A COMPARISON OF METHODS FOR ITEM ANALYSIS AND DIF USING CLASSICAL TEST THEORY, ITEM RESPONSE THEORY, AND GENERALIZED LINEAR MODEL

by

YOUNGSOON SOHN

(Under the Direction of Seock-Ho Kim)

ABSTRACT

Using three different theoretical frameworks – classical test theory, item response theory, and generalized linear model – this study analyzed items and searched for DIF on a nationwide English exam for students in the third grade of middle school in South Korea. The results from the three approaches were compared. For item analysis, the estimates of item difficulty and item discrimination were examined, and the correlations between indices for item parameters were calculated. The item parameter indices from classical test theory, item response theory, and generalized linear model behaved very similarly, and the correlations were high ranging from .882 to 1.0. Thus, the results of this study indicated that indices for item parameters under three different approaches are very comparable and may be substitutable for each other. Unlike item analysis, there were large differences in detecting DIF under the different frameworks. The IRT-LR procedure for item response theory and the logistic regression model under generalized linear model were more general and flexible than the Mantel-Haneszel statistic in detecting DIF.

INDEX WORDS: Classical test theory, Differential item functioning (DIF), Generalized linear model, Item analysis, Item difficulty, Item discrimination, Item response theory, Logistic regression, Mantel-Haenszel statistic.

A COMPARISON OF METHODS FOR ITEM ANALYSIS AND DIF USIGN CLASSICAL TEST THEORY, ITEM REPONSED THEORY, AND GENERALIZED LINEAR MODEL

by

YOUNGSOON SOHN

B.A., Seoul National University, South Korea, 1993

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF ARTS

ATHENS, GEORGIA

2009

A COMPARISON OF METHODS FOR ITEM ANALYSIS AND DIF USIGN CLASSICAL TEST THEORY, ITEM RESPONSE THEORY, AND GENERALIZED LINEAR MODEL

by

YOUNGSOON SOHN

Major Professor:	Seock-Ho Kim
Committee:	Karen Samuelsen Jonathan Templin

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2009

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

1.1 OVERVIEW

The ultimate goal of test construction is to create a high quality test that exhibits an adequate degree of reliability and validity for the purpose of the test. Reliability is the degree to which individuals' deviation scores remain relatively consistent over repeated administration of the same test or alternative test forms (Crocker & Algina, 1986). If the test has the acceptable reliability, it is assumed that the results could be replicated and extended to a population and to a variety of conditions.

Validity refers to the degree to which the instrument measures what it intends to measure. It can be thought of as the accuracy of the measurement. Although a variety of sources for validity evidence exists, test items are a fundamental source of validity evidence, because the individual item has always been the basic building block of a test (Wainer, 1989).

A review of the historical development of the concept of validity in testing provides additional support for the importance of test items as sources of validity evidence. Messick (1995) who unified traditional types of validity – criterion, content and construct validity – noted, "Validity is not a property of the test or assessment as such, but rather of the meaning of the test scores, and emphasized score meaning and social values in test interpretation and test use" (p. 741). He also mentioned item analysis as an example of content aspect evidence which is one of his six distinguishable aspects of construct validity: "We can directly probe the ways in which

individuals cope with the items or tasks, in an effort to illuminate the processes underlying item response and task performance" (Messick, 1989, p. 6).

Although Messick's unified framework for validity has some positive contributions such as a broad focus on an array of issues related to the interpretations and uses of test scores, his unitary concept of validity has faced criticism, because of its impracticality due to the many different kinds of validity evidence he described and the ambiguity of the conceptual boundaries of different validity evidences.

In order to make the implications of test scores clear, Kane (1992) organized Messick's validity framework by focusing attention on the details of the interpretation. He described using an interpretive argument in order to justify test use and interpretation with appropriate evidence, and he emphasized that the most basic evidence for supporting inference is procedural like the standardization of questions and scoring keys. Kane's argument also stressed that validity evidence must be collected during the test construction process, and not only after complete building of the test. The argument-based approach, which focuses attention on specific aspects of measurement procedures, can make negative evidence (e.g., many items are biased) easily be found and revised.

Dissatisfaction with Messick has influenced many critics to search for the locus of evidence for validity in the essentially internal portion of the test, instead of external portion such as the correlation between test scores and other measures. Lissitz and Samuelsen (2007) suggested that content validity and reliability are the critical descriptors of the test, and they emphasized the importance of items on a test as evidence for content validity and reliability. They presented diverse questions with regard to test construction which should be considered, including a) "Are the items appropriate for the purpose of the assessment?," b)"When matched on ability, do

students from different racial groups perform similarly? Genders?," and c) "Does the test provide the same information on different occasions?" (Lissitz & Samuelsen, 2007, p. 441).

Based on the review above, it is noteworthy that the results of item analysis are always regarded as essential evidence for validity, irrespective of extremely diverse views on validity. Thus, for the construction of a test with higher level of reliability and validity, all the items should be examined and revised from the results of formal item review during the constructive process of the test, so as to eliminate flawed items and to select the items that maximize the predictive power of the test.

1.2 ITEM ANALYSIS AND DIF

To construct a test of minimum length with maximum reliability and validity for the purpose of the test, we can select a subset of items that maximally contribute to the reliability and validity out of a large pool of items through the process of item analysis. Item analysis, which examines whether there is any statistical difference in examinees' responses to an individual test item (Crocker & Algina, 1986) assists us in constructing an optimal test with sufficient degree of reliability and validity for the purported uses of the test.

The representative item parameters are item difficulty, item discrimination and distracter. The definition and meanings under theoretical frameworks for item analysis are explained in detail in Chapter 2 of this study.

Differential item functioning (DIF) is present when examinees from different groups have differing likelihoods of success on an item, after they have been matched on the ability of interest (Clauser & Mazor, 1998). It is important to match groups, since the comparison should establish a distinction between differences in item responses from divergences between two groups. For

example, in a mathematics test which requires calculation ability and English reading comprehension, consider examinees with the same calculation ability. However, one group (e.g., native English speakers) is more proficient in English reading comprehension than the other group (e.g., examinees who speak English as a second language). If the two groups exhibit differences in the probability of answering some of the items of the test correctly, due solely to differences in English proficiency, the items can be said to present DIF.

It is necessary for test developers or test users to investigate whether items influence examinees' performance in systematically biased ways for some particular subgroups due to any extraneous sources of variance. Thus, if there are DIF items, it means that nuisance factors which probably have effects on the responses, but are not interested in, are driving the responses beyond the latent variable that is purportedly measured (Ackerman, 1992). If some items function unfavorably over specific groups, the interpretations made from the test cannot be thought of as valid and fair.

DIF may be divided into two types; uniform DIF and nonuniform DIF. Uniform DIF means that the correct response probability for one group is uniformly greater than that for the other group across all ability levels. Thus, uniform DIF exhibits when there is no interaction between the level of ability and group membership. Nonuniform DIF occurs when interaction between ability and group membership is checked. That is, nonuniform DIF means that the probability of getting items right for two groups does not follow a consistent pattern across all levels of ability (Swaminathan & Rogers, 1990).

1.3 THE PURPOSE OF THE STUDY

Many methods and indices have been developed in order to sort and select proper items during the test development process. Although numerous indices have been devised and used, most of these methods rely on the item parameters of classical test theory (CTT) framework and item response theory (IRT) framework.

This study examines the examinees' actual responses and detects DIF on a practical test with both dichotomously- and polytomously-scored items, using not only CTT and IRT models, but also generalized linear model (GLM) such as Swaminathan and Rogers (1990) used for DIF detection.

Specifically, the first purpose of this study is to estimate whether the items of the test are well constructed for the intended uses of the test under CTT, IRT, and GLM methods of item analysis. Also, these three frameworks are compared to determine whether there is any difference in the information given by the different approaches of item analysis.

The second purpose is to examine whether there are problematic items with respect to DIF in different conditions such as gender and the degree of urbanization. The pattern of examinees' responses are reviewed according to gender and the degree of urbanization, and it is determined which items on the test function advantageously over any particular group.

CHAPTER 2

THEORETICAL BACKGROUD

The second chapter presents theoretical frameworks that underlie CTT, IRT, and GLM, along with a review of the literature comparing these approaches with regard to item analysis and DIF. In the theoretical frameworks section, the basic concepts and assumptions of the three approaches are presented as well as the theoretical methods for analyzing items including DIF. The literature review presents information about the distinct and relative features of the frameworks, including a comparison of the item parameter properties, in order to easily to understand the results of this study from the three different approaches.

2.1 THEORETICAL FRAMEWORKS

2.1.1 CLASSICAL TEST THEORY (CTT)

An essential premise of CTT is that any observed score for person j on a test is the composite of a true score and a random error for that person on the test. The model is presented in the form such as

$$X_j = T_j + E_j, \tag{1}$$

where X_j is the observed score for person j on a test, T_j is the true score , and E_j is the error.

The true score for person j on a test is defined as

$$T_j = E(X_j) = \mu_{x_j}. \tag{2}$$

That is, person j 's true score can be considered as the average of all the observed scores which

the person *j* may obtain over repeated testing. The variance of observed score can be interpreted as a composite of the variance of true score and the error variance. This is because the variance of a composite test score is created by summing two or more subtest scores (i.e., the sum of the variance term plus the sum of covariance term), and the sum of covariance term is assumed to be zero. The variance of observed score is expressed in the form as

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2 . \tag{3}$$

Reliability, the desired consistency of test scores, refers to the degree to which test scores are free of measurement error. The reliability index can be presented as the correlation between true scores and observed scores, and can be expressed as

$$\rho_{XT} = \frac{\sigma_T}{\sigma_X} . \tag{4}$$

However, because true scores are not observable, we can just use the reliability coefficient which is defined as the correlation between observed scores on parallel tests. The reliability coefficient can be mathematically calculated as the ratio of the variance of true scores to the variance of observed scores, when two tests are parallel. Hence, the reliability coefficient is

$$\rho_{XX'} = \frac{\sigma_T{}^2}{\sigma_X{}^2} . \tag{5}$$

Two tests can be defined as parallel when each examinee has the same true score on the two tests and the error variance is homogeneous for two tests.

In CTT, the main concern of item analysis is to describe the statistical characteristics of each item. The total score of a test is considered the sum of scores on the individual items, and the individual item is of interest through its effect on the total test score (Lord & Novick, 1968).

Thus, item analysis in CTT is focused on the degree to which each item influences the whole measurement.

Item Parameters in CTT. For dichotomously scored items, the classical item difficulty is a ratio of the number of examinees getting the item right to the total number of examinees. The range of this proportion is always from .00 to 1.00. The observed value of item difficulty is affected by both the examinees' true score and the effect of their guessing. Item true score variance is maximized when half of the examinees answer correctly based on knowledge. Thus, the expected item difficulty with maximum true score variance is

$$P_0 = .50 + .50/m, \tag{6}$$

where m is the number of choices on the item (Crocker & Algina, 1986). For polytomously-scored items, either item mean or rescaled mean based on the range of the item score can be used as the item difficulty.

Item discrimination is an index of how effectively the item separates examinees who vary in their degree of knowledge tested and their ability to use it. The discrimination index (D) is one of the most useful methods for dichotomously scored items, due to its computational simplicity. The value of D is the difference between the proportion getting the item correct in the upper group who scored above a standard which test developers set according to the purpose of a test, and the proportion answering the item right in the lower group who scored below the standard. Kelly (1939) suggested that the value of item discrimination can be stable, using the upper 27% and the lower 27% of examinees, if no distinction is made among the members of each group separately. Values of item discrimination may be given as between -1.00 to 1.00.

To calculate item discrimination for dichotomous items, the correlation coefficient is also used. The formula of point biserial correlation is presented as the Pearson correlation between responses to a particular item and scores on the total test and defined as

$$\rho_{pbis} = \frac{(\mu_{+} - \mu_{X})}{\sigma_{X}} \sqrt{p/q}\,, \tag{7}$$

where μ_{+} is the mean score for examinees who get the item right, μ_{X} is the mean score for all examinees who take the test, σ_{X} is the standard deviation, p is item difficulty, and q is $(1-p)$. If the latent variable underlying item performance has normal distribution and test users want to select items with an extreme value of difficulty, the biserial correlation coefficient is recommended. This formula can be derived by substituting (p/Y), where Y is the Y ordinate of the standard normal curve at the z-score associated with the item difficulty, for $\sqrt{p/q}$ in the formula for point biserial correlation.

In general, items are considered appropriate when they exhibit the proper difficulty level and discrimination value in terms of the intended purpose of the test. However, in the reverse case such as when items have inappropriate difficulty or discriminative values, the items can be improved through the distracter analysis. Distracter analysis is the process for evaluating whether alternative responses to each item effectively function. Considering the ideal responses to the alternatives of an item, Hills (1981, p. 81) suggested four patterns: (1) at least one examinee should select every distracter, (2) the right answer should be selected much more frequently by the examinees in the upper group than those in the lower group, (3) each decoy must be much more chosen by the lower-scoring examinees than the higher-scoring ones, and (4) it is desirable that the difficulty level of each item is similar to the optimal proportions.

DIF under CTT. The Mantel-Haenszel (MH) statistic, one of the representative methods using analysis of contingency tables, is considered the DIF method under CTT, because it matches examinees on an observable variable like total test score. In the MH method, counts of right and wrong responses on a studied item are compared in both the focal group which is the focus of analysis and the reference group that is a counterpart in the comparison. Then, using a chi-square test, the null DIF hypothesis is examined. This hypothesis can be expressed as

$$H_0 : [R_{rm} / W_{rm}] = \alpha [R_{fm} / W_{fm}] \quad m = 1,..., M \text{ and } \alpha = 1, \tag{8}$$

where R_{rm}, the count of right in the reference group; W_{rm}, the count of wrong in the reference group; R_{fm}, the count of right in the focal group; W_{fm}, the count of wrong in the focal group; m, the number of score levels; α, constant odds ratio which is the same value for all m.

The a chi-square test associated with the MH approach is defined as

$$\chi^2 = \left[\left| \sum_m R_{rm} - \sum_m E(R_{rm}) \right| - \frac{1}{2} \right]^2 / \sum_m Var(R_{rm}), \tag{9}$$

where,

$$E(R_{rm}) = E(R_{rm} \mid \alpha = 1) = N_{rm} R_{tm} / N_{tm},$$
$$Var(R_{rm}) = Var(R_{rm} \mid \alpha = 1) = [N_{rm} R_{tm} N_{fm} W_{tm}] / [N_{tm}^2 (N_{tm} - 1)]$$

If the null hypothesis is rejected as the result of chi-square significance test, we can say that the odds of answering the item correctly at a given level of the matching variable is not the same in the two groups, and the item exhibits DIF (Dorans & Holland, 1993).

MH statistic, which assumes the homogeneity of the odds ratios can be generalized to $I \times J \times K$ tables, and thus can be used for polytomously-scored items (Kim et al., 2007; Meyer et al.; 2004, & Yanagawa et al., 1994). Generalized MH may be divided into three versions based on whether both, one, or neither of the predictor and the response variable are ordinal. In the case

that only the response variable is ordinal such as the polytomous items in this study, scores are associated with only for levels of the response variable, and thus, the examinees' responses within a given row can be summarized by the mean of scores on the response variable. To test the null hypothesis of conditional independence, the I rows are compared using the statistic which examines differences among true mean values, based on the variation among the I averaged row mean responses (Agresti, 2007).

2.1.2 ITEM RESPONSE THEORY (IRT)

Unlike the focus on an aggregate of item responses such as a test score under CTT, IRT is primarily focused on the individual items of a test and whether an examinee answers each item correct or not (Baker, 2001). One of main assumptions of IRT is that an examinee responding to each item has some amount of the underlying ability or latent trait denoted by θ_j, and a test score places him or her somewhere on the ability scale. To present how an examinee's response depends on level of ability, the item characteristics curve (ICC) is used. An ICC plots the probability of responding correctly to an item, denoted by $P(\theta)$,as a function of the latent trait underlying performance on the items on the test (Crocker & Algina, 1986). Each item in a test has its own ICC.

Dichotomous Models for ICC. Models that mathematically relate θ and $P(\theta)$ have been utilized to determine the ICC that best fits the observed proportions of correct response. The normal ogive model, which may be any continuous cumulative frequency curve was used in the early work on latent trait theory. Today, however, the logistic models are predominant due to their simpler computations. The three most popular unidimensional IRT models for dichotomous tests are the one parameter logistic (Rasch) model, the two parameter logistic model, and the

Bimbaum's three parameter model (Hambleton et al., 1991). The Rasch model uses only one item parameter, item difficulty, to define the ICC. The Rasch model is expressed as

$$P_i(\theta_j) = \frac{e^{(\theta_j - b_i)}}{1 + e^{(\theta_j - b_i)}}, \tag{10}$$

where e is the constant 2.718, θ_j is the ability level, and b_i is the difficulty parameter for an individual item. The item difficulty parameter b_i is the location parameter, since it represents the ability level at which half of examinees answer the item correctly. Although the value of b_i theoretically ranges from $-\infty$ to ∞, the typical range is considered from -3 to +3. The two parameter logistic model add one item parameter to the Rasch model, the discrimination parameter, and is defined as

$$P_i(\theta_j) = \frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}}, \tag{11}$$

where a_i is the discrimination parameter, without the scaling constant 1.7, which defines the slope or the steepness of the ICC. Although the theoretical range of a_i is $-\infty < a_i < \infty$, the value of a_i is usually positive for the correct response to an item, and the value seen in practice are typically less than 2.5 (Baker & Kim, 2004). A relatively flat ICC and a low value of a_i means that the item is ineffective in discriminating between different ability levels. The equation for the three parameter model is

$$P_i(\theta_j) = c_i + (1 - c_i)\frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}}, \tag{12}$$

where c_i is the pseudo-guessing parameter that represents the probability of getting the item correct by guessing alone.

Polytomous Models for ICC. IRT can facilitate handling polytomously scored items, because the only necessary change is models for ICC, in the transition from dichotomously scored items to polytomous ones (Thissen et al., 2001). Samejima's graded model which may be the earliest among IRT polytomous models assumed that the probability of an observation in category k is the probability of observing category k or higher minus the probability. Thus, the graded model can be generalized such as

$$T(\mu = k) = T^*(k) - T^*(k+1) = \frac{1}{1+\exp\left[-a_i(\theta_j - b_{ik})\right]} - \frac{1}{1+\exp\left[-a_i(\theta_j - b_{ik+1})\right]}, \quad (13)$$

where ordered responses $\mu = k$, k = 0, 1,… $m-1$. In this equation, a_i is the slope, and b_{ik} is the point on the θ_j axis where the probability of responses for category k or higher is larger than 50%.

Alternative polytomous models of IRT can be a rating scale, a partial credit, or a generalized partial credit model. The rating scale model and partial credit models belong to the Rasch family of models, and they have the assumption of uniform discriminating power of test items. The equation is expressed as

$$P_{x_i}(\theta_j) = \frac{\exp\sum_{k=0}^{x_i}[\theta_j - (b_i + d_{ik})]}{\sum_{w=0}^{M_i}\exp\sum_{k=0}^{w}[\theta_j - (b_i + d_{ik})]}, \quad (14)$$

where k is the number of score categories, b_i is average difficulty of all responses in item i, and d_{ik} represents thresholds (i.e., deviations from the item difficulty). In the rating scale model, all items have the same thresholds, while the partial credit model assumes that the thresholds differ for all of the items. The generalized partial credit model is more flexible than the partial credit model by setting different discrimination for all items. In the equation for generalized partial

credit model below, if the value of a_i is 1, the generalized partial credit become the partial credit model.

$$P_{x_i}(\theta_j) = \frac{\exp\sum_{k=0}^{x_i} a_i[\theta_j - (b_i + d_{ik})]}{\sum_{w=0}^{M_i}\exp\sum_{k=0}^{w} a_i[\theta_j - (b_i + d_{ik})]}. \quad (15)$$

Estimate of Item Parameters. The basic task for estimating item parameters is to find an ICC which best fits the data, that is, the observed proportions of correct response, using the maximum likelihood procedure (Baker & Kim, 2004). For this, any initial values for item parameters are firstly established. Then, using those estimates, the value of $P(\theta_j)$ is computed at each ability level by means of appropriate equation for ICC models which are explained above. If the observed value of $p(\theta_j)$ and the computed value of $P(\theta_j)$ is the same or the deviation is so small, the values of the item parameters are selected. However, if the difference between two values is great, the process of adjustment may continue.

Then, using the selected values of item parameters and the equation for the ICC, $P(\theta_j)$ is computed and the ICC is plotted. By the chi-square goodness-of-fit index defined below, the agreement between the observed value of $p(\theta_j)$ and the value of $P(\theta_j)$ computed by the fitted ICC for each item is computed. The chi-square goodness-of-fit index is represented as

$$\chi^2 = \sum_j^J m_j \frac{\left[p(\theta_j) - P(\theta_j)\right]^2}{P(\theta_j)Q(\theta_j)}, \quad (16)$$

where J is the number of ability groups, θ_j is the ability level of group j, m_j is the number of examinees with ability θ_j, $p(\theta_j)$ is the observed proportion of correct response for group j,

and $P(\theta_j)$ is the probability of correct response for group j, measured from the fitted ICC using the selected estimates of item parameters.

If the value of the obtained index is smaller than a criterion value, the fitted ICC using the values of item parameters fits the data, and vice versa (Baker, 2001).

DIF under IRT. In the context of IRT, the value of the trace line (i.e., the item characteristic curve, ICC) at each level of latent ability is the varying probability of a correct response across the ability continuum. When there is any difference in responses for the focal and reference groups on an item, that is, the item shows DIF, the ICCs for the two groups are not coincident.

The existence of DIF can also be estimated by the comparison of item parameters between the groups, since the item parameters in IRT models determine the shape of the ICC for an individual item. After adjusting possible differences in the distribution of latent ability within the two groups, we can find evidence of DIF through a statistical test of whether between-group differences in item parameters are significant (Thissen, Steinberg, & Wainer, 1993).

The differences in item parameters between the two groups can be identified by diverse approaches. One is to quantify the area between the ICCs for the focal and reference groups on each item. Raju (1988) offered formulas for computing the exact area between ICCs for the one-, two-, and three-parameter IRT models explained above. Another approach suggested by Lord (1980) is to compare the difference in an item parameter between two groups with its standard error under the assumption that the ability parameter is known. In the case of the item difficulty parameter, the formula can be explained as

$$SE(\hat{b}_{i1} - \hat{b}_{i2}) = \sqrt{Var(\hat{b}_{i1}) + Var(\hat{b}_{i2})}\,. \qquad (17)$$

This procedure can be applied for the item discrimination parameter, and for both item difficulty and item discrimination parameters.

Cohen, Kim, and Baker (1993) presented the extension of Raju's and Lord's methods to polytomous items. And Thissen, Steinberg, and Wainder (1993) suggested that a model of IRT can be fitted to the data for the two groups, and that likelihood ratios can be used to evaluate if the differences in item parameter estimates are significant. If there is any significant difference, the item can be said to exhibit DIF.

2.1.3 GENERALIZED LINEAR MODEL (GLM)

GLM is considered a flexible generalization of ordinary least squares regression. This broad class of models includes ordinary regression and analysis of variance (ANOVA) models for continuous responses as well as models for categorical responses. All generalized linear models include three components: a random component, a systematic component, and a link function (Agresti, 2007).

The random component is the response variable Y and the assumed distribution for this response variable. The systematic component identifies the explanatory variables for the model. The link function which connects the random and systematic components through function $g(\cdot)$. The function linearly or nonlinearly relates a set of fitted value from the model to predictors. According to the fitted values, the model can be considered ordinary regression $[g(\mu)=\mu]$, loglinear $[g(\mu)=\log(\mu)]$, or logistic regression $[g(\mu)=\log[\mu/(1-\mu)]]$. In this study, the model used is logistic regression which is the most popular model.

Logistic regression model (Logit model). What distinguishes a logistic regression model from the linear regression model is that the response variable in logistic regression has a binomial or dichotomous distribution (Hosmer & Lemeshow, 2000). When logistic distribution is used, we can think of $\pi(x)$, a "success" probability at the value of an explanatory variable x,

[i.e., $\pi(x) = p(Y = 1 | x)$] as $E(Y | x)$, the conditional mean of Y given x in ordinary regression. The logistic regression model which has linear form for the logit of $\pi(x)$ is defined as

$$\text{logit}[\pi(x)] = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x. \quad (18)$$

This formula means that the increase of a unit of x results in the logit increase by β. If $\beta = 0$, $\pi(x)$ is the same across all values of x and the binary response variable Y is independent of X.

The mathematical function which is implied in the logistic regression model can be presented in the form of

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}. \quad (19)$$

Multicategory logit models. These models handle the case where the response variable consists of two categories. Two types of multicategory logit models, multinominal logit models and cumulative logit models, are differentiated by the characteristics of the response variable. Multicategory logit models have a nominal response variable, and cumulative logit models have an ordinal response variable.

Logit models for nominal response variables where the order of categories is unimportant pair each category with a baseline category, often the last one or the most common one. The goal is to simultaneously represent the odds that the response falls in one category relative to the case that it falls in another category for all pairs of categories. The baseline-category logit model is

$$\log\left(\frac{\pi_j}{\pi_J}\right) = \alpha_j + \beta_j x, \qquad j = 1, \ldots J-1. \quad (20)$$

This model describes the effects of a predictor x on these $J-1$ logits at the same time. The effects vary according to which category in the response variable is paired with the baseline. The

generalized model of the baseline-category logit, a discrete choice model, allows predictors to take different values for different categories in the response variable.

The logits of the cumulative probabilities and the model are like

$$\text{logit}\left[P(Y \le j)\right] = \log\left[\frac{P(Y \le j)}{1 - P(Y \le j)}\right] = \log\left[\frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_J}\right] = \alpha_j + \beta x, \quad j = 1, ..., J-1, \qquad (21)$$

where the final category is not used, because the sum for cumulative probabilities is always 1. To calculate the estimated cumulative probability for an explanatory variable, the model expression below can be used:

$$P(Y \le j) = \frac{\exp(\alpha_j + \beta x)}{1 + \left[\exp(\alpha_j + \beta x)\right]}. \qquad (22)$$

DIF under GLM. Swaminathan and Rogers (1990) suggested that the logistic regression procedure can be used as an effective alternative to the contingency table methods (e.g., Mantel-Haenszel method) and IRT – based methods. Logistic regression considers the total score as a continuous variable, while the contingency table methods classify groups by discrete score categories (Clauser & Mazor, 1998). The formula of DIF analysis for any item i and the two groups of interest can be defined as

$$P(u_i = 1) = \frac{e^{z_i}}{\left[1 + e^{z_i}\right]}, \qquad (23)$$

where U_i is the response to the item i, and

$$z_i = \tau_0 + \tau_1\theta_i + \tau_2 g + \tau_3(\theta_i g), \qquad (24)$$

where τ_0 is the intercept parameter, τ_1 is the performance difference based on the matching ability (θ), τ_2 corresponds to the group difference in performance, g is the group coded 0, 1, and τ_3 is the interaction effect between group and ability. It can be noted that in a usual

application of logistic regression to the DIF analysis, θ will be replaced with the total score, yielding $z_i = \tau_0 + \tau_1 x_i + \tau_2 g + \tau_3 (x_i g)$, instead of the equation 24.

If $\tau_2 \neq 0$ and $\tau_3 = 0$, the item has uniform DIF, while the item exhibits nonuniform DIF, when $\tau_3 \neq 0$, ignoring whether the value of τ_2 is 0. The parameters for each item can be estimated by the method of maximum likelihood. The likelihood function of the observed performances is represented as

$$L(\text{Data} \mid \tau) = \prod_{i=1}^{N} P(u_i)^{u_i} \left[1 - P(u_i)\right]^{1-u_i}, \tag{25}$$

where N is the total subjects, which equals the sum of the number of group 1 and the number of group 2.

2.2. LITERATURE REVIEW

2.2.1 THE COMPARISON OF ITEM INDICES

One of major characteristics of item indices under CTT and IRT models is whether they are sample dependent or sample invariant. The item parameters under CTT are regarded as sample dependent, because they are based on the total score of the test which is the person parameter in CTT and has a variant attribute. Another way of saying this is that the values of the item parameters are different across the samples collected for the test. This characteristic may be a threat to the reliability of the test. So, in order to generalize the results of the test, random sampling is assumed for CTT.

The item parameters under IRT, however, are not considered to be dependent upon the ability level of the examinees responding the item (Baker, 2001). In other words, the item parameters are regarded as sample invariant. If an item measures the same latent trait for groups, the

estimated item parameters are assumed to be the same. Because the item difficulty parameter under IRT is independent of the samples, it is considered easier to interpret than that under CTT.

Baker and Kim (2004) argued that the concepts of item difficulty in CTT and the location parameter b_i in IRT are not completely interchangeable. Under CTT, an easy item is defined by a low ratio of the correct response to an item in the total population. On the other hand, under IRT, an item is defined as easy when the magnitude of the item difficulty parameter is less than the average level of ability. Considering item discrimination parameter, however, it is regarded as the parameter which makes it possible to establish a distinction among examinees' different ability under both CTT and IRT.

Thus, IRT has been considered to hold advantages over CTT at least in terms of theoretical point of view. Lord (1980) argued that IRT provides the methods of optimally discriminating items in the scope of a passing score. However, practical researches comparing CTT and IRT have not shown there is consistent superiority of IRT measurement statistics.

Cook, Eignor, and Taft (1988) examined the stability of item parameter estimates based on CTT and IRT, using two different test administrations. They showed that the item difficulty statistics of both CTT and IRT were more stable in more homogeneous samples. Their conclusion was that the parameter estimates depend on the study samples. Contrary to general expectation, the stability of the item difficulty value from IRT was slightly lower than one from CTT. The result of the study was well summarized in their statement: "It is equally apparent that neither classical test theory nor item response theory is sufficiently robust to provide viable item analysis or equating results when faced with a lack of parallelism such as that exhibited by the 58 common items given to the spring and fall groups in this study" (Cook, Eignor, & Taft, 1998, p. 43-44).

Shannon and Cliver (1987) found in their research which analyzed 10,266 broker and 32,538 salesperson candidates' scores on a test for licensure in the field of real estate sales, that the classical indices like phi-coefficient as item discrimination parameter may be substitutable for the measurement counterpart under IRT. They suggested that the conventional item discrimination indices for a criterion-referenced test may be comparable to item information functions under IRT. Lawson (1991) showed that the correlation coefficient (r = -.9949) of the level of item difficulty through CTT and IRT (Rasch model) were very high. This result indicated that item difficulty estimates behaved very similarly in two different approaches, CTT and IRT.

CTT and IRT have also been compared under simulated conditions. Macdonald and Paunonen (2002) used Monte Carlo simulation to examine the invariance and accuracy of item statistics from CTT and IRT across examinee samples are. Their result demonstrated that the item difficulty indices generated by CTT and IRT were very accurate and highly comparable across all conditions. In the case of the item discrimination estimates, however, the statistic under IRT accurately estimated the discrimination value in all conditions. On the other hand, the item discrimination value under the CTT framework obtained an accurate estimate only when the potential item pool had a narrow range of item difficulty levels.

Regardless of the degree of the measurement accuracy of CTT and IRT approach, CTT holds some practical advantages over IRT, in that CTT methods are computationally simpler, less expensive, and require smaller samples sizes.

2.2.2 DIFFERENT STANDARDS FOR ITEM PARAMETERS

It is also important to review what values of item parameters are acceptable according to the purpose of a test. In general, tests can be classified into two categories: a norm-referenced test

(NRT) and a criterion-referenced test (CRT). In the NRT, the quality of an examinee's performance is compared to the behavior of others (Thorndike, 2005). We want to know where an examinee's score falls in relation to other examinees. In NRT, it is easier to compare performance when the distribution of examinees' scores is spread out as far as possible. That is to say, with NRT, it is essential to maximize test score variance.

As briefly mentioned earlier in the section describing item parameters under CTT, the score variance is maximized when half of the examinees can get an item correctly with knowledge. Therefore, a test of a medium difficulty level is considered appropriate for general NRT. For this, uniform item difficulty is considered more useful than a mixture of items of varying difficulty, because items at the 50% difficulty level have maximum potential of discrimination (Worthen et al., 1999). Cronbach and Warrington (1952) also said that test reliability and variance are maximized when items are at the same difficulty level. However, the intended purpose of some tests, such as a test to select the most qualified teachers, necessitates having relatively high difficulty levels across items.

In the case of CRT, the interpretation of item parameters is quite different than under NRT. The focus of CRT is to determine if an examinee reaches a performance standard or threshold on the specific skill if interest. Maximizing score variance is not an objective in this case, therefore the standards of item parameters are differently established.

The acceptable level of item difficulty on the CRT depends on the rigor of the standards which test developers set. If examiners expect that most examinees should reach the mastery level, the difficulty level will be closer to 1.0. In most cases, however, items that exhibit a medium difficulty level are acceptable, such as for items that measure instructional effectiveness (Crocker & Algina, 1986).

Item discrimination parameters are also interpreted differently under NRT and CRT. If an item is found to have a zero value of discrimination, the item on a NRT is eliminated or revised. However, on a CRT, such an item can provide evidence that an instructional objective has been achieved (Worthen et al., 1999).

Distracters are interpreted similarly under CRT and NRT. In both cases, the information from distracters can help test developers detect which parts of items are confusing regardless of the content of the test.

The particular index for analyzing items on a CRT is a measure of sensitivity to instructional effects. The value of instructional sensitivity can be calculated by subtracting the proportion of people who get the item right on a pretest before instruction from the proportion who answer the item correctly on a posttest after instruction. This value will fall within the range of -1.00 to 1.00. Because the index of instructional sensitivity estimates the effectiveness of instruction, values closes to 1.00 are desirable.

2.2.3. THE COMPARISON OF APPROACHES FOR DIF

The Mantel-Haenszel (MH) statistic, IRT methods, and logistic regression (LR) are methods for identifying items that exhibit DIF. These various methods have some similarities and differences.

The various IRT approaches for identifying DIF are similar in that they use an estimate of latent ability as a matching variable. A strength of IRT methods is the ICC, which provides a vivid graphical display of both uniform and nonuniform DIF. However, there are some drawbacks of IRT methods. As Clauser and Mazor (1998) indicated, the data should satisfy the assumptions of the IRT model selected, and generally large samples are needed. Thus, IRT approaches are considered to be more time-consuming than other approaches. It is also difficult

to examine the significance of the tests through indices like the area between ICCs (Swaminathan & Rogers, 1990).

Some researchers have compared the LR procedure with the MH statistic which may be regarded as a specific type of LR that has a discrete ability variable and no interaction term. The MH method provides a solution for some of the drawbacks to IRT approaches, because it can provide a test significance using the chi-square distribution, and it requires a relatively small sample size. The major disadvantage of MH approach, however, is that the detection rate for nonuniform DIF is very low.

Swanminathan and Rogers (1990) examined the detection rates of uniform and nonuniform DIF using the MH procedure and the LR approach under six varying test length and sample size conditions. The results of their research showed that very similar rates of uniform DIF detection under both the MH approach and the LR method were found. For nonuniform DIF, the LR method found it out with 50% accuracy in a small sample size (250 per group) regardless of test length (40, 60, and 80 items), while it could hardly be detected by the MH procedure under any condition.

In addition to powerful detection of nonuniform DIF, the other important benefit of the LR procedure is that it is a model-based approach. So, relevant factors like an ability variable can be included in the equation of LR procedure, and the effects may be examined (Swaminathan & Rogers, 1990). Related to the criteria of the ability estimate in LR approach, Crane et al. (2006) suggested that an IRT-based ability estimate can be used for better ability measurement instead of total score which was suggested by Swaminathan and Rogers (1990).

CHAPTER 3

PROCEDURE

3.1 INSTRUMENTATION

The data used for this study is the subset of the results of a nationwide English exam in South Korea. The exam, called ‘A Nationwide Measurement of Academic Achievement,’ has been administered by the Korean Institute for Curriculum and Evaluation (KICE) every October beginning in 1998. The purpose of the evaluation is to examine students’ achievement compared to educational standards, to identify problems in educational processes, and to present helpful references for the improvement of courses of study. Thus, it can be seen as a criterion-referenced test.

The examination covers five subjects: Korean language, social studies, science, mathematics, and English. The examinees are about 3% of all students in the following grades: the sixth grade of elementary school, the third grade of middle school, and the first grade of high school. This study considers only middle school data from the English portion of the exam conducted in 2007.

Based on their test score, students are placed into one of four categories: excellent, satisfactory, basic and below basic. A student who correctly answers at least 80% of the test items is placed in the “excellent” category. Students who correctly respond to 50% to 80% of the test content are placed in the “satisfactory” category. Students who respond correctly to less than 20% of the test items fall into the “below basic” category. This standard was set using the modified Angoff methods.

The English exam is composed of four sub-domains: listening, speaking, reading, and writing. The composition of the sub-domain is shown in Table 1 below. The items can be divided into dichotomously-scored (32 items, 1.5 points for each one) and polytomously-scored (10 items, 2 points for 8 items and 4 points for 2 ones) items. The total score is 72 points.

Table 1

The Composition of Four Sub-domains of the Test.

	Type	Number	Points		
			Each item	Sub-total	total
Listening	Dichotomous	12	1.5	28	
	Polytomous	5	2 (partial point: 1)		
Speaking	Dichotomous	2	1.5	3	
	Polytomous	0			
Reading	Dichotomous	18	1.5	37	72
	Polytomous	4	3 items: 2 (partial point: 1)		
			1 item: 4 (partial point: 2)		
Writing	Dichotomous	0		4	
	Polytomous	1	4 (partial point: 2)		

3.2 SAMPLE

The original sample for the nationwide English exam (about 3% of the third grade students in Korean middle schools) was selected by a two-stage stratified cluster sample design. To collect examinees which are representative of the third grade of the middle school, the population was classified into several strata according to the factors related to students' educational achievement,

such as the degree of urbanization and the size of schools. Among each stratum, schools were randomly selected and then, classes were selected within schools. That is to say, primary sampling unit is school, and the secondary sampling unit is class.

This study uses a subset of data from the results of the original sample. The data used in this study was randomly collected from the original sample based on the degree of urbanization and gender, because one of the specific purposes of this study is to examine whether there is any difference in the level of students' achievement according to gender and the region where they live. The total sample was 1,801. Of these, 912 examinees came from 12 middle schools in 'city' regions and 889 middle school students came from 14 schools in 'rural' regions. In total, there are 903 girls and 898 boys in the sample.

The mean and standard deviation of the test score of this sample are presented in Table 2.

Table 2

The Mean and Standard Deviation of the Test Score.

			Urbanization		Gender		Total
	points		City	Rural	Male	Female	
Listening	28	*M*	16.88	14.69	14.70	16.90	15.80
		SD	7.85	6.60	7.39	7.13	7.34
Speaking	3	*M*	1.87	1.68	1.63	1.91	1.78
		SD	1.18	1.17	1.18	1.16	1.18
Reading	37	*M*	18.48	15.68	15.24	18.94	17.09
		SD	11.21	9.07	10.18	10.10	10.30
Writing	4	*M*	2.37	2.03	1.94	2.47	2.20
		SD	1.72	1.63	1.71	1.61	1.68
Total	72	*M*	39.60	34.07	33.51	40.22	36.87
		SD	20.42	16.53	18.68	18.32	18.80

Note. *M* = Mean. *SD* = Standard Deviation.

3.3 COMPUTER PROGRAMS

Several computer programs are used in this study. In order to analyze items and identify DIF, IRTLRDIF, Lertap 5, Multilog 7, SAS, and SPSS programs are used.

The version 5 of the Laboratory of Educational Research Test Analysis Package (Lertap 5, upgraded by Curtain University of Technology in 2007) is a classical item and test analysis system and is based on the Excel program. To analyze the item based on CTT, Lertap 5 is mainly used. Item analysis and DIF under IRT are examined by the use of Multilog 7 (Thissen, Chen, & Bock, 2003) and IRTLRDIF (Thissen, 2001). Multilog 7 handles all the major models under IRT including 1, 2, and 3 parameter logistic models, and the graded response model. SAS is used to calculate the Mantel-Haenszel statistic for DIF detection under CTT, as well as GLM-derived item analysis and DIF. Additionally, Statistical Package for the Social Science (SPSS) is also used to obtain various statistical results related to this study.

CHAPTER 4

RESUTLS

4.1 ITEM ANALYSIS

To analyze items of a test and to search for DIF items, three different theoretical frameworks, CTT, IRT, and GLM, were used. Before analyzing the items, the weights of the values of polytomously-scored items were reassigned. Specifically, the original weights of items P9 and P10 (i.e., 0, 2, and 4) were shifted to be consistent with the weights for other polytomous items (i.e., 0, 1, and 2). This change was justified, because the original weighting system was based on academic decision rather than statistical concerns. Dichotomously- and polytomously-scored items, were analyzed separately, since different methods for examining item parameter estimates were used according to the item format. Both item formats were considered using the same criteria, subtest total score and entire-test total score.

4.1.1 DICHOTOMOUS ITMES

To analyze dichotomous items under CTT, *p*-value, item difficulty index and item discrimination indices including the point-biserial correlation, biserial correlation, and *D* for 27% and 50% upper and lower groups were obtained with the criteria of total score of dichotomous items and total score of entire test, respectively. The biserial correlation, r_{bis}, was used as the representative discrimination index under CTT for comparing with other discrimination indices under IRT and GLM, as Wainer (1989) suggested.

The results of item analysis under CTT are presented in Table 3. In conformity with the rule of thumbs (Cangelosi, 1990), Items D1, D2, and D12 were relatively easy ($P \geq .75$), while *p*-values of all items were larger than .25 which is regarded as the standard of difficulty items. Regarding to item discrimination, almost all of items were considered as acceptable. Items D1 and D4, however, were checked to have poor discriminative power ($D \leq .29$), according to Ebel (1965)'s guideline for interpretation of the index of discrimination, D. To compare the patterns of item difficulty and item discrimination across items, graphs were created (see Figure 1 and Figure 2).

For item analysis using IRT, at first, dichotomously-scored items were separately calibrated with 2PL and 3PL models by the computer program Multilog 7. With 3PL, however, the pseudo-guessing parameter was not fully estimated on some items, particularly on easy items. Thus, the item difficulty parameter (b_i) and item discrimination parameter (a_i) with 2PL were only presented in Table 4. And items were simultaneously calibrated with the criterion of entire total score.

Using both criteria, items which had relatively lower value of item difficulty were listed as D1, D2, D10, and D12, while items D4, D6, and D20 were difficult items. Items D4 and D9 had ineffective discriminative power. In other hand, items D2, D12, D15, D19, and D25 effectively discriminated between examinees with different ability levels. Easily to grasp characteristics of items, the ICCs of items D2 and D4 are presented in Figures 3 and 4, respectively. These items are noteworthy because item D2 exhibited a low level of item difficulty and high discriminative power, while item D4 exhibited a high difficulty value and low level of discrimination.

To get item parameter estimates under GLM, the logistic regression model in the SAS program was used with the two criteria that were also used for CTT and IRT approaches. In a

logistic regression equation, the parameter β determines the rate of increase or decrease of the logistic regression curve for the probability of correct response. As $|\beta|$ increases, the rate of change in the curve also increases (Agresti, 2007). Thus, the parameter β can be used as an item discrimination index. In a logistic regression model, the x value at which half of examinees get the item right can be represented as $x = -\alpha / \beta$. This value can considered the index of item difficulty under GLM, which corresponds to the item location parameter under IRT. Although α and β were estimates that it should be noted with carets ($\wedge$), the notation was not used.

The results presented in Table 5 indicate that items D1, D2, and D12 were easy, while items D4 and D6 were very difficult based on either the subtest total score or the entire-test total score. In terms of item discrimination, Items D2, D15, D19, and D25 were very effective, and items D4, D9, and D23 had little discriminative power. In Figures 5 and 6, the logistic regression curves for items D2 and D4 are presented. Figures 7 and 8 show the comparison of the discriminative patterns across items under IRT and GLM.

Table 3

Item Statistics for Dichotomously-scored Items Based on CTT Framework

	Criterion									
	Subtest total score					Entire-test total score				
Item	P	r_{pbis}	r_{bis}	$D_{27\%}$	$D_{50\%}$	P	r_{pbis}	r_{bis}	$D_{27\%}$	$D_{50\%}$
D1	.80	.41	.59	.47	.30	.80	.44	.63	.45	.29
D2	.75	.53	.71	.62	.42	.75	.56	.76	.62	.43
D3	.52	.49	.62	.69	.45	.52	.52	.65	.66	.44
D4	.34	.27	.34	.36	.22	.34	.31	.39	.35	.21
D5	.57	.58	.73	.75	.57	.57	.62	.78	.76	.56
D6	.34	.43	.55	.55	.33	.34	.47	.60	.53	.33
D7	.60	.53	.68	.71	.50	.60	.57	.72	.71	.49
D8	.46	.52	.66	.72	.48	.46	.56	.71	.72	.47
D9	.53	.36	.45	.52	.34	.53	.40	.51	.52	.33
D10	.70	.48	.63	.60	.42	.70	.51	.67	.58	.41
D11	.61	.46	.58	.63	.43	.61	.49	.62	.60	.40
D12	.76	.49	.67	.60	.37	.76	.52	.72	.58	.37
D13	.65	.52	.66	.68	.48	.65	.55	.70	.67	.47
D14	.57	.55	.69	.73	.52	.57	.58	.73	.70	.52
D15	.52	.64	.80	.84	.61	.52	.66	.83	.84	.61
D16	.55	.50	.63	.67	.48	.55	.53	.66	.65	.46
D17	.43	.57	.71	.75	.51	.43	.59	.75	.74	.50

Table 3 (continued)

Item Statistics for Dichotomously-scored Items Based on CTT Framework

	Criterion									
	Subtest total score					Entire-test total score				
Item	P	r_{pbis}	r_{bis}	$D_{27\%}$	$D_{50\%}$	P	r_{pbis}	r_{bis}	$D_{27\%}$	$D_{50\%}$
D18	.45	.53	.67	.70	.50	.45	.56	.70	.67	.48
D19	.62	.60	.76	.77	.59	.62	.62	.78	.74	.58
D20	.32	.42	.55	.52	.35	.32	.46	.59	.51	.34
D21	.43	.54	.69	.71	.49	.43	.57	.72	.70	.48
D22	.68	.50	.65	.63	.44	.68	.51	.67	.60	.42
D23	.39	.41	.52	.54	.36	.39	.44	.55	.52	.35
D24	.56	.50	.63	.68	.49	.56	.53	.67	.67	.48
D25	.68	.57	.74	.71	.53	.68	.59	.78	.69	.53
D26	.51	.60	.75	.77	.55	.51	.61	.77	.76	.54
D27	.61	.47	.60	.63	.42	.61	.49	.63	.61	.39
D28	.46	.51	.64	.70	.44	.46	.54	.68	.68	.44
D29	.69	.50	.65	.64	.44	.69	.52	.68	.62	.41
D30	.47	.50	.62	.69	.44	.47	.52	.66	.65	.44
D31	.39	.48	.61	.63	.41	.39	.50	.63	.61	.40
D32	.49	.52	.65	.70	.48	.49	.55	.69	.69	.46

Figure 1. Item difficulty and biserial correlations for dichotomous Items under CTT based on subtest total score and entire-test total score.

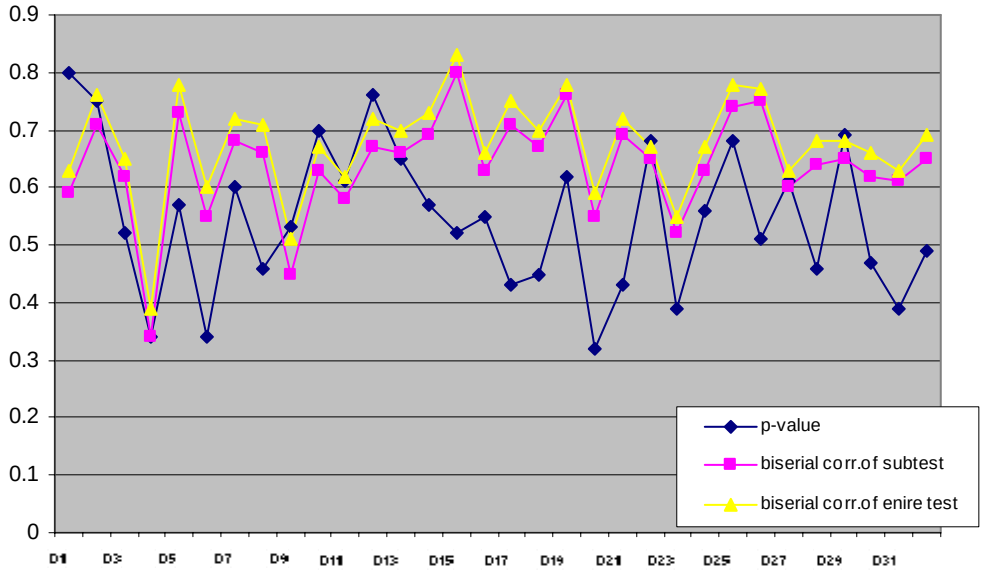


Figure 2. Item difficulty and item discrimination index (*D*) for dichotomous items under CTT based on subtest total score and entire-test total score.

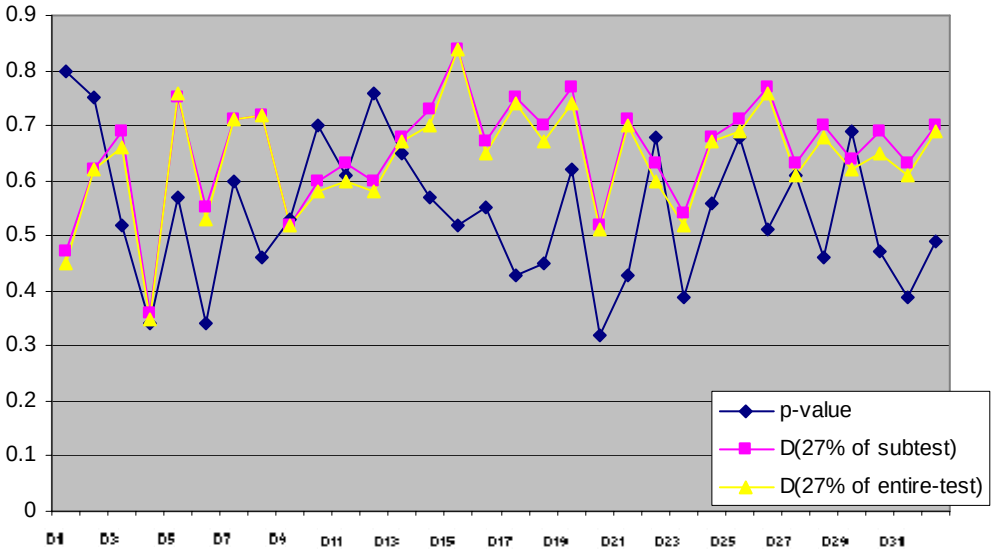


Table 4

Item Statistics for Dichotomously-scored Items from Separate (2PL) and Simultaneous (2PL & GR) Calibrations Based on IRT Framework

Item	Separate Calibration		Simultaneous Calibration	
	a_i	b_i	a_i	b_i
D1	1.73	-1.14	1.64	-1.18
D2	2.58	-0.79	2.41	-0.81
D3	1.42	-0.09	1.35	-0.07
D4	0.66	1.13	0.65	1.16
D5	1.99	-0.26	1.96	-0.25
D6	1.20	0.73	1.19	0.75
D7	1.82	-0.37	1.76	-0.36
D8	1.54	0.13	1.53	0.15
D9	0.87	-0.16	0.88	-0.14
D10	1.71	-0.73	1.61	-0.75
D11	1.33	-0.46	1.27	-0.46
D12	2.18	-0.87	2.10	-0.88
D13	1.79	-0.53	1.68	-0.53
D14	1.80	-0.27	1.71	-0.26
D15	2.47	-0.12	2.33	-0.09
D16	1.48	-0.22	1.39	-0.21
D17	1.75	0.21	1.68	0.24

Table 4 (continuted)

Item Statistics for Dichotomously-scored Items from Separate (2PL) and Simultaneous (2PL & GR) Calibrations Based on IRT Framework

Item	2PL		2PL & GR	
	a_i	b_i	a_i	b_i
D18	1.52	0.17	1.47	0.20
D19	2.42	-0.41	2.15	-0.41
D20	1.16	0.81	1.13	0.84
D21	1.65	0.23	1.56	0.26
D22	1.81	-0.65	1.59	-0.68
D23	1.06	0.50	0.99	0.54
D24	1.50	-0.25	1.44	-0.24
D25	2.52	-0.58	2.26	-0.60
D26	2.04	-0.09	1.89	-0.06
D27	1.43	-0.43	1.32	-0.44
D28	1.50	0.12	1.42	0.15
D29	1.83	-0.67	1.66	-0.69
D30	1.44	0.09	1.36	0.11
D31	1.32	0.47	1.24	0.51
D32	1.54	0.02	1.48	0.05

Figure 3. Item characteristic curve for dichotomous item D2 under IRT.

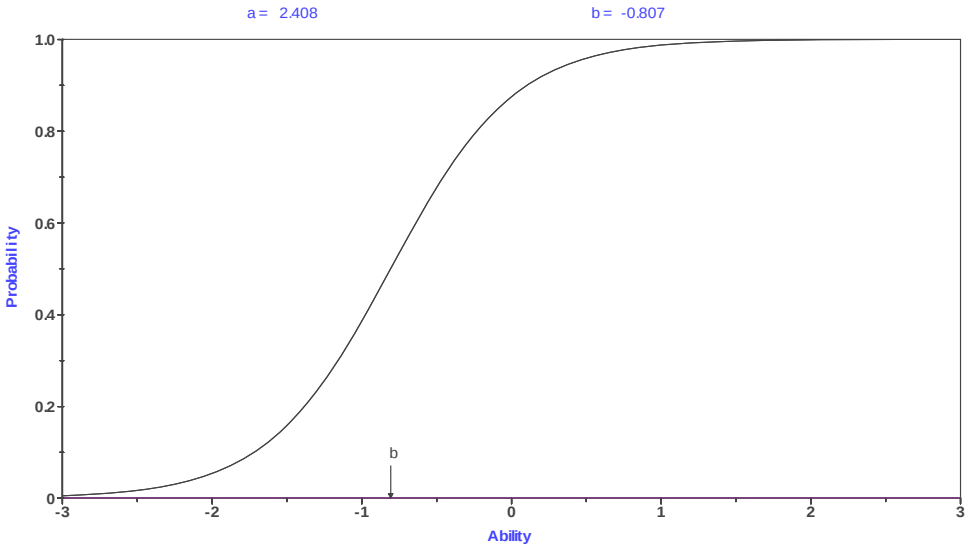


Figure 4. Item characteristic curve for dichotomous item D4 under IRT.

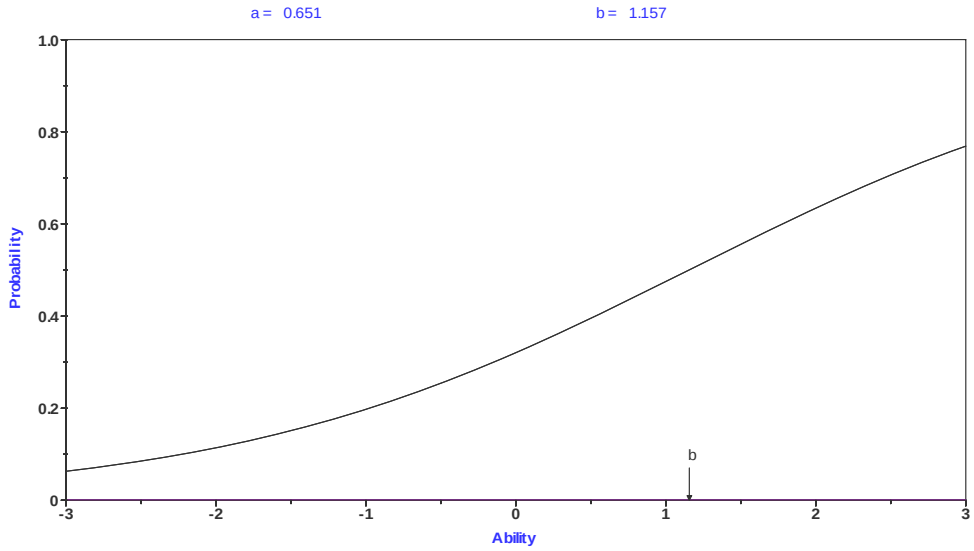


Table 5

Item Statistics for Dichotomously-scored items Based on GLM Framework

	Criterion					
	Subtest total score			Entire-test total score		
Item	α_i	β_i	$-\alpha_i/\beta_i$	α_i	β_i	$-\alpha_i/\beta_i$
D1	-1.3819	0.1968	7.0219	-1.0903	0 .1173	9.2950
D2	-2.6923	0.2722	9.8909	-2.2806	0.1619	14.0865
D3	-2.7230	0.1639	16.6138	-2.4019	0.0955	25.1508
D4	-2.2332	0.0854	26.1499	-2.0622	0.0500	41.2440
D5	-3.2725	0.2172	15.0668	-3.0198	0.1334	22.6372
D6	-3.3214	0.1403	23.6736	-3.1103	0.0864	35.9988
D7	-2.7496	0.1962	14.0143	-2.4861	0.1189	20.9092
D8	-3.2630	0.1752	18.6244	-3.0364	0.1068	28.4307
D9	-1.7875	0.1116	16.0170	-1.6242	0.0672	24.1696
D10	-2.0360	0.1908	10.6709	-1.7268	0.1128	15.3085
D11	-2.0918	0.1572	13.3066	-1.7977	0.0916	19.6256
D12	-2.1979	0.2371	9.2699	-1.8858	0.1437	13.1232
D13	-2.4980	0.1990	12.5528	-2.1593	0.1171	18.4398
D14	-2.9639	0.1983	14.9466	-2.6460	0.1180	22.4237
D15	-4.1974	0.2545	16.4927	-3.7821	0.1513	24.9974
D16	-2.6474	0.1710	15.4819	-2.2899	0.0985	23.2477
D17	-3.8243	0.1966	19.4522	-3.4894	0.1170	29.8239

Table 5 (continued)

Item Statistics for Dichotomously-scored items Based on GLM Framework

	Criterion					
	Subtest total score			Entire-test total score		
Item	α_i	β_i	$-\alpha_i/\beta_i$	α_i	β_i	$-\alpha_i/\beta_i$
D18	-3.3675	0.1775	18.9718	-3.0354	0.1045	29.0469
D19	-3.4129	0.2518	13.5540	-2.9066	0.1445	20.1149
D20	-3.3827	0.1396	24.2314	-3.1227	0.0826	37.8051
D21	-3.6287	0.1851	19.6040	-3.2555	0.1081	30.1156
D22	-2.2444	0.1972	11.3813	-1.8044	0.1107	16.2999
D23	-2.8138	0.1307	21.5287	-2.4897	0.0743	33.5088
D24	-2.5891	0.1708	15.1587	-2.3145	0.1017	22.7581
D25	-3.1219	0.2633	11.8568	-2.6379	0.1525	17.2977
D26	-3.6711	0.2194	16.7325	-3.2024	0.1262	25.3756
D27	-2.2025	0.1630	13.5123	-1.8610	0.0935	19.9037
D28	-3.1486	0.1695	18.5758	-2.8451	0.1002	28.3942
D29	-2.2848	0.2030	11.2552	-1.8579	0.1152	16.1276
D30	-2.9719	0.1630	18.2325	-2.6442	0.0949	27.8630
D31	-3.3585	0.1565	21.4601	-2.9806	0.0895	33.3028
D32	-3.0610	0.1731	17.6834	-2.7748	0.1031	26.9137

Figure 5 . Logistic regression curve for dichotomous item D2 under GLM.

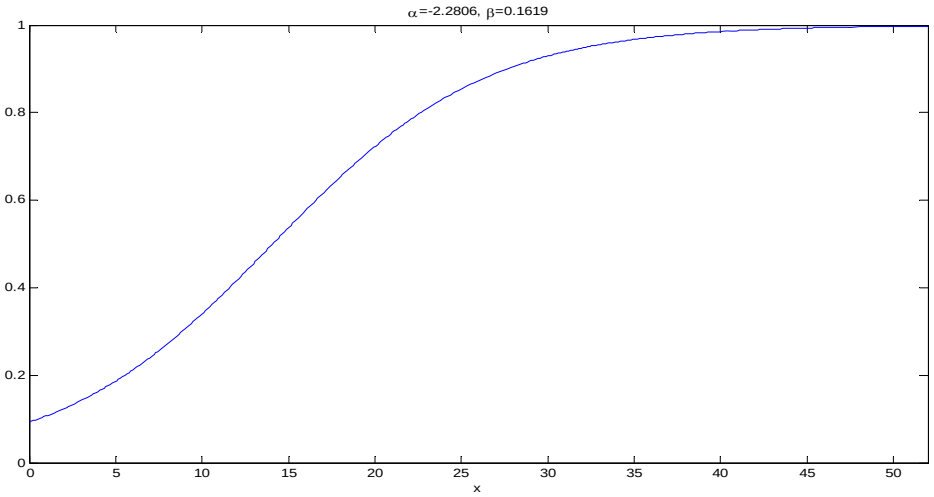


Figure 6. Logistic regression curve for dichotomous item D4 under GLM.

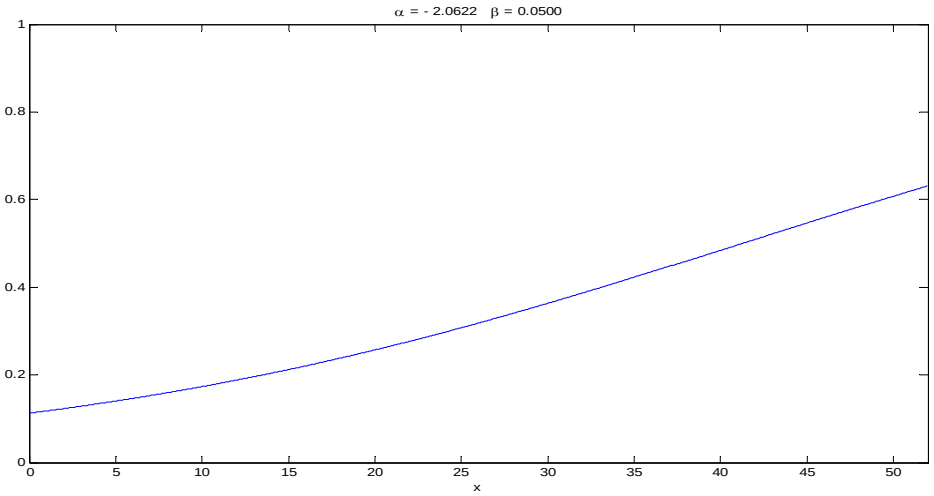


Figure 7. The pattern of item discrimination for dichotomous items under IRT based on subtest total score and entire-test total score.

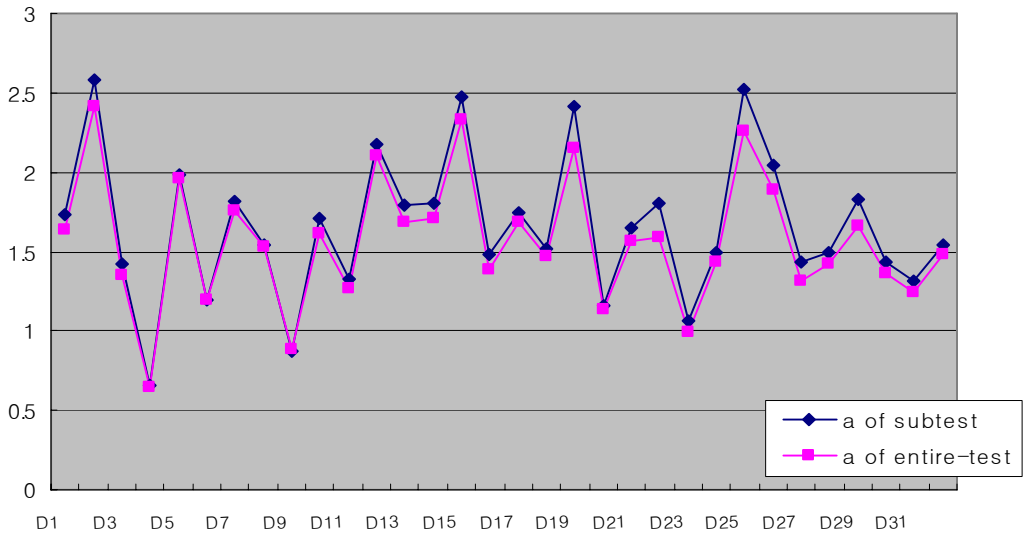
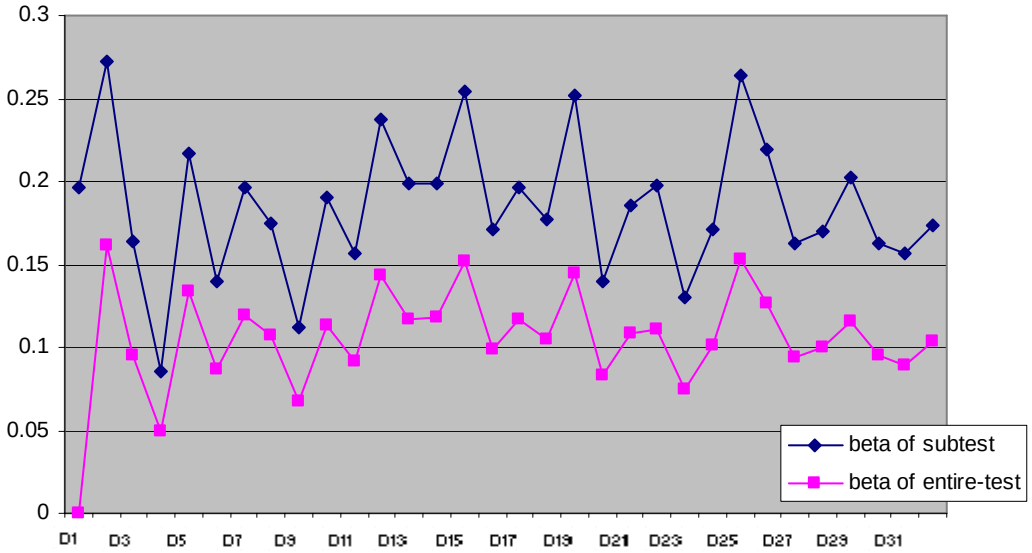


Figure 8. The pattern of item discrimination for dichotomous items under GLM based on subtest total score and entire-test total score.



For dichotomous items, the Pearson and the Spearman correlations between item difficulty indices were calculated for the values obtained under the three frameworks (see Table 6). In the case of IRT and GLM, lower values correspond to easier items, while under CTT, higher values represents easier ones. Therefore, negative correlations were shown. The absolute range of correlations was .979 to 1.0.

Table 6

The Pearson (Upper Triangle) and the Spearman Correlations (Lower Triangle) of Item Difficulty Indices for Dichotomous Items Between CTT, IRT, and GLM

		CTT		IRT		GLM	
		$p_{subtest}$	$p_{entire-test}$	$b_{separate}$	$b_{simul\,taneous}$	$-(\alpha/\beta)_{subtest}$	$-(\alpha/\beta)_{entire-test}$
CTT	$p_{subtest}$	_	1.0	-.979	-.981	-.989	-.988
	$p_{entire-test}$	1.0	_	-.979	-.981	-.989	-.988
IRT	$b_{separate}$	-.997	-.997	_	1.0	.998	.998
	$b_{simul\,taneous}$	-.998	-.998	1.0	_	.999	.999
GLM	$-(\alpha/\beta)_{subtest}$	-.998	-.998	1.0	1.0	_	1.0
	$-(\alpha/\beta)_{entire-test}$	-.998	-.998	1.0	1.0	1.0	_

The correlations between item discrimination indices also are presented in Table 7. The range of correlations was .882 to .998

Table 7

The Pearson (Upper Triangle) and the Spearman Correlations (Lower Triangle) of Item Discrimination Indices for Dichotomous Items Between CTT, IRT, and GLM

		CTT		IRT		GLM	
		$r_{bis/subtest}$	$r_{bis/entire-test}$	$a_{separate}$	$a_{simultaneous}$	$\beta_{subtest}$	$\beta_{entire-test}$
CTT	$r_{bis/subtest}$	_	.994	.882	.888	.887	.882
	$r_{bis/entire-test}$	.992	_	.891	.906	.896	.901
IRT	$a_{separate}$	.894	.900	_	.994	.998	.992
	$a_{simultaneous}$	.908	.922	.988	_	.993	.998
GLM	$\beta_{subtest}$	.883	.886	.989	.982	_	.995
	$\beta_{entire-test}$	.892	.906	.984	.997	.982	_

4.1.2. POLYTOMOUS ITEMS

To analyze polytomously-scored items under CTT, means and standard deviations were estimated, along with correlations between each item and the two criteria, the subtest-total score and the entire-test-total score. These analyses are presented in Table 8. If an item has a high mean, a small standard deviation, and low correlations with either of the criteria, the item can be said to have low discriminative power. In other hand, the item which has the average score of the mid range, relatively a large standard deviation, and correlates highly with the criteria can be regarded to be a good item in terms of effectively separating examinees who vary in the degree

of knowledge tested. Based on this standard, item P5 was considered to have high discriminating power.

Table 8

Item Statistics for Polytomously-scored Items Based on CTT Framework

Item	Mean	Standard Deviation	Pearson Correlation Coefficient	
			vs. Subtest total	vs. Entire-test total
P1	0.79	0.978	.55	.48
P2	0.59	0.797	.75	.75
P3	1.36	0.933	.63	.57
P4	1.63	0.776	.57	.52
P5	0.93	0.826	.82	.80
P6	1.16	0.930	.75	.71
P7	0.29	0.704	.52	.49
P8	0.74	0.965	.74	.72
P9	0.50	0.717	.69	.68
P10	1.10	0.842	.75	.73

For item analysis using IRT and GLM, items P1, P3, P7, and P8 were treated to have only two categories, because no examinee practically had partial credit. That is, there were just two responses (correct and wrong) on those items. For those four items, only one item difficulty index - b_1 under IRT, and $-\alpha_i/\beta_i$ under GLM – was examined.

Similar to the analysis of dichotomously-scored items, polytomous items under IRT were separately calibrated with GR model, and simultaneously calibrated with 2PL and GR models

(see Table 9). Under two different criteria, the most difficult item was P7, and the easiest items were P3, P4, and P10. Item P1 had the lowest value in terms of discriminative power, while P5 exhibited the highest value for discrimination.

For polytomous items under GLM, the parameters β and $-\alpha/\beta$ were estimated (see Table 10). The item which had highest value for the parameter $-\alpha/\beta$ under both of criteria was P7. In other words, P7 was considered to be the most difficult item for examinees. In terms of item discrimination, item P8 had the largest discrimination value under the criterion of subtest-total score, while under entire-test-total score, P5 got the highest value. The difference, however, was not large. The item which had the lowest discriminative power was P1 under two criteria.

Thus, the results of item analysis for polytomously-scored items under IRT and GLM were considered as moderately similar. The ICC and logistic regression curve for item P5 which effectively discriminated examinees are presented in Figure 9 to 10.

Table 9

Item Statistics for Polytomously-scored Items from Separate (GR) and Simultaneous (2PL & GR) Calibrations Based on IRT Framework

Item	Separate Calibration			Simultaneous Calibration		
	a_1	b_1	b_2	a_1	b_1	b_2
P1	1.12	0.48		1.00	0.54	
P2	3.25	0.32	0.96	2.54	0.32	1.10
P3	1.83	-0.64		1.80	-0.64	
P4	2.27	-1.14	-1.13	2.31	-1.08	-1.07
P5	3.48	-0.33	0.57	2.75	-0.39	0.62
P6	2.60	-0.39	-0.05	2.40	-0.42	-0.09
P7	2.02	1.38		1.76	1.56	
P8	3.14	0.40		2.57	0.39	
P9	2.39	0.39	1.35	2.13	0.39	1.52
P10	2.37	-0.59	0.31	2.18	-0.61	0.31

Table 10

Item Statistics for Polytomously-scored Item Based on GLM Framework

	Criterion									
	Subtest total score					Entire-test total score				
Item	α_{1i}	α_{2i}	β_i	$-\alpha_{1i}/\beta_i$	$-\alpha_{2i}/\beta_i$	α_{1i}	α_{2i}	β_i	$-\alpha_{1i}/\beta_i$	$-\alpha_{2i}/\beta_i$
P1	-2.8783		0.2496	11.5317		-2.7962		0.0848	32.9741	
P2	-5.6258	-7.7995	0.5085	11.0635	15.3382	-5.9298	-7.9890	0.1907	31.0949	41.8930
P3	-2.0909		0.3936	5.3122		-2.3426		0.1375	17.0371	
P4	-1.4090	-1.4463	0.5353	2.6321	2.7018	-1.9992	-2.0290	0.1839	10.8711	11.0332
P5	-3.5351	-6.8332	0.5312	6.6549	12.8637	-3.9798	-6.9528	0.1950	20.4092	35.6554
P6	-2.8829	-3.8502	0.4474	6.4437	8.6057	-3.2909	-4.1409	0.1651	19.9328	25.0812
P7	-7.5866		0.4494	16.8816		-7.1837		0.1553	46.2569	
P8	-6.2878		0.5437	11.5648		-6.0303		0.1861	32.4036	
P9	-4.6250	-7.0330	0.4042	11.4424	17.3998	-4.9304	-7.2365	0.1539	32.0364	47.0208
P10	-2.0833	-4.3870	0.3985	5.2279	11.0088	-2.6356	-4.7700	0.1535	17.1700	31.0749

Figure 9. Item characteristic curve for polytomous item P5 under IRT.

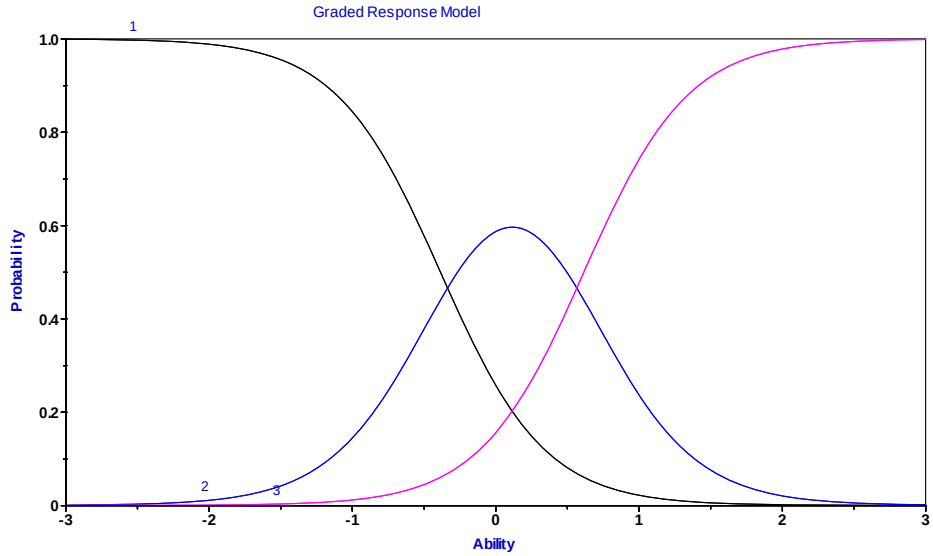
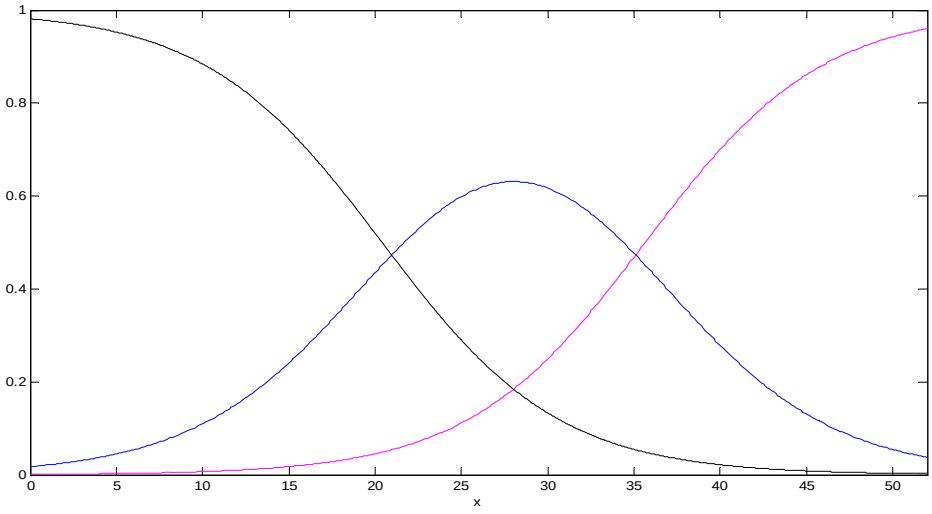


Figure 10. Logistic regression curve for polytomous item P5 under GLM.



4.2 DIFFERENTIAL ITEM FUNCTIONING (DIF)

DIF analyses were conducted to determine whether items function unfavorably across gender. Males were identified as the focal group, and females were regarded as the reference group. For easier understanding the results of DIF, the *p*-values and r_{bis} for dichotomous items, and means and correlations between each item and the entire-test total score for polytomous items were presented in terms of gender and the degree of urbanization in Table 11 and Table 12.

MH statistics for CTT framework were applied. Through SAS, the Cochran-Mantel-Haenszel (CMH) statistic, and the value and 95% confidence interval of common odds ratio were estimated for each item. For items favoring the focal group, common odds ratios of the items take values between zero to one, while for items functioning advantageously to the reference group, the range of common odds ratios is one to infinity (Clauser & Mazor, 1998).

Some items, however, did not yield the common odds ratio, because the partial contingency table contained cells with no case present. For polytomous items with more than two categories, the reported DIF values were the Mantel test statistics (i.e., 'Row Mean Scores Differ' in the SAS output). The degree of freedom of the Mantel test statistics was 1. Without loss of generality, CMH was used for both dichotomous and polytomous items in the subsequent presentation. Since CMH has a large-sample chi-squared null distribution with $df = 1$, the null hypothesis was examined using a critical value of 3.84 at $\alpha = .05$. 11 items were identified as DIF items.

To determine whether any of gender groups has advantage of answering each item under IRT, the computer program IRTLRDIF was used. A total of 22 items were identified as statistically significant DIF items (11 uniform DIF items and 11 nonuniform DIF items). The significance of the likelihood ratio statistic G^2 was determined using the .05 critical value from a chi-square distribution with one degree of freedom. Typically, the degrees of freedom for each item is two

for dichotomous items, or three for polytomous ones. However, in this case, one degree of freedom was used in accordance with Thissen (2001), because any problematic items in terms of DIF can be more easily identified with smaller degrees of freedom.

To check for DIF under GLM, the logistic regression equation noted in Swanminathan and Rogers (1990) was applied. Four parameter estimates of the equation were calculated using SAS: the intercept parameter, the parameter for the performance difference based on total score as the representative of the matching ability, the parameter for the performance difference across groups, and the parameter for the interaction effect between group and ability. Two uniform DIF and 12 nonuniform DIF items were yield. For detecting either uniform or nonuniform DIF, the critical value under the chi-square distribution with one degree of freedom at $\alpha = .05$ was used like the analysis under IRT.

DIF analyses were also conducted to determine whether examinees' responses differed according to the region (i.e., the degree of urbanization) where they live. The focal group was 'city', and the reference group was 'rural'. The methods which were used for DIF in this case, such as the critical value and the computer programs, were the same as described above. Using the MH approach, 19 items were checked as statistically significant DIF items. The IRTLRDIF program yielded 18 uniform DIF and 9 nonuniform DIF items which were statistically significant. Using the logistic regression model, three uniform DIF and nine nonuniform DIF items were identified as statistically significant.

The results of the DIF analyses for both gender and the degree of urbanization under CTT, IRT, and GLM are presented in Tables 13 to 18. And the ICC and the logistic regression curve for city and rural examinees in item D6 are presented in Figure 11 and Figure 12.

Table 11

The p-values and r_{bis} for Dichotomous Items in terms of Gender and the Degree of Urbanization

Item	Gender				The degree of urbanization			
	Male		Female		City		Rural	
	p	r_{bis}	*p*	r_{bis}	*p*	r_{bis}	*p*	r_{bis}
D1	.77	.61	.83	.64	.83	.71	.77	.55
D2	.70	.73	.79	.77	.79	.83	.71	.69
D3	.48	.64	.56	.66	.58	.73	.46	.55
D4	.33	.27	.34	.53	.40	.42	.28	.33
D5	.54	.74	.59	.81	.63	.83	.51	.71
D6	.33	.62	.35	.60	.39	.72	.28	.41
D7	.54	.71	.66	.72	.65	.79	.55	.64
D8	.40	.64	.52	.76	.53	.77	.39	.62
D9	.49	.50	.57	.50	.55	.56	.51	.44
D10	.64	.66	.75	.65	.71	.77	.69	.56
D11	.57	.59	.66	.63	.62	.68	.61	.55
D12	.71	.70	.80	.72	.78	.80	.73	.64
D13	.61	.66	.69	.74	.69	.79	.61	.61
D14	.52	.69	.62	.76	.63	.81	.51	.63
D15	.46	.85	.58	.81	.56	.90	.48	.76
D16	.52	.68	.59	.64	.59	.66	.52	.67
D17	.38	.74	.48	.75	.48	.79	.38	.68

Table 11 (Continued)

The p-values and r_{bis} for Dichotomous Items in terms of Gender and the Degree of Urbanization

Item	Gender				The degree of urbanization			
	Male		Female		City		Rural	
	p	r_{bis}	*p*	r_{bis}	*p*	r_{bis}	*p*	r_{bis}
D18	.42	.69	.48	.71	.51	.72	.39	.66
D19	.56	.74	.68	.82	.66	.84	.58	.73
D20	.28	.56	.37	.60	.38	.69	.27	.43
D21	.39	.66	.47	.77	.50	.80	.36	.59
D22	.62	.66	.74	.66	.69	.75	.67	.58
D23	.35	.57	.44	.53	.41	.59	.38	.51
D24	.52	.70	.60	.63	.60	.74	.52	.59
D25	.61	.75	.74	.78	.69	.82	.67	.74
D26	.46	.78	.57	.74	.56	.82	.47	.70
D27	.54	.62	68	.61	.59	.72	.63	.55
D28	.42	.63	.50	.73	.52	.74	.40	.60
D29	.62	.67	.75	.67	.66	.75	.71	.63
D30	.43	.63	.51	.67	.51	.68	.43	.62
D31	.32	.65	.45	.60	.41	.72	.36	.52
D32	.44	.63	.53	.74	.52	.75	.46	.62

Table 12

The Mean and the Pearson Correlation Between Each Item and Entire- test Total Score for Polytomous Items in terms of Gender and the Degree of Urbanization

	Gender				The degree of urbanization			
	Male		Female		City		Rural	
Item	Mean	*r*	Mean	*r*	Mean	*r*	Mean	*r*
P1	.78	.48	.79	.49	.80	.52	.78	.44
P2	.50	.75	.67	.74	.75	.77	.42	.69
P23	1.27	.60	1.45	.53	.1.39	.59	1.33	.56
P4	1.50	.53	1.76	.47	1.62	.56	1.63	.48
P5	.81	.80	1.05	.78	1.04	.81	.81	.77
P6	.99	.73	1.33	.66	1.21	.75	1.11	.66
P7	.23	.48	.35	.50	.38	.55	.20	.38
P8	.67	.73	.80	.71	.94	.76	.53	.65
P9	.40	.68	.61	.67	.60	.70	.40	.64
P10	.97	.76	1.23	.68	1.19	.76	1.01	.68

Table 13

Differential Item Functioning for Gender Based on CTT Framework (Mantel-Haenszel Statistic)

Item	CMH	Common odds ratio	95% CI		(Who does item favor)
Dichotomous					
D1	0.2876	0.9264	0.7008	1.2248	
D2	0.1450	0.9484	0.7203	1.2488	
D3	0.7035	0.9081	0.7254	1.1369	
D4	2.5821	0.8390	0.6765	1.0407	
D5	10.9109	0.6499	0.5037	0.8387	DIF (Male)
D6	5.1837	0.7612	0.6003	0.9654	DIF (Male)
D7	0.3184	1.0703	0.8450	1.3557	
D8	2.6504	1.2110	0.9610	1.5262	
D9	0.0502	0.9761	0.7910	1.2046	
D10	1.0627	1.1380	0.8905	1.4544	
D11	0.0226	1.0172	0.8129	1.2729	
D12	0.2250	0.9351	0.7095	1.2324	
D13	0.3682	0.9280	0.7287	1.1818	
D14	0.0390	0.9766	0.7713	1.2366	
D15	0.0042	1.0086	0.7797	1.3045	
D16	2.7025	0.8277	0.6603	1.0376	
D17	0.0226	1.0187	0.7997	1.2977	
D18	5.9423	0.7452	0.5880	0.9444	DIF (Male)
D19	1.1185	1.1475	0.8895	1.4804	
D20	1.8481	1.1791	0.9296	1.4956	
D21	0.7463	0.8966	0.7010	1.1467	
D22	2.8776	1.2293	0.9679	1.5614	

Note. *CMH* = Cohan-Mantel-Haenszel Test. *CI* = Confidence Interval.

Table (continued)

Differential Item Functioning for Gender Based on CTT Framework (Mantel-Haenszel Statistic)

Item	CMH	Common odds ratio	95% CI		(Who does item favor)
Dichotomous					
D23	0.3498	1.0686	0.8577	1.3315	
D24	0.7933	0.9014	0.7173	1.1327	
D25	1.7214	1.1966	0.9180	1.5599	
D26	0.5249	1.0949	0.8564	1.3998	
D27	6.6137	1.3419	1.0714	1.6807	DIF (Female)
D28	0.0449	0.9749	0.7713	1.2324	
D29	4.1693	1.2896	1.0098	1.6469	DIF (Female)
D30	0.0031	0.9935	0.7901	1.2493	
D31	5.0260	1.2988	1.0336	1.6319	DIF (Female)
D32	0.0736	1.0331	0.8171	1.3064	
Polytomous					
P1	18.9352	0.6038	0.4805	0.7586	DIF (Male)
P2	0.7439				
P3	1.7887	0.8403	0.6492	1.0876	
P4	12.3556				DIF
P5	0.2039				
P6	12.0549				DIF
P7	0.1365	1.0694	0.7497	1.5255	
P8	8.1999	0.6533	0.4865	0.8773	DIF (Male)
P9	6.5893				DIF
P10	1.9935				

Table 14

Differential Item Functioning for Gender Based on IRT Framework

Item	G^2			
	H_0 : all equal	H_0 : a equal	H_0 : b equal	
Dichotomous				
D1	0.2			
D2	0.3			
D3	1.0			
D4	28.6	25.1	3.5	Non-uniform
D5	16.5	3.7	12.8	Uniform
D6	9.2	0.0	9.2	Uniform
D7	0.5			
D8	11.5	10.0	1.5	Non-uniform
D9	0.1			
D10	1.1			
D11	1.6			
D12	0.4			
D13	5.8	4.6	1.2	Non-uniform
D14	4.9	4.6	0.3	Non-uniform
D15	1.4			
D16	3.4			
D17	0.7			
D18	6.4	0.5	5.9	Uniform
D19	5.4	5.0	0.5	Non-uniform
D20	1.9			
D21	10.9	9.3	1.7	Non-uniform
D22	2.0			

Table (continued)

Differential Item Functioning for Gender Based on IRT Framework

Item	G^2			
	H_0 : all equal	H_0 : a equal	H_0 : b equal	
Dichotomous				
D23	0.5			
D24	5.1	4.1	0.9	Non-uniform
D25	2.3			
D26	0.8			
D27	7.6	0.0	7.6	Uniform
D28	7.3	6.2	1.0	Non-uniform
D29	4.4	0.0	4.4	Uniform
D30	1.2			
D31	5.8	0.5	5.2	Uniform
D32	9.7	9.3	0.4	Non-uniform
Polytomous				
P1	11.4	0.0	11.4	Uniform
P2	1.8			
P3	5.4	4.9	0.5	Non-uniform
P4	14.5	1.6	12.9	Uniform
P5	1.7			
P6	17.1	2.8	14.3	Uniform
P7	0.1			
P8	10.7	0.3	10.4	Uniform
P9	19.0	1.1	18.0	Uniform
P10	17.5	13.9	3.6	Non-uniform

Table 15

Differential Item Functioning for Gender Based on GLM Framework

	Ability (Total Score)		Group		Interaction		
Item	Estimate	Wald	Estimate	Wald	Estimate	Wald	
Dichotomous							
D1	0.1134	23.8407	-0.0970	0.1100	0.0029	0.0381	
D2	0.1516	30.6231	-0.1742	0.2623	0.0074	0.1699	
D3	0.0875	32.6823	-0.2499	0.8412	0.0059	0.3605	
D4	-0.0102	0.6710	-1.4595	28.3667	0.0427	25.7453	Non-uniform
D5	0.1037	27.9234	-0.9621	8.7121	0.0230	3.1767	Uniform
D6	0.0827	31.0898	-0.4139	1.7378	0.0029	0.0902	
D7	0.1093	36.5633	-0.0464	0.0258	0.0061	0.2744	
D8	0.0569	13.4934	-0.7643	6.2390	0.0339	10.5383	Non-uniform
D9	0.0668	25.3167	0.0043	0.0003	0.0002	0.0006	
D10	0.1168	36.6724	0.1944	0.4982	-0.0034	0.0752	
D11	0.0765	24.4220	-0.2101	0.6689	0.0101	1.0178	
D12	0.1533	35.8868	0.0543	0.0298	-0.0061	0.1415	
D13	0.0832	20.4554	-0.6359	4.8745	0.0240	3.8547	Non-uniform
D14	0.0829	22.7023	-0.6287	4.4816	0.0244	4.4570	Non-uniform
D15	0.1700	60.2560	0.3055	0.7343	-0.0123	0.8244	
D16	0.1144	49.1872	0.0703	0.0676	-0.0097	0.9130	
D17	0.1016	36.2166	-0.2956	0.8151	0.0104	0.9052	
D18	0.0896	31.8356	-0.5857	3.6883	0.0114	1.2354	
D19	0.0950	20.8425	-0.6237	3.5543	0.0341	5.7213	Non-uniform
D20	0.0621	18.1028	-0.2901	0.8663	0.0133	2.0144	
D21	0.0604	14.7346	-1.0723	11.0973	0.0335	10.0278	Non-uniform
D22	0.1097	35.1702	0.1814	0.4410	-0.0002	0.0002	

Table (continued)

Differential Item Functioning for Gender Based on GLM Framework

	Ability (Total Score)		Gender		Interaction		
Item	Estimate	Wald	Estimate	Wald	Estimate	Wald	
Dichotomous							
D23	0.0788	32.5972	0.1337	0.2415	-0.0032	0.1304	
D24	0.1292	58.2801	0.3120	1.3118	-0.0175	2.8331	
D25	0.1284	30.2512	-0.1480	0.2016	0.0158	1.0355	
D26	0.1358	53.5818	0.1779	0.3191	-0.0065	0.3074	
D27	0.0875	31.7872	0.2471	0.9221	0.0028	0.0789	
D28	0.0609	16.1900	-0.8472	8.1645	0.0275	7.4545	Non-uniform
D29	0.1094	33.4394	0.2073	0.5557	0.0027	0.0480	
D30	0.0780	27.2993	-0.3520	1.5629	0.0116	1.4495	
D31	0.0946	39.8378	0.3927	1.7509	-0.0042	0.2033	
D32	0.0586	14.5836	-0.8667	8.6578	0.0309	9.0210	Non-uniform
Polytomous							
P1	0.0758	27.2940	-0.6734	5.1942	0.0084	0.8004	Uniform
P2	0.1801	82.2033	-0.3449	0.6234	0.0074	0.3509	
P3	0.1943	64.6525	0.5771	3.5109	-0.0360	6.1981	Non-uniform
P4	0.1380	17.6804	0.0565	0.0222	0.0307	1.7602	
P5	0.2036	144.1253	0.2344	0.6236	-0.0060	0.3416	
P6	0.2090	97.4249	1.0852	11.8039	-0.0301	5.4555	Non-uniform
P7	0.1332	19.3260	-0.5332	0.4576	0.0145	0.5689	
P8	0.1833	47.9036	-0.6014	1.2395	0.0045	0.0701	
P9	0.1922	108.1358	1.1722	9.0707	-0.0250	5.2530	Non-uniform
P10	0.2040	153.8732	0.9872	14.4592	-0.0335	11.6658	Non-uniform

Table 16

Differential Item Functioning for the Degree of Urbanization Based on CTT Framework (Mantel-Haenszel Statistic)

Item	CMH	Common odds ratio	95% CI		(Who does item favor)
Dichotomous					
D1	4.6238	0.7354	0.5567	0.9716	DIF (City)
D2	8.9305	0.6583	0.4990	0.8686	DIF (City)
D3	3.8263	0.8006	0.6411	0.9998	DIF (City)
D4	6.0113	0.7624	0.6138	0.9469	DIF (City)
D5	5.9425	0.7382	0.5779	0.9429	DIF (City)
D6	0.4844	0.9189	0.7245	1.1654	
D7	3.5661	0.7987	0.6315	1.0103	
D8	4.6451	0.7763	0.6162	0.9780	DIF (City)
D9	0.5399	1.0823	0.8770	1.3357	
D10	3.2129	1.2457	0.9790	1.5851	
D11	2.4769	1.1966	0.9561	1.4976	
D12	3.2605	0.7767	0.5908	1.0211	
D13	1.3912	0.8658	0.6819	1.0994	
D14	4.5618	0.7746	0.6127	0.9792	DIF (City)
D15	2.6100	1.2390	0.9571	1.6040	
D16	0.0030	0.9937	0.7933	1.2447	
D17	0.0386	1.0247	0.8038	1.3063	
D18	2.6349	0.8233	0.6519	1.0398	
D19	0.3456	0.9272	0.7209	1.1926	
D20	0.0012	0.9958	0.7853	1.2627	
D21	1.3112	0.8679	0.6819	1.1046	
D22	4.0517	1.2720	1.0055	1.6091	DIF (Rural)

Note. *CMH* = Cohan-Mantel-Haenszel Test. *CI* = Confidence Interval.

Table (continued)

Differential Item Functioning for the Degree of Urbanization Based on CTT Framework (Mantel-Haenszel Statistic)

Item	CMH	Common odds ratio	95% CI		(Who does item favor)
Dichotomous					
D23	4.9519	1.2931	1.0330	1.6188	DIF (Rural)
D24	0.0701	0.9700	0.7739	1.2158	
D25	7.2548	1.4265	1.1000	1.8498	DIF (Rural)
D26	0.0660	1.0323	0.8092	1.3170	
D27	26.5484	1.8204	1.4466	2.2907	DIF (Rural)
D28	0.2453	0.9434	0.7488	1.1884	
D29	26.8757	1.8978	1.4850	2.4253	DIF (Rural)
D30	1.2028	1.1367	0.9046	1.4284	
D31	4.9284	1.2992	1.0307	1.6377	DIF (Rural)
D32	3.8982	1.2643	1.0014	1.5962	DIF (Rural)
Polytomous					
P1	9.8790	1.4390	1.1456	1.8074	DIF (Rural)
P2	17.1712				DIF
P3	0.1438	1.0509	0.8125	1.3592	
P4	0.1830				
P5	4.6308				DIF
P6	4.3042				DIF
P7	0.1025	1.0632	0.7333	1.5414	
P8	22.3065	0.4955	0.3690	0.6655	DIF (City)
P9	1.0726				
P10	0.5945				

Table 17

Differential Item Functioning for the Degree of Urbanization Based on IRT Framework

Item	G^2			
	H_0 : all equal	H_0 : *a* equal	H_0 : *b* equal	
Dichotomous				
D1	4.6	0.2	4.3	Uniform
D2	8.8	0.7	8.1	Uniform
D3	4.9	1.4	3.5	-
D4	8.4	0.0	8.3	Uniform
D5	5.3	0.1	5.2	Uniform
D6	14.9	13.2	1.7	Non-uniform
D7	3.0			
D8	5.7	0.2	5.5	Uniform
D9	1.1			
D10	5.8	2.2	3.6	-
D11	6.1	0.6	5.5	Uniform
D12	1.8			
D13	2.0			
D14	7.0	2.5	4.4	Uniform
D15	2.9			
D16	14.9	14.7	0.1	Non-uniform
D17	0.4			
D18	5.2	3.0	2.2	-
D19	1.1			
D20	6.3	5.5	0.8	Non-uniform
D21	7.7	4.4	3.3	Non-uniform
D22	3.8			

Table (continued)

Differential Item Functioning for the Degree of Urbanization Based on IRT Framework

Item	G^2			
	H_0 : all equal	H_0 : *a* equal	H_0 : *b* equal	
Dichotomous				
D23	4.9	1.1	3.8	Uniform
D24	0.1			
D25	12.5	6.3	6.2	Non-uniform
D26	0.1			
D27	30.7	0.0	30.7	Uniform
D28	1.3			
D29	32.8	2.4	30.5	Uniform
D30	2.4			
D31	7.4	1.6	5.8	Uniform
D32	2.4			
Polytomous				
P1	12.2	0.1	12.1	Uniform
P2	13.7	4.7	9.0	Non-uniform
P3	9.1	7.9	1.2	Non-uniform
P4	4.7	1.4	3.3	-
P5	6.1	5.3	0.8	Non-uniform
P6	6.7	0.1	6.7	Uniform
P7	1.1			
P8	19.8	0.0	19.8	Uniform
P9	19.4	7.0	12.4	Non-uniform
P10	5.8	1.3	4.5	Uniform

Table 18

Differential Item Functioning for the Degree of Urbanization Based on GLM Framework

	Ability (Total score)		Group		Interaction		
Item	Estimate	Wald	Estimate	Wald	Estimate	Wald	
Dichotomous							
D1	0.1450	36.1024	0.0552	0.0372	-0.0180	1.4497	
D2	0.2053	45.9269	0.0862	0.0662	-0.0270	2.2081	
D3	0.1225	64.3207	0.2353	0.7627	-0.0191	3.8429	Non-uniform
D4	0.0516	18.0226	-0.2649	1.0469	-0.0026	0.0975	
D5	0.1461	55.1585	-0.1085	0.1211	-0.0089	0.5058	
D6	0.1452	90.9416	1.0556	11.3472	-0.0427	19.7737	Non-uniform
D7	0.1406	57.7933	0.1045	0.1334	-0.0149	1.6248	
D8	0.1242	60.9102	0.0212	0.0050	-0.0129	1.5732	
D9	0.0740	33.3812	0.2041	0.7260	-0.0045	0.2779	
D10	0.1485	58.1042	0.7050	6.6050	-0.0240	3.8028	Uniform
D11	0.0942	38.7151	0.2733	1.1404	-0.0010	0.0106	
D12	0.1727	43.2633	0.1811	0.3435	-0.0190	1.3716	
D13	0.1473	58.8521	0.2864	1.0498	-0.0204	2.8832	
D14	0.1556	70.8549	0.3120	1.1539	-0.0255	4.9409	Non-uniform
D15	0.1775	67.0634	0.5707	2.5146	-0.0171	1.5858	
D16	0.0567	14.5115	-0.7123	6.7990	0.0301	8.4128	Non-uniform
D17	0.1225	53.7161	0.0492	0.0226	-0.0041	0.1419	
D18	0.0904	35.2773	-0.4737	2.4822	0.0092	0.8027	
D19	0.1453	45.6392	-0.1114	0.1225	-0.0007	0.0022	
D20	0.1283	73.7144	0.8149	6.7171	-0.0325	11.5734	Non-uniform
D21	0.1503	82.8819	0.5464	3.0591	-0.0297	8.1538	Non-uniform
D22	0.1267	47.7123	0.4366	2.5880	-0.0107	0.8006	

Table (continued)

Differential Item Functioning for the Degree of Urbanization Based on GLM Framework

	Ability (Total score)		Gender		Interaction		
Item	Estimate	Wald	Estimate	Wald	Estimate	Wald	
Dichotomous							
D23	0.0749	31.1892	0.1578	0.3268	0.0005	0.0029	
D24	0.1184	55.2572	0.2192	0.6574	-0.0116	1.2783	
D25	0.1109	23.8323	-0.2644	0.6548	0.0300	3.7053	
D26	0.1351	56.3365	0.0873	0.0771	-0.0063	0.2974	
D27	0.1084	48.6637	0.7930	9.0803	-0.0079	0.5775	Uniform
D28	0.1112	53.5310	0.0414	0.0205	-0.0082	0.6891	
D29	0.1014	30.0469	0.4232	2.2422	0.0124	0.9352	
D30	0.0837	33.2689	-0.2213	0.6204	0.0079	0.6536	
D31	0.1253	68.3840	0.8921	8.4958	-0.0234	5.8945	Non-uniform
D32	0.1149	54.3108	0.3277	1.2842	-0.0077	0.5705	
Polytomous							
P1	0.0945	42.5032	0.5261	3.1471	-0.0044	0.2109	
P2	0.1638	75.1033	-1.0411	5.4174	0.0171	1.7777	Uniform
P3	0.0952	21.1605	-0.4508	2.1789	0.0301	4.4731	Non-uniform
P4	0.1593	22.9571	-0.0242	0.0045	0.0166	0.5679	
P5	0.1926	139.9172	-0.1849	0.3871	0.0012	0.0125	
P6	0.1800	78.3306	0.4491	2.0139	-0.0094	0.5363	
P7	0.2422	51.3170	2.2312	7.1596	-0.0585	8.3124	Non-uniform
P8	0.1832	51.8402	-0.7908	2.1568	0.0011	0.0046	
P9	0.1791	106.5350	0.5948	2.3617	-0.0171	2.4302	
P10	0.1743	128.0257	0.2992	1.3488	-0.0143	2.1671	

Figure 11. Item characteristic curve for the degree of urbanization in dichotomous item D6 under IRT.

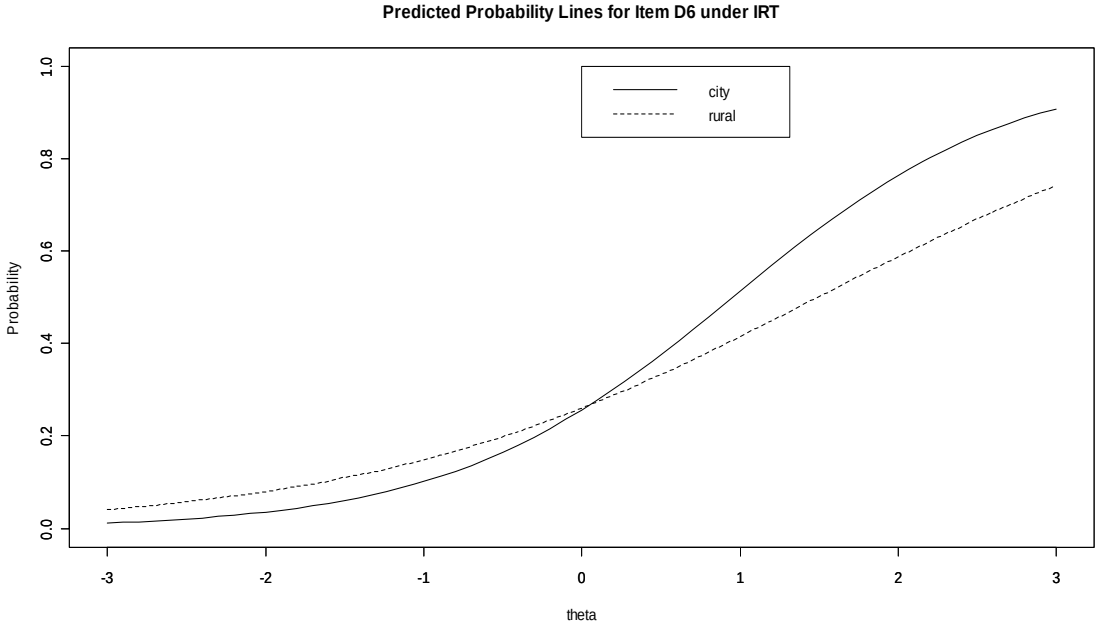
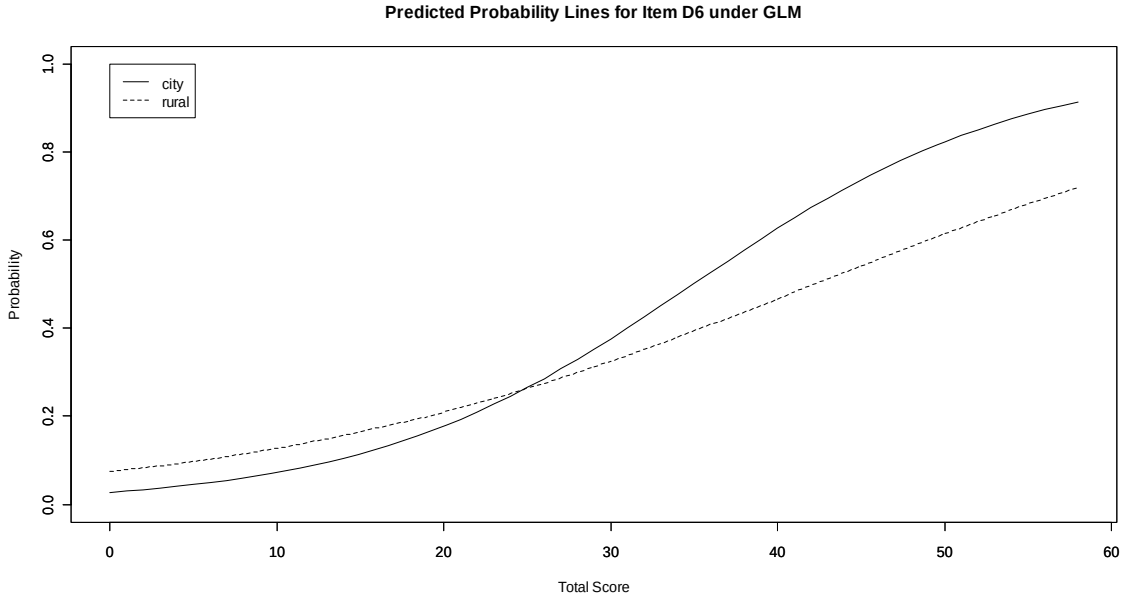


Figure 12. Logistic regression curve for the degree of urbanization in dichotomous item D6 under GLM.



CHAPTER 5

SUMMARY AND DISCUSSION

5.1 SUMMARY

The first objective of this study was to analyze each item of a test using CTT, IRT, and GLM frameworks and to compare the results. Examinees' responses on a nationwide English exam for the third grade students in Korean middle schools were examined. Dichotomous and polytomous items were separately analyzed, and the estimates of item difficulty and item discrimination as representative item parameters were calculated with two criteria, subtest total score and entire-test total score.

With regard to item difficulty estimates for dichotomously-scored items, under CTT, the same values of *p*-value were obtained with both of criteria, because the *p*-value based on the subtest total score and the entire-test total score were calculated by the same formula – the ratio of the number of examinees who got each item right to the total number of examinees. In the case of the item difficulty parameter under IRT and GLM, there was different in item difficulty levels based on subtest total score and entire-test total score. The difficulty patterns based on two criteria, however, were very similar across the items.

The high correlation range of .979 to 1.0 showed that three different frameworks yielded almost the same results of item difficulty estimates across all items. Particularly, the correlations of item difficulty index between IRT and GLM were extremely high: the Pearson correlation between two theoretical frameworks based on entire-test total score was .999, and the Spearman

correlations between IRT and GLM based on the two criteria were 1.0. This indicates that item difficulty indices, b_i under IRT, and $-\alpha/\beta$ under GLM function in the same way as the item location parameter.

In connection with the item discrimination parameter for dichotomous items, using CTT, r_{pbis}, r_{bis}, $D_{27\%}$, and $D_{50\%}$ yielded roughly similar discriminative patterns over items, although their relative magnitudes of item discrimination were different. Thus, it could be inferred that the difference in absolute values between item discrimination indices was created by the differences in the correlation formula, not by real differences in discriminating items.

Under IRT, the values of item discrimination based on separate and simultaneous calibration were very close, while there was a relatively large deviation between two criteria, using GLM (see Figure 7 and 8). Also the discriminative power based on the subtest was generally larger than on the entire test, because within the subtest, each item makes up a larger portion of the total score.

The discriminative patterns between CTT, IRT, and GLM were very similar, as is evident in the Figures. The correlations of item discrimination indices under the three frameworks also showed the same result. There was high correlation across all three frameworks, with a range was .882 to .998. The correlation between IRT and GLM was especially higher than those based on CTT.

The items which had extreme item difficulty and item discrimination values based on either of the two criteria were quite similarly selected from the three approaches used. The items which were difficult were D4, D6, and D20, while the easy items were D1, D2, and D12. The items which had good discriminative power were D2, D15, D19, and D25. In other hand, D4 and D9 were considered to have very low level of discrimination.

In terms of polytomously-scored items, the comparison of the results of item analysis was not easier than the case of dichotomously-scored items, because different methods such as the comparison of means and standard deviations were used for CTT approach. Under the three frameworks, the most difficult items were items P7 and P9, while items P3 and P4 were regarded as easy. Related to item discrimination, the three different approaches commonly identified that item P1 had the lowest value, while item P5 had the highest value. As in the case of dichotomous items, the same results were also obtained from CTT, IRT, and GLM in analyzing polytomous items.

The second objective of this study was to examine if there are any items which unfavorably function in different conditions such as gender and the degree of urbanization. In terms of gender, the male group had a lower probability of correct response across all items. Using MH approach, 11 items (D5, D6, D18, D27, D29, D31, P1, P4, P6, P8, P9) exhibited DIF. Nonuniform DIF items were not found through the MH approach as Clauser and Mazor (1998) indicated. Also, common odds ratios for some items could not be obtained, since some of the partial tables had empty cells due to a wide-range of distribution of total scores (i.e., 53 different total scores in this study).

Under IRT, the male group presented the mean of almost -0.35 with a standard deviation of 0.98, when the mean of female group set at 1 for each item. 22 items which were identified as statistically significant were listed: 11 uniform DIF items (D5, D6, D18, D27, D29, D31, P1, P4, P6, P8, P9), and 11 nonuniform DIF items (D4, D8, D13, D14, D19, D21, D24, D28, D32, P3, P10). The logistic regression model under GLM yielded two uniform DIF and 12 nonuniform DIF items. Uniform DIF items were items D5 and P1, and nonuniform DIF items were items D4, D8, D13, D14, D19, D21, D28, D32, P3, P6, P9, and p10.

The results of comparisons of DIF for gender between CTT, IRT, and GLM are presented in Table 19. Different detection methods did not identify identical items as exhibiting DIF. Common uniform DIF items from three approaches were two (D5, P1), while common nonuniform DIF items from IRT and GLM were 10 ones (D4, D8, D13, D14, D19, D21, D28, D32, P3, P10). There were only four items that were commonly identified as exhibiting DIF across the three theoretical frameworks: D5, P1, P6, and P9.

Item D5 was a question of listening part. The value of CMH was 10.9109, and the likelihood ratio statistic G^2 under IRT was 16.5. And the Wald statistics for group and interaction were 8.7121 and 3.1767, respectively. The CMH statistic of item P1 which was in listening sub-domain was 18.9352. The value of G^2 was 11.4, and the Wald statistics of group and interaction were 5.1942 and 0.8004, respectively. Items P6 and P9 were ones for assessing the ability of English reading. The value of CMH of item P6 was 12.0549, and G^2 statistic was 17.1. The values of the Wald test were 11.8039 and 5.4555 for group and interaction effect, respectively. The CMH statistic of P9 was 6.5893, and G^2 statistic was 19.0. Item P9 had the Wald test statistics for group and interaction, 9.0707 and 5.2530, respectively. Related to contents of these common DIF items, however, it was difficult to find out the obvious source of differential functioning over gender.

Differences in results obtained using the CTT, IRT, and GLM approaches were also found in the analyses for the degree of urbanization. The examinees who live in ‘cities’ showed a higher probability of getting each item right than examinees in ‘rural’ regions.

Under the MH statistic, 19 items were identified as DIF items: D1, D2, D3, D4, D5, D8, D14, D22, D23, D25, D27, D29, D31, D32, P1, P2, P5, P6, P8. Under IRT, across all items, the mean for examinees in the ‘city’ region was almost 0.41 with a standard deviation of 1.49, while the

mean for examinees in ‘rural’ locations was standardized as 1. The computer program IRTLRDIF for IRT framework found out 15 uniform and nine nonuniform DIF items. 15 uniform DIF items were D1, D2, D4, D5, D8, D11, D14, D23, D27, D29, D31, P1, P6, P8, and P10. The nonuniform DIF items were D6, D16, D20, D21, D25, P2, P3, P5, and P9. Using the logistic regression model, a total of 12 DIF items were identified: three uniform DIF items (D10, D27, P2), and nine nonuniform DIF items (D3, D6, D14, D16, D20, D21, D31, P3, P7).

There were just one common uniform DIF, item D27, through CTT, IRT, and GLM, while five nonuniform DIF items were checked out between IRT and GLM (D6, D16, D20, D21, P3). Table 19 presents the results of comparisons of DIF analyses between the three approaches over the degree of urbanization. The DIF items which were identified across all three frameworks were D14, D27, D31, and P2.

The CMH statistic of item D14 which assessed the ability of speaking was 4.5618. The value of G^2 was 7.0, and the Wald statistics of group and interaction were 1.1539 and 4.9409, respectively. Items D27, D31, and P2 belonged to the reading section. Item D27 had CMH statistic and G^2 statistic, 26.5484 and 30.7, respectively. And the Wald test statistics for gender and interaction effect of that item were 9.0803 and 0.5775. The value of CMH of item D31 was 4.9284, and the G^2 statistic was 7.4. And the Wald statistics of that item were 8.4958 and 5.8945 for group and interaction effect. Item P2 got the CMH statistic, 17.1712. The value of G^2 statistic was 13.7, and the Wald test statistics for group and interaction effect were 5.4174 and 1.7777, respectively.

Unlikely common DIF items which did not present the obvious content sources for differential functioning, item D3 of which the content was how to deal with a state of emergency in elevator showed the possibility that examinees who live in rural region might have

disadvantage, because they might get less experience related to elevator in their lives than examinees in city.

Based on the results of the DIF analyses, many items appeared to put males and examinees who live in ‘rural’ regions at a disadvantage. And it was noted that the IRT-LR procedure for IRT was very powerful in detecting DIF. And the logistic regression was very sensitive to nonuniform DIF due to inclusion of the interaction term between group membership and ability. Also there were substantial differences across methods in terms of identifying items that exhibit DIF unlike the results of item analysis.

Table 19

The Comparisons of DIF for Gender Between CTT, IRT, and GLM

		CTT	
		DIF	NON-DIF
IRT	DIF	11 items (D5, D6, D18, D27, D29, D31, P1, P4, P6, P8, P9)	11 items
	NON-DIF	0 items	20 items

		CTT	
		DIF	NON-DIF
GLM	DIF	4 items (D5, P1, P6, P9)	10 items
	NON-DIF	7 items	21 items

		IRT	
		DIF	NON-DIF
GLM	DIF	14 items (D4, D5, D8, D13, D14, D19, D21, D28, D32, P1, P3, P6, P9, P10)	0 items
	NON-DIF	8 items	20 items

Table 20

The Comparisons of DIF for the Degree of Urbanization Between CTT, IRT, and GLM

		CTT	
		DIF	NON-DIF
IRT	DIF	16 items (D1, D2, D4, D5, D8, D14, D23, D25, D27, D29, D31, P1, P2, P5, P6, P8)	8 items
	NON-DIF	3 items	15 items

		CTT	
		DIF	NON-DIF
GLM	DIF	5 items (D3, D14, D27, D31, P2)	7 items
	NON-DIF	14 items	16 items

		IRT	
		DIF	NON-DIF
GLM	DIF	10 items (D6, D10, D14, D16, D20, D21, D27, D31, P2, P3)	2 items
	NON-DIF	14 items	16 items

5.2 DISCUSSION

Comparing item parameters under different theoretical frameworks such as CTT and IRT has been a major topic of researches related to item analysis. The results of this study were similar to previous studies on item analysis.

Macdonald and Paunonen (2002) suggested that item difficulty indices generated through CTT and IRT are very comparable, and the item difficult indices in this study behaved very similarly under the three different approaches. This result was also supported by very high correlations between item difficulty indices. Item discrimination indices from CTT, IRT, and GLM, also showed similar patterns, and there were high correlations between them.

Thus, item difficulty and item discrimination parameters under CTT, IRT, and GLM might be interchangeable under general conditions. Especially, it can be said that the alternative possibility between item difficulty parameters are very larger than that of item discrimination, since the correlations between item difficulty indices were higher than ones between item discrimination parameters, as the study of Lawson (1991) indicated.

Item statistics generated in this study also informed us of incorrect performance on each item as well as correct performance. In revising items appropriately and in selecting proper items, it is very essential to find out the inclination and reasons of incorrect performance, as Wainer (1989) indicated that one of many purposes of item analysis is to detect performance anomalies. Although incorrect performances of all items are not analyzed in this context, take item D4 which was in listening section and had the lowest discriminative power among dichotomous items as example. The item D4 is presented in Figure 13.

Using item statistics under CTT, *p*-value and r_{bis} of item D4 were the same as .34, and $D_{27\%}$ was .36. The proportions that all examinees chose each distracter was 24(A), .19(B), 14(C),

34(D), and .07(E). And the proportions that examinees in 27% upper group selected each decoy was .21(A), .12(B), .07(C), .55(D), and .04(E), while the proportions for 27% lower group were .20(A), .26(B), .20(C), .20(D), and .09(E).

It could be checked that the distracter A was problematic, because the proportions that both of examinees with high ability and those with low ability selected the distracter A were almost the same, although distracters which are much more chosen by the lower-scoring examinees than the higher-scoring ones are generally considered desirable. The reason might be that the distracter A included both of 'February', and '20th' represented in the listening material, and it might cause examinees to be confused. Thus, in revision of this item, item D4 can have higher discriminative power and proper difficulty level by changing distracter A such as May 20th or October 12th.

Figure 13. The question and the listening material of dichotomous item D4.

Item D4:

Q : Select proper answer after listening conversation

A. February 20th B. April 20th C. June 12th D. August 20th E. September 12th

Listening Material:

G: When is your birthday, Min-ho?

B: August 20th. When is yours?

G: February 12th . What's a good message for a birthday card?

B: I don't know. "Happy birthday," I guess.

G: Sounds good.

Question (M): When is Min-ho's birthday?

The analysis of incorrect performance, in fact, is easier using graphs which can be generated under IRT and GLM. Regarding the effectiveness of the methods for expressing the results of item analysis, this study replicated the results reported in previous studies. As Thissen, Steinberg, and Fitzpatrick (1989) asserted, the IRT models were preferable to the classical test model using numerical summaries, due to the ability to graphically present the trace lines such as ICCs which were associated with distracters on each category of each item. The logistic regression model under GLM also could present the visual graphic lines. This type of graphical data is essential, because the graphs easily reveal flaws in terms and help to determine whether distracters or categories are constructed to meet the specific purpose of the test. (Wainer, 1989).

In this study, the effects about the degree of sample dependency of item parameters under different measurement theories were not analyzed. In the future, it would be beneficial for researchers to further study what theoretical frameworks can generate the most stable item parameters in diverse situations using empirical and simulated data.

In terms of screening DIF items, many researchers have studied the methods for the detection and elimination of DIF. This is an important field of inquiry, because elimination of DIF items can increase the validity of the test for all subgroups (Thissen, 2001).

In this study, DIF detection methods showed different characteristics. The MH approach based on analysis of contingency tables was not able to identify items that displayed nonuniform DIF. This result was consistent with previous studies that documented this as a representative shortcoming of the MH method. Also the MH statistic was demonstrated to be unsuitable, when the observable criterion on which examinees are matched, such as total score, is widely distributed. In that situation, the result of DIF detection may be unreliable, because data in contingency tables based on the range of total score may be sparse.

The IRT-LR procedure which used the likelihood ratio test under IRT was highly powerful in detecting not only uniform DIF but also nonuniform DIF items, and had several other advantages over the other methods. That is, it separately detected DIF items which were caused by differential difficulty or discrimination. Additionally, tests of hypotheses for uniform and nonuniform DIF could be performed at the same time. For this IRT-LR procedure, however, the procedure of model-data fit was needed, which is a common drawback of IRT-derived approaches. Also the IRT-LR approach has the limitation that only 2PL, 3PL, and Samejima's graded model can be implemented.

The method using the LR equation under GLM also demonstrated high nonuniform DIF detection similar to the IRT-LR procedure. The LR procedure which Swaminathan and Rogers (1990) suggested allowed DIF items to be screened quickly and easily without the model-fitting procedure and problems due to sparse data. Due to the flexibility of the logistic regression procedure as a model-based approach, showed were the possibilities that the model can simultaneously incorporate two variables, gender and the degree of urbanization, in the equation. This allows multiple relationships between important factors related to DIF to be examined.

Overall, in this study there was considerable variation in the rates of DIF detection across the different approaches, as Finch and French (2007) showed. An explanation for this result may be that each procedure is vey susceptible to different conditions. Thus, accurately to grasp strong and weak points of DIF detection methods, further studies should be accomplished in various conditions such as diverse sample sizes and different data forms.

Given these results, it may not be sufficient to check for DIF only with tests for statistical significance. When deciding if items exhibit DIF, it is recommended to use both an effect size and a statistical test, because the effect size can prevent flagging unimportant differences in large

samples, and the test of significance can prevent flagging noise in small samples (Monahan et al., 2007). Kim, Oswalt, and Cohen (2005) suggested that it may be necessary to consider effect sizes with descriptive measures with the results of DIF detection, such as model-based impact indices for IRT models and *R*-squared measures for logistic regression. Educational Testing Service (ETS) classifies the magnitudes of DIF as three categories: negligible, moderate, and large magnitude. To decide what values of the effect size represent three magnitudes, ETS uses thresholds of 1.0 and 1.5 on the absolute value of the delta metric, which are equivalent to odds ratios greater than 1.53 and greater than 1.89, respectively.

Also to improve the effect of identifying DIF, an iterative application of the methods for DIF detection has been recommended (Clauser & Mazor, 1998). One such iterative procedure, the purification of criterion, includes the repetitive omission of DIF items and recalculation of an internal criterion like total score. The iterative approach has been reported to produce results equal or superior to ones for a nonpurified criterion.

The results of this study which analyzed dichotomous and polytomous items and detected DIF using CTT, IRT, and GLM will provide test developer and measurement specialists with information for the test construction of high quality, although, from practical point of view, the interpretation and application of statistical evidence of item analysis and DIF can be eventually the portion of policy judgment.

REFERENCES

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

Agresti, A. (2007). *An introduction to categorical data analysis*. Hoboken, NJ: Wiley-interscience.

Baker, F. B. (2001). The basics of item responses theory. Retrieved January 20, 2009, from http://eric.ed.gov/ERICDocs/data.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Marcel Dekker.

Cangelosi, J. S., (1990). *Designing tests for evaluating student achievement.* New York: Longman.

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

Cohen, A. S., Kim, S.-H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement, 17,* 335-350.

Cook, L. L., Eignor, D. R., & Taft, H. L. (1988). A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameter estimates. *Journal of Educational Measurement, 25,* 31-45.

Crane, P. K., Gibbons, L. E., Jolly, L., & Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques – DIF detect and difwithpar. *Medical Care, 44*(Suppl. 3), 115-123.

Crocker, L., & Algina, J. (1986). *Introduction to classical & modern test theory*. New York: Holt, Rinehart & Winston.

Cronbach, L. J., & Warrington, W. G. (1952). Efficacay of multiple choice tests as a function of spread of item difficulties. *Psychometrika, 17,* 127-147.

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: Lawrence Erlbaum.

Ebel, R. L. (1965). *Measuring educational achievement*. Englewood Cliffs, N.J.: Prentice-Hall.

Finch, W. H., & French, B. F. (2007). Detection of crossing differential item functioning: A comparison of four methods. *Educational and Psychological Measurement, 67,* 565-582

Fraenkel, J. R., & Wallen, N. E., (2003). *How to design and evaluate research in education* (5th ed.). New York: McGraw Hill.

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

Herrera, A.-N., & Gomez, J. (2008). Influence of equal or unequal comparison group sample sizes on the detection of differential item functioning using the Mantel-Haenszel and logistic regression techniques. *Quality & Quantity, 42,* 739-755.

Hills, J. R. (1981). *Measurement and evaluation in the classroom*. Columbus, OH: Charles E. Merrill.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression*. New York: Wiley-interscience.

Kane, M. T. (1992). An argument based approach to validity. *Psychological Bulletin, 112,* 527-535.

Kelly, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology, 30,* 17-24.

Kim, S.-H., Cohen, A. S., Alagoz, C. & Kim, S. W. (2007). DIF detection and effect size measures for polytomously scored items. *Journal of Educational Measurement, 44,* 93-116.

Kim, S.-H., Oswalt, S. B., & Cohen, A. S. (2005, April). *An investigation of DIF detection methods for likert-type items*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36,* 437-448.

Lawson, S. (1991). One parameter latent trait measurement: Do the results justify the effort? In B. Thompson (Ed.), *Advances in educational research: Substantive findings, methodological developments* (Vol. 1, pp. 159-168). Greenwich, CT: JAL.

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Macdonald, P., & Paunonen, S. V. (2002). A Monte Carlo comparison of item and person statistics based item response theory versus classical test theory. *Educational and Psychological Measurement, 62,* 921-943.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50,* 741-749.

Meyer, J. P., Huynh, H., & Seaman, M. A. (2004), Exact small-sample differential item functioning methods for polytomous items with illustration based on an attitude survey. *Journal of Educational Measurement, 41,* 331-334.

Monahan, P. O., McHomey, C. A., Stump, T. E., & Perkins, A. J. (2007). Odds ratio, delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression. *Journal of Educational and Behavioral Statistics, 32,* 92-109.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 54,* 495-502.

Shannon, G. A., & Cliver, B. A. (1987). An application of item response theory in the comparison of four conventional item discrimination indices for criterion-referenced tests. *Journal of Educational Measurement, 24,* 347-356.

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

Thissen, D. (1991). MULTILOG: Multiple, categorical item analysis and test scoring using item response theory (Version 6.0) [Computer program]. Chicago: Scientific Software.

Thissen, D. (2001). IRTLRDIF v.2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning [Computer program]. University of North Carolina at Chapel Hill: L. L. Thurstone Psychometric Laboratory.

Thissen, D., Chen, W-H, & Bock, R. D. (2003). MULTILOG (version 7) [Computer software]. Lincolnwood, IL: Scientific Software International.

Thissen, D., Nelson, L., Rosa, K., & McLeod, L. D. (2001). Item response theory for items scored in more than two categories. In D. Thisssen & H. Wainer (Eds.), *Test scoring* (pp. 141-186). Mahwah, NJ: Lawrence Erlbaum.

Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-choice models: the distractors are also part of the item. *Journal of Educational Measurement, 26*, 161-176.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response model. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-114). Hillsdale, NJ: Lawrence Erlbaum.

Thorndike, R. M. (2005). *Measurement and evaluation in psychology and education*. Upper Saddle River, NJ: Pearson Merrill Prentice Hall.

Wainer, H. (1989). The future of item analysis. *Journal of Educational Measurement, 26,* 191-208.

Worthen, B. R., White, K. R., Fan, X., & Sudweeks, R. R. (1999). *Measurement and assessment in schools*. New York: Longman.

Yanagawa, T., Fujii, Y., & Mastuoka, J. (1994). Generalized Mantel-Haenszel procedures for 2×J tables. *Environmental Health Perspectives, 102*(Suppl. 8), 57-60.

APPENDICES

A. LERTAP 5 Input File for Item Analysis

English Test in 2007

&

*col (c1-c32)

*sub Res=(1,2,3,4,5), Name=(English), Title=(dichotomous), per.

*key 35242 55435 21345 31221 12433 35443 54

&

*col (c33-c42)

*sub Res=(R, P, W), Name=(English), Title=(polytomous), per.

*key RRRRR RRRRR

*mws Call, 2, 1, 0

B. MULTILOG 7 Input File for Separate Calibration of Dichotomous Items

```
>PROBLEM RANDOM,
        INDIVIDUAL,
        DATA = 'C:\thesis\irt.dat',
        NITEMS = 32,
        NGROUPS = 1,
        NEXAMINEES = 1801,
        NCHARS = 4,
>TEST ALL,
     L2;
>END ;
2
01
11111111111111111111111111111111
N
(4A1, T5, 32A1)
```

C. MULTILOG 7 Input File for Separate Calibration of Polytomous Items

```
>PROBLEM RANDOM,
        INDIVIDUAL,
        DATA = 'C:\thesis\irt.dat',
        NITEMS = 10,
        NGROUPS = 1,
        NEXAMINEES = 1801,
        NCHARS = 4;
>TEST ALL,
    GRADED,
    NC = (2, 3, 2, 3, 3, 3, 2, 2, 3, 3);
>END ;
3
012
1111111111
2222222222
0303330033
(4A1, T5, 10A1)
```

D. MULTILOG 7 Input File for Simultaneous Calibration

```
>PROBLEM RANDOM,
        INDIVIDUAL,
        DATA = 'C:\thesis\irt.dat',
        NITEMS = 42,
        NGROUPS = 1,
        NEXAMINEES = 1801,
        NCHARS = 4;
>TEST ALL,
    GRADED,
    NC = (2(0)33, 3, 2, 3, 3, 3, 2, 2, 3, 3);
>END ;
3
012
11111111111111111111111111111111111111111
22222222222222222222222222222222222222222
00000000000000000000000000000000303330033
(4A1, T5, 42A1)
```

E. SAS Input for Analysis of Dichotomous Items

```
data ialr;
 infile 'C:\thesis\ialr.dat';
 input @1(D1-D32)(1.) @33(P1-P10)(1.);
 total1=sum(of D1 - D32); /*SUM OF DICHOTOMOUS*/
 proc logistic descending; model D1 = total1 /
    scale=none aggregate;
 proc logistic descending; model D2 = total1 /
    scale=none aggregate;
 proc logistic descending; model D3 = total1 /
    scale=none aggregate;
                •
                •
                •
 proc logistic descending; model D30 = total1 /
    scale=none aggregate;
 proc logistic descending; model D31 = total1 /
    scale=none aggregate;
 proc logistic descending; model D32 = total1 /
    scale=none aggregate;
 run;
```

F. SAS Input for Analysis of Dichotomous and Polytomous Items

```
data ialr;
infile 'C:\thesis\ialr.dat';
input @1(D1-D32)(1.) @33(P1-P10)(1.);
total1=sum(of D1 - D32); /*SUM OF DICHOTOMOUS*/
total2=sum (of P1 - P10); /*SUM OF POLYTOMOUS*/
total=total1+total2; /*SUM OF TOTAL*/
proc logistic descending; model D1 = total /
  scale=none aggregate;
proc logistic descending; model D2 = total /
  scale=none aggregate;
                •
                •
                •
proc logistic descending; model D32 = total /
  scale=none aggregate;
proc logistic descending; model P1 = total /
  scale=none aggregate;
                •
                •
                •
proc logistic descending; model P10 = total /
  scale=none aggregate;
run;
```

G. SAS Input for DIF Detection for Gender by the MH Statistic

```
data difmh;
infile 'C:\thesis\dif.dat';
input urban 1 gender 2 @3(D1-D32)(1.) @35(P1-P10)(1.);
total1=sum(of D1 - D32); /*SUM OF DICHOTOMOUS*/
total2=sum(of P1 - P10); /*SUM OF POLYTOMOUS*/
total=total1+total2;
proc freq data=difmh; tables total*gender*D1 /
   norow nocol nopct chisq expected cmh;
proc freq data=difmh; tables total*gender*D2 /
   norow nocol nopct chisq expected cmh;
                    •
                    •
                    •
proc freq data=difmh; tables total*gender*D32 /
   norow nocol nopct chisq expected cmh;
proc freq data=difmh; tables total*gender*P1 /
  norow nocol nopct chisq expected cmh;
                    •
                    •
                    •
proc freq data=difmh; tables total*gender*P10 /
   norow nocol nopct chisq expected cmh;
run;
```

H. IRTLRDIF Input File for DIF Detection for Gender

1801

42

22222222222222222222222222222222323332233

dif.dat

0

2

3-44

difgender.out

I. SAS Input for DIF Detection for Gender

```
data diflr;
infile 'C:\thesis\dif.dat';
input urban 1 gender 2 @3(D1-D32)(1.) @35(P1-P10)(1.);
total1=sum(of D1 - D32); /*SUM OF DICHOTOMOUS*/
total2=sum(of P1 - P10); /*SUM OF POLYTOMOUS*/
total=total1+total2;
proc logistic descending; model D1 = total gender total*gender /
   scale=none aggregate rsquare;
proc logistic descending; model D2 = total gender total*gender /
   scale=none aggregate rsquare;
                    •
                    •
                    •
proc logistic descending; model D32 = total gender total*gender /
  scale=none aggregate rsquare;
proc logistic descending; model P1 = total gender total*gender /
  scale=none aggregate rsquare;
                    •
                    •
                    •
proc logistic descending; model P10 = total gender total*gender /
  scale=none aggregate rsquare;
  run;
```