, for example, the monotonously decreasing line, because it doesn't assume any parametric constraints.

The TestGraf program (Ramsay, 2000) was suitable to graphically examine the DIF items. The TestGraf program estimates the probability of item correct as a function of ability level. The probability of item correct, $P_i(\theta)$, is estimated by *smoothing* the

relationship between the 0-1 item variable values, y_{ia}, and the standard normal quantiles, z_a. In order to get the standard normal quantile, z_a, the numbers of correct answers for the examinees were sorted, with ranks within tied values assigned randomly. Then, the a^{th} examinee by order of size of the number of item correct was assigned the a^{th} quantile of the standard normal distribution, z_a. Smoothing is a type of local averaging, in which for any ability level, θ, the probability of item correct $P_i(\theta)$ at that level is a weighted average of the values of y_{ia} for examinees. The smoothing technique used is a kernel smoothing operation that uses a *Gaussian kernel* with a smoothing parameter that is given a default value by TestGraf. In this study 51 smoothing points were used.

Since TestGraf uses the rank order and quantiles to estimate the ability parameter, it can provide non-parametric graphs, which are simply a picture of the smoothed means of each group for each category of standard normal quantiles. These nonparametric graphs make no assumptions at all about the regression of probability correct on raw score. They were used to confirm that a DIF item flagged from both methods shows discrepancy between two groups.

In this section, all 5 representative graphs are shown. The first two graphs are the typical DIF for uniform DIF (Figure **5-1**) and non-uniform DIF (Figure **5-2**) identified by both methods. Both methods show low level of *p*-values for either group coefficient or interaction coefficient. Both items belong to the *Vocabulary* sub-test. It can be said that Item 27 in Figure 5-1 is obviously favorable for the *Male* group, since the gap between two groups' lines is "huge" (more than a 0.2 probability difference in passing the item in many places throughout the scales) and looks large and in the same direction across all range of ability levels. In contrast, Item 17 in Figure 5-2 shows that as the

ability level goes up, there is a reversal in the pattern of which group is favored, which is called “crossing DIF”.

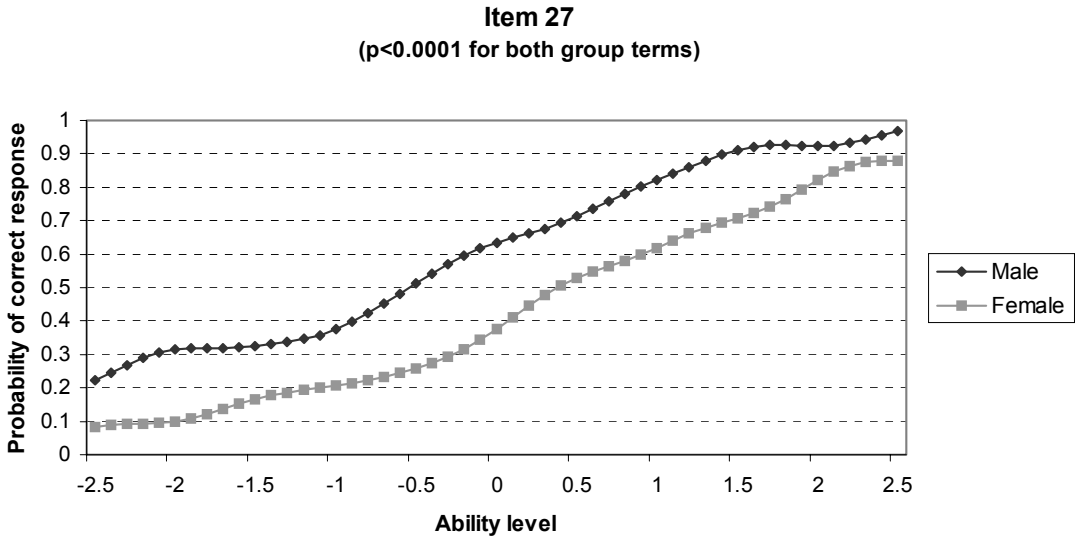


Figure **5-1**: The example of uniform DIF item identified by both methods

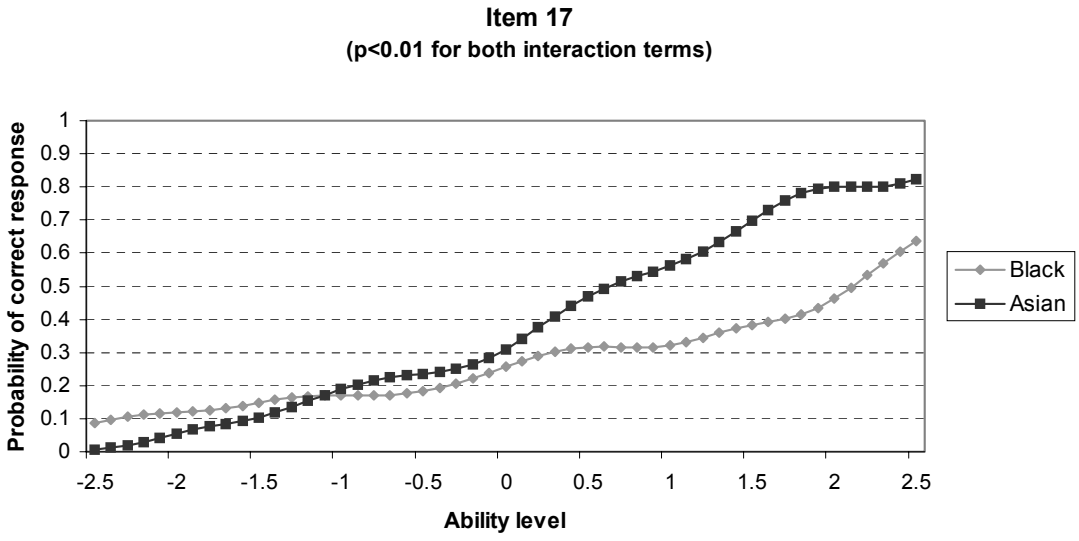
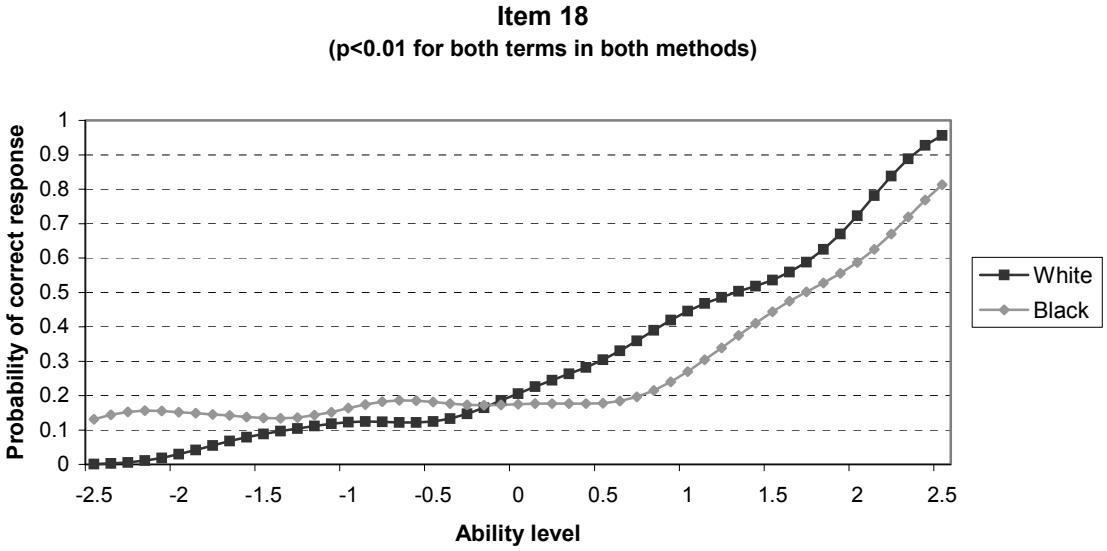


Figure **5-2**: The example of non-uniform DIF item identified by both methods

These items were identified as having DIF from both methods with p<0.0001 level for group term and p< 0.01 level for the interaction term, respectively.

Item 18 shown below in Figure **5-3** also shows agreement of flagging from both methods, and is an example of crossing DIF. However, in this case both the group and interaction terms were significant at the p<0.01 level.

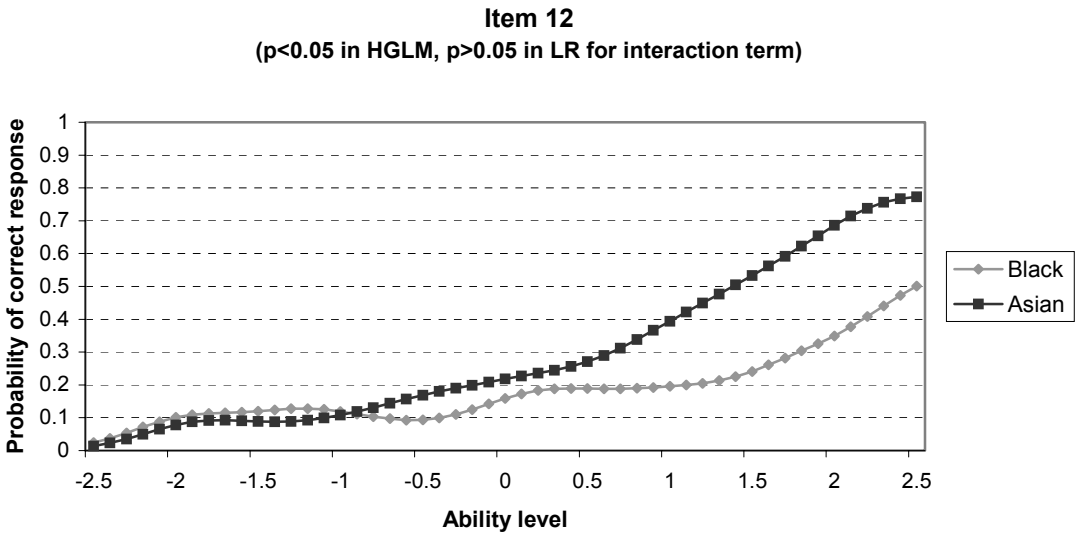


Place Figure **5-3**: The example of DIF item being significant in both group and interaction terms

When the interaction term is significant, the lines do not always cross; there can be another type of non-uniform DIF called the ordinal non-uniform DIF. Examples are Item 12 in *Figure* 5-4, and Item 35 in *Figure* 5-5 below.

As shown in the previous section, most discrepancies in flagging DIF items between the two methods occurred in the identification of non-uniform DIF. With the p

value of 0.05 criterion, four items were identified as DIF items differently in both two methods. For three items among them, HGLM DIF method is more sensitive than LR DIF. That means that HGLM DIF might identify three DIF items which might not be by the LR DIF method. Figure **5-4** for Item 12 is an example of the item showing such discrepancy where HGLM was more sensitive. *Figure* 5-5 demonstrates Item 35 which was the one item where the LR DIF method was more sensitive to non-uniform DIF. It is speculated that the non-uniform DIF characteristic in *Figure* 5-4 and the non-monotonous curve in Figure **5-5** may cause the different *p*-values between two DIF methods.



Place

Figure **5-4**: The example of non-uniform DIF item identified by only the HGLM DIF method

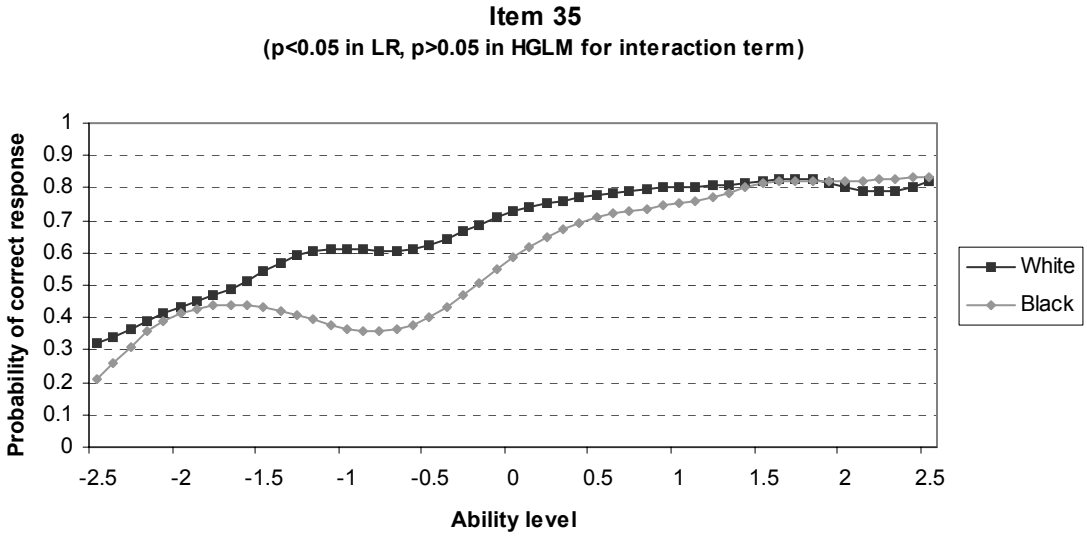


Figure **5-5**: The example of non-uniform DIF item identified by only the logistic regression DIF method

Mathematical Similarities and Discrepancies between Two Methods

These resemblances between the two models were not expected before the analysis so this result is enough to conclude that they are interchangeable in terms of results, even though the results of the DIF analyses with the real data set from the two methods are not perfectly identical. In order to further examine their equivalence, the mathematical relationship between the two models was examined. What should be mentioned first is that both methods are based upon a similar parametric form. They have almost identical terms including matching variable, group variable and interaction term. This structural similarity also is reflected in the mathematical expression of the two

models. When the HGLM DIF model, which is originally expressed in a two-level format, is displayed in the combined equation format, the similarity becomes clearer. The equation is the results of combining the equation for the level-1 model and that for the level-2 model, as shown in Equation (5.1) below:

$$\ln\left(\frac{P_{ij}}{1-P_{ij}}\right) = \gamma_{00} + u_{oj} + \gamma_{q0} + \gamma_{q1}X_j + \gamma_{q2}G_j + \gamma_{q3}(XG)_j \qquad (\ 5.1\ )$$

where q is the conventional symbol to indicate item number. The coefficient of γ_{q1} indicates the change of logit in relation to the change of 1 unit of ability level. The coefficient γ_{q2} and γ_{q3} are the part of the equation, which indicate uniform DIF and non-uniform DIF, respectively.

In comparison, here is the equation of the logistic regression model of DIF analysis.

$$\ln\left(\frac{P_{ij}}{1-P_{ij}}\right) = \beta_{i0} + \beta_{i1}X_j + \beta_{i2}G_j + \beta_{i3}(XG)_j \qquad (\ 5.2\ )$$

As shown, both models have almost the same structure and components. In particular, group coefficients and interaction coefficients of both models take exactly the same form. As for the effect of the matching variable, the coefficients of Kamata's ability estimation variable, γ_{q1}, in Equation (5.1) is equivalent to β_{i1}, the coefficient of the total observed score in the logistic regression in Equation (5.2). In addition, $\gamma_{00} + \gamma_{q0}$ in Equation (5.1) is equivalent to the value of β_{i0} in the logistic regression model in

Equation (5.2). This structural similarity explains the high closeness between the two results of DIF analysis shown in this study.

Based on the empirical evidence that the HGLM DIF result is similar to the logistic regression DIF and the structural mathematical resemblance between these two models, it is possible to assert that the HGLM DIF can be used as a DIF method.

However, there are noteworthy differences between the HGLM DIF and the logistic regression DIF method from a mathematical perspective. These differences include the existence of the additional residual term (u_{0j}) in the HGLM DIF model, the number of equations, overall model testing and the different matching variables (Kamata's ability estimation vs. Total observed scores).

The Existence of the residual Term

The existence of the residual term in the HGLM DIF method is the most obvious difference from the logistic regression DIF method. In the original Kamata model, the residual term (u_{0j}), indicates examinee's ability estimation. The equivalence of this term with Rasch model estimation has been confirmed in this study as shown in Chapter 4. However, in the HGLM DIF model, the meaning and its interpretation have to be changed; the residual term is no longer the estimate of ability parameter. It is the resultant residual after explaining the variance of the dependent variable by the variables of ability, group, and interaction terms; it cannot represent the ability estimation parameter any longer. Since most of the variance would be explained by the three terms in the HGLM DIF model, the u_{0j} has very small variance with zero mean. With this

characteristic as a residual among *j* examinees, it may be used as a tool to examine the residual variation explained by the given input variables.

The Number of Equations

Another difference between the HGLM DIF method and the logistic regression DIF method is the number of equations in the model and in the process of estimation. While the logistic regression needs one equation per item, which results in, for example, 60 equations for all the test items in the FTCAP English test in this study, the HGLM DIF method needs only one equation for all test items. This difference in the number of equations may have a consequence in the estimation process. The logistic regression process does not consider the information from other items, but the HGLM DIF method uses the additional information from other items.

Overall Model Testing

Further, the HGLM DIF method can produce a statistical test for overall effect of group differences or interaction across all items. This enables one to say that there is at least one statistically significant DIF when considering all items together. This is possible through examining the difference of the log-likelihood value between two nested models in which the compact model is deficient from the augmented model in either (1) group variables or (2) group variables and interaction variables. The difference of the two models is asymptotically distributed as a chi-squared random variable with 60

degrees of freedom. An equivalent counter part in the logistic regression cannot be made because of the separate analyses done for each item. However, the logistic regression can provide other useful DIF indices, for example R-square, for each item, which is not practical in the case of the HGLM DIF method. Since the HGLM DIF estimates the coefficients of items in a single equation, it is difficult to get the R-square changes for each item between nested models. This can be considered a drawback of the HGLM DIF.

Features of HGLM DIF method and its Applications

Now that the empirical and mathematical similarities and discrepancies between the HGLM DIF method and the logistic regression DIF method have been discussed, it is useful to draw attention to the features of the HGLM DIF as a DIF method. According to the DIF method classification framework (Potenza & Dorans, 1995) discussed in Chapter 2, the HGLM DIF method can be classified into the parametric and the latent matching variable category. As the methods in the latent matching variable and parametric approach, the HGLM DIF shares its features with other methods in the same category.

First, like other DIF methods in the same category, for example, the IRT DIF method, the HGLM DIF assumes specific linkage to test theory. While the observed score approaches, such as the Mental-Haenszel DIF method and the standardization DIF method, do not make any assumptions about the classical test theory decomposition of scores, the HGLM DIF method is closely linked to a test theory that decomposes an

observed score into a systematic true score and a stochastic error score (Potenza & Dorans, 1995).

Second, the HGLM DIF uses the latent variable as the matching variable as the IRT DIF method does. While the logistic regression method also assumes a parametric relationship between the item score and the total observed score, it is an observed score approach, not a latent variable approach. The fact that the HGLM DIF used the latent variable as the matching variable makes it possible to address the purification process of the matching variable discussed when the IRT DIF method was introduced in Chapter 2, even though this is not examined in this study. However, since the HGLM DIF method utilizes only the item difficulty parameter in the estimation process, the result of the purification of matching variable will be different from the ones using the 2-parameter or 3-parameter IRT based DIF method. Even so, the iterative process for the selection of non-DIF items seems to be possible.

In addition to the common features with the latent matching variable and parametric approach, there are other features the HGLM DIF possesses in relation to the issues discussed in Chapter 2: the effect size as the measure of DIF index, the non-uniform DIF analysis, and the application to the polytomous data. Even though five issues relating to five representative DIF methods were discussed in Chapter 2, the purification of matching variable and the dimensionality issue are not included in this section, since they are discussed in the section of matching variable issue.

First, the HGLM DIF method provides the effect size as the measure of DIF index. The DIF measure index is another useful criteria proposed in Potenza and Dorans (1995) that can be used to assess DIF procedures. To be used effectively, a DIF detection

technique needs an interpretable measure of the amount of DIF. In the case of the logistic regression DIF method, Zumbo (1999) provided two kinds of effect size: R-square change between nested models, and coefficients of group and interaction variables. Similar to the logistic regression DIF method, the HGLM DIF method can provide the logit coefficient that can be transformed into the odds-ratio measures as the measure of DIF. However, it is difficult to provide the R-square change for each item in the HGLM DIF.

Second, the HGLM DIF method can deal with the non-uniform DIF as the logistic regression DIF can do, since both method have very similar structural formulation as shown in previous discussion in this chapter.

Third, because of increasing use of Likert scale and partial credit items in educational testing, the capability of dealing with the polytomous item in DIF procedure become important. It is possible for the HGLM DIF to apply for the polytomous data. Usually, the polytomous item can be assumed having either approximately continuous distribution or multinomial distribution. According to the assumption, related polytomous DIF procedures were developed. Since the logistic regression DIF procedure has successfully extended to the polytomous item situation and the HGLM DIF share the structural components and features with it except for the different type of matching variable, it is anticipated that the HGLM DIF can be extended to the polytomous DIF procedure. However, this study didn't address that issue.

In general, the common issues about DIF procedure can be dealt with the HGLM DIF procedure as discussed above. However, it may not be strongly recommended for ordinary DIF analysis in practical application because of its difficulties in maintaining

data file and embedding the procedure in computer program. The HGLM DIF procedure requires more effort and more time than the traditional DIF procedures like the logistic regression procedure does.

In order to appreciate the implication of the HGLM DIF procedure, it is necessary to consider the additional features of it. Beyond the features discussed in Chapter 2 and previous section of this chapter, three more points can be made as the unique application of the HGLM DIF: the different item number for each examinee, the possibility of using the residual term (u_{0j}) as an index of person-fit or adjusted ability estimates, and multiple matching variables or multiple grouping variables.

Different Item Number for Each Examinee

A merit of HGLM DIF is that it can handle testing situations in which examinees receive different items. This is possible because the item dummy variable in the HLM data set can determine whether a certain item is included in the ability estimation. If all items are included in estimation of ability, the corresponding item dummy variable will be equal to 1 for the included items, and zero for those excluded from the estimation of ability. Based upon the item dummy variable, estimated ability will result from the Empirical Bayesian estimation process in Kamata's model, and will be produced and used in the HGLM DIF model. This is useful characteristic when DIF analysis is applied to the computer adaptive testing, in which the number of administered items is different for each examinee. In this case, the total observed score couldn't be used as the matching

variable and the logistic regression DIF method would not be practical to use; however, the HGLM DIF method can be applied.

The Possibility of using the residual term as an Index of Person-fit

As mentioned earlier in this chapter, the existence of the u_{0j} term needs more attention. In Kamata's model, the u_{0j} is equivalent to the ability estimates of the Rasch model. However, in the HGLM DIF model, it is the residual after being explained by the matching variable, group variable, and interaction term. If the residual term for each examinee is greater than zero, this indicates that there are yet unexplained factors in the examinee's responses, beyond matching variable, group variable, and/or interaction variable of the model. Residuals that result from the explanation of the matching variable might be analogous to the person-fit index, even though they are expressed in different units. The person-fit index is concerned with an examinee whose responses do not fit the typical response pattern (Bracey, Gerald & Rudner, 1992; Drasgow, Levine & Zickar, 1996). Similar to the person-fit index, the u_{0j} with only the ability estimate explained might indicate abnormal response of examinee when it is greater than zero. Furthering this logic, a u_{0j} of greater than zero in the DIF model (with ability estimates and group variable explained) might indicate the usefulness of another DIF analysis with a different or additional group variable.

In the discussion of the u_{0j} term, it is worth mentioning the implication of using the u_{0j} term as the matching variable. Luppescu (2002) developed a very similar DIF procedure to the one presented in this paper. The main difference, however, is that he did

not include the matching variable term in the model. In order to explain the difference, it is necessary to discriminate the u_{0j} in Kamata's model from the u_{0j} in Luppescu's DIF model and from that of the HGLM DIF method proposed in this study. There are three residual terms mentioned in this context. The first u_{0j} from Kamata's model, which is the ability estimation, was used as the matching variable in the HGLM DIF model. First of all, the HGLM DIF model (Equation 2-6, 2-7, and 2-8) has the u_{0j} term, but it has different meaning from the u_{0j} term in Kamata model. Apart from the matching variable, which is the same as the u_{0j} in the Kamata model, the u_{0j} still remains in the HGLM DIF model with different meaning from the u_{0j} of the Kamata model. Since the ability variable and group membership variable and their interaction variables explained the variance, the u_{0j} term in the HGLM DIF method indicates the residuals, not the ability estimation. In the case of Luppescu's DIF model, the u_{0j} term is neither the ability estimation nor the residual term in the HGLM DIF method. It cannot be the ability estimation because it is kind of residual term resulting from conditioning the group membership variable. In addition, it is not the same residual term in the HGLM DIF method, because his DIF model didn't include the matching variable and interaction term. That is, in his DIF model, as the group membership effect was explained, the resultant residual random effect, u_{0j}, may be a group membership adjusted estimation of person ability.

In addition, it should be noted that the group coefficient in Luppescu's DIF model cannot be a DIF index, because eventually there is no matching variable in his model. Even though the u_{0j} looks like it is functioning as a matching variable when the mean and

variance of the matching variable of the two groups are very similar, the group coefficient is not the DIF index.

Multiple Matching Variables or Multiple Grouping Variables

Finally, it is possible to include multiple matching variables in the HGLM DIF models, because of its flexibility. When it is desirable to use multiple test scores the HGLM DIF can provide a mean to do so. According to research on using multiple test scores, using multiple test scores can reduce the Type I error rate more so than when only one matching variable is used (Clauser, Nungester, & Swaminathan, 1996). In addition, it is also possible to include multiple group variables in the HGLM DIF model. Even though it cannot be classified as a traditional DIF study because it does not specify specific focal and referent groups, it is possible to imagine the situation in which we need adjusted ability estimation after conditioning multiple group effects (Kim, Cohen, & Park, 1995).

Future Research

It is worth to examine the applicability of the HGLM DIF into the situations described in the previous section such as CAT situation, person fit index, and multiple matching and/or group variables. However, for future research, it seems to be necessary to suggest the need for simulation study of the HGLM DIF method focusing on power

and the Type I error rates, and also the difference in DIF results from different types of matching variables.

The results presented above for the comparison between the HGLM DIF and the logistic regression DIF with a data set from real situation cannot be used to conclude which model is better for DIF research. This is partly because the empirical research showed that they are essentially interchangeable. The advantage of HGLM DIF addressed in this chapter is mainly based on the theoretical and practical implications. The absolute and relative performance of the methods for situations involving different item, ability, and group membership variables could be further studied using simulation date set with known parameters such as the number of DIF items and the amount or type of DIF.

A DIF method should have high statistical power. That is, the probability of an item being flagged as DIF item when it is a DIF item should be high. But it is also necessary to have reasonably low Type I error rate, because the power of detecting DIF may be decreased by the high Type I error rate with all other conditions same. In many cases, test developers are confronted with items that show DIF for no apparent reason (Angoff, 1993). It could be that the items are not truly functioning differently, but the methods are merely incorrectly flagging items as having DIF, because of the high Type I error rate of the method.

It is impossible to determine the value of the Type I error rate of a specific DIF method with this kind of comparison study. This study shows that both methods identified relatively large number of items as having DIF. But it is impossible to figure out whether the flagged DIF items are real or not. Moreover, if the two methods show

quite different results, it would be difficult to say which method is better. The HGLM DIF in this case seems to be acceptable, but it cannot be guaranteed because there is no such criterion in the comparison study.

Usually, simulation study scheme for DIF procedure uses simulated data set with determined item parameters. Difficulty parameter and discrimination parameters should be selected and assigned appropriately so that can cover reasonable range of values and have good item fit statistics. Same number of sample is assigned into either reference or focal group. For each group, same item parameter values are given except for one or some DIF item(s). DIF item can be programmed to be different in difficulty parameter and/or discrimination parameter. Replication for the simulation experiment usually 100 so that it can easily show the percentage of detection power and Type I error rate. For example, the comparison with the logistic regression DIF procedure, using the replication of 100 simulation data can show which procedure has higher percentage of detecting DIF items and Type I error rate, that is, false DIF detecting percentage. As in the other previous simulation studies, the simulation study of the HGLM DIF may focus on the effect of the varying the number of items, sample size, and amount or type of DIF. Various sample size will affect the power of the HGLM DIF procedure through its effect on estimation. In small samples, the asymptotic results may not hold and hence the test statistic may not be a valid indicator of the presence of DIF. Amount and/or type of DIF are also important concern because of its nature of the DIF (uniform or nonuniform DIF). However, the number of item should be considered as the most important variation in comparison study with the logistic regression DIF with simulation data, because it affects the accuracy of total score as a measure of ability. In general, the longer the test, the more

reliable the total score. Given that the HGLM DIF uses the latent variable as the matching variable, while the logistic regression DIF uses the observed total score as the matching variable, the variation of the number of item will reveal the performance of both procedures explicitly. More concern about the difference in type of matching variable will come next.

What to do next for future research about the HGLM DIF methods concerns issue of the type of matching variable effect on the DIF results. The necessity of this future research comes from the fact that the HGLM DIF method seems feasible to use with any kind of matching variable. Even the latent matching variable, the Kamata's ability estimates used in this study can be justified theoretically because the HGLM DIF method itself is based on Kamata's model. What is more important is the fact that the DIF results can be changed by using different kinds of matching variables either observed total score or latent variable score within specific method (Ackerman & Evans, 1994; Clauser, Nungester, Mazor, & Ripkey, 1996).

If focused on the difference between the latent variable and observed variable, the different results due to the different matching variables seem to relate to the reliability of total observed score and its sequent distributional differences between the two types of matching variables. Generally, the variance of true score variable or latent variable is smaller than the one of the observed score variable, since the observed variance is the combination of the true score variance and the random sampling variance. However, in terms of statistical power, DIF results from two types of matching variables have shown that they are not so different. According to one simulation study about the effect of different matching variable types on the DIF results, using an examinee's estimate of

ability as the matching variable is comparable to using total test score in terms of power (Walker, Beretves, & Ackerman, 2001). Similarly, additional analysis of the HGLM DIF method in this study, which has examined the difference of two types of matching variables on the DIF results and was not reported here, shows almost the same results, that is, both analyses identified almost the same number of items as having DIF. This result looks consistent with the above simulation study. However, the issue of statistical power cannot be discussed without consideration of the Type I error rate, because of the fact that the statistical power can be raised by increasing the Type I error rater when other conditions are same.

The theoretical relationship between DIF using the latent matching variable and an observed matching variable has been discussed (Holland and Thayer, 1988; Zwick, 1990; Meredith & Millsap, 1992). According to Meredith and Millsap (1992), methods that rely exclusively on observed variables are not diagnostic of measurement bias or lack of bias. However, the observed matching variable may replace the latent matching variable, when the observed matching variable is a highly reliable measure of the latent matching variable. So, if the observed matching variable is a total score, it might be expected that the reliability of the observed matching variable be increased with the increase of the number of items. The low reliability of the sum score as matching variable can produce some problems in DIF procedure.

Related to the issue of the matching variable, more simulation studies are required to investigate the issue of the lower limit for reliability of the observed matching variable that will yield near invariance of the conditional distribution of the item scores, and the usage of the regression correction procedure in considering the reliability.

Conclusions

Since a test cannot be perfect, scores on any tests are subject to have, to some degree, biased items against individuals with certain characteristics such as gender, ethnicity, and socio-economic background. Any DIF procedures are developed to detect such items and provide statistical evidence for expert's review of such items, aiming to reduce any threat to validity of tests. This study is also dedicated to the development of a DIF procedure by extending Kamata's multilevel item analysis model, which has different perspective on the item response model. This study has shown the capability of providing satisfying DIF results comparing to the logistic regression DIF procedure. In addition, the HGLM DIF procedure provides possibility to apply to specific situation (CAT, multiple group membership and/or matching variables), and to pursue somewhat different but related research area (person-fit index research), as explained earlier. This significance mainly stems from the multilevel formulation of item response model that enables to include person characteristic variables in one-step analysis.

However, it is necessary to point out some drawbacks of the HGLM DIF procedure. First, it is obvious that the HGLM DIF procedure cannot cope with the dimensionality issue sufficiently. It is partly because the definition of DIF is different to the one of SIBTEST procedure which can fully address the issue. Second, there is doubt that it is efficient in terms of managing data file and running program. Since items scores are analyzed in a single equation, the congregated data file would be huge and the estimation process would be delayed.

In conclusion, this study has shown that the HGLM DIF extended from the Kamata's two-level item analysis model can detect DIF items, by comparing the results with ones from the logistic regression DIF method. Group effects from two methods show an almost perfectly same rank order of the coefficients. This is also true for the effects of interaction term which is suppose to indicate the non-uniform DIF item, even though the intensity of similarity of the interaction term effect does not as strong as the case of the group effects. Comparison of the *p*-values which is used directly for flagging DIF item also revealed very close similarity. Although particular method shows more sensitivity to particular items, no systemic sensitivity of particular method was found, that is, the number of items above the diagonal equals that of items below the diagonal. The HGLM DIF has almost the same equations as the logistic regression DIF, but it is potentially better in terms of the type of matching variable, model flexibility, and its simultaneous estimation process.

Bibliography

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. Journal of Educational Measurement, 29(1), 67-91.

Ackerman, T. A., & Evans, J. A. (1994). The influence of conditioning scores in performing DIF analysis. Applied Psychological Measurement, 18, 329-342.

Adams, R. J., & Wilson, M. (1996). Formulating the Rasch model as a mixed coefficients multinomial logit. In G. Englhard & M. Wilson (Eds.), Objective measurement: Theory and practice (Vol. 3, pp. 143-166). Norwood, NJ: Ablex.

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. Applied Psychological Measurement, 21(1), 1-23.

Adams, R. J. Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. Journal of Educational and Behavioral Statistics, 22(1), 47-76.

Agresti, A. (1990). Categorical data analysis. New York: Wiley.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P.W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 3-23). Hillsdale, NJ: Erlbaum.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. Psychometrika, 46, 443-459.

Bryk, A. S., & Raudenbush, S. W. (1992). Hierarchical linear models. Thousand Oaks, CA: Sage.

Bryk, A. S., Raudenbush, S. W., & Congdon, R. (1996). HLM: Hierarchical linear and nonlinear modeling with the HLM/2L and HLM/3L programs [Computer program]. Chicago, IL: Scientific Software International.

Bracey, Gerald & Rudner, Lawrence M. (1992). Person-fit statistics: High potential and many unanswered questions. Practical Assessment, Research & Evaluation, 3(7).

Camilli, G., & Shepard, L. A. (1987). The inadequacy of ANOVA for detecting test bias. Journal of Educational Statistics, 12, 87-99.

Camilli, G., & Shepard, L. A. (1994). Methods for identifying biased test items. Thousand Oaks, CA: Sage.

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. Applied Psychological Measurement, 12, 253–260.

Clauser, B. E., Nungester, R. J., Mazor, K. M., & Ripkey, D. (1996). A comparison of alternative matching strategies for DIF detection in tests that are multidimensional. Journal of Educational Measurement, 33, 202-214.

Clauser, B. E., Nungester, R. J., & Swaminathan, H. (1996). Improving the matching for DIF analysis by conditioning on both test score and an educational background variable. Journal of Educational Measurement, 33, 453-464.

Drasgow, F., Levine, M. V., & Zickar, M. J. (1996). Optimal detection of mismeasured individuals. Applied Measurement in Education, 9, 47–64.

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

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. Journal of Educational Measurement, 23, 355-368.

Faggen, J. (1987). Golden rule revisited: Introduction. Educational Measurement: Issues and Practice, 6, 5-8.

French, A. W. & Timothy, R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. Journal of Educational Measurement, 33, 315-32.

Goldstein, H. (1995). Multilevel statistical model. (2nd Edition). London: Arnold and New York: John Wiley and Sons.

Hambleton, R. K., Kanjee, A. (1995). Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptations. European Journal of Psychological-Assessment, 11(3), 147-157.

Hedeker, D. & Gibbons, R. D. (1993). MIXOR: A computer program for mixed-effects ordinal probit and logistic regression analysis [Computer program]. University of Illinois at Chicago.

Hong, S., & Roznowski, M. (2001). An investigation of the influence of internal test bias on regression slope. Applied Measurement in Education, 14(4), 351-68.

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

Jodoin, M. G. & Gierl, M. J. (2001). Evaluating Type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. Applied Measurement in Education, 14(4), 329-49.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. Journal of Educational Measurement, 38(1), 79-93.

Kim, S. H., Cohen, A. S., & Park, T. H. (1995). Detection of differential item functioning in multiple groups. Journal of Educational Measurement, 32, 261-276.

Luppescu, S. (2002). DIF detection in HLM item analysis. Paper presented at the Annual meeting of the American Eductional Research Association, New Orleans.

Linn, R. L., & Drasgow, F. (1987). Implications of the Golden Rule settlement for test construction. Educational Measurement, 6, 13-17.

Longford, N. T. (1995). Hierarchical models and social sciences. Journal of Educational and Behavioral Statistics, 20, 205-9.

Lord, F. M. (1980). Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Erlbaum.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.

Mazor, K. M., Hambleton, R. K., & Clauser, B. E. (1998). Multidimenstional DIF analyses: The effect of matching on unidimensional subtest scores. Applied Psychological Measurement, 22(4), 357-67.

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, 32,131-44.

McCullagh, P., & Nelder, J. A. (1989). Generalized linear models. (2nd edition ed.). London: Chapman and Hill.

Meredith, W., & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. Psychometrika, 57(2). 289-311.

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. Journal of Educational Measurement, 30, 107-22.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. Applied Psychological Measurement, 17(4), 297-334.

Mislevy, R. J., & Bock, R. D. (1984). BILOG IX maximum likelihood item analysis and test scoring: Logistic model [Computer program]. Mooresville, IN: Scientific Software.

Monahan, P. (2000). The effect of unequal variance in the ability distributions on the Type I error rate of the Mantel-Haenszel chi-square test for detecting DIF. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Narayanan, P. Swaminathan, H. (1996). Identification of items that show nonuniform DIF. Applied Psychological Measurement, 20, 257-274.

Park, D., & Rautenshlager, G. J. (1990). Improving IRT Item Bias Detection with Iterative Linking and Ability Scale Purification. Applied Psychological Measurement, 14(2), 163-73.

Pennock-Roman, M., (1986). New directions for research on Spanish-language tests and test-item bias. In Michael A. Olivas (Ed.), Latino College Students, NY: Teachers College Press.

Penny, J., & Johnson, R. L. (1999). How group difference in matching criterion distribution and IRT item difficulty can influence the magnitude of the Mantel-Haenszel chi-square DIF index. Journal of Experimental Education, 67(4), 343-66.

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomous scored items: A framework for classification and evaluation. Applied Psychological Measurement, 19(1), 23-37.

Ramsay (2000). TestGraf program: A Program for the Graphical Analysis of Multiple Choice Test and Questionnaire Data [Computer program]. Retrieved January 25, 2001 from ftp://ego.psych.mcgill.ca/pub/ramsay/testgraf/

Raudenbush, Bryk, Cheong, & Congdon, (2001). HLM5: Hierarchical linear and nonlinear modeling [Computer program]. Lincolnwood, IL: Scientific Software International.

Reise, S. P. (2000). Using multilevel logistic regression to evaluate person-fit in IRT models. Multivariate Behavioral Research, 35(4), 53-68.

Roussos, L., & Stout, W. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. Journal of Educational Measurement, 33, 215-230.

Raju, N. S. (1988). The area between two item characteristic curves. Psychometrika, 53, 495-502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. Applied Psychological Measurement, 14, 197-207.

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters in SIBTEST and Mantel-Haenszel type I error performance. Journal of Educational Measurement. 33(2), 215-30.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. Psychometrika. 58(2), 159-94.

SPSS. (1997). SPSS for Windows [Computer program. Chicago, IL: SPSS.

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

Teresi, J. A. (2001). Statistical methods for examination of differential item functioning (DIF) with applications to cross-cultural measurement of functional, physical and mental health. Journal of Mental Health and Aging. 7(1), 31-40.

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

Tissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. Holland & H. Wainer (Eds.), Differential item functioning (pp. 67-1130. Hilsdale, NJ: Erlbaum.

Walker, C. M., Beretvas, S. N. & Ackerman, T. (2001). An examination of conditioning variables used in computer adaptive testing for DIF analyses. Applied Measurement in Education, 14(1), 3-16.

Whitney, D. J. & Schimitt, N. (1997). Relationship between culture and responses to biodata employment items. Journal of Applied Psychology, 82(1), 113-129.

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

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? Journal of Educational Statistics, 15(3), 185-97.

Appendix A

Item analysis statistics and reliability analysis of the FTCAP English test

TCAP English test	Mean	SD	Cases	Item-total correlation	Alpha if item deleted
Item 1	0.6397	0.4802	1654	0.3875	0.884
Item 2	0.2044	0.4034	1654	0.2821	0.8853
Item 3	0.6626	0.473	1654	0.3434	0.8845
Item 4	0.5218	0.4997	1654	0.3418	0.8846
Item 5	0.5774	0.4941	1654	0.4143	0.8836
Item 6	0.4426	0.4968	1654	0.3627	0.8843
Item 7	0.3688	0.4826	1654	0.3559	0.8844
Item 8	0.6276	0.4836	1654	0.3502	0.8844
Item 9	0.4758	0.4996	1654	0.4061	0.8837
Item 10	0.52	0.4998	1654	0.3439	0.8845
Item 11	0.2703	0.4442	1654	0.2677	0.8855
Item 12	0.2243	0.4172	1654	0.2759	0.8853
Item 13	0.2467	0.4312	1654	0.2131	0.8861
Item 14	0.2878	0.4529	1654	0.3048	0.885
Item 15	0.3319	0.471	1654	0.2874	0.8852
Item 16	0.1312	0.3377	1654	0.2291	0.8858
Item 17	0.3029	0.4597	1654	0.2848	0.8853
Item 18	0.2787	0.4485	1654	0.3302	0.8847
Item 19	0.4661	0.499	1654	0.308	0.885

Item 20	0.2158	0.4115	1654	0.2673	0.8854
Item 21	0.1378	0.3448	1654	0.1641	0.8864
Item 22	0.3356	0.4723	1654	0.2448	0.8858
Item 23	0.4135	0.4926	1654	0.3408	0.8846
Item 24	0.2424	0.4287	1654	0.2242	0.8859
Item 25	0.1324	0.339	1654	0.1844	0.8862
Item 26	0.6481	0.4777	1654	0.308	0.885
Item 27	0.5121	0 . 5	1654	0.3553	0.8844
Item 28	0.1614	0.368	1654	0.1755	0.8863
Item 29	0.2044	0.4034	1654	0.278	0.8853
Item 30	0.4172	0.4932	1654	0.2676	0.8855
Item 31	0.7975	0.402	1654	0.2797	0.8853
Item 32	0.2455	0.4305	1654	0.3296	0.8847
Item 33	0.4571	0.4983	1654	0.3879	0.8839
Item 34	0.5405	0.4985	1654	0.3848	0.884
Item 35	0.6245	0.4844	1654	0.2669	0.8855
Item 36	0.2146	0.4107	1654	0.2693	0.8854
Item 37	0.4462	0.4972	1654	0.3488	0.8845
Item 38	0.6463	0.4783	1654	0.3998	0.8838
Item 39	0.6941	0.4609	1654	0.3767	0.8841
Item 40	0.6252	0.4842	1654	0.3568	0.8844
Item 41	0.4988	0.5001	1654	0.3913	0.8839
Item 42	0.2503	0.4333	1654	0.2126	0.8861
Item 43	0.6505	0.4769	1654	0.4051	0.8837
Item 44	0.5302	0.4992	1654	0.465	0.8829
Item 45	0.5816	0.4934	1654	0.4287	0.8834

Item 46	0.578	0.494	1654	0.4658	0.8829
Item 47	0.6052	0.489	1654	0.4348	0.8833
Item 48	0.2926	0.4551	1654	0.3582	0.8844
Item 49	0.2715	0.4448	1654	0.2171	0.8861
Item 50	0.0816	0.2739	1654	0.1093	0.8867
Item 51	0.3839	0.4865	1654	0.1968	0.8865
Item 52	0.4401	0.4966	1654	0.4694	0.8828
Item 53	0.3495	0.4769	1654	0.3288	0.8847
Item 54	0.2183	0.4132	1654	0.3677	0.8843
Item 55	0.2461	0.4309	1654	0.2293	0.8859
Item 56	0.3603	0.4802	1654	0.3916	0.8839
Item 57	0.0635	0.2439	1654	0.14	0.8864
Item 58	0.2709	0.4445	1654	0.3144	0.8849
Item 59	0.2739	0.4461	1654	0.3453	0.8845
Item 60	0.1632	0.3697	1654	0.3174	0.8849

Appendix B

An example of the data set used in the HGLM DIF procedure

Level-1 data

	Item01	Item02	Item03	Item04	Item05		Item58	Item59	Y
Person0231	1	0	0	0	0		0	0	1
Person0231	0	1	0	0	0		0	0	0
Person0231	0	0	1	0	0		0	0	0

Person0231	0	0	0	1	0		0	0	1
Person0231	0	0	0	0	1		0	0	0
Person0231	0	0	0	0	0		0	0	0
Person0231	0	0	0	0	0		0	0	0
Person0231	0	0	0	0	0		0	0	1
Person0231	0	0	0	0	0		0	0	1
Person0231	0	0	0	0	0		0	0	0
Person0231	0	0	0	0	0		0	0	1
Person0231	0	0	0	0	0		0	0	0
Person0231	0	0	0	0	0		0	0	1
Person0231	0	0	0	0	0		0	0	0
Person0231	0	0	0	0	0		0	0	0
Person0231	0	0	0	0	0		0	0	0
Person0231	0	0	0	0	0		0	0	1
Person0231	0	0	0	0	0		0	0	0
Person0231	0	0	0	0	0		1	0	1
Person0231	0	0	0	0	0		0	1	1
Person0231	0	0	0	0	0		0	0	0

Level-2 Data

id	ftcap	item01	item02	item03	item04	item05	item57	item58	item59
Person023	0	1	0	0	0	0	0	0	0
Person023	1	0	1	0	0	0	0	0	0
Person023	0	0	0	1	0	0	0	0	0

Person023	0	0	0	0	1	0	0	0	0
Person023	0	0	0	0	0	1	0	0	0
Person023	1	0	0	0	0	0	0	0	0
Person023	0	0	0	0	0	0	0	0	0
Person023	0	0	0	0	0	0	0	0	0
Person023	0	0	0	0	0	0	0	0	0
Person023	0	0	0	0	0	0	0	0	0
Person023	0	0	0	0	0	0	0	0	0
Person023	0	0	0	0	0	0	0	0	0
Person023	0	0	0	0	0	0	0	0	0
Person023	0	0	0	0	0	0	0	0	0
Person023	0	0	0	0	0	0	1	0	0
Person023	0	0	0	0	0	0	0	1	0
Person023	0	0	0	0	0	0	0	0	1
Person023	0	0	0	0	0	0	0	0	0

Wonsuk Kim

Doctoral candidate of Pennsylvania State University
301 Jack's Mill DR. #5
Boalsburg, PA 16827
Phone: (814) 466-0529
Email: wxk130@psu.edu

Education

Doctoral Student of Pennsylvania State University 1998-present
MA degree in Seoul National University (SNU) 1997
BA degree in Seoul National University 1990

Experience

Research assistant of the Engineering College in PSU Fall 2002

- Program evaluation by using the multiple-observation data analysis of Entrepreneurship minor program using web-based survey system

Summer intern of ETS Summer 2002

- Comparability research of TOEFL CBT Essay Prompts for Gender groups by using the logistic discriminant function analysis

Research assistant of the Methodology Center in PSU 2001-2002

- Power analysis of hierarchical linear modeling
- Polytomous DIF analysis for affective test for the NIDA project

Research assistant for a test development and validation project 2001-2002

- Psychometric analysis for POINT (Parents' Observations of Infants and Toddlers) an early childhood screening test: sponsor by the First Point, Inc. in AZ

Research assistant of the Department of Educational Psychology in PSU 1999-2000

- Evaluation of staff development workshop in Pennsylvania: sponsor by the Tuscarora Intermediate Unit

Assessment and statistical Skills

- Classical test theory analysis (SPSS, ITEMAN, GENOVA)
- Most Item response theory and/or Rasch model analysis programs
- Structural equation modeling (EQS, AMOS)
- Multivariate and multilevel data analysis using most statistical packages

Publications

- Kim, W., Lee, Y., & Breland, H. (in process). Comparability of TOEFL CBT Essay Prompts for Gender groups
- Kim, W. & Pennock-Roman, M. (in process). Polytomous differential item functioning of affective test for preventive factors in the context of drug abuse prevention.
- Suen, H. K., & Kim, W.S. (2000). An impact evaluation of the Tuscarora Intermediate Unit Statewide Standards network Seminars. McVeytown, PA: Tuscarora Intermediate Unit.