ED 392 821 TM 024 470

AUTHOR	Lane, Suzanne; And Others
TITLE	Gender-Related Differential Item Functioning on a Middle-School Mathematics Performance Assessment.
SPONS AGENCY	Ford Foundation, New York, N.Y.
PUB DATE	Apr 95
CONTRACT	890-0572
NOTE	52p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22, 1995).
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Educational Assessment; Intermediate Grades; *Item Bias; Junior High Schools; *Mathematics Tests; Middle Schools; *Performance Based Assessment; Problem Solving; *Scoring; *Sex Differences; *Test Construction; Test Reliability; Thinking Skills
IDENTIFIERS	*Middle School Students; QUASAR Cognitive Assessment Instrument; QUASAR Project (Mathematics Education)

ABSTRACT

This study examined gender-related differential item functioning (DIF) using a mathematics performance assessment, the QUASAR Cognitive Assessment Instrument (QCAI), administered to middle school students. The QCAI was developed for the Quantitative Understanding: Amplifying Student Achievement and Reading (QUASAR) project, which focuses on reasoning, problem solving, and communication. By including two QCAI administration occasions, it was possible to examine the stability of gender-related DIF over time. On one occasion, 1,947 students took 4 forms of the QCAI, and on the other 1,999 students took the 4 forms. Student responses were rated by middle school mathematics teachers trained in the scoring procedure. Results indicated that 4 of the 36 tasks favored females and 2 favored males with respect to uniform DIF. The tasks that favcred females were set in a real-world context, suggesting that some of the features that have been associated with gender-related DIF in the past may not hold true for performance assessments. For example, male students may have been disadvantaged in providing explanations for their answers in this performance assessment. (Contains 4 figures, 13 tables, and 43 references.) (SLD)

Gender-Related Differential Item Functioning on a Middle-School Mathematics Performance Assessment

by

Suzanne Lar.e, Ning Wang, and Maria Magone
University of Pittsburgh

Preparation of this paper was supported by a grant form 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. This paper was presented at the annual meeting of the American Educational Research Association San Francisco, CA, April 1995.

Gender-Related Differential Item Functioning on a Middle-School Mathematics Performance Assessmenτ

As espoused by mathematics educators, the new vision of mathematics curriculum, instruction and assessment encompasses the need to continue to address equity-related issues (National Research Council, 1989; National Council of Teachers of Mathematics, 1989). Although there has been an abundance of research examining gender differences in mathematics performance for students who typically receive traditional forms of instruction, there are few studies that have examined the extent to which gender differences are exhibited when students have had the opportunity to receive instruction that focuses on high-level mathematical thinking and reasoning. Moreover, the majority of studies are based on high school students' performances on tests consisting of multiple-choice items. With the continued efforts to reform mathematics curriculum and instruction, it is imperative to examine differential performance among male and female students as they begin to have the opportunity to receive instruction that focuses on reasoning and problem-solving rather than rote memorization and computation. Such studies will require the use of performance assessments that measure students' mathematical problem solving and reasoning and will need to not only involve high school students but middle-school and elementary students. Further, with the increasing use of performance assessments at the local, state, and national level, evidence is needed to ensure that inferences made from the measures are equally valid for different subgroups in the population (Linn, Baker, & Dunbar, 1991).

The differential representation of women and men in scientific and technical fields was the impetus for research examining gender differences in mathematics performance. Sells (1980) pointed out the need to focus on mathematics as a critical filter in limiting women's entry into advanced technological and scientific work. This led to studies and interventions involving high school mathematics, and more recently elementary and middle school mathematics.

In recent years, meta-analyses of studies examining overall gender differences in mathematics performance have been undertaken. The results of these studies indicate that the magnitude of the gender differences have declined over the years, but the differences are not consistent from preschool to high school (Friedman, 1989; Hyde, Fennema, & Lamon, 1990). An examination of age trends indicate that female students performed slightly better in computation in elementary and middle school, no differences existed at any age level in understanding mathematical concepts, no differences existed in problem solving in elementary and middle school, and male students outperformed female students in problem solving in high school and college (Hyde, Fennema, & Lamon, 1990). In general, these two meta-analysis studies were based on data from published national standardized achievement tests consisting of multiple-choice items.

In addition to the examination of overall differences in male and female student performance, the examination of gender-related differential item functioning (DIF) is important in evaluating differences among male and female student mathematics performance. Differential item functioning refers to items that do not function the same after groups have been matched with respect to the attribute being measured (Holland & Thayer, 1986). Differential item functioning, however, is a statistical finding and may not necessarily warrant removal of items that are flagged as DIF when the content quality of the assessment may be jeopardized (Angoff, 1993; Doolittle & Cleary, 1987), but rather items that exhibit DIF may have implications for curriculum and instructional changes (Harris & Carlton, 1993). Thus, the examination of gender-related DIF in mathematics for students who are attending schools that are aligned with the new vision of mathematics (National Council of Teachers in Mathematics, 1988), which places an emphasis on mathematical problem solving, reasoning, and communication, may have important implications for implementing innovative mathematics curriculum and instruction.

In examining gender-related differential item functioning on mathematics achievement and aptitude tests consisting of multiple choice items, researchers have attempted to identify item features (e.g., mathematical content, item format, item context) that are related to differential performance by male and female students (e.g., Doolittle & Cleary, 1987; Harris & Carlton, 1993; O' Neil & Mcpeek; 1993; Ryan & Fan, 1994). For example, high school female students, as compared to their matched male students, tend to perform relatively poorer on geometry and algebraic/arithmetic reasoning items (Doolittle & Cleary, 1987); on items involving ratios, proportions, and percents (Jackson & Braswell, 1992); on items embedded in a 'real world' context (Harris & Carlton, 1993; O'Neil & Mcpeek, 1993); and on items that involve the use of solution strategies that are generally not taught in school (Harris & Carlton, 1993; Gallagher & Lisi, 1992; O'Neil & Mcpeek, 1993). The focus of most gender-related DIF studies, however, has been on tests consisting of multiple-choice items (some exceptions include Green, Fitzpatrick, Candell, and Miller (1992); Noble (1992); Wang & Lane (in press)) and the population of interest has been primarily high school students (an exception is Ryan & Fan, 1994). Moreover, with the increased use of open-ended mathematics items, the opportunity now exists to examine differences in male and female student performances with respect to their thinking and reasoning not only with respect to the task features.

Review of Gender-Related DIF

The identification of characteristics of mathematics items that are related to differential performance by matched male and female students has been undertaken (Doolittle & Cleary, 1987; Harris & Carlton, 1993; O'Neill & McPeek, 1993). Doolittle and Cleary (1987) examined gender-related DIF with respect to content categories. In their study, they controlled the effect of differential instruction in examining gender-related DIF on the ACT Mathematics Usage Test (ACTM). Their results indicated that, even when the effect of differential instruction is minimized, high school female students still tended to perform relatively poorer on geometry and arithmetic/algebraic reasoning (word problem) items, as compared with matched male students. Whereas, intermediate algebra/arithmetic and algebraic operations items tended to be relatively more difficult for male students than female students. Because the geometry items, in which many contained diagrams, favored males, they suggest that male students may be more proficient with respect to some types of spatial skills. They also found that for tasks that assess the same mathematical concepts, male students performed better when the task was related to a 'real world' situation, whereas, female students performed better when the task involved the application of an explicit operation and was not embedded in a 'real world' context. From this finding they suggest that male students may be more proficient in mathematical reasoning, whereas, female students may be more proficient in solving algorithmic or computational tasks. They further suggest that it may not be enough to balance on the basis of the high school curriculum because differences in instruction or background may have been firmly established prior to high school. As they point out, this is consistent with Fennema and Sherman's (1977) position that background needs to be defined more broadly to include other relevant experiential factors.

Harris and Carlton (1993) identified several item features which were related to differential performance of matched male and female high school students on the mathematics section of the Scholastic Aptitude Test (SAT). Using the Mantel-Haenzel procedure (Holland & Thayer, 1988), they found that geometry and arithmetic/geometry items tended to be relatively more difficult for female students than matched male students, whereas, arithmetic/algebra and miscellaneous (related to number sets, number systems, etc.) items tended to be relatively easier for female students than matched male students. With respect to item context, male students performed relatively better than female students on items embedded in a 'real world' context, whereas male students performed relatively poorer than female students on items that were not embedded in a 'real world' context. They indicate that this finding may lend support to the position that because males students view mathematics as more valuable or applicable in their lives. they are more proficient in mathematics (Fennema & Sherman, 1977). Significant statis.ical differences were also found on other categorizations of items, but these factors contributed marginally to the variance in DIF; therefore, practical differences were deemed questionable. These results indicated that male students performed statistically better than matched female students on items requiring a high-level of cognitive processing, requiring a computed solution rather than a general solution, and containing a table, graph or figure, whereas female students performed statistically better than matched male students on items that were similar to those found in textbooks.

The studies mentioned thus far have focused on gender-related DIF for high school students; few studies have examined gender-related DIF for students in elementary or middle-school. An exception is the study conducted by Ryan & Fan (1994) in which they investigated whether the relationship between DIF and item features for junior high school male and female students is similar to the relationship obtained for junior and senior female and mal high-school students, using data from a representative sample of male and female eighth grade students from the Second International Mathematics Study (1985). Consistent with the results at the high-school level they found that applied arithmetic items were relatively more difficult for female students than matched male students.

Moreover, few studies have examined gender-related DIF on mathematics performance assessments (for exceptions, see Green, Fitzpatrick, Candell, & Miller, 1992; Noble, 1992). One exception was Noble's (1992) use of logistic discriminant functional analysis (Miller & Spray, 1993) for examining gender-related DIF on two forms of the ACT Assessment Mathematics Test which consists of multiple-choice, gridded-response, and open-ended items. Three of the eleven open ended items were flagged for DIF and two of them favored female students in the lower score ranges.

Purpose of the Study

This study examined gender-related differential item functioning for middle-school students on a mathematics performance assessment (QUASAR[1] Cognitive Assessment Instrument (QCAI)) consisting of open-ended tasks (Lane, 1993). To examine gender-related DIF a procedure based on logistic discriminant function analysis (Miller & Spray, 1993) was employed on data from two administration occasions (Spring of 1993 and 1994). Advantages of this technique include the capability of examining nonuniform as well as uniform DIF and an accompanying post hoc procedure that examines the severity of DIF and at what score levels DIF is occurring. By including two administration occasions, it was possible to examine the stability of gender-related DIF over time.

For a subset of items that exhibited gender-related differential item functioning, an analytical analysis of student responses was undertaken to uncover potential differences in male and female students' solution strategies, mathematical explanations, and/or mathematical errors. Because most of the studies that have examined gender-related DIF in mathematics performance involved the use of multiple-choice items, in-depth analyses of differences in male and female students' thinking and reasoning have not been undertaken in these studies. The present study attempts to provide more detailed information about differences in male and female students' mathematical thinking and reasoning which may have more direct implications for assessment, curriculum, and instructional changes.

Methodology

Assessment Instrument

At the time the project was initiated in 1989, there were no existing assessment instruments for middle school mathematics that were aligned with key features of the reform-oriented conceptualization of mathematical proficiency (e.g., problem solving,

[1] QUASAR (Quantitative Understanding: Amplifying Student Achievement and Reasoning) is a national project that seeks to demonstrate that it is feasible to implement instructional programs in the middle-school grades that promote the acquisition of thinking and reasoning skills in mathematics (Silver, 1994). The project is directed at students attending schools in economically disadvantaged communities.

reasoning, communicating) and that had sufficient reliability and validity evidence to support their use. Therefore, the project developed and validated its own assessment instrument: the QUASAR Cognitive Assessment Instrument (QCAI).

The QCAI is designed to measure student outcomes and growth in mathematics, and to help evaluate attainment of the goals of the mathematical instructional programs (Lane, 1993). The QCAI consists of a set of open-ended tasks that assess students' mathematical problem solving, reasoning, and communication. Throughout the development process, steps were taken to ensure that the QCAI assesses students' knowledge of a broad range of mathematical content, understanding of mathematical concepts and their interrelationships, and capacity to use high-level thinking and reasoning processes to solve complex mathematical tasks (NCTM, 1989). Figure 1 provides examples of QCAI tasks[2].

Insert Figure 1 about here

The 6th/7th grade version of the QCAI was used for this study. It consists of 36 open-ended tasks, which are distributed into four forms, each containing nine tasks (Lane, Stone, Ankenmann, & Liu, 1994). Although the forms are not considered to be parallel, the tasks were distributed systematically across the forms to help ensure that the forms were as similar as possible with regard to content, processes, modes of representation (text, picture, graph, tables), context, and difficulty.

Assessment Specifications

The specification of the QCAI includes four major components: mathematical content, cognitive process, mode of representation, and task context. The content areas

[2] The QCAI is secure. The decision to keep the QCAI secure was based in part on the belief that evidence obtained from the assessment regarding student performance and program accountability would be more credible if the tasks were kept secure and also in part on the impractical and technical demands of developing a large number of tasks each year for assessing change in student performance. The items appearing in Figure 1 are tasks that appeared on the QCAI during the period 1990-1993 but that are now released and longer part of the current versions of the QCAI.

that were specified are number and operation, estimation, patterns, pre-algebra, geometry, measurement, probability, and statistics. These content areas are crossed by cognitive processes including understanding and representing mathematical problems; discerning mathematical relationships; organizing information; using strategies, procedures, and heuristic processes; formulating conjectures; evaluating the reasonableness of answers; generalizing results; justifying answers or procedures; and communicating mathematical ideas. The types of representations include text, pictorial, graphic, and arithmetic and algebraic expressions. Lastly, some of the tasks are embedded in "real world" contexts, while others are not. The components and categories within the components are interrelated; therefore, the framework allows for an individual task to assess topics in more than one content area and to assess a variety of processes[3].

Administration of the QCAI and Sample

The QCAI is administered within one class period (i.e., approximately 40-45 minutes). The data analyzed in this study were collected during QCAI administrations in the Spring of 1993 and 1994. Students received a different form of the QCAI on each administration[4]. The number of students who responded to each form in the Spring of 1993 was 469 for form A, 496 for form B, 506 for form C, and 476 for form D; and the number of students who responded to each form in the Spring of 1994 was 497 for form A, 506 for form B, 528 for form C, and 468 for form D.

Scoring Student Responses

A focused holistic scoring method was used for scoring the student responses to each task. This was accomplished by first developing a general scoring rubric that reflected

[3] Lane (1993) provides further detail regarding the conceptual framework for the QCAI.

[4] The forms were randomly distributed within each class in the schools participating in QUASAR in the fall of 1990, and thereafter each student received a different form on each administration occasion (Lane, Stone, Ankenmann, & Liu, 1994). The use of this sampling approach allows for the assessment of students in a relatively short time frame, thereby keeping interruptions to the instructional process minimal; minimizes the occurrence of practice effects; avoids the problems associated with sampling only a small number of tasks (e.g., Mehrens, 1992); and affords valid generalizations about students' mathematical proficiency at the school leve1.

the conceptual framework used for constructing the assessment tasks (Lane, 1993). The general scoring rubric incorporates three interrelated components: mathematical conceptual and procedural knowledge, strategic knowledge, and communication. In developing the general scoring rubric, criteria representing the three interrelated components were specified for each of five score levels (0-4). Five score levels were used to facilitate capturing various levels of student understanding.

Based on the specified criteria at each score level a specific rubric was developed for each task. The emphasis on each component for a specific rubric is dependent on the cognitive demands on the task. The criteria specified at each score level for each specific rubric is guided by theoretical views on the acquisition of mathematical knowledge and processes assessed by the task, and the examination of actual student responses to the task. This scoring procedure allows the assessment of differential levels of students' mathematical proficiency.

Student responses were rated by middle school mathematics teachers. The raters scored the student responses after they were formally trained. First, the general rubric was presented and discussed. Then a specific rubric and pre-scored student responses were presented and discussed. The raters then practiced scoring student responses, and their scores were discussed in relation to the scores previously assigned by the assessment team. Finally, the raters scored the actual student responses. Each response was scored independently by two raters. If the raters disagreed by more than one point, the assessment team rated the student response and it was this score that was used in subsequent analyses.

In addition to scoring the student responses as described above, the responses to some of the tasks that showed gender-related DIF were scored using an analytical procedure. This analysis provided information on appropriate strategies used by students, the correctness of numerical answers, the quality of the explanations, and misconceptions displayed in student responses. Additional information on this analysis will be presented in the Results and Discu.sion section.

Validity Evidence for the QCAI

If valid inferences are to be drawn from the scores on an assessment to the broader construct domain, both logical and empirical evidence to support such inferences is required (Linn, Baker, & Dunbar, 1991; Messick, 1989). The construct domain of mathematics, the task specifications, and the scoring rubric specifications were explicitly delineated to ensure that the tasks and scoring rubrics reflected the construct domain. The specification of the theoretical processes that can account for task performance provides information for construct validation. Empirical evidence, however, is also needed to ensure that the tasks evoke cognitively complex performances, that generalizations from the derived scores to the construct domain are valid.

Content Quality and Cognitive Complexity. Researchers have stressed the need for supplementing expert opinions regarding content quality and cognitive complexity with empirical evidence of the cognitive complexity of open-ended performance assessments (Linn et al., 1991). The development of the QCAI tasks includes logical analysis and expert judgment of the tasks in terms of the content quality, cognitive complexity, and fairness as well as empirical evidence of the underlying cognitive processes and content knowledge required for solutions. Student responses to both the pilot and operational tasks are analyzed in terms of the quality and nature of students' mathematical knowledge, solution strategies, representations, and communication. The assessment tasks are piloted with students from the participating schools and with students who have similar backgrounds to the students from the participating schools to help ensure that the tasks allow for the various representations, strategies, and ways of thinking that are common across the schools and that may be unique to one or more schools. In addition, the development of the scoring rubrics are based on both theoretical views underlying students' mathematical proficiency and empirical evidence obtained through the analysis of the student responses. It should also be mentioned that throughout the development process, the QCAI framework, tasks, and scoring rubrics were reviewed by teams of mathematics educators, mathematicians, cognitive psychologists, psychometricians, and multicultural educators, thereby ensuring that the QCAI blended considerations of mathematical content quality, current conceptualizations of mathematical proficiency, contemporary perspectives on student learning and understanding, as well as important equity and psychometric considerations[5] .

More extensive analyses of content quality and cognitive complexity were also undertaken for selected QCAI tasks (Magone, Cai, Silver, & Wang, 1994; Magone, Wang, Cai, & Lane, 1993). These analyses involved a detailed examination of task content, cognitive features, and student responses. The analyses of student responses focused on the representations and strategies used by students, the errors made, and the level and nature of communication used by students. The results of these efforts indicated that the tasks involved mathematical content appropriate for the intended grades, that the content being sampled is mathematically important at these grade levels, and that the tasks are cognitively complex.

Generalizability Evidence. One validity aspect of performance assessments that has received much attention from measurement specialists is the issue of generalizability of performance (Dunbar & Witt, 1993; Linn, Baker, & Dunbar, 1991). Of interest is the extent to which inferences or generalizations can be made from test scores based on a sample of tasks within the domain to the more broadly defined domain. As indicated by Dunbar, Koretz, and Hoover (1991), in order to make valid score inferences it is not only

[5] Because the QCAI includes a number of task formats, a variety of representations, strategies, and processes can be elicited from students. In addition, students can select representations and strategies that best facilitates them in solving the tasks. The QCAI also attempts to capture the types of task formats used at each of the schools, and consequently, this helps ensure a valid assessment of all students. Also, multiple variants of tasks are piloted to examine the best way to word and format tasks to help ensure that all students have the opportunity to display their mathematical thinking and reasoning. Our work has indicated that careful attention is needed in examining the relationship between the format and wording of a task and the nature of the student responses that the task engenders. Additional discussion on the use of a variety of task formats to help ensure the complexity of the domain of mathematics is captured by the assessment can be found in Lane, Parke, & Moskal, 1992; Parke and Lane, 1993.

necessary to examine the error that may be due to potential unreliability of raters but also to examine the error that may be due to the sampling of tasks. The generalizability of the derived QCAI scores was assessed through the use of generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). Both intertask and interrater consistency were addressed in examining the generalizability of the derived scores[6].

To examine the generalizability of a student's score for each form person x task generalizability studies were conducted (Lane, Liu, Ankenmann, & Stone, in press). These analyses are most pertinent for the purpose of this study since the examination of differential item functioning is based on student-level scores. The generalizability coefficients ranged from .71 to .84 when the number of tasks was equal to 9.

Interrater variance is a major source of potential measurement error for performance assessments. However, Dunbar, Koretz, and Hoover (1991) have argued that the effects of rater variance can be controlled with well-specified scoring rubrics and appropriate training of raters. To examine the errors in measurement due to potential unreliability of raters, person x task x rater generalizability studies were conducted (Lane, Stone, Ankenmann, & Liu, 1994; Lane, Liu, Ankenmann, & Stone, in press). The results of these analyses indicated that the variance components that included the rater source (i.e., rater, rater x person, rater x task, and rater x task x person variance components) were relatively small, suggesting that the use of one rater instead of two or more raters would

[6] A task by person-nested-within-school (t x (p:s)) generalizability study was conducted on data from each form to examine the extent to which generalizations to the larger domain of mathematics for school-level scores are valid (Lane, Liu, Ankenmann, & Stone, in press). For the purpose of examining the impact of the instructional programs on student performance, this generalizability study is most relevant. The obtained generalizability coefficients are for absolute decisions rather than relative decisions, which is appropriate for our purposes, since we are not interested in rank ordering the schools according to their performance on the QCAI. The coefficients were based on either 100 or 350 students so as to reflect the school with the smallest number of students and the school with the largest number of students at a particular grade level. Thirty-six tasks were used to reflect the number of tasks in the QCAI. When the number of persons is equal to 350, the generalizability coefficients ranged from .80 to .97 depending upon grade level and form. When the number of persons is equal to 100, they ranged from .71 to .95. These results provide support for using the QCAI to make decisions about schools' absolute scores. It should be noted that all but one school has a least 200 students per grade level.

have very little effect on the generalizability of the scores[7]. This result indicates that the validity of QCAI score inferences is minimally affected by rater inconsistency.

Dimensionality of the QCAI. A confirmatory factor analysis (CFA) was conducted on each of the four QCAI forms (Lane, Stone, Ankenmann, & Liu, in press). This procedure provides a test for the hypothesis that one latent variable accounts for the interrelationships among the tasks within each form. Note that since individual students were not administered all the tasks across the four forms, it was not possible to ascertain the interrelationships among all the tasks in one CFA.

Using LISREL version 7.1 (Joreskog & Sorbom, 1989), the results indicated that a 1-factor model fit the data for each form; thus, providing evidence of the unidimensionality of each form. While the dimensionality of the entire QCAI may still be in question, it is important to consider the way in which the forms were constructed. As previously mentioned, the tasks were distributed systematically across the forms to help ensure that the forms were as similar as possible with regard to content, process, context, representation, and difficulty level. Consequently, it may be reasonable to assume that the entire assessment is approximately unidimensional given unidimensionality in each form.

Results and Discussion

Descriptive Statistics for QCAI Forms

Table 1 provides the mean, standard deviation, and skewness for the test scores on each form for both female and male students. It is apparent from inspection of Table 1 that the mean test scores are not distributed normally. In particular, Form B is considerably skewed. It should also be noted that, in general, the plots of the frequency distributions for item scores are not distributed normally. However, as indicated by

[7] Additional empirical evidence for the validity, generalizability, and scaling of the QCAI is reported in Lane, Liu, Ankenmann, & Stone(in press), and Lane, Stone, Ankenmann, & Liu (in press).

Miller and Spray (1993) the LDFA procedure does not assume the normality of the independent variables.

Insert Table 1

Logistic Discriminant Function Analysis

Logistic Discriminant Function Analysis

The DIF detection procedure used in this study is the logistic discriminant function analysis (LDFA) (Miller & Spray, 1993). The form of the function is:

$$P(g \mid x, u) = \frac{e^{(1-g)[-a_0 - a_1x - a_2u - a_3(xu)]}}{1 + e^{[-a_0 - a_1x - a_2u - a_3(xu)]}} \qquad (1)$$

The function predicts the probabilities of group memberships (0 for focal group and 1 for reference group), given an item score and an observed total test score. 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 levels; 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 a_i ($i = 0, 1, 2, 3$) are the discriminant function coefficients to be estimated for each item.

Once the estimates of the four coefficients for an item are obtained from test responses, the likelihood ratio chi-square tests of significance of a_2 and a_3, can be conducted to address questions concerning uniform and nonuniform DIF of the item, respectively. The null hypothesis for detecting DIF is $a_2 = a_3 = 0$. An item shows uniform DIF if $a_2 \neq 0$ and $a_3 = 0$ with one degree of freedom, and shows nonuniform DIF if $a_3 \neq 0$ (whether or not $a_2 = 0$) with one degree of freedom. In this procedure, an item is flagged if for at least one of the item score levels, the probability of group membership differs significantly from that which would be predicted from the observed score alone, given that item score and observed score (Miller & Spray, 1993).

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 u. These confidence bands are then compared with the estimated P(g I x) under the null model that only contains the matching variable x as the predictor (i.e., let $a_2 = a_3 = 0$ and estimate a_0 and a_1 in equation (2)). If the confidence bands include the estimated P(g I 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).

The.Logistic Procedure in SAS was used in applying the LDFA procedure. 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 and the interaction between the two predictors); the next hierarchical model containing just the first two predictors in the full model, but not the interaction; and the last hierarchical model, the null model, containing only the mean test score as a predictor. In addition to the estimated coefficients, the program also provided the value of the log-likelihood function for each model. The statistics,

G = - 2[log-likelihood function from the second model
- log-likelihood function from the full model]; and
G = - 2[log-likelihood function from the null model
- 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, G is distributed as a chi-square with one degree of freedom.

Results of the LDFA procedure

For the Spring 1993 and 1994 administration of the QCAI, a total of 42 tasks were examined with respect to gender-related DIF. This is because between the two administrations some of the tasks were released and consequently, replaced by new tasks. Table 2 indicates that out of the 29 tasks that were used in both Spring 1993 and Spring 1994, 3 tasks exhibited uniform DIF in both years, and for at least one of the years it was significant at the .01 level (PRP4 and PGE4 were in favor of male students and PNS1 was in favor of female students). Out of these 29 tasks, no tasks exhibited uniform DIF in just one of the years at the .01 level of significance. Out of the 6 tasks that were used in the Spring 1993 only, 2 tasks exhibited uniform DIF at the .01 level of significance (RES1 and RPC1 favored female students); and out of the 7 tasks that were used in the Spring 1994 only, 1 task exhibited uniform DIF at the .01 level of significance (RPC1C favored female students).

The tasks that favored males students at the .01 level of significance, PGE4 and PRP4, were in the content areas of geometry and ratio/proportion, respectively. Both of the tasks included a figure; and one of the tasks (PRP4) was set in a real world context, whereas, the other (PGE4) was not. Neither of these tasks required a written verbal response: PGE4 asked students to provide reflections for a given figure and PRP4 asked students to show their solution strategies.

The four tasks that favored female students, PNS1, RES1, RPC1, and RPC1C, were in the content areas of number sense, estimation, patterns, and ratio/proportion, respectively. It should be noted that RPC1C was a much easier task than PRP4 and although both tasks are considered ratio/proportion tasks, the appropriate solution strategy is much more apparent for RPC1C than PRP4. None of the tasks that favored female students included a figure, but each task was set in a real world context. All four tasks required students to show their solution strategy and tasks PNS1 and RES1 required a written verbal response (e.g., an explanation).

With respect to nonuniform DIF, Table 1 indicates that out of the 29 tasks that were used in both of the years, 1 task exhibited nonuniform DIF in both years, and for at least one of the years it was significant at the .01 level (PCO2); whereas, no task exhibited nonuniform DIF in just one of the years at the .01 level. Out of the 6 tasks that were used in the Spring 1993 only, no task exhibited nonuniform DIF at the .01 level; and out of the 7 tasks that were used in the Spring 1994 only, no task exhibited nonuniform DIF.

Insert Table 2

For each item that was flagged simultaneous 95% Scheffé type confidence bands around the estimated logistic discriminant function (1), 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 only one item (PGE4 for the Spring 1994) was of particular concern. Female students were more likely to obtain lower scores on the item in contrast to matched male students. Figure 2 shows the plot for this item.

Insert Figure 2

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 2 shows that at item score levels 0 and 1, the 95% confidence bands do not include the estimated probability under the null model for mean test scores ranging from approximately 1 to 4. This indicates that for examinees with mean test scores within this range, the probability of a 0 or 1 score level would be higher for female students than male students. At the two most proficient score levels (3 and 4), the 95% confidence bands do not include the estimated probability under the null model for mean test scores ranging from 0 to approximately

2.5. This indicates that for examinees with mean test scores within this range, the probability of a 3 or 4 score level would be lower for female students than male students. Thus, females who have higher mean test scores are at a disadvantage in solving this item, whereas male students who have lower mean test scores have an advantage. In other words, the item tends to favor male students.

Analytic Scoring Analyses

The three tasks that were consistent in exhibiting gender-related uniform DIF across the two years were chosen for the analytic analysis (PGE4, PNS1, and PRP4). These tasks were flagged for DIF at the .01 level of significance for at least one of the two years; however, task PGE4 was the only task that indicated serious DIF in the post hoc analysis. The fourth task that was chosen for the analytic analysis was the one task that was used in the Spring 1994 only and was flagged for uniform DIF at the .01 level of significance (RPC1C). This task was revised between the two years and its earlier version (RPC1) was flagged for uniform DIF at the .01 level of significance in the Spring 1993. It should be noted that analytic analyses on the other tasks that were flagged as DIF are underway.

The student responses to the tasks were analyzed with respect to the use of an appropriate solution strategy, obtainment of the correct numerical answer, completeness of response, errors in understanding, and quality of explanation. In coding the student responses rater agreement ranged from 86% to 100%.

Analyses and Results for Task PNS1

This task assesses a student's proficiency in evaluating the reasonableness of an obtained answer. To solve this task, students need to not only correctly choose an appropriate solution strategy, but also make sense of the computational result according to the context of the problem. More specifically, the task indicates that a certain size canister can hold only a certain number of objects. The student is asked to identify the least number of canisters needed to hold a certain total number of objects, and to provide an explanation to the answer. A key element of this item is that there are not enough objects to completely fill one of the canisters. Task PNS1 is similar to a task that appeared on the Mathematics portion of the Third National Assessment of Educational Progress (NAEP, 1983): "An army bus holds 36 soldiers. If 1,128 soldiers are being bused to their training site, how many buses are needed." In addition to the context being different in Task PNS1, the numbers are smaller than the NAEP task to allow for students to partition using a diagram.

The sample consisted of 460 6th and 7th grade students who responded to the task in Spring 1994. There were 250 female student responses and 210 male student responses.

Each student response was coded with respect to four features: identification of an appropriate solution strategy, correct execution of the solution strategy, obtainment of the correct numerical answer, and quality of the explanation of the numerical answer. Parts of the qualitative analytic scheme for this task were adapted from schemes developed for a similar task (Cai, in press; Silver, Shapiro, and Deutsch, 1993).

Solution strategies and execution of strategies. To solve the numerical part of this problem successfully students need to identify an appropriate solution strategy that involves partitioning the objects into groups and then they need to apply the strategy correctly. A variety of appropriate solution strategies such as repeated addition, repeated subtraction, multiplication, division, or partitioning the objects using a drawing were used to solve this task. Table 3 provides the strategies used by female and male students. A chi-square analysis indicated that there was no significant gender differences in type of strategy used, χ^2 (4, N= 460) = 6.58, p = .160. It should be noted that an analysis was also conducted that included a category indicating that no solution strategy was shown; however, it was not significant at the .01 level, χ^2 (5, N= 498) = 12.263, p = .031.

Insert Table 3 about here

In addition to examining whether there were gender differences in identifying an appropriate strategy, the extent to which the strategies were executed correctly was examined. Table 4 provides the distributions indicating whether the execution of the strategies was correct for female and male students. The table indicates that out of the 218 female students and 187 male students who used an appropriate strategy, a larger percentage of female students (83%) than male students (73%) executed the strategy correctly although it was not significant at the .01 level, χ^2 (1, N=405) = 6.291, p = .012.

Insert Table 4 about here

Numerical answer. If an appropriate solution strategy is executed correctly, the result is a whole number with a remainder. However, the question posed to the student requires the student to map this numerical result back to the problem situation. Studies have indicated that middle-school students have a great deal of difficulty in relating computational results to the problem situation (e.g., Silver, et al., 1993). In this task, the student needs to recognize that the correct numerical answer (i.e., 4) involves rounding the obtained number (3 and a remainder) to the next whole number. Table 5 provides the distributions of each type of numerical answer provided by female and male students.

Insert Table 5 about here

Although a larger percentage of females (58%) provided the correct answer of 4 than males (50%), the difference in the distributions was not significant, χ^2 (3, N=460) = 7.708, p = .052.

Explanation of numerical answers. Students were asked not only to provide their numerical answers and to show their solution processes, but also to explain their answers in order to determine whether students were logically mapping their numerical answer back to the problem situation. Three categories were used to code students' explanations: conceptual explanation, procedural e.xplanation, and inappropriate or no explanation. The basis of a conceptual explanation or justification would be that a whole number is needed because one cannot have, for example, 3 and 1/2 canisters. A procedural explanation would be a description of the execution of the solution strategy. In the focused holistic scoring rubric a conceptual explanation is expected at the higher score levels.

Table 6 provides the distributions of explanations given by female and male students. The difference between the distributions is significant, χ^2 (2, N= 460) = 11.776, p = .003. The table indicates that a larger percentage of females (48%) than males (33%) provided conceptual explanations. In fact, a chi-square analysis examining whether there were gender differences in providing a conceptual explanation versus a procedural or no explanation was significant, χ^2 (1, N=460) = 11.373, p = .001.

Insert Table 6 about here

Relatic ship between the analytic and the DIF results. The result from the DIF analysis indicates that when male and female students are matched on mean test score, female students have a higher probability of obtaining the more proficient score levels (3 and 4) and have a lower probability of obtaining the less proficient score levels (0 and 1). Thus, the item tends to favor female students. To explore possible factors which could be related to the gender-based differential item functioning, a link between the results from the analytic analysis and the results based on the holistic scoring procedure is necessary.

The holistic scoring criteria for the most proficient score level (4) indicate that an appropriate solution strategy is used and executed correctly, and the conc3ptual explanation is correct and complete. The criteria for a score level of 3 is similar to the criteria for a score level of 4 except it allows for a minor error in the solution process and/or explanation. Thus, the criteria for score levels of 3 and 4 stress the need to not only execute the solution strategy correctly, but also to map the obtained answer to the problem situation in arriving at the final answer and to support the final answer with a conceptual explanation. The criteria for a score level of 2 indicate that an appropriate solution strategy is identified and that it may or may not be executed correctly, but the explanation is poor. A student would receive a score level of 1 if only a very limited understanding of the problem was demonstrated, and a student would receive a score level of 0 if there was no evidence of understanding the problem.

The results from the analytic analysis indicate that there is a significant gender difference with respect to the type of explanation provided, with female students providing more conceptual explanations than male students. In addition, although not significant at the .01 level, there was a larger percentage of female students than male students who executed the solution strategy correctly and then mapped the obtained answer back to the problem situation to arrive at the final answer. The differences of student performance on these factors appear to contribute to the gender-related differential item functioning.

Analyses and Results for Item RPC1C

This item assesses a student's proficiency in determining a proportional relationship between pairs of numbers in a problem situation. The student is asked to find the missing number in a pair and describe or show how the answer was obtained. The task is shown in Figure 3.

Insert Figure 3 about here

The sample consisted of 457 6th and 7th grade students who responded to the task in Spring 1994. There were 244 female student responses and 213 male student responses.

Each student response was coded with respect to three features: use of an appropriate solution strategy, obtainment of a numerical answer, and omissions in the response. The analytic analysis and results for this task are described in more detail by Magone (in preparation).

Solution strategies. To correctly solve the numerical part of this problem students need to identify an appropriate solution strategy and then apply the strategy correctly. The solution strategy that was displayed in the student responses and that would allow for obtaining the correct answer involved determining that the factor of four described the proportional relationship between the pairs of numbers. The other solution strategies that were displayed in the student responses would not lead to the correct answer (e.g., unjustified manipulation of numbers). Student responses were coded into 3 categories: use of an appropriate solution strategy, use of an inappropriate solution strategy, and no solution strategy shown. Figure 3 shows an example of the use of an appropriate solution strategy.

Table 7 provides the distributions of strategy types used by female and male students. A chi-square analysis indicated that there was a significant gender difference in use of solution strategy, χ^2 (2, N=457) = 22.418, p < .001. In general, a larger percentage of male students (31%) than female students (13%) did not display their solution strategies, and a larger percentage of female students (68%) than male students (51%) used an appropriate solution strategy. An analysis examining whether there was a difference between male and female students with respect to showing their solution strategy regardless of its appropriateness versus showing no solution strategy was significant, χ^2 (1, N=457) = 21.742, p < .001. This result indicates that female students were more likely to show their solution strategies; whereas, males were more likely to show no work. However, an analysis examining, for those students who showed their work, whether there was a gender difference in use of an appropriate strategy versus an inappropriate strategy was not significant, χ^2 (1, N=358) = .695, p = .409. For those students who showed their work, 74% of the male students used an appropriate strategy and 77% of the female students used an appropriate strategy.

Insert Table 7 about here

Numerical answer. Ta'ole 8 provides the distributions indicating whether a correct or incorrect answer was obtained for female students and male students regardless of the extent to which they showed their solution processes.

Insert Table 8 about here

Although a larger percentage of female students (67%) than male students (59%) provided the correct answer, the difference was not significant at the .01 level, χ^2 (1, N= 457) = 3.185, p = .074.

Omissions. To further explore whether gender differences existed with respect to the extent to which students displayed their solution strategies, for those students who obtained the correct answer, their responses were coded as whether they contained no omissions, minor or a moderate level of omissions, or many omissions or no work. Table 9 provides the distributions of omissions by female and male students. The difference between the distributions of female and male students was significant, χ^2 (2, N=289) = 21.526, p < .001. This table indicates that a larger percentage of female students (93%) than male students (77%) provided complete or nearly complete work; whereas, male students more often than female students provided very little work or no work to support their answer.

Insert Table 9 about here

Relationship between the analytic and the DIF results. The result from the DIF analysis indicates that when male and female students are matched on mean test score, female students have a higher probability of obtaining the more proficient score levels (3 and 4) and have a lower probability of obtaining the less proficient score levels (0 and 1). Thus, the item tends to favor female students.

The holistic scoring criteria for the most proficient score level (4) indicate that an appropriate solution strategy is used and executed correctly. For example, the student may indicate that the number of small tiles is four times the number of large tiles, or that 8 x 4 =32 and 3 x 4 = 12, so 9 x 4 =36. It should be noted that if the student relies on calculations, at least two of the three given pairs must be examined to receive a 4 score level. The criteria for a score level of 3 is similar to the criteria for a score level of 4 except it allows for a minor error or omission in the solution process. Thus, the criteria for score levels of 3 and 4 stress the need to be explicit in showing the solution strategy that was used. The criteria for a score level of 2 indicate that an appropriate solution strategy is identified, but the work is very incomplete. A student would receive a score level of 1 if only a very limited understanding of the problem was demonstrated, and a student would receive a score level of 0 if there was no evidence of understanding the problem.

The results from the analytic analysis indicate that although there is not a significant difference between female and male students with respect to obtaining the correct answer, a larger percentage of female students (67%) than male students (59%) provided the correct answer. Moreover, there is a significant gender difference with respect to the extent to which male and female students show their solution process. Overall, female students showed more work than male students. To obtain a high level score on the holistic scoring rubric, students need to be explicit in showing their solution strategy; thus, the differences in the extent to which the genders display their solution strategies contribute to the gender-related differential item functioning.

Analyses and Results for Task PRP4

This task assesses a student's proficiency in solving a problem that involves ratio and proportion. The student needs to demonstrate an understanding of a proportional relationship between two pairs of scores on different scales. This is accomplished by showing how the missing value in one of the pairs is obtained. The task is shown in Figure 4. It should be noted that this task is more difficult than Task RPC1C. By providing a series of paired numbers, Task RPC1C allows for the student to recognize that the first number in the pair just needs to be multiplied by 4 in order to obtain the second number in the pair. In contrast, the strategy for solving Task PRP4 is less apparent.

Insert Figure 4 about here

The sample consisted of 370 6th and 7th grade students who responded to the task in Spring 1994. There were 178 female student responses and 192 male student responses.

Similar to Task RPC1C, each student response was coded with respect to three features: use of an appropriate solution strategy, obtainment of a numerical answer, and omissions in the response. A more in-depth description of the analytic analysis and results for this task is provided by Magone (in preparation).

Solution strategies. To correctly solve the numerical part of this problem students need to identify an appropriate solution strategy and then apply the strategy correctly. The solution strategy that was displayed in the student responses and wou'd allow for obtaining the correct answer involved finding a factor that described the proportional relationship between the two pairs of numbers. The other solution strategies that were displayed in the student responses would not lead to the correct answer (e.g., unjustified manipulation of numbers). Student responses were coded into 3 categories: use of an appropriate solution strategy, use of an inappropriate solution strategy, and no solution strategy displayed. Figure 4 shows an example of the use of an appropriate solution strategy.

Table 10 provides the distributions of strategy types used by female and male students. A chi-square analysis indicated that there was a significant gender difference in the display of a solution strategy, χ^2 (2, N=369) = 15.895, p < .001. In general, a larger percentage of female students (62%) than male students (43%) used an inappropriate strategy rather than an appropriate strategy. Further, a larger percentage of male students (17%) than female students (7%) did not display their solution strategies. An analysis examining whether there was a difference between male and female students with respect to showing their solution strategy regardless of its appropriateness versus showing no solution strategy was significant, χ^2 (1, N=369) = 15.895, p = .0004. This analysis indicated that females were more likely to show their solution strategies; whereas, males were more likely to show no work. An analysis examining, for those students who showed their work, whether there was a gender difference in use of an appropriate strategy versus an inappropriate strategy was not significant at the .01 level, χ^2 (1, N=324) = 6.339, p = .012. However, for those students who showed their work, a larger percentage of male students (47%) than female students (34%) used an appropriate strategy to solve the problem.

Insert Table 10 about here

Numerical answer. Table 11 provides the distributions indicating whether a correct or incorrect answer was obtained for female students and male students regardless of the extent to which they showed their solution processes.

Insert Table 11 about here

The chi-square analysis was not significant at the .01 level, χ^2 (1, N=369) = 5.722, p = .017. However, a larger percentage of male students (30%) than female students (19%) provided the correct answer.

Omissions. To further explore whether gender differences existed with respect to the extent to which students displayed their solution strategies, for those students who obtained the correct answer, their responses were coded as whether they contained no omissions or some omissions/no work. Table 12 provides the distributions of omissions by female and male students. The difference between the distributions of female and male students was significant, χ^2 (1, N= 88) = 6.723, p = .009). This table indicates that a larger percentage of females (82%) than males (55%) provided complete work.

Insert Table 12 about here

Relationship between the analytic and DIF results. The result from the DIF analysis indicates that when male and female students are matched on mean test scores, male students have a higher probability than female students of obtaining the two most proficient score levels (3 and 4); whereas, female students have a higher probability of obtaining the two least proficient score levels (0 and 1). Thus, the task tends to favor male students.

In general, the holistic scoring criteria for the most proficient score level (score level 4) on this task indicate that an appropriate solution strategy was used and executed correctly. For example, the factor of 4 relationship between the distance from Martinsburg to Grantsville and the distance from Martinsburg to Rivertown needs to be explicitly shown and used to determine the distance from Martinsburg to Rivertown (e.g., 12/3 = 4, so 54 x 4 = 216), or the factor of 18 relationship between miles and centimeters is explicitly shown and used to determine the distance from Martinsburg to Rivertown (e.g., 54/3 = 18, so 1 centimeter equals 18 miles and 18 x 12 = 216 miles). The criteria for a score level of 3 is similar to the criteria for a score level of 4 except it allows for a minor error in the solution process. The criteria for a score level of 2 indicate that an appropriate solution strategy is identified, but the work is very incomplete. A student would receive a score level of 1 if a very limited understanding of the problem was indicated and would receive a score level of 0 there was no evidence of understanding the problem.

The results from the analytic analysis indicate that although there is a significantly larger percentage of female students (93%) than male students (83%) that showed their solution strategy regardless of its appropriateness, for those who showed their work, a larger percentage of male students (47%) than female students (34%) used an appropriate strategy. Consequently, a larger percentage of male students (30%) than female students (19%) obtained the correct answer. In other words, although female students showed more work than male students, male students were more likely to use an appropriate solution strategy and, consequently, to obtain a correct answer. Thus, these differences with respect to the genders help explain why this task favors male student than female students. Further, a plausible reason for why Task PRP4 favors male students and Task RPC1C favors females students may be because the solution process needed to solve Task PRP4 is not readily apparent; whereas, the solution process needed to solve Task RPC1C is more readily apparent.

Analyses and Results for Task PGE4

This task assesses a student's understanding of two concepts of geometric transformations. Students are asked to draw transformations of two given figures using the concepts of symmetry and conservation of area. In particular, in Part A of the task a figure is presented and the task specifies "A paper shape is folded in half. The folded shape looks like the figure below. The dark line is the fold. Use the figure to draw what the paper shape looks like when it is unfolded". Part B of the task presents a different figure printed three times on the paper and the task specifies "Three other paper shapes are folded in half. The folded shapes look like the figures below. Use each figure to draw what the paper shape would look like when it is unfolded. Make three different shapes."

The sample consisted of 467 6th and 7th grade students who responded to the task in Spring 1994. There were 243 female student responses and 224 male student responses.

For Part A, each student response was coded according to whether a correct reflection of the figure was provided and if it was incorrect, the type of error that was displayed. For Part B, each student response was coded according to whether no, 1, 2, or 3 correct reflections were provided; and for those responses that contained errors, the errors were coded. A more in-depth description of the analytic analysis and results for this task is provided by Magone (in preparation).

Drawing of figure. For Part A, student responses were coded as either providing the correct reflection or not. A chi-square analysis indicated that there was no significant gender difference in providing a correct reflection, χ^2 (1, N=467) = 2.610, p = .106. However, a larger percentage of male students (70%) than female students (62%) provided a correct drawing. For Part B, student responses were coded according to the number of correct reflections: 0, 1, 2, or 3. The distribution of correct drawings for female and male students is provided in Table 13. A chi-square analysis was significant, χ^2 (3, N=467) = 15.31, p = .001. A larger percentage of male students (31%) than female students (18%) provided three correct reflections; whereas, a smaller percentage of male students (27%) than female students (41%) provided no correct reflections.

Insert Table 13 about here

Errors. Nineteen error types were identified for this task. For both parts of the task, as might be expected given the results above, females tended to make a larger percentage of errors within each error type.

Relationship between the analytic and DIF results. The result from the DIF analysis indicates that when male and female students are matched on mean test scores, male students have a higher probability than female students of obtaining the two most proficient score levels (3 and 4); whereas, female students have a higher probability of obtaining the two least proficient score levels (0 and 1). Thus, the task tends to favor male students.

In general, the holistic scoring criteria reflect the extent to which students can be flexible in providing different reflections of the figure in Part B. For example, to receive a score level of 4, the student needs to provide a correct reflection for the figure in Part A and provide three correct, but different, reflections in Part B, and to receive a score level of 3, the student needs to provide three correct reflections. Thus, part of the analytic analysis was very similar to the holistic analysis. Overall, males were more proficient in providing different reflections for the figures and females were more likely to make more errors than males (e.g., the figure provided by the student may be congruent to the figure in the task, but it may not be a reflection).

Summary

The examination of gender-related differential item functioning in the present study was set in a context in which middle-school students are receiving mathematics instruction with a focus on reasoning, problem solving and communication. Gender-related DIF was examined with respect to a mathematics performance assessment consisting of open-ended tasks that ask students to show their solution strategies and/or explain their reasoning. This is in contrast to most of the literature that has examined gender-related DIF, in that, DIF was identified for performance on multiple-choice mathematics items (e.g., Doolittle & Clear, 1987; Harris & Carlton, 1993). Moreover, an examination of student performance allowed for the identification of plausible reasons for DIF on a subset of the tasks.

The results in this study indicate that four tasks favored female students and two task favored male students with respect to uniform DIF. The two tasks that favored male students included a figure; whereas, the four tasks that favored female students did not include a figure. This is consistent with Harris and Carlton's (1993) finding that indicates there is a tendency for male than female high school students to perform better on tasks that include a figure. The one task that showed severe DIF, PGE4, favored male students. This task is in the area of geometry, but more specifically it assesses whether students can provide different reflections for a given figure. The results indicated that males students are more proficient at providing a number of different reflections for the same figure. This finding is consistent with other studies that have indicated that male high school students perform relatively better on geometry items which usually contain figures, as compared to matched female students, and have suggested that this may be due to males being more proficient with some types of spatial skills (e.g., Doolittle & Cleary, 1987).

For the two tasks that assess ratio/proportion and were flagged as DIF, one favored female students (RPC1C) and the other favored male students (PRP4). In general, the literature has indicated that male students as compared to matched female students perform better on tasks that involve ratios and proportions (e.g., Jackson & Braswell, 1992). However, as mentioned previously, RPC1C is an easier task than PRP4 in that the appropriate solution strategy and answer for RPC1C is more apparent than it is for PRP4. Moreover, for these two tasks in order to receive one of the two most proficient score levels students need to explicitly show their solution strategies. The analytic analysis indicated that for Task RPC1C there was not a significant gender difference with respect to selecting an appropriate solution strategy nor obtaining the correct numerical answer; however, there was a significant gender difference with respect to the extent to which work was shown, with male students showing less work than female students. Thus, it appears that a critical aspect that is contributing to gender-related DIF on this relatively easy task is related to the extent to which male and female students were providing their work. In contrast, Task PRP4, which favored male students, is more difficult and the numerical answer and appropriate solution strategy is less apparent than it is for RPC1C. Although not significant at the .01 level, for those students who showed their work, a larger percentage of male students (47%) than female students (34%) used an appropriate solution strategy and a larger percentage of male students (30%) than female students (19%) obtained the correct numerical answer. Thus, although male students were less likely to show their work, when they did their work demonstrated the use of an appropriate solution strategy more often than female students. Thus, they were more likely to receive one of the two most proficient score levels based on the holistic scoring procedure.

Task PNS1, which favored female students, requires students to provide a numerical answer and an explanation for their numerical answer. Although a larger percentage of female than male students executed their solutions successfully and provided the correct numerical answer, it was not significant. However, there was a significant difference with respect to providing conceptual explanations, with female students providing more conceptual explanations than male students. These factors, and in particular, the fact that female students provided more conceptual explanations appear to contribute to the gender-related DIF. This result is consistent with research that has indicated that males tend to prefer nonverbal modes of representations; whereas, females prefer verbal modes (Clements and Battista, 1992). Futher, when solving mathematics tasks, females more often than males use written accounts of their solution (Fennema & Tartre, 1985; Tartre, 1990).

Lastly, research has indicated that male students as compared to matched female students, perform better on tasks that are set in a real world context (e.g., Harris & Carlton, 1993; O'Neil & Mcpeek, 1993). In this study, however, the four tasks that favored female students were set in a real world context. In fact, the majority of the tasks on the QCAI are set in a real world context which reflects the current thinking in the mathematics education reform movement. Moreover. the nature of the instruction at these schools places an emphasis on mathematics thinking and reasoning, and being able to "do" mathematics in real world contexts. The reason why the context of the problem was not related to DIF may be because both male and female students in this study have had the opportunity in their classrooms to solve applied problems that are set in a realistic context. Thus, some of the features that have been associated with gender-related DIF in mathematics may not hold when the studies involve the use of performance assessments and students who are receiving instruction that focuses on reasoning and problem solving. However, other features may be associated with gender-related DIF when using open-ended assessment tasks. For example, in this study male students as compared to matched females students may have been at a disadvantage on a few tasks because they were not complete in showing their solutions processes and providing explanations for their numerical answers.

References

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

Cai, J. (in press). A cognitive analysis of U.S. and Chinese students' mathematical performance on tasks involving computation, simple problem solving, and complex problem solving. Journal for Research in Mathematics Education, Research Monograph.

Clements, D. H., & Battista, D. H. (1992). Geometry and spatial reasoning. In D. A. Grouws (Ed.) Handbook of research on mathematics teaching and learning, (420-464). New York: Macmillan Publishing Company.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurement. New York: Wiley.

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

Dunbar, S. B., & Witt, E. A. (1993). Design innovations in measuring mathematics achievement. In National Research Council, Mathematical Sciences Educational Board (Ed.), Measuring what counts: A conceptual guide for mathematics assessment (pp. 175-200). Washington, DC: National Academy Press.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D.. (1991). Quality control in the development and use of performance assessments. Applied Measurement in Education, 4, 289-304.

Fennema, E., & Sherman. J. (1977). Sex-related difference in mathematics achievement, spatial visualization and affective factors. American Educational Research Journal, 14, 51-71. Friedman, L. (1989). Mathematics and the gender gap: A meta-analysis of recent studies on sex differences in mathematical tasks, Review of Educational Research, 59, 185-214.

Fennema, E., & Tartre, L. (1985). The use of spatial visualization in mathematics by girls and boys. Journal for Research in Mathematics Education. 16, 184-206.

Friedman, L. (1989). Mathematics and the gender gap: A meta-analysis of recent studies on sex differences in mathematical tasks, Review of Educational Research, 59, 185-214.

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.

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.

Green, D. R., Fitzpatrick, A. R., Candell, G., & Miller, E. (1992, April). Bias in performance assessment. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco.

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.

Holland, P. W. & Thayer, D. T. (1986). Differential item performance and the Mantel-Haenszel procedure. (Research Report No. 86-31) Princeton: Educational Testing Service.

Hyde, J. S., Fennema, E., & Lamon, S. (1990). Gender differences in mathematics performance: A meta-analysis. Psychological Bulletin, 107(2), 139-155.

Jackson & Braswell (1992, April). An analysis of factors causing differential item functioning on SAT-Mathematics items. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Joreskog, K. G., & Sorbom, D. (1989; second edition). LISREL 7: A guide to the program and applications. Mooresville, IN: Scientific Software, Inc.

Lane, S. (1993). The conceptual framework for the development of a mathematics assessment instrument for QUASAR. 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., Parke, C., & Moskal, B. (1992). Principles for developing performance assessments. Paper presented at the annual meeting of the American Educational research Association, San Francisco, CA.

Lane, S., Stone, C. A., Ankenmann, R. D., & Liu, M. (in press). Examination of the assumptions and properties of the graded item response model: An example using a mathematics performance assessment. Applied Measurement in Education, 8(4).

Lane, S., Stone, C. A., Ankenmann, R. D., & Liu, M. (1994). Reliability and validity of a mathematics performance assessment. International Journal of Educational Research, 21(3), 247-262.

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

Magone, M. (in preparation). Gender differences in a mathematics performance assessment consisting of open-ended tasks.

Magone, M. E., Cai, J., Silver, E. A., & Wang, N. (1994). Validity evidence for cognitive complexity of performance assessments: An analysis of selected QUASAR tasks. 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. A paper presented in the symposium "Assessing performance assessments: Do they withstand empirical scrutiny?" at the 1993 Annual Meeting of the American Educational Research Association, Atlanta, GA.

Mehrens, W. A. (1992). Using performance assessment for accountability purposes. Educational Measurement: Issues and Practice, 11(1), 3-9, 20.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational Measurement (3rd ed.) (pp. 13-104). New York: American Council on Education.

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.

National Assessment of Educational Progress (1983). The third national mathematics assessment: Results, trends and issues. Denver, CO: Author.

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

National Research Council (1989). Everybody counts. Washington, DC: National Academy of Sciences.

Noble, A. C. (1992, April). Differential item functioning in innovative-format mathematics items. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

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

Parke, C. & Lane, S. (1993). Designing performance assessments: An examination of changes in task structure on student performance. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Ryan, K. E. & Fan, M. (1994, April). Gender differences on a test of mathematics: Multidimensionality or differential test functioning. Paper presented at the Annual Meeting of the National Council of Measurement in Education, New Orleans.

Second International Mathematics Study. (1985). Summary report for the United States. Champaign, IL: Stipes.

Sells, L. (1980). The mathematics filter and the education of women and minorities. In L. H. Fox, L. Brody, & D. Tobin (Eds.), Women and the mathematical mystique (pp. 66-75). Baltimore, MD: Johns Hopkins University Press.

Silver, E. A. (1994, April). Building capacity for mathematics instructional reform in urban middle schools: Contexts and constraints in the QUASAR project. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Silver, E. A., Shapiro, L. J., & Deutsch, A. (1993). Sense-making and solution of division problems involving remainders: An examination of students' solution processes and their interpretations of solutions. Journal for Research in Mathematics Education, 24(2), 117-135.

Tartre, L. (1990). Spatial skills, gender and mathematics. In E. Fennema, E., & G. Leder (Eds.). Mathematics and gender, (pp. 27-59). New York: Teacher's College Press.

Wang, N. & Lane, S. (in press). Detection of gender-related differential Item Functioning in a mathematics performance assessment. Applied Measurement in Education.

Table 1

Descriptive Statistics for the Test Scores on Each QCAI Form for Female and Male Students

Date	Form	Gender	Mean	Std. Dev.	Skewness	N
Sp 93	A	F	1.432	.808	0.506	248
		M	1.358	.780	0.496	221
	B	F	1.170	.815	1.037	283
		M	1.215	.955	1.129	213
	C	F	1.649	.861	.139	247
		M	1.751	.876	.070	259
	D	F	1.201	.852	.791	238
		M	1.260	.864	.589	238
Sp 94	A	F	1.500	.921	0.494	264
		M	1.591	.920	0.395	233
	B	F	1.265	.816	0.944	254
		M	1.294	.927	0.846	252
	C	F	1.728	.860	-0.021	278
		M	1.596	.908	0.192	250
	D	F	1.442	.887	0.421	233
		M	1.448	.941	0.414	235

Table 2

DIF Statistics for the LDFA Procedure

Items	Uniform DIF		Nonuniform DIF	
Form A	S93	S94	S93	S94
RPG1	.010	1.575	.265	.935
PPA1	2.564	3.522	1.117	.255
RES1	10.652***		.001	
PST8		2.876		.586
PST1	.770	3.376	.534	.018
PGE1	3.285	5.218*	4.372*	.927
RNS3	.317	.643	1.767	.771
PRP2	1.764	.488	.393	.688
PME1	.245	1.416	.090	.022
PST2	.165		.007	
PPA4		.568		1.283
Form B				
RPN1	.255	.290	3.986*	.673
PCO4	.742	.112	4.143*	1.759
PST4	.233	.019	5.421*	1.667
PCO2	.129	.308	9.102***	5.924*
PGE3	.046	.141	1.744	.004
RNS1	1.341		4.705*	
PME4		.010		.749
PNS3	1.934	4.884*	.227	2.152
PES3		.894		2.940
PRP1	.098	2.672	3.323	2.269
Form C				
PES1	4.392*	.004	.000	1.023
PNS4	5.760*	2.922	.270	.455
PCO3	.188	.362	.330	3.412
PCO5	1.380	.143	1.180	.268
PGE4	4.281*	39.296***	.545	.080
RLO1	1.355	.316	.187	2.780
PNS1	8.332**	6.238*	.099	.053
RPC1	8.922**		.038	
RPC1C		7.165*		.106
RNS2	3.123	1.628	.291	.752
Form D				
PCO1	2.898	2.687	.645	2.070
RPN2	.932	.489	.767	.049
PES2	1.144		.827	
RPA3		.475		1.159
PME2	1.416	2.571	1.574	2.398
PGE2	.673	.831	.031	.607
RPA2	.028		.028	
PNS7		.776		2.844
PST3	.004	6.352*	.746	.000
PNS5	1.282	1.352	.789	2.838
PRP4	7.946**	6.205*	2.565	.012

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

Table 3

Distributions of Female and Male Students' Solution Strategies for Task PNS1

Solution Strategy	Females (n=250)	Males (n=210)
Division	70 (28%)	70 (33%)
Repeated addition or multiplication	95 (38%)	60 (29%)
Repeated subtraction	9 (4%)	7 (3%)
Partitioning using drawings	44 (18%)	50 (24%)
Inappropriate strategies	32 (13%)	23 (11%)

Table 4

Distributions of Female and Male Students' Execution of Appropriate Strategies for Task PNS1

Strategy Execution	Females (n=218)	Males (n=187)
Correct execution of a strategy	182 (83%)	137 (73%)
Incorrect execution of a strategy	36 (17%)	50 (27%)

Table 5

Distributions of Female and Male Students' Numerical Answers to Task PNS1

Numerical Answer	Females (n=250)	Males (n=210)
4	146 (58%)	105 (50%)
3	25 (10%)	20 (10%)
3 and a remainder	18 (7%)	31 (15%)
Other answer or no answer	61 (24%)	54 (26%)

Table 6

Distributions of Female and Male Students' Explanations for Task PNS1

Explanation	Females (n=250)	Males (n=210)
Conceptual explanation	121 (48%)	69 (33%)
Procedural explanation	79 (32%)	81 (39%)
Inappropriate or no explanation	50 (20%)	60 (29%)

Table 7

Distributions of Female and Male Students' Solution Strategies for Task RPC1C

Solution Strategy	Females (n=244)	Males (n=213)
Appropriate solution strategy	165 (68%)	109 (51%)
Inappropriate solution strategy	48 (20%)	39 (18%)
No solution strategy displayed	31 (13%)	65 (31%)

Table 8

Distributions of Female and Male Students' Correctness of Numerical Answer to Task RPC1C

Numerical Answer	Females (n=244)	Males (n=213)
Correct numerical answer	164 (67%)	126 (59%)
Incorrect answer or no answer	80 (33%)	87 (41%)

Table 9

Distributions of Female and Male Students' Omissions for Task RPC1C

Omission	Females (n=164)	Males (n=125)
No omissions	104 (63%)	72 (58%)
Minor or moderate level of omissions	49 (30%)	24 (19%)
Many omissions/no work	11 (7%)	29 (23%)

Table 10

Distributions of Female and Male Students' Solution Strategies for Task PRP4

Solution Strategy	Females (n=178)	Males (n=191)
Appropriate solution strategy	56 (31%)	75 (39%)
Inappropriate solution strategy	110 (62%)	83 (43%)
No solution strategy displayed	12 (7%)	33 (17%)

Table 11

Distributions of Female and Male Students' Correctness of Numerical Answer to Task PRP4

Numerical Answer	Females (n=178)	Males (n=191)
Correct numerical answer	34 (19%)	57 (30%)
Incorrect answer or no answer	144 (81%)	134 (70%)

Table 12

Distributions of Female and Male Students' Omissions for Task PRP4

Omission	Females (n= 33)	Males (n= 55)
No omissions	27 (82%)	30 (55%)
Some omissions /no work	6 (18%)	25 (45%)

Table 13

Distributions of Female and Male Students' Correctness of Drawings for Task PGE4

Drawing	Females (n= 242)	Males (n= 223)
No correct drawing	100 (41%)	60 (27%)
1 correct drawing	72 (30%)	68 (30%)
2 correct drawings	26 (11%)	25 (11%)
3 correct drawings	44 (18%)	70 (31%)

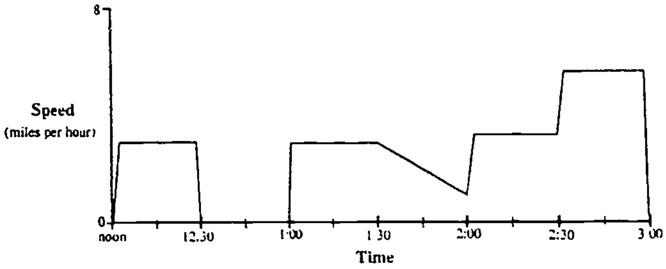
Figure 1

Four QCAI Release Tasks

QCAI Graph Interpretation Task

Use the following information and the graph to write a story about Tony's walk.

At noon, Tony started walking to his grandmother's house. He arrived at her house at 3:00. The graph below shows Tony's speed in miles per hour throughout his walk.



Write a story about Tony's walk. In your story, describe what Tony might have been doing at the different times.

QCAI Decimal Value Task

Circle the number that has the greatest value

.08 .8 .080 00800

Explain your answer

QCAI Number Theory Task

Yolanda was telling her brother Damian about what she did in math class.

Yolanda said, "Damian, I used blocks in my math class today. When I grouped the blocks in groups of 2, I had 1 block left over. When I grouped the blocks in groups of 3, I had 1 block left over. And when I grouped the blocks in groups of 4, I still had 1 block left over."

Damian asked, "How many blocks did you have?"

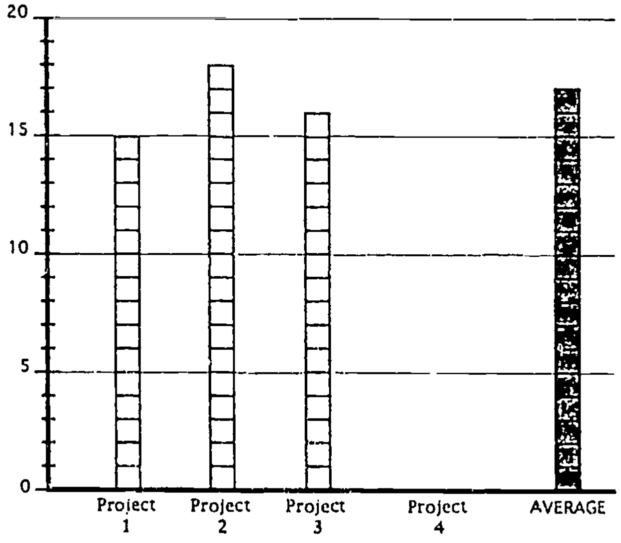
What was Yolanda's answer to her brother's question?

Show your work.

Answer: ______________

QCAI Average Score Task

Anita has four 20-point projects for science class. Anita's scores on the first 3 projects are shown below



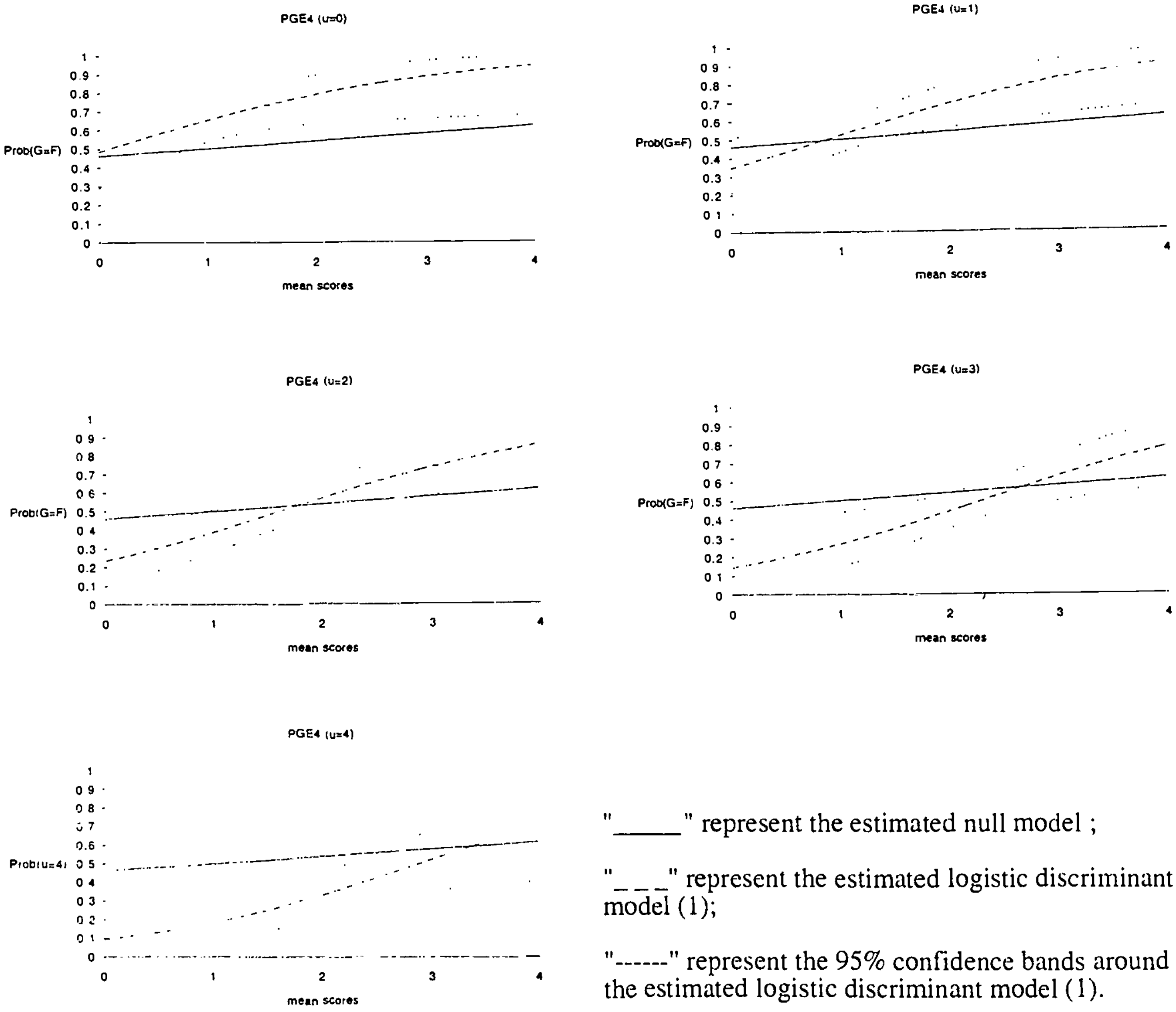
A What score must Anita get on Project 4 so that her average for the four projects is 17?

Answer ________________ You may draw your answer on the graph

B Explain how you found your answer

Figure 2

LDFA Confidence Bands for Task PGE4



"______" represent the estimated null model ;

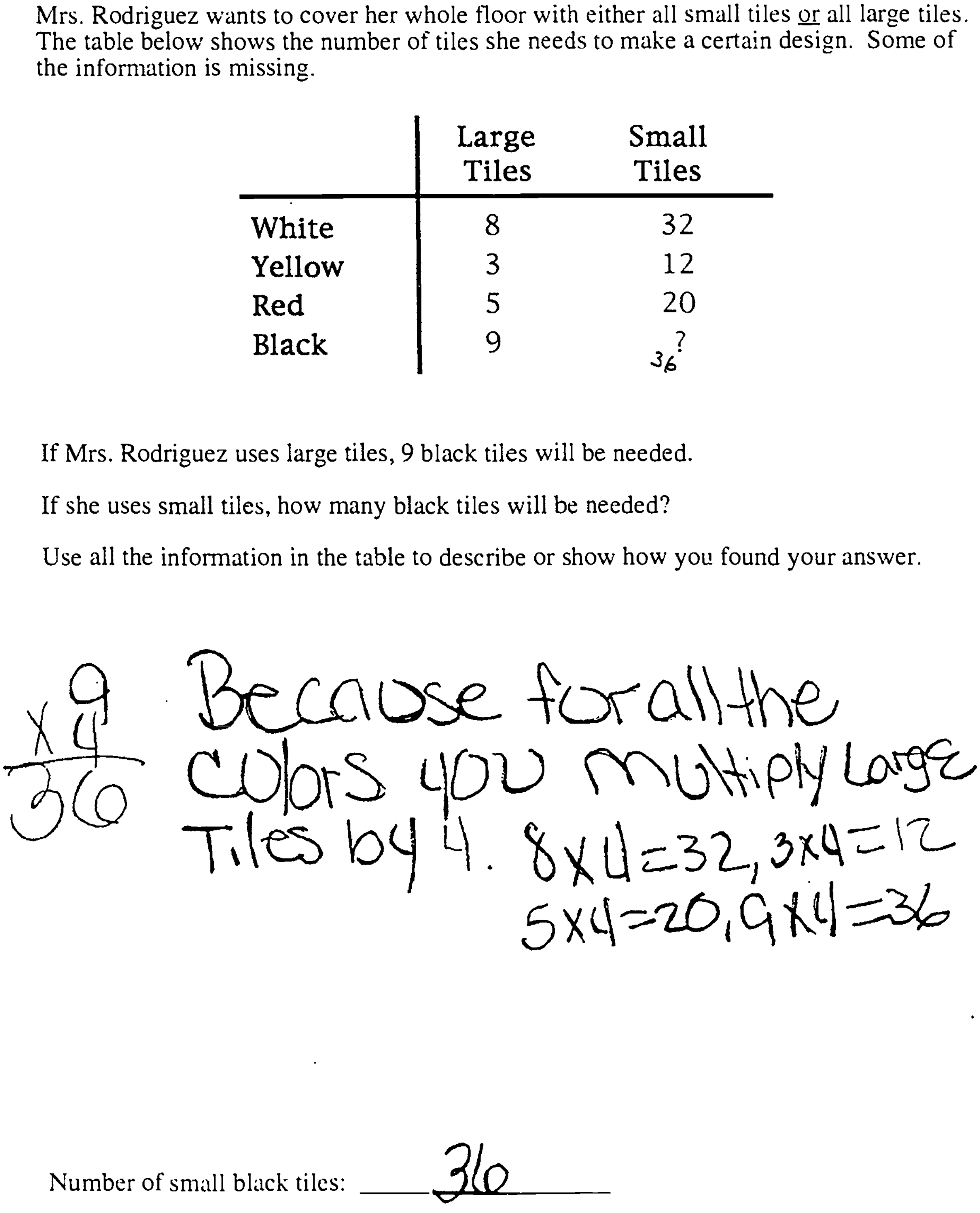
"_ _ _" represent the estimated logistic discriminant model (1);

"------" represent the 95% confidence bands around the estimated logistic discriminant model (1).

BEST COPY AVAILABLE

Figure 3

Task RPC1C and an Appropriate Solution Strategy



Mrs. Rodriguez wants to cover her whole floor with either all small tiles or all large tiles. The table below shows the number of tiles she needs to make a certain design. Some of the information is missing.

	Large Tiles	Small Tiles
White	8	32
Yellow	3	12
Red	5	20
Black	9	? 36

If Mrs. Rodriguez uses large tiles, 9 black tiles will be needed.

If she uses small tiles, how many black tiles will be needed?

Use all the information in the table to describe or show how you found your answer.

$$\begin{array}{r} 9 \\ \times 4 \\ \hline 36 \end{array}$$

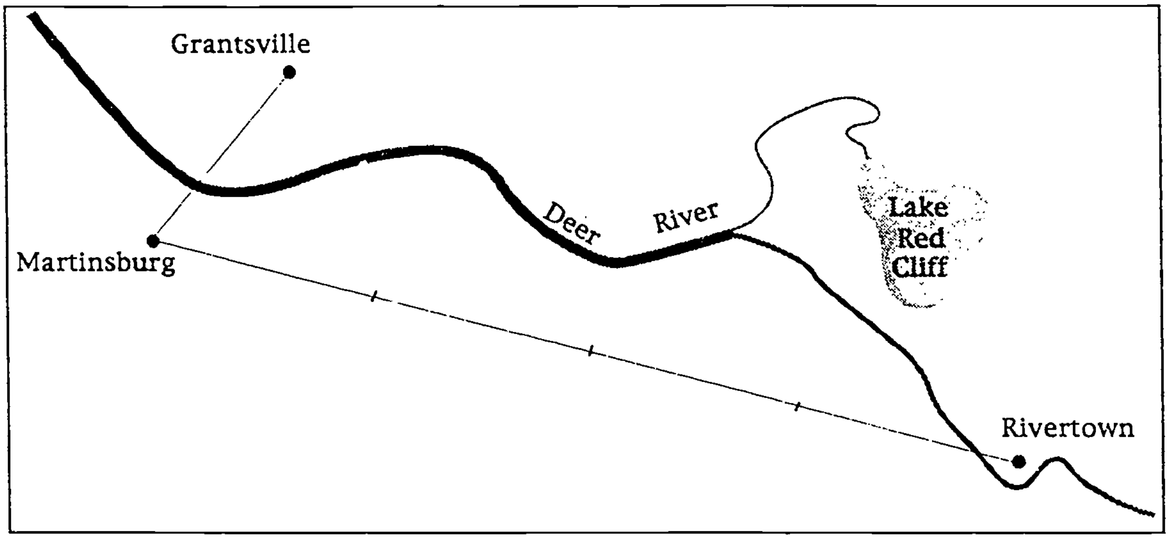
Because for allthe colors you multiply Large Tiles by 4. 8x4=32, 3x4=12 5x4=20, 9x4=36

Number of small black tiles: 36

Figure 4

Task PRP4 and an Appropriate Solution Strategy

The map below shows the locations of three cities.



The actual distance between Grantsville and Martinsburg is 54 miles. On the map, Grantsville and Martinsburg are 3 centimeters apart. On the map, Martinsburg and Rivertown are 12 centimeters apart.

What is the actual distance between Martinsburg and Rivertown?

$$\begin{array}{r} {}^{1} \\ 54 \\ 54 \\ 54 \\ +54 \\ \hline 216 \end{array} \qquad \begin{array}{r} 54 \\ 4\overline{)216} \\ -20 \\ \hline 16 \\ -16 \\ \hline 0 \end{array}$$

I meature Grantsville + Martinburg
And that was 54 then I meature
the space between Martinsburg + Rivertown
and then it was four space between it
So I added it all together and I came
out with 216 then I divided 4 into 216
I got my answer

Answer: 216 miles.

BEST COPY AVAILABLE