DOCUMENT RESUME

ED 377 247 TM 022 515

AUTHOR Wang, Ning; Lane, Suzanne
TITLE Detection of Gender-Based Differential Item Functioning in a Mathematics Performance Assessment.
SPONS AGENCY Ford Foundation, New York, N.Y.
PUB DATE 94
CONTRACT 890-0572
NOTE 32p.; Version of a paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 3-7, 1994).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Educational Assessment; *Elementary School Students; Grade 6; Grade 7; *Identification; Intermediate Grades; *Item Bias; Junior High Schools; Junior High School Students; *Mathematics; *Sex Differences; Test Items; Thinking Skills
IDENTIFIERS *Performance Based Evaluation; Polytomous Scoring; *QUASAR Project (Mathematics Education)

ABSTRACT

This study used three different differential item functioning (DIF) procedures to examine the extent to which items in a mathematics performance assessment functioned differently for matched gender groups. In addition to examining the appropriateness of individual items in terms of DIF with respect to gender, an attempt was made to identify factors that may be related to DIF. The QUASAR Cognitive Assessment Instrument (QCAI) is designed to measure mathematical thinking and reasoning skills through open-ended questions. In this study, 33 polytomously scored QCAI items, from 4 test forms completed by 1,782 6th and 7th graders, were evaluated for gender-related DIF. Results suggest that DIF may not be serious for 31 of the 33 QCAI items. Explanations are suggested for the DIF of the remaining two items, one of which is of particular concern. Three figures and six tables present study findings. (Contains 29 references.) (SLD)

Detection of Gender-Based Differential Item Functioning in a Mathematics Performance Assessment

Ning Wang and Suzanne Lane

University of Pittsburgh
Pittsburgh, PA 15260

An earlier version of this paper was presented at the 1994 Annual Meeting of the National Council on Measurement in Education, New Orleans, LA. Preparation of this paper was partially supported by a grant from the Ford Foundation (Grant number 890-0572) for the QUASAR project. Any opinions expressed herein are those of the authors and do not necessarily represent the views of the Ford Foundation.

Detection of Gender-Based Differential Item Functioning in a Mathematics Performance Assessment

Abstract

This study used three different DIF detection procedures to examine the extent to which items in a mathematics performance assessment functioned differently for matched gender groups. In addition to examining the appropriateness of individual items in terms of DIF with respect to gender, an attempt was made to identify factors (e.g., content, cognitive processes, differences in ability distributions, and so forth) that may be related to DIF. The QUASAR Cognitive Assessment Instrument [QCAI] is designed to measure students' mathematical thinking and reasoning skills and consists of open-ended items that require students to show their solution processes and provide explanations for their answers. In this study, 33 polytomously scored items, which were distributed within four test forms, were evaluated with respect to gender-related DIF. The data source was 6th and 7th grade student responses to each of the four test forms administrated in the spring of 1992 at all six school sites participating in the QUASAR project. The sample consisted of 1782 students with approximately equal numbers of females and males. The results indicated that DIF may not be serious for 31 of the 33 items (94%) in the QCAI. For the two items that were detected as functioning differently for males and females, several plausible factors for DIF were discussed. The results from the secondary analyses, which removed the mutual influence of the two items, indicated that DIF in one item, PPP1, which favored females rather than their matched males, was of particular concern. These analyses suggested that the detection of DIF in the other item in the original analysis may have been due to the influence of item PPP1 since they were both in the same test form.

Detection of Gender-Based Differential Item Functioning in a Mathematics Performance Assessment

In recent years, educators have been redefining the goals of instruction and learning to include increased attention to high-level thinking skills (e.g., NCTM, 1989). At the same time, educators and psychometricians have been re-evaluating how best to assess students' thinking and reasoning skills. Consequently, there has been an increase interest in the use of performance assessments since they have the potential for allowing students to display their solution processes and reasoning. However, evidence is needed to ensure reliable and valid assessments of students' high-level thinking skills. In particular, evidence is needed to ensure that inferences made from performance assessments are equally valid for different subgroups in the population (Linn, Baker, & Dunbar, 1991); therefore, the detection of differential item functioning (DIF) is important in addressing issues regarding the quality of the assessment instrument. As indicated by Dorans and Holland (1993), "DIF refers to differences in item functioning *after* groups have been matched with respect to the ability or attribute that the item purportedly measures. . . DIF is an *unexpected* difference among groups of examinees who are supposed to be comparable with respect to the attribute measured by the item and the test on which it appears" (p. 37).

One problem with DIF detection studies is that it is relatively difficult to determine the actual factors that result in a significant DIF statistic. Identifying what factors are associated with DIF would significantly contribute to the development of valid assessment instruments. Some researchers have made efforts in this direction for dichotomously scored items. For example, in the mathematics domain, O'Neil and McPeek (1993) investigated several factors, such as content and format, in relationship to DIF in the SAT (Scholastic Aptitude Test), GRE (Graduate Record Examinations), and other admissions tests. They found that when matched on mathematics test scores women perform better on algebra items than men and men perform better on geometry and mathematics problem-solving items than women. Moreover, they noted, in mathematics tests, generally, "men perform better on the word problems than do their matched female counterparts, and women perform better on the more abstract pure mathematics items than do their matched

counterparts" (p. 268). Jackson (1992) conducted a study to detect gender-based DIF on items assessing understanding of percents on the SAT. The results showed that the concept of a percent being greater than 100% can be problematic for females compared to their matched males and items with unusual answers (e.g., 2000%) tended to favor males. Further, research has indicated that the extent to which the items are similar to textbook items is related to gender-based DIF. Females outperform their matched males when the items are similar to problems that appear in textbooks, whereas males outperform their matched females when the items evoke some solution strategies that are generally not taught in school (Harris & Carlton , 1993; Gallagher & Lisi, 1992).

There is limited research, however, that attempts to identify the possible factors that may be related to gender-based DIF for polytomously scored items. Therefore, studies are needed not only for collecting empirical evidence of the presence of DIF in performance assessments, but also for identifying factors associated with DIF.

The QUASAR[1] Cognitive Assessment Instrument [QCAI], a mathematics performance assessment, has been developed to assess students' mathematical reasoning, mathematical problem-solving, and mathematical communication (Lane, 1993). It consists of a relatively large set of open-ended mathematics tasks, which allow for opportunities for students to construct their solutions and provide visible records of their reasoning processes. It is designed to assess the impact of innovative instructional programs on students mathematical thinking and reasoning; and thus, it provides programmatic rather than individual student level information and is administered in the fall and spring of each school year. A series of studies have been conducted to provide evidence for the reliability and validity of the QCAI (Lane, Liu, Ankenmann, & Stone, in press; Lane, Stone, Ankenmann, & Liu, 1994; Magone, Cai, Silver, & Wang, 1994; Magone, Wang, Cai, & Lane, 1993; Stone, Ankenmann, Lane, & Liu, 1993). However, no studies have examined the extent to which the QCAI items function similarly across gender groups that are comparable with respect to mathematics proficiency.

[1]QUASAR (Quantitative Understanding: Amplifying Student Achievement and Reasoning) is a national project designed to improve mathematics instruction for students attending middle schools (grades 6 - 8) in economically disadvantaged communities (Silver, 1993).

The main purpose of this study is to examine the extent to which the QCAI items function differently for matched gender groups using three different DIF detection procedures. In addition to examining the appropriateness of individual items in terms of DIF with respect to gender, an attempt is made to identify content, cognitive processes, and/or other factors (e.g., differences in ability distributions) that may be related to DIF.

Methods

Administration and Scoring of the QCAI

In the 1991-1992 school year, the QCAI consisted of 36 open-ended items which were divided into four forms, each containing 9 different items. The four forms (A - D) were randomly distributed within each 6th and 7th grade class in schools participating in the QUASAR project. A focused holistic scoring method was used for scoring the student responses to each item (Lane, 1993). This was accomplished by first developing a general scoring rubric that reflected the conceptual framework used for constructing the items. In developing the general scoring rubric, criteria were specified for each of five score levels (0 - 4). Based on the specified criteria at each of the score levels, a specific rubric was developed for each item. Each student's response to each item was scored by two trained raters. If the raters disagreed by more than one point, a member of the assessment team adjudicated and rated the student's response; and if the raters disagreed by only one point, one of the two scores was randomly selected (see Lane et al., in press).

In this study, the data source is the 6th and 7th grade student responses to each of the four forms administrated in the spring of 1992 at all six schools participating in QUASAR. The sample consists of 1782 students with approximately equal numbers of females and males. The number of students who responded to each form was 458 for form A, 446 for form B, 426 for form C, and 446 for form D.

Generalizability and Dimensionality of the QCAI

In addition to the actual factors associated with DIF, significant DIF statistics may also be a function of a particular psychometric characteristic of the test or the nature of the sample distributions. For example, a significant Mantel-Haenszel DIF statistic for dichotomous items might be due to the differences of ability distributions for groups and/or the low reliability of a test

rather than the difference between item response functions, conditional on the ability estimate (Zwick, 1990). Also, as Angoff (1993) indicated, "all the methods and techniques that have been developed to identify DIF in the items assume that the group of items, or the test that contains the item, is homogeneous and unidimensional" (p. 14). Thus, careful examination of the properties of the test and the sample distributions is needed to ensure that it is appropriate to apply a DIF procedure.

The generalizability studies conducted by Lane and her colleagues (in press) for each form provided information about the intertask and interrater reliability of the QCAI. The results from the person by task (p × t) design, used to examine differential student performance across tasks, revealed that the generalizability coefficients were .76, .83, .76, and .82 for Forms A, B, C, and D, respectively, given 8 items on Forms A, B, and D and 9 items on Form C. These coefficients indicate moderate generalizability of the QCAI student level scores[2]. As mentioned previously, the QUASAR project is interested in assessing mathematics achievement at the program level not the individual student level. Thus, for the purpose of the project, the results from person nested within school by task ((p:s) × t) generalizability studies were of main concern. The coefficients from these studies ranged from .80 to .97, indicating that valid generalizations based on school-level scores can be made.

The confirmatory factor analyses conducted by Stone and his colleagues (1993) indicated that a one-factor model fits the data for each of the four QCAI forms. Thus, the unidimensionality of each form of the QCAI is tenable. Descriptive statistics of the sample distributions on each form for each group will be provided in the result section of the present study.

DIF Detection Procedures

Psychometric procedures that detect DIF for use with dichotomously scored test items have been developed, but they are not all easily applied to open-ended items which are scored polytomously. Although these procedures can be extended to polytomously scored items, there are complications or limitations in doing so. Consequently, several DIF detection procedures for

[2]The generalizability coefficients obtained separately for 6th and 7th grade student responses ranged from .71 to .83 for the p × t design. Since the analyses in the present study are using the combined 6th and 7th grade student responses, the p × t generalizability coefficients shown here are based on the results from the combined data, which were provided in the earlier version of Lane and her colleagues' article presented at the 1993 AERA annual meeting.

polytomously scored items have recently been developed and demonstrated using both simulated data and actual data (Miller & Spray, 1993; Miller, Spray, & Wilson, 1992; Miller & Welch, 1993; Welch & Hoover, 1993). For the purpose of this study, three DIF detection procedures recommended for use with polytomously scored items are employed to detect DIF for the QCAI items. They are the logistic discriminant function analysis (LDFA) procedure (Miller & Spray, 1993), the HW3 method (Welch & Hoover, 1993), and the logistic regression (LOG) procedure for polytomously scored items (Miller & Spray, 1993; Wilson, Spray, & Miller, 1993).

Logistic regression procedure and its extension

The logistic regression (LOG) procedure uses a logistic regression model to study DIF (Swaminathan & Rogers, 1990). In this approach, the logistic regression model,

$$P(u \mid x, g) = \frac{e^{(1-u)[-\beta_0 - \beta_1 x - \beta_2 g - \beta_3 (xg)]}}{1 + e^{[-\beta_0 - \beta_1 x - \beta_2 g - \beta_3 (xg)]}}, \qquad (1)$$

is used for predicting the probabilities of correct and incorrect responses to each dichotomously scored item, given an observed total test score and its associated group membership. In equation (1), variable g represents group membership (0 for focal group, e.g., female, and 1 for reference group, e.g., male); x is the matching variable(e.g., the observed total test score); u represents the item response value (0 for an incorrect answer and 1 for a correct answer); the term xg is the product of the two independent variables x and g and represents the interaction between the matching variable and the group variable; and β_0, β_1, β_2, and β_3 are the parameters to be estimated.

Once the estimates of the four coefficient parameters, β_0, β_1, β_2, and β_3, for an item are obtained from a sample of test responses, the usual likelihood ratio chi-square tests of significance of the estimates of β_2 and β_3 are conducted to examine if DIF exists. The null hypothesis is that $\beta_2 = \beta_3 = 0$. An item shows uniform DIF if $\beta_2 \neq 0$ and $\beta_3 = 0$ with one degree of freedom, and nonuniform DIF if $\beta_3 \neq 0$ (whether or not $\beta_2 = 0$) with one degree of freedom (Swaminathan & Rogers, 1990).

Extensions of the LOG procedure for use with polytomously scored items have been developed. One that is commonly recommended is the continuation ratio logit analysis for ordinal

responses. In this procedure, the polytomous item responses are recoded as dichotomous based on the following rule (given K ordered item response categories) (Wilson, Spray, & Miller, 1993):

Analysis 1: recode response k = 1 as 0, k = 2, ..., K as 1;

Analysis 2: recode response k = 1, 2 as 0, k = 3, ..., K as 1;

. . . . . .

Analysis K-1: recode response k = 1, 2, ..., K-1 as 0, k = K as 1.

K-1 logistic regression analyses according to equation (1) are then performed on the above recoded dichotomous item responses. Two separate sums of the K-1 chi-square statistics for testing the overall effects of the coefficients, $ß_2$ and $ß_3$, are used to detect uniform DIF and nonuniform DIF, respectively. The degrees of freedom are K - 1 for each of the two tests.

Logistic discriminant function analysis procedure

The logistic discriminant function analysis (LDFA) DIF detection procedure uses the logistic discriminant function (Miller & Spray, 1993),

$$P(g \mid x, u) = \frac{e^{(1-g)[-\alpha_0 - \alpha_1 x - \alpha_2 u - \alpha_3 (xu)]}}{1 + e^{[-\alpha_0 - \alpha_1 x - \alpha_2 u - \alpha_3 (xu)]}}, \qquad (2)$$

to predict the probabilities of group memberships (0 for focal group and 1 for reference group), given an item score and an observed total test score. This function is a modification of equation (1). As in equation (1), g represents group membership; x represents the matching variable; u is the item response variable that could have more than two item score categories; xu is the product of the two independent variables x and u and represents the interaction between the matching variable and the item response variable; and α_i (i = 0, 1, 2, 3) are the discriminant function coefficients to be estimated for each item.

Similar to the LOG procedure, once the estimates of the four coefficients for an item are obtained from test responses, likelihood ratio chi-square tests of significance of α_2 and α_3 can be conducted to address questions concerning uniform and nonuniform DIF, respectively. The null hypothesis for detecting DIF is $\alpha_2 = \alpha_3 = 0$. An item shows uniform DIF if $\alpha_2 \neq 0$ and $\alpha_3 = 0$ with one degree of freedom, and shows nonuniform DIF if $\alpha_3 \neq 0$ (whether or not $\alpha_2 = 0$) with one degree of freedom. In this procedure, the interpretation of DIF is that "for at least one of the

item score levels or categories, the probability of group membership, given that item score and observed score, differs significantly from that which would be predicted from the observed score alone" (Miller & Spray, 1993, p. 111).

A well-known problem associated with the chi-square test is that if the sample size is large enough, the null hypothesis will be rejected. Thus, it may be difficult to judge the practical significance of the results. In order to inspect the actual severity of DIF and to identify which group an item favors, it is suggested that for those items with significant DIF, simultaneous 95% Scheffé type confidence bands need to be constructed around the estimated logistic discriminant function (2) for each item score value υ. These confidence bands are then compared with the estimated P(g | x) under the null model that only contains the matching variable x as the predictor (i.e., let $\alpha_2 = \alpha_3 = 0$ and estimate α_0 and α_1 in equation (2)). If the confidence bands include the estimated P(g | x) under the null model for most values of x at every item score level, then the actual severity of DIF for that item may not be of particular concern (Miller & Spray, 1993).

HW3 method

The HW3 statistic (Welch & Hoover, 1993) tests a hypothesis about the difference between two mean values summed across M-independent tests for each item, where M refers to the maximum category level of stratification on a matching variable. At each level of the matching variable, a separate t-statistic between focal and reference group for each item is computed. The formula used is

$$t_i = \frac{(\overline{X}_f - \overline{X}_r)}{\sqrt{\frac{(S_f^2 n_f + S_r^2 n_r)}{(n_f + n_r - 2)}\left(\frac{1}{n_f} + \frac{1}{n_r}\right)}}, \qquad (i = 1, 2, \ldots, M),$$

with $df = n_f + n_r - 2$. Where, $\overline{X}_f$ and $\overline{X}_r$ are the mean scores of the item for focal and reference groups, respectively; S_f^2 and S_r^2 are the sample variances of item scores in focal and reference groups, respectively; and n_f and n_r are the number of examinees in focal and reference groups, respectively. When there are different sample sizes of the two subgroups at each of the M-ability levels, a weighting procedure is used to calculate the effect size, d_i $(i = 1, 2, \ldots, M)$, from its corresponding statistic t_i. Given $S_1, S_2, \ldots, S_M$ represent M estimated standard errors of the independent effect size, the weighted average and its standard error of the M effect sizes are

$$D = \frac{(d_1/S_1) + \ldots + (d_M/S_M)}{1/S_1 + \ldots + 1/S_M} \text{ and } S_d = \frac{1}{\sqrt{1/S_1^2 + \ldots + 1/S_M^2}}.$$

The HW3 statistic for testing the null hypothesis that there is no DIF between the focal group and the reference group in the item is

$$Z = \frac{D}{S_d}$$

which is normally distributed (N(0, 1)). If Z is significantly different from 0, DIF exists (Welch & Hoover, 1993).

Advantages of using each of the DIF detection procedures

In choosing a procedure for a DIF detection study samples sizes and available computing facilities need to be considered. In general, the procedures using a latent ability variable, such as techniques based on IRT (item response theory), are theoretically preferred since the groups are matched on the variable of real interest (i.e., the true ability). However, "From a practical point of view, the latent trait approach is not always the most appropriate because large samples and sophisticated and expensive computing procedures are needed. Conditioning on the observed score yields methods that usually are easier to apply" (Mellenbergh, 1982, p. 107).

As mentioned previously, not all procedures for use with dichotomously scored items can be easily extended to polytomously scored items. Both the LDFA procedure and the HW3 method have been developed primarily for and recommended for use with polytomously scored items. The HW3 method is a relatively straightforward procedure for detecting DIF in polytomously scored items; it is easy to calculate and easy to interpret (Welch & Hoover, 1993). As Miller and Spray (1993) have indicated, the main advantage of the LDFA procedure is that "it treats the item response as an independent variable, thus requiring only a single regression for each item" (p. 118). In other words, it is able to handle any type of item response, including polytomously scored items, using one logistic discriminant function, instead of using several regressions as in the LOG procedure for polytomously scored items.

DIF procedures also differ with respect to which they are sensitive in detecting uniform and nonuniform DIF. An item shows *uniform* DIF when the item performance for one group is better than the other group uniformly across all matched proficiency levels, that is, there is no interaction between proficiency level and group membership; whereas, *nonuniform* DIF exists when there is

an interaction between proficiency level and group membership, that is, the item performance for one group is not the same as that for the other group at all matched proficiency levels (Mellenbergh, 1982). If the interaction between proficiency level and group membership is of concern, the LOG procedure for polytomously scored items and the LDFA procedure are more appropriate since they are designed for and do show their power in detecting both uniform and nonuniform DIF for polytomously scored items (Miller & Spray, 1993; Miller & Welch, 1993; Swaminathan & Rogers, 1990). The comparisons among the LDFA, the LOG procedure, and the Mantel-Haeszel (MH) method using simulated data show that both the LDFA and the LOG procedures are as powerful as the MH procedure in identifying uniform DIF and more powerful than the MH in identifying nonuniform DIF (Miller, Spray, & Wilson, 1992; Swaminathan & Rogers, 1990). Moreover, an in-depth, post hoc analysis approach is available in the LDFA procedure so that the practical significance of DIF and the direction of DIF items can be further investigated.

The HW3 method is not designed for and may not be powerful in detecting nonuniform DIF. However, Welch and Hoover (1993) compared the HW3 statistic with Mantel's χ^2 statistic (an extension of the MH method) under several simulated testing conditions and indicated that the HW3 index was as good as Mantel's for controlling the type I error rate and the HW3 tended to be more powerful than Mantel's in detecting DIF when the sample sizes were equal to or larger than 500 per group. Additionally, even though the performance of the Mantel's χ^2 statistic seemed less affected than the HW3 when sample sizes decreased from 500 per group to 250 per group, "By combining ability levels until the necessary sample sizes were obtained, . . ., the HW3 became a more powerful statistic than . . . the MH with sample sizes of 250:250" (p. 14). Furthermore, the HW3 is a nonparametric approach and does not use an iterative algorithm in data analysis; therefore, it is relatively inexpensive in terms of computing time. If the computing expenditure is of particular concern and/or a nonparametric procedure is preferred, the HW3 method is more appropriate than those procedures using parametric functions in modeling test responses (as do the LDFA and the LOG procedures).

Data Analysis

In this study, student observed mean test score on the assessment form (A, B, C, or D) served as the matching variable to detect DIF for each individual item. The mean score was computed by dividing a student's observed total test score by the number of items the student reached. It should be noted that one item in three of the forms (A, B, and D) was excluded from the analyses because too few students responded to these items at the 3 and 4 score levels.

The Logistic Procedure in SAS was used in applying the LDFA procedure and the LOG procedure for polytomously scored items. To test the DIF hypotheses, the coefficients for each of three hierarchical models were estimated by maximizing the likelihood function obtained from each model (Hosmer & Lemeshow, 1989). The three hierarchical models were: the full model containing three predictors (the mean test score, item score for the LDFA procedure or group membership for the LOG procedure, and the interaction between the two predictors); the second model containing just the first two predictors in the full model, but not the interaction; and the third model, the null model, containing only the mean test score as a predictor. In addition to the estimated coefficients, the computer program also provided the value of the log-likelihood function for each model. The statistics,

$$G = -2[\text{log-likelihood function from the second model} - \text{log-likelihood function from the full model}]$$

and

$$G = -2[\text{log-likelihood function from the null model} - \text{log-likelihood function from the second model}],$$

were then computed to test the null hypotheses for nonuniform DIF and uniform DIF, respectively, in each analysis. Under the null hypothesis of the LDFA procedure, G is distributed as a chi-square with one degree of freedom. For the LOG procedure for polytomously scored items, the recoding procedure as described previously was conducted to obtain four subsets of dichotomous response data for each item (each item is scored for 5 score levels, thus, K - 1 = 4) before estimating the coefficients for each model. Then, for each subset of the response data, regression coefficients were estimated for the three hierarchical models, respectively. In the LOG procedure for polytomously scored items, the statistics used to test the null hypotheses of nonuniform and

uniform DIF were respectively generated by summing the four G statistics obtained from the four analyses of the subsets of the dichotomous response data. For each test, the statistic is distributed as a chi-square with four degrees of freedom.

Unlike the above two procedures, to apply the HW3 method the matching variable had to be divided into categorical levels. At each level of the matching variable, a separate t-statistic between focal and reference group for each item was computed. In this analysis, every student's observed mean test score was divided into seven categories: 0 - 0.5, 0.5 - 1.0, 1.0 - 1.5, 1.5 - 2.0, 2.0 - 2.5, 2.5 - 3.0, and 3.0 - 4.0. The difference between two adjacent category levels is .5 except that the levels of 3.0 - 3.5 and 3.5 - 4.0 were combined as one category to obtain necessary sample sizes. A computer program in SAS was written to calculate the HW3 statistic for each item. As mentioned earlier, this statistic is normally distributed as N(0, 1) under the null hypothesis that there is no DIF between males and females.

For those items which were flagged by at least two procedures including the LDFA procedures, at the .05 level of significance, simultaneous 95% Scheffé type confidence bands were constructed around the estimated logistic discriminant function (2) for each item score level u. A comparison with the estimated probability under the null model was then provided. This was accomplished by first setting group membership g to be the focal group (g = 0) for the estimated function (2) and then plotting the estimated probability under the null model, along with the 95% confidence bands around the estimated function (2). This post hoc analysis allowed for examination of the severity of DIF.

Results and Discussion

Descriptive Statistics for Forms

Table 1 provides the mean test scores on each form for females and males, respectively. From Table 1, it is apparent that the mean test score in this study is not distributed normally. Also, the plots of the frequency distributions of each item response indicated that item responses are not necessarily distributed normally. As Miller and Spray (1993) noted, the LDFA procedure does not assume the normality of the independent variables. Thus, the test data appear to be suitable for the LDFA procedure. For the other two DIF procedures used in this study, although the normal ability

distributions with and/or without mean differences between the reference group and the focal group had been simulated in researchers' studies, no assumptions of normality for independent variables have been proposed and limited research has been conducted to inspect the robustness of these procedures with non-normal data (Miller, Spray, & Wilson, 1992; Welch & Hoover, 1993; Wilson, Spray, & Miller, 1993).

Insert Table 1 about here

DIF Statistics and Agreement of the Three Procedures

Table 2 shows the DIF statistics of the three procedures for each of the 33 items. The HW3 procedure flagged 7 items at the .05 level of significance (2 of them were in favor of females and the others were in favor of males) and only 2 items at the .005 level (one favored females and another favored males). The LDFA procedure flagged 1 item as nonuniform and flagged 6 items as uniform DIF at the .05 level of significance. At the .005 level, only 2 items were flagged as uniform DIF and no items were flagged as nonuniform DIF by the LDFA procedure. No items were flagged by the LOG procedure as nonuniform DIF regardless of the significance level and 9 items were flagged by this procedure as uniform DIF at the .05 level of significance. At the .005 level of significance, only 3 items were flagged as uniform DIF by the LOG procedure.

Insert Table 2 about here

Table 2 indicates that among the 33 items, 23 items were not flagged by all three procedures and 6 items were flagged by all three procedures at least at the .05 level of significance. Thus, the percent of agreement among the three procedures is 88%, which is relatively high. It is interesting to note that only one item was flagged for nonuniform DIF. This was detected at the .05 level using the LDFA procedure.

In order to inspect the consistency between any two of the three procedures in detecting DIF for the QCAI items, the percent of pairwise agreements among the three procedures were computed. Table 3 summarizes the consistency in which the LDFA and the HW3 procedures flagged the items at the .05 level. The percent of agreement between the LDFA and the HW3 is

94%. Table 4 summarizes the consistency in which the LOG and the HW3 procedures flagged the items at the .05 level. The percent of agreement between the LOG and the HW3 is 88%. Table 5 summarizes the consistency in which the LDFA and the LOG procedures flagged the items at the .05 level as uniform and nonuniform DIF, respectively. The percent of agreement between the LDFA and the LOG on uniform DIF is 94%. The percent of agreement between the LDFA procedure and the LOG on nonuniform DIF is 97%; however, this may be an artifact of the data in that only one item was detected as nonuniform DIF by the LDFA procedure.

Insert Tables 3 - 5 about here

In summary, the percentages of agreement among the three procedures in detecting DIF for the QCAI items are relatively high. The range is from 88% to 97% for detecting uniform and nonuniform DIF at the .05 level of significance.

A Post Hoc Analysis

For each of the 7 items which were flagged by at least two of the three DIF procedures, including the LDFA procedures, at the .05 level of significance simultaneous 95% Scheffé type confidence bands around the estimated logistic discriminant function (2), along with the estimated probability under the null model, were plotted for females ($g = 0$) at each item score level u ($u = 0$, 1, 2, 3, or 4). The results indicate that two items, PPP1 and RPN1, are of particular concern. Females with low mean test scores outperformed their matched males on item PPP1 at item score level 4; whereas on item RPN1, females with high mean test scores are more likely to obtain lower scores on the item in contrast to their matched males. Figures 1 and 2 show the plots for these two items, respectively.

Insert Figures 1 and 2 about here

As mentioned earlier, if the confidence bands include the estimated probability under the null model for most values of x at every item score level, the actual DIF for the item may not be serious. Otherwise, DIF for the item is of particular concern. Figure 1 indicates that at the item score levels 0 - 3 for PPP1, the 95% confidence bands include the estimated probability under the

null model for most values of the mean test score. However, at item score level 4, the estimated probability under the null model is located below the lower confidence band at the 0 - 1.7 range on the matching variable. This means that for examinees with mean test scores within the range of 0 to 1.7, the probability of a level 4 response to PPP1 would be larger for females than for males. In other words, this item tends to favor females who have low mean test scores at item score level 4.

Figure 2 indicates that at item score levels 2, 3, and 4 for item RPN1, the 95% confidence bands include the estimated probability under the null model for most values of the mean test score. However, at item score levels 0 and 1, the estimated probability under the null model is located below the lower confidence band at some values of the mean test score. The probability of a response of score level 0 to RPN1 would be larger for females than for males at mean test scores above 1.15 points. The probability of a score level 1 response to RPN1 is larger for females than for males who have the mean test scores above 1.5 points. In other words, the females who have high mean test scores would be at a disadvantage in responding to item RPN1 than their matched male counterparts.

An inspection of Table 2, which provides the DIF statistics, indicates that items PPP1 and RPN1 are the only items that had been flagged by all three procedures at the .005 level of significance. It should be noted that a total of 10 items were flagged by at least one of the three DIF procedures at the .05 level. The post hoc analysis was conducted for 7 out of the 10 items (these 7 items are all items flagged by the LDFA procedure) and indicated that only 2 of them are of particular concern. This result is similar to Miller and Spray's (1993) finding that among 9 items flagged at the .001 level of significance by the LDFA procedure, the post hoc analysis indicated that only 2 items showed serious DIF.

Factors that Affect DIF

Several factors that may contribute to differential performance for females and their matched males on items PPP1 and RPN1 will be discussed. The factors that may have influenced the DIF statistics for both items are considered as well as possible factors that are unique to each of the items.

One factor that may have resulted in significant DIF statistics might have been the mutual influence of the two items. Since both PPP1 and RPN1 are in the same test form (Form B) and students' obtained mean test scores on the form were used as the matching variable, the DIF statistic for one item may have been affected by existing DIF or a DIF tendency in the other item. Moreover, since DIF in the two items, as discussed earlier, is in the opposite direction, it is plausible that the significant DIF statistics for the items could have resulted from three possibilities: 1) both items have DIF but in different directions; 2) one item has DIF in one direction and another item has no DIF, but the latter has a DIF tendency in the opposite direction to the former; 3) both items have no DIF, but have DIF tendency in the opposite direction. To identify which possibility is most likely, the mutual influence on the DIF statistics needs to be removed, and then, the DIF procedures can be employed again without the mutual influence. A solution to remove the mutual influence and the findings will be discussed in the next section.

A significant DIF statistic may also be a result of a particular psychometric characteristic of the test or the nature of the sample distribution. Differences in mean test score between females and males would result in significant DIF statistics even though no difference exists between two item response functions, conditional on mean test scores. From Table 1, which shows the descriptive statistics of the mean test scores for females and males, it is apparent that the mean test score distributions on Form B for females and males are different: The mean scores differ by .18 points over a maximum of 4 points and the distributions tend to be more positively skewed for males than for females. Thus, the distribution differences might be another factor that contributed to the identification of PPP1 and/or RPN1 as DIF.

Besides the factors discussed above, several other factors that may affect DIF in items PPP1 and RPN1 are plausible. These are described separately for each item. Item PPP1 is the only mathematical problem-posing task in the QCAI. In this task, students are asked to pose three questions that could be answered based on information given in the following story: "Jerome, Elliott, and Arturo took turns drive home from a trip. Arturo drove 80 miles more than Elliott. Elliott drove twice as many as Jerome. Jerome Drove 50 miles." According to the scoring rubric, if a student posed three mathematical questions that could be answered from the information given

in the problem situation or by any additional information explicitly provided by the student, the student response was scored as 4. For example, the following two questions would be considered appropriate: "How many miles did Jerome drive?" and "How many miles did the three boys drive in all?". Obviously, the second question is more complex than the former. However, since the task did not indicate a requirement of complexity to the students, a student response was scored as 4 if the student posed any three mathematical questions regardless of their complexity. In other words, the complexity of questions posed by a student was not a factor in assigning a score to the response based on the scoring rubric; therefore, the nature of the scoring criteria could have been a potential factor that contributed to DIF in item PPP1. It should be noted that item PPP1 is no longer in the QCAI.

An alternative way of analyzing student responses for PPP1 has also been conducted (Silver & Cai, 1993). In Silver and Cai's study, a qualitative analysis framework was used to categorize the complexity of students' posed mathematical questions. They found that students who performed well on Form B appear to pose more complex problems. Thus, the level of complexity of posed problems may reflect students' level of mathematical thinking. They also found that females generated more questions than males, but males generated more complex mathematical questions than females. As mentioned above, since the task did not indicate a requirement of complexity to the students, the complexity of questions was not considered as a factor in scoring student responses in the QCAI scoring rubric. These differences are not visible in the scores assigned by the scoring rubric.

Item RPN1 is a pattern task in which four decimal numbers are listed with the same difference between any two consecutive numbers. This task was designed to assess students' proficiency in identifying the underlying mathematical regularities of an increasing decimal pattern. Students were asked to extend the pattern using these regularities and to communicate the regularities. Gallagher and Lisi (1992) have found that females outperformed their matched male counterparts when tasks invoked school-taught algorithmic strategies, while males outperformed matched females when tasks invoked estimation or logical inference strategies. In the present study, females who had high mean test scores were at a disadvantage in responding to item RPN1

with respect to their matched male counterparts. Since item RPN1 is a task which invokes a mathematical reasoning(or logical inference) strategy, the cognitive skills required for solving the task might be a possible factor that affects DIF in this item. However, item RPN2 which is a decreasing whole number pattern and requires the same type of reasoning abilities as RPN1 was not flagged as DIF. Thus, the nature of the numbers (i.e., decimal numbers) may be a factor contributing to DIF in item RPN1.

Finally, it should be noted that all the factors discussed above are not causal factors which influence DIF in items PPP1 and RPN1. They are just plausible factors based on the testing circumstance and the review of previous research results. To provide more evidence on the factors that affect DIF for these items, additional studies that provide a qualitative analysis of male and female student responses conditional on the mean test score are needed.

Removal of the Mutual Influence of DIF Items

One plausible factor that may be related to DIF is the mutual influence of items PPP1 and RPN1. To remove this mutual influence on DIF statistics one of the two items is excluded from the DIF analysis and the DIF detection procedure is applied to the rest of the test items (Angoff, 1993; Holland & Thayer, 1988). In this study, first item PPP1 was excluded from the analysis and the mean test score was based on the other seven items (including item RPN1) in Form B and the three DIF procedures were applied to these seven items. The same procedure was then applied after excluding RPN1 from the item set (PPP1 was excluded for this analysis).

Table 6 shows the DIF statistics for the three DIF detection procedures for items PPP1 and RPN1, using the refined matching variable in the analyses[3]. These two items were flagged by all three procedures at the .005 level once again.

Insert Table 6 about here

The LDFA post hoc analysis was conducted for each of the two items. The results indicate that DIF in item RPN1 may not be serious (i.e., the confidence bands around the estimated probability under the full model includes the estimated probability under the null model for most

[3]For the other seven items on Form B, the results from the analyses using two different refined matching variables (one excluded item PPP1 and the other excluded item RPN1) were consistent with that in the original analyses.

values of the refined mean test score at every item score level of RPN1). However, the results from the post hoc analysis for item PPP1 using the refined mean test score are about the same as those in the original analysis. At the item score levels 0 - 3 the 95% confidence bands include the estimated probability under the null model for most values of the matching variable, but at item score level 4, the estimated probability under the null model is located below the lower confidence band at the 0 - 1.4 range on the refined mean test score. The plot for item PPP1 at item score level 4 is shown in Figure 3.

Insert Figure 3 about here

The analysis using the refined mean test score provides more evidence that DIF in item PPP1 is of particular concern, and DIF in item RPN1 may not be as serious. Caution is needed, however, in interpreting these results because rather than 8 items, 7 items were used in these analyses.

Summary

This study provides some preliminary empirical evidence of the presence of gender-based differential item functioning in the QUASAR Cognitive Assessment Instrument which consists of mathematics open-ended items. The logistic discriminant function analysis post hoc procedure indicated that DIF is of particular concern for 2 of the 33 items. Thus, DIF may not be serious for 31 of the 33 items (94%) in the QCAI. For the two items that were detected as functioning differently for males and females, several plausible factors for DIF have been discussed. One of them is the mutual influence of the two items. In order to remove this mutual influence, the three DIF detection procedures and the post hoc analyses were conducted again, excluding the two items one at a time. These secondary analyses indicated that DIF in item PPP1, which favors female students rather than their matched male students, is of particular concern. This finding suggests that the detection of DIF in item RPN1 in the original analysis may have been due to the influence of PPP1 since they were both in the same test form. However, the analyses to detect the mutual influence of the items was based on a total of 7 items; and therefore, caution is needed in interpreting this result.

Several studies that have examined gender-based DIF found that male students perform relatively better than female students on items that were embedded in a real life context (e.g., Doolittle & Cleary, 1987; Harris & Carlton, 1993). Consequently, Harris and Carlton (1993) suggested that prior to including more real life items in a test, changes in mathematics curricular are needed to ensure that females are exposed to these types of problems. Zwick (1994) has also pointed out that some of the features of performance assessments that are considered to be advantageous may actually have negative consequences from an equity perspective. The students in this study are from six middle-schools across the country in which innovative mathematics instructional programs, which emphasize making sense of mathematics in real world contexts, are being implemented. Thus, a potential reason for not observing more items that function differentially for matched male and female students may be that both male and female students are receiving the same type of curriculum and instruction.

In examining gender-based differential item functioning on mathematics achievement and aptitude tests consisting of multiple-choice items, researchers have attempted to identify item features (e.g., content, format, and context) that are related to differential performance by male and female students (e.g., Doolittle & Cleary, 1987; Harris & Carlton, 1993; O'Neil & McPeek; 1993). With the increasing use of open-ended items, there is now an opportunity to examine differences in male and female student performances with respect to their thinking and reasoning not only with respect to the task features. For items in which matched male and female students perform differently, a qualitative analysis of the student responses can be undertaken to examine whether differences between male and female students are related to their solution strategies, representations, mathematical explanations, and/or mathematical errors. Such an analysis will most likely contribute to our understanding of performance differences on mathematics items for male and female students.

References

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

Doolittle, A. E. & Cleary, T. A. (1987). Gender-based differential item performance in mathematics achievement items. *Journal of Educational Measurement*, 24(2), 157-166.

Gallagher, A. M. & Lisi, R. D. (1992). *Gender differences in mathematics problem solving strategies.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Harris, A. M. & Carlton, S. T. (1993). Patterns of gender differences on mathematics items on the Scholastic Aptitude Test. *Applied Measurement in Education,* 6(2), 137-151.

Hauck, W. W. (1983). A note on confidence bands for the logistic response curve. *The American Statistician,* 37(2), 158-160.

Hosmer, D. W. & Lemeshow, S. (1989). *Applied Logistic Regression.* New York, NY: John Willey & Sons.

Jackson, C. A. (1992). *An analysis of factors related to male/female differential item functioning on percent questions on the SAT.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Lane, S. (1993). The conceptual framework for the development of a mathematics performance assessment instrument. *Educational Measurement: Issues and Practice,* 12(2), 16-23.

Lane, S., Liu, M., Ankenmann, R. D., & Stone, C. A.(in press). Generalizability and validity of a mathematics performance assessment. *Journal of Educational Measurement.*

Lane, S. Stone, C. A., Ankenemann, R. D. & Liu, M. (1994). Reliability and validity of a mathematics performance assessment. *International Journal of Educational Research,* 21(3), 247-266.

Linn, R. L., Baker, E., & Dunbar, S. B. (1991). Complex performance-based assessment: Expectations and validation criteria. *Educational Researcher,* 20(8), 15-21.

Magone, M. E., Cai, J., Silver, E. A., & Wang, N. (1994). Validating the cognitive complexity and content quality of a mathematics performance assessment. *International Journal of Educational Research,* 21(3), 317-340.

Magone, M. E., Wang, N., Cai, J., & Lane, S. (1993). *An analysis of the cognitive complexity of QUASAR's performance assessment tasks and their sensitivity to measuring changes in students thinking.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Marshall, S. P., Barthuli, K. E., Brewer, M. A., & Rose, F. E. (1989). *Story problem solver: A schema-based system of instruction.* Technical Report No. 89-01. ONR Contract N00014-85-K-0661.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics*, 7, 105-108.

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement* , 30(2), 107-122.

Miller, T. R., Spray, J. A., & Wilson, A. (1992). *A comparison of three methods for identifying nonuniform DIF in polytomously scored test items.* Paper presented at the 1992 psychometric society meeting, Columbus, OH.

Miller, T. R., & Welch, C. J. (1993). *Issues and problems in assessing differential item functioning in performance assessments* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

National Council of Teachers of Mathematics. (1989). Curriculum and evaluation standards for school mathematics. Reston, VA: Author.

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

Scheuneman, J. D., & Bleistein, C. A. (1989). A consumer's guide to statistics for identifying differential item functioning. *Applied Measurement in Education*, 2(3), 255-275.

Silver, E. A. (1991). *Quantitative Understanding: Amplifying Student Achievement and Reasoning*. Pittsburgh, PA: Learning Research and Development Center.

Silver, E. A., & Cai, J. (1993). *Mathematical problem-posing and problem-solving by middle school students*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Stone, C. A., Ankenmann, R. D., Lane, S. & Liu, M. (1993). *Scaling QUASAR's performance assessment.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27(4), 361-370.

Welch, C. J., & Hoover, H. D. (1993). Procedures for extending item bias Detection techniques to polytomously scored items. *Applied Measurement in Education*, 6(1), 1-19.

Wilson, A., Spray, J. A., & Miller, T. R. (1993). *Logistic regression and its use in detecting nonuniform differential item functioning in polytomous items*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics*, 15(3), 185-191.

Zwick (1994). *Differential item functioning in new modes of assessment: opportunities for experimental and methodological research.* Paper presented at the annual meeting of the National Council on Measurement in Education invited symposium, "Research directions in educational measurement: the dissertation and beyond", New Orleans, LA.

Table 1

Descriptive Statistics of the Mean Test Scores on Each QCAI Test Form for Females and Males

Form	Gender	Mean	Std Dev	Skewness
A	F	1.299	0.772	0.837
A	M	1.301	0.791	0.807
B	F	1.598	0.942	0.687
B	M	1.412	0.947	0.743
C	F	1.652	0.817	0.292
C	M	1.633	0.809	0.274
D	F	1.376	0.868	0.796
D	M	1.335	0.927	0.728

Table 2

DIF Statistics for HW3, LDFA, and LOG Procedures

(Form A: 8 Items; Form B: 8 Items; Form C: 9 Items; Form D: 8 Items)

Items	HW3 (Z statistic)	Uniform DIF LDFA (χ^2 statistic) (df=1)	Uniform DIF LOG (χ^2 statistic) (df=4)	Nonuniform DIF LDFA (χ^2 statistic) (df=1)	Nonuniform DIF LOG (χ^2 statistic) (df=4)
Form A					
RPG1	-0.007	0.041	1.286	1.559	5.263
PPA1	1.481	2.258	5.893	0.797	9.310
PST1	-1.669 *	2.125	4.367	1.290	2.885
PGE1	-1.471	2.581	6.674	0.871	3.761
RNS3	1.126	1.404	5.442	0.020	4.248
PRP2	-0.466	0.168	0.858	0.252	2.752
PNS2	1.828 *	4.235 *	11.407 *	0.118	3.076
PST2	-2.175 *	7.106 **	21.817 ***	0.222	5.388
Form B					
RPN1	-3.563 ***	14.033 ***	26.950 ***	2.845	[illegible]
PCO4	1.003	0.573	10.475 *	0.538	6.924
PST4	0.262	0.959	1.255	0.765	1.798
PCO2	-0.557	0.837	2.389	0.415	3.176
PGE3	-0.735	1.133	5.805	0.208	5.538
PNS3	1.293	2.461	5.753	0.687	0.601
PPP1	3.487 ***	16.503 ***	44.497 ***	3.507	7.254
PRP1	-0.292	1.213	0.911	0.983	3.797
Form C					
PES1	1.605	3.275	8.864	0.012	3.555
PNS4	0.047	0.021	2.760	1.986	7.517
PCO3	-0.418	0.095	6.408	0.053	3.871
PCO5	-0.664	0.676	2.069	0.016	1.325
PGE4	-2.199 *	5.261 *	12.053 *	0.257	1.248
RLO1	0.620	0.281	2.98	0.741	4.890
PNS1	1.164	1.610	5.024	0.010	3.416
PGE1A	-2.178 *	6.231 *	14.409 **	1.223	7.534
RNS2	1.522	2.983	8.391	0.617	1.056
Form D					
PCO1	1.101	2.128	12.603 *	2.554	3.717
RPN2	0.594	0.379	6.703	0.825	2.248
RLO1A	1.154	1.750	9.489	3.542	0.985
PGE2	0.399	0.105	2.336	3.884	4.297
RPA2	-1.346	4.456 *	12.590 *	5.824 *	8.506
PST3	0.319	0.497	2.373	0.037	3.872
CNS1	-1.445	0.432	7.092	2.199	6.092
PCO2A	-1.168	2.725	3.743	0.575	4.550

*p < .05; **p < .01; ***p < .005

Table 3

Agreement Between LDFA and HW3 Procedures at the .05 Level of Significance

		Results from LDFA		
		# of non-flagged items	# of flagged items	Marginal totals
Results from HW3	# of non-flagged items	25	1	26
	# of flagged items	1	6	7
	Marginal totals	26	7	33

Table 4

Agreement Between LOG and HW3 Procedures at the .05 Level of Significance

		Results from LOG		
		# of non-flagged items	# of flagged items	Marginal totals
Results from HW3	# of non-flagged items	23	3	26
	# of flagged items	1	6	7
	Marginal totals	24	9	33

Table 5

Agreement Between LDFA and LOG Procedures for Uniform and Nonuniform DIF at the .05 level of significance

		Results from LDFA					
		Uniform			Nonuniform		
		# of non-flagged items	# of flagged items	Marginal totals	# of non-flagged items	# of flagged items	Marginal totals
Results from LOG	# of non-flagged items	24	0	24	32	1	33
	# of flagged items	2	7	9	0	0	0
	Marginal totals	26	7	33	32	1	33

Table 6

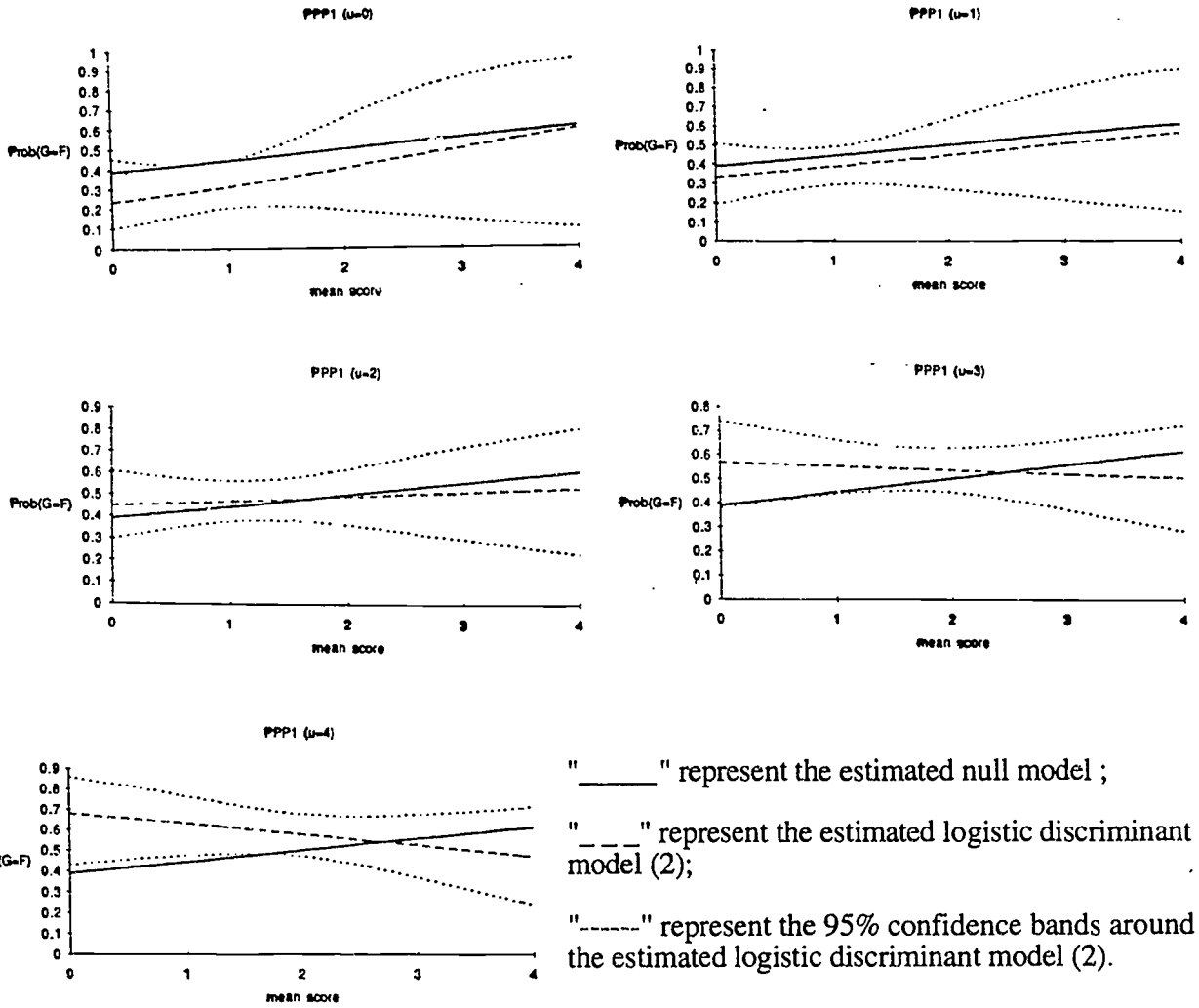
DIF Statistics for Items PPP1 and RPN1 Using the Refined Matching Variable

		Uniform DIF		Nonuniform DIF	
	HW3	LDFA	LOG	LDFA	LOG
	(Z statistic)	(χ^2 statistic)		(χ^2 statistic)	
		(df=1)	(df=4)	(df=1)	(df=4)
Items					
RPN1	-2.901 ***	9.700 ***	18.194 ***	1.113	4.111
PPP1	3.082 ***	13.979 ***	36.860 ***	3.534	6.918

***p<.005

Figure 1

Plots From the Post Hoc Analysis in the LDFA Procedure for Item PPP1



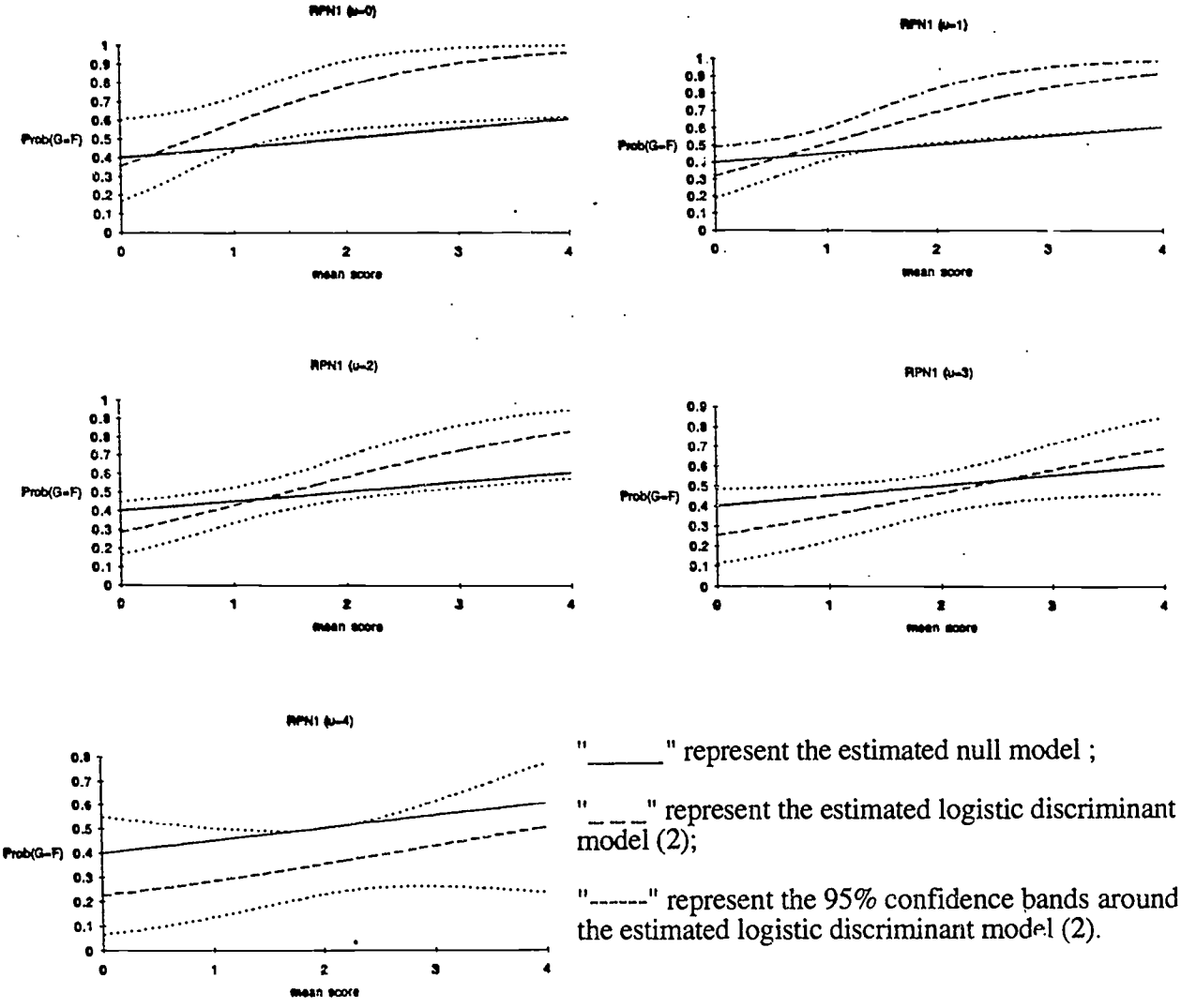
"_____" represent the estimated null model ;

"_ _ _" represent the estimated logistic discriminant model (2);

"------" represent the 95% confidence bands around the estimated logistic discriminant model (2).

Figure 2

Plots From the Post Hoc Analysis in the LDFA Procedure for Item RPN1



"_____" represent the estimated null model ;

"_ _ _" represent the estimated logistic discriminant model (2);

"------" represent the 95% confidence bands around the estimated logistic discriminant model (2).

Figure 3

Plot From the Post Hoc Analysis in the LDFA Procedure for Item PPP1 at Item Score Level 4, Using the Refined Matching Variable

"_____" represent the estimated null model ;

"_ _ _" represent the estimated logistic discriminant model (2);

"-------" represent the 95% confidence bands around the estimated logistic discriminant model (2).