Gender and Achievement
-- Understanding Gender Differences and Similarities in Mathematics Assessment

Liru Zhang, Delaware Department of Education
Jon Manon, University of Delaware

Background of the Study

Differences between females and males in mathematics have long been observed and discussed among educators and researchers. Presumably, because of the complexity of gender-related issues and the multi-faceted nature of test performance, results reported from a variety of studies are inconsistent and often even contradictory (Hoover and Han, 1995). Evidence from several comprehensive meta-analytic studies (Willingham and Cole, 1997; Hyde & Linn, 1988; Linn & Hyde, 1989; Hyde, Fennema & Lamon, 1990; Hyde, 1991; Cleary, 1992) suggest that such contradictory results may be accounted for by disentangling effects of different cohorts (different populations by age and grade), construct (knowledge and skills assessed by the test), and selectivity of the sample (self-selected, representative, or available). Cleary (1992) examined gender differences in test performance on commonly used achievement tests across three subject areas including mathematics. The effective size was compared at the levels of quantitativeness of the test and task complexity for age groups, selective and unselective samples, and students at the 10^{th}, 50^{th}, and 90^{th} percentile ranks. She concluded that between-group differences are smaller when compared to differences within each group, but females are generally scored lower relative to males. This disadvantage becomes greater as age increases, task complexity increases, and the quantitativeness of the test increases. Factors such as course selection, test-taking strategies, social environment, and interest patterns have also been investigated as an alternative to explain gender differences in mathematics.

Although evidence support the similarities of overall mathematical performance between females and males in many reports, researchers have increasingly recognized the effects of differences in variability in order to understand better the gender-related issues in mathematics. Recent analyses seem generally consistent, finding greater variability for males than females on many mathematics achievement tests (Willingham and Cole, 1997; Fan et al, 1997; Wang, 1995; Feingold, 1992; Benbow, 1992; Cleary, 1992). As a consequence, there are more males than females scoring in the highest and lowest ranges. When the differences in variability are combined with mean differences in test scores, the gender differences among students in the top 10% of a test can be substantial. In light of the obvious importance of variability, Willingham and Cole (1997) used the female/male standard deviation ratio (SDR) and the ratio of number of females to number of males (F/M) in addition to the standard mean difference (D) in their outstanding study to describe gender differences.

A number of writers investigated the grade or age trend of gender differences and similarities in mathematics. A consistent conclusion over the past two to three decades has been the generally similar achievement between females and males in mathematics in the early grades (Zhang, Wilson & Manon, 1999; Wilson and Zhang, 1998; Willingham and Cole, 1997; Cleary, 1992; Hyde, Fennema & Lamon, 1990). Willingham and Cole (1997), in a summary of recent literature on gender and performance, indicate a general agreement that there is little overall gender difference in mathematics test scores in elementary school, some differences appear in middle school, and average differences favoring males are more common in high school, particularly on problem-solving. Males tended to gain in math scores relative to females as they move through the school grades, especially in high-level thinking tasks (Willingham and Cole, 1997; Cleary, 1992; Hyde et al, 1990; Maccoby and Jacklin, 1974). Data from Project TALENT in the 1960s and NAEP in the 1980s support the grade trend that the greater gain by males on problem solving and reasoning occurred in high school, particularly in grades 11 and 12.

Over the years, research data indicate that the nature of differences in mathematics performance between females and males can be masked or distorted by looking at total score alone and paying no attention to the construct components of the test (Cole, 1997; Lane, Wang & Magone, 1996; Wang & Lane, 1996; Ryan & Fan, 1994; Cleary, 1992; Doolittle & Cleary, 1987). Many analyses have attempted to identify item features, such as content, format, context, and cognitive process required, that are related to differential performance. For example, some reported that males outperformed females in the content areas of geometry, ratio, proportion, and percent, and arithmetic/algebraic reasoning involving a real-world context; females tended to perform better in algorithmic or computational skills; but showed little difference on general knowledge of mathematics.

The recent educational reform of curriculum in mathematics focuses on high-level thinking, such as problem solving and reasoning rather than rote memorization and computation. With the increasing use of performance assessment (e.g., constructed-response items) in the local, state, and national assessment programs, evidence is needed to ensure that inferences made from the measures are equally valid for different sub-groups in the population (Linn, Baker & Dunbar, 1991). In examining the format effects, Willingham and Cole (1997) reviewed various types of tests that provided both multiple-choice and constructed-response items. The data indicate a format effect as a discrepancy in the gender difference on a CR (constructed-response) test compared to the gender difference on a MC (multiple-choice) test in the same subject for the same sample of students. However, consequential format effects were seldom found in mathematics. Comparing three statewide mathematics assessments, the results suggest that there were slight format effects favoring females on CR items in Kentucky and a notably larger format effect in the same direction in Kansas. The authors called attention to the characteristics of the particular CR test when only a few items are used for a given test. "It seems wiser to regard such results as illustrative rather than the basis for conclusion about format effects." Constructed-response, free-response, and open-ended items are used interchangeably throughout this paper.

Purposes of the Study

The primary objective of this study was to investigate the overall patterns of gender differences and similarities of test performance in mathematics. To achieve that objective, observed test scores on the Delaware standard-based assessment were analyzed to examine (1) gender differences and similarities across grades 3, 5, 8, and 10 over two years; (2) the existence and magnitude of the effects of item format on gender performance; and (3) the existence and magnitude of the effects of test construct, content standards and cognitive processes, on gender performance.

Methodology of the Study

<u>*Assessment Instrument*</u> The Delaware Student Testing Program (DSTP) mathematics assessment is designed to measure students' progress toward the Delaware Content Standards (See *Appendix A: Delaware Content Standards in Mathematics*) in grades 3, 5, 8, and 10. Multiple-choice (MC), short answer (SA), and extended constructed-response (ECR) items are used to measure concepts and procedures in computation, estimation, number sense, patterns, algebra, and functions, geometry, probability and statistics on three cognitive processes, conceptual knowledge, procedural knowledge, and mathematical process (or problem solving) (See *Appendix B: Definitions of Cognitive Processes*). The Stanford Achievement Series, 9^{th} edition (SAT9) abbreviated version in mathematical problem solving is administered to the same grades as part of the DSTP. SAT9 scores are used to compare Delaware students' achievement with the nation and combined with the Delaware developed items to derive a standard-based score in mathematics. Number of items, item format, and the maximum possible points (MP) in the categories of content standards and cognitive processes are presented in Tables 1a to 1d by grade and test form. About 30% of the Delaware developed items were replaced in the 1999 test form from the 1998 test form.

<u>*Item Format and Scoring Rubrics*</u> The DSTP mathematics assessment consists of three item formats, multiple-choice, short answer, and extended constructed-response items (See *Appendix C: Item Format*). The traditional multiple-choice item format has four response options and is scored dichotomously. Short answer items require students to reflect on a situation and communicate briefly the reasoning behind their solutions with a couple of words or sentences. The extended constructed-response item format requires students to consider a situation within or across the content strand areas, solve the problem, and provide evidence of their solution strategies. General and specific scoring rubrics (See *Appendix D: General Rubrics and Sample Items*) are developed for scoring short answer and extended constructed-response items on a 2-point scale and a 4-point scale respectively. One trained reader was used to evaluate each student's response in both 1998 and 1999.

<u>*Test Administration*</u> The DSTP mathematics assessment was given in three sessions. The SAT9 was administered in a 28-minute session of the first day under standardized testing conditions followed by an approximately 60-minutes session for the Delaware developed items. A 10-minute for the rest period between the two sessions was recommended. In day 2, students took another 60-minute session for the Delaware portion. Commonly used mathematical formulas for grades 8 and 10 were provided as a reference during testing (See *Appendix E: Mathematics Reference Sheet*). Calculators, graphing calculators for grade 10, were allowed for one session of the Delaware portion.

Sample of Subjects Students in grades 3, 5, 8, and 10 who received a valid score on the 1998 and 1999 SAT9 and the Delaware portion were included for data analyses in this study.

Data Analysis Descriptive statistics, means, standard deviations, and frequency distributions of test scores on the SAT9 and the DSTP, were calculated for the total group, females, and males by grade and year of testing. Data analyses focused mainly on three measures of gender difference and similarity: the standard mean difference (D = (female mean - male mean)/pooled standard deviation), the female and male ratio (F/M = number of females/number of males), and the standard deviation ratio (SDR = female standard deviation/male standard deviation) by grade, test construct, item format, and student achievement. The Point Ratio (PR = average score/maximum points) as an indicator of the average test difficulty was reported for selected comparisons.

Results and Discussion

Means, standard deviations, and frequency distributions of test scores were calculated for each comparison group. The index D is referred to as the standard mean difference. The measure varies about zero. That is, when there is no gender difference, D is zero; if females have the higher mean score, D is positive; if males have the higher mean score, D is negative. D is comparable from measure to measure, even if the score scales on the measures are different. A widely used criterion for the interpretation of D-values proposed by Cohen (1988) evaluates .20 - .49 is a small difference, .50 - .79 is medium, and .80 or higher is large. Many researchers, however, strongly encouraged using these criteria as reference only and considering the particular situation of a given test to avoid misleading (Willingham and Cole, 1997; Glass et al, 1981). In this study, D-values are interpreted in the latter way. SDR is the ratio of female to male standard deviation. The value of SDR indicates the difference in variability of test scores between genders. A value of SDR greater than 1.00 means that females' test scores are more variable than males'; a value of SDR smaller than 1.00 means that males' test scores are more variable than females'. The F/M is the ratio of number of females to males among the 10% of high and low scores on the SAT9 mathematical problem solving test and the DSTP mathematics assessment.

1. Descriptive statistics and the indexes of D, SDR, and F/M on the 1998 and 1999 DSTP mathematics assessments are summarized by gender and grade in Tables 2a and 2b. The standard difference means (D) ranged from -.06 to .03 across grades 3, 5, 8, and 10 over two years. A slight male advantage in mean score in grades 3 (D = -.06) and 8 (D = -.04) was found in 1999. The near zero D-value indicates that there is no gender difference in overall performance. The SDRs show a consistent pattern that female standard deviations are about 95%, on the average of the four grades, as large as that of male standard deviations across years. The value of SDR appears to decline from .99 to .87 from grade 3 to grade 10 in 1999.

Similar patterns are observed from the analyses for the SAT9 mathematical problem-solving test (Tables 3a and 3b). The D-value ranging from -.11 to .01 suggests that there is no significant gender difference on the overall performance across grades 3, 5, 8, and 10. Males performed, however, a little better than females in grades 8 (D = - .09) and 10 (D = -

.11) in 1999. In viewing the two years data, variability of test scores for males is consistently larger than that for females. On the average, female standard deviations are about 94% as large as that of male standard deviations. Similarly, the size of SDR drops slightly from grade 3 to grade 10. For example, the SDR of .98 for grade 3 and .88 for grade 10 in 1999 simply indicates greater gender difference of variability in higher grades.

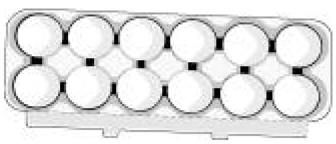
2. The current educational reform has great impacts on teaching and learning; however, the implementation of the new curriculum usually takes a long time and the stage of the implementation may vary from school to school. From such considerations, the PR (the ratio of the mean score to the possible maximum points) was calculated on the Delaware developed portion and SAT9 to explore the possible interactive effects of test difficulty with gender performance. Data presented in Tables 4a and 4b show the results of analyses for the same group of students. It is apparent that the Delaware portion is more difficult (PR = .57 to .62 in grade 3; PR = .35 to .27 in grade 10) than SAT9 (PR = .63 to .67 for grades 3 and 5; PR = .43 to .50 for grades 8 and 10) within each grade. What is the corresponding relationship between test difficulties with gender performance? Can gender difference be masked by the difficult test, especially in grades 8 and 10 in this case? No sufficient evidence or clear pattern, however, was found in this study for any reasonable explanations.

3. Gender differential performance in the top and the bottom 10% illustrates an interesting, but different picture from the overall performance on the DSTP (Tables 2a and 2b). First, the D-values are noticeably more variable in the range of -.26 to .16 across the four grades. In grade 3, a positive D of .16 and .10 indicates a higher mean score for females at the bottom and the top 10%, respectively. Similarly, female mean scores are higher than males among both low-achieving and high-achieving students in grade 5 in 1999. In grades 8 and 10, females continued outperforming to males but only at the lower end (D = .14 for grade 8 in 1999; D = .09 for grade 10 in 1999); while males outperformed females at the higher end (D = -.06 for grade 8 and D = -.16 for grade 10 in 1998; D = -.26 for grade 10 in 1999). Second, greater male variability is present through grades at the top and the bottom of the score distributions with only a few exceptions. On the average, female standard deviations are about 95% as large as corresponding male standard deviations. Third, the F/M ratios show a pattern of considerably more males than females among the 10% of low-achieving and high-achieving students in each grade. This pattern becomes clearer in grades 8 and 10 (F/Ms range from .82 to .99; F/Ms range from .68 to .85 for grades 8 and 10). This quite different picture at the mean and in the tails well illustrates why it can be insufficient to examine gender difference only using the average score (Willingham and Cole, 1997).

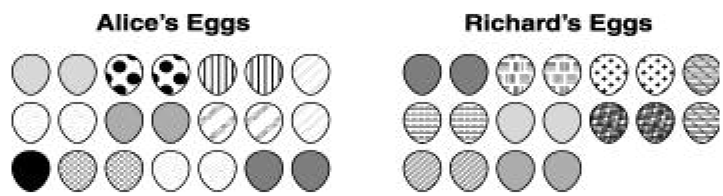
There are several interesting features to the overall pattern of gender performance on the SAT9 mathematical problem-solving test. First, males consistently performed better than females at the top 10% with a D in the range of -.02 to -.16 in 1998 and a D in the range of -.05 to -.20 in 1999; whereas females performed better than males at the bottom 10% with a D in the range of .00 to .17 for most grades (Tables 3a and 3b). Second, the SDR ranging from .87 to near 1.00 and the F/M ranging from .57 to .89 suggest a greater variability of test scores with an excess of males found at the top and the bottom of the distributions through grades and years.

4. The DSTP mathematics assessment consists of three item formats, multiple-choice (MC), short answer (SA), and extended constructed-response (ECR) items. MC items require students to choose the correct answer from the four responses provided; both SA and ECR items typically require students to reflect on a real-world situation. SA items require students to explain their solutions with a couple of words or sentences; ECR items require extended written response and evidence to support their solution strategies in problem solving.

Below is an example of a released ECR item for grade 3 (Appendix D). This item measures several components of number sense. Students may use different strategies to solve the problem in the context involving a physical representation of number. A correct answer with clear verbal or pictorial explanations is required to receive a 4-point score.



Alice and Richard are coloring eggs and putting them back into the egg cartons. Alice colored 21 eggs and Richard colored 18 eggs. (There are 12 eggs in a carton.)



How many cartons will they need to hold all the eggs? Explain how you got your answer.

Tables 5a and 5b show the results of analyses of gender differences by item format. First, all negative D-values, ranging from -.11 to -.01, indicate a pattern of slight male advantage in mean score for MC items. This result is generally similar to previous findings. Gender differences for ECR items tended to slightly favor females in grades 3 (D = .10) and 5 (D = .09) in 1998 and in grades 5 (D = .18) and 10 (D = .07) in 1999. The largest D-value is .04 for SA items that suggests no gender difference on short answer items. It is also noticed that the DSR increases slightly from MC to SA to ECR items for most cases within each grade of 5, 8, and 10 across years. Further, D_{ECR-MC}, the difference of D-values for ECR from MC, was calculated and compared. Positive values of ECR-MC indicate that females, compared to males, do relatively better on extended constructed-response items than on multiple-choice items. Here, all the values of D_{ECR-MC} in the same direction (D_{ECR-MC} = .11, .11, .04, and .03 in 1998; D_{ECR-MC} = .03, .23, .10, and .18 in 1999 for grades 3, 5, 8, and 10 respectively) present a consistent pattern of slight format effect favoring females on ECR items. Apparently, gender differences on different item formats are not uniform in direction or size of Ds. It is worth repeating based on the evidence from the current analyses that males appear to do somewhat better than females on multiple-choice; females appear to do somewhat better than males on extended constructed-response items; females and males do nearly equally well on short answer items. These results support early findings that item format involving the ability of writing (especially extended writing) may favor females. Such connection further suggests that different item formats may measure different constructs, where format effects associated with writing skills contribute to differential gender performance.

5. The results compiled in Tables 6a to 6c give a picture of differential gender performance by content standards. Unlike the categories used in early studies, the present analyses were based on broader content areas specified in each standard (See *Appendix A: Delaware Content Standards in Mathematics*). The negative values of Ds suggest that males consistently outperformed females slightly on estimation, measurement and computation across grades and years. These differences became more noticeable in 1999 than the previous year (D = -.09 for grade 3; D = -.06 for grade 5; D= -.07 for grade 8; D = -.11 for grade 10). For the standard of number sense, females and males performed equally well in grade 3, a slight advantage on the mean score favored females in grade 5, but in grade 8, males outperformed females (D = -12 in 1998; D = -.15 in 1999). Performance between females and males presented a mixed picture on spatial sense and geometry, where males received a higher mean score in grade 3 (D = -.15 in 1998) and grade 8 (D = -.11 for 1998; D = -.09 for 1999), but almost no gender differences in grades 5 and 10 (Ds are ranged from -.01 to .04). Further, the near zero D-values for algebra, statistics and probability, and patterns, relationship and functions suggest no gender differences exist in these areas. Although male variability is still larger than female's in most cases, the SDR is near 1.00 in grades 3, 5, and 8 across algebra, statistics and probability, and patterns, relationship and functions. Note also that the size of SDR drops in grade 10 across the five content standards, indicating larger variability for males at a higher-grade level.

These mixed patterns and inconsistencies may be explained by three possible reasons. First, each standard contains broad content domains which are measured with a limited number of items. The question here is whether the standard scores provide a reliable and sufficient picture of gender performance. Second, instead of focusing on a single theme, constructed-response items tend to measure more complex knowledge and skills at the higher thinking level and usually have more than one definite answer. One item may be assigned to more than one category. In examining the effects of test constructs, it seems important to categorize the item in an appropriate and consistent way to reflect the content and cognitive process measured by the item. Some suggest categorizing the item on all the knowledge and skills measured; some suggest categorizing the item on the dominant area only. Third, some items may assess irrelevant constructs.

6. The Mathematics Content Standards provide the framework of what types of knowledge and skills are to be measured in the assessment. Data reported in Tables 7a and 7b show test performance of females and males on three cognitive processes: Conceptual Knowledge, Procedural Knowledge, and Mathematical Processes (or Problem Solving). As can be seen in the tables, standard mean differences vary widely across the three cognitive categories from -.11 in grade 5 to .08 in grade 3 on mathematical processes. Even within the cognitive category, there is some variation in gender differences. Data in Table 7s also show larger male standard deviations across cognitive categories, grades, and years of testing with only a few exceptions. The size of the SDRs between females and males drops from grades 3 and 5 to grade 8, especially in grade 10.

Findings of the Study

The primary objective of this study was to examine the overall patterns of gender differences and similarities of test performance in mathematics. To achieve the objective, gender performance was analyzed using the observed test scores on the 1998 and 1999 Delaware statewide assessments by item format, student achievement, content standards, and cognitive processes across grades 3, 5, 8, and 10. Four main findings are summarized below based on the current analyses. These findings, however, must be viewed and interpreted within the limitations of the design and analyses of the current study.

1. The similarities of the overall test performance and distributions of test scores between females and males indicate no gender differences on the DSTP mathematics assessment and on the SAT9 mathematics problem-solving test. Data, however, reveal a tendency to somewhat greater variability in test scores of males than females in mathematics across grades. Consistent with previous findings, a relative increase in the variability of male scores from grade 3 to grade 10 was observed in this study.

2. Evidence from the current analyses suggest that gender differences do exist when the significant differences in variability are combined with mean differences among students scoring at the top and the bottom 10% of the DSTP and the SAT9. In grades 8 and 10, females tended to outperform males among the low-achieving students; while males tended to outperform females among the high-achieving students.

3. The results of data analyses indicate a slight, but consistent format effect across grades, that males performed better than females on multiple-choice; females performed better than males on extended constructed-response items. All positive $D_{ECR\text{-}MC}$-values (from .03 to .23) further provide evidence to support the notion that item format involving the ability of writing, particularly extended constructed-response format in this study, favor females. Such connection suggests that different item formats may measure different constructs, where format effects associated with writing skills contribute to differential gender performance.

4. Unlike a variety of early studies, no clear patterns of test performance between females and males have been found in this study by the content and cognitive categories in mathematics. A mixed picture shows that males consistently outperformed females slightly on estimation, measurement and computation across grades; males also somewhat performed better than females on number sense in grade 8 and on spatial sense and geometry in grade 3 and 5. These results may be due to the broader content areas (or knowledge and skills) covered by each standard but measured with a limited number of items.

Educational Importance of the Study

Test performance of males and females has been one of the major research topics of historical interest in education, and it continues engaging scholars, educators, researchers, and society at large. In designing a test, the fundamental concern is validity, whether the test measures what it is intended to measure and whether it provides all students with an equal opportunity to demonstrate what they know and are able to do. Many educators point out that standardized tests, particularly traditional norm-referenced, multiple-choice tests are usually designed to assess generic skills with indirect measures, that are decoupled from any specific curriculum or instructional goals of the schools (Linn, 1993; Burton, 1996). Although there have been an abundance of research inspecting gender differences and similarities, few studies have examined gender performance on assessments measuring high-level thinking and reasoning in mathematics. With the increased use of performance assessment in large-scale testing programs in recent years, there is a great need for evidence to ensure that inferences made from the measures are equally valid for different sub-groups in the population (Linn, Baker & Dunbar, 1991). The challenge is understanding gender differences and similarities, questions such as whether there is possible format effect; how item format changes the test construct in ways that tend to favor either male or female; and how the content standard and cognitive process of a given test change the patterns of gender performance. Such challenges become critical when test performance is used as the primary indicator for high-stakes decisions, such as retention, summer school, and graduation. The educational implications of the current study can be considered as following:

First, the results of this study provide evidence to confirm the dominant pattern of similarity of test performance between females and males on a statewide mathematics assessment.

Second, concerns are raised regarding the differential gender performance in the high-scoring and low-scoring regions of distributions because the test scores serve as the primary indicator for high-stakes decisions of most statewide assessment programs. Consequently, there are more male than female students at the extreme score levels -- in the top selected group of students for rewards and in the bottom selected group of students for sanctions when multiple cut-off scores are used even if there is no mean gender difference.

Third, with the growing use of performance assessment in national, state, and local large-scale assessments, research evidence concerning the format effects on differential gender performance is needed to develop a better understanding of the relationship between item format, particularly constructed-response items, and its impact on test constructs.

References

Bielinski, J., & Davidson, M. L. (1997). Gender differences by item difficulty interactions in multiple-choice mathematics items. Paper presented at the Annual Meeting of AERA, Chicago, IL.

Burton, N. (1996). Have changes in the SAT affected women's mathematics performance? *Educational Measurement: Issues and Practice*, Winter 1996.

Carpenter, T. P. & Fennema E. (1998). New perspectives on gender differences in mathematics: An Introduction. *Educational Research*, 27(5), 4-5.

Cleary, T. A. (1992). Gender differences in aptitude and achievement test scores. Paper presented at the 1991 ETS Invitational Conference on Sex Equity in Educational Opportunity, Achievement and Testing. Princeton, NJ.

Cohen, J. (1988). *Statistical power analysis for the behavioral science (2nd ed.).* Hillsdale, NJ: Lawrence Erlbaum Associates.

Delaware Student Testing Program -- State Summary Report, 1998 Administration (January, 1999). Delaware Department of Education

Delaware Student Testing Program -- State Summary Report, 1999 Administration (January, 2000). Delaware Department of Education

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

Droegemueller, L., Poggio, J. P., Glasnapp, D. R., and Capps, L. R. (1992). Report of results from the 1991 Kansas Mathematics Pilot Assessment. Paper presented as part of the Symposium titled "Statewide Assessment of the New Mathematics Standards: Issues, Procedures, Results and Policy Implications' at the 1992 NCME Annual Meeting, San Francisco, CA.

Englehart, Jr., G. (1990). Gender differences in performance on mathematics items: Evidence for the United States and Thailand. *Contemporary Educational Psychology*, 15, 13-26.

Fan X., Chen, M., & Matsumoto, A. (1997). Gender differences in mathematics achievement: Findings from the National Education Longitudinal Study of 1988. *The Journal of Experimental Education.* 65(2), 229-242.

Fennema E. Carpenter T. P., Jacobs V. R., Franke, M. L. & Levi L. W. (1998). A longitudinal study of gender differences in young children's mathematical thinking. *Educational Research* 27(5), 6-11.

Fennema E. Carpenter T. P., Jacobs V. R., Franke, M. L. & Levi L. W. (1998). New perspectives on gender differences in mathematics: A reprise. *Educational Research* 27(5), 19-21.

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(2), 195-213.

Glass, G. V. et al (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Hoover, H. D. and Han, L. (1995). The effect of differential selection on gender differences in college admission test scores. Paper presented at the 1995 AERA Annual Meeting, San Francisco, CA.

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

Lane, S., Wang N., & Magone M. (1996). Gender-related differential item functioning on a middle-school mathematics performance assessment. *Educational Measurement: Issues and Practice*, Winter, 1996.

Linn, M. C. and Petersen, A. C. (1986). A meta-analysis of gender differences in spatial ability: Implications for mathematics and sciences achievement. In J. S. Hyde & M. C. Linn (Ed.), *The Psychology of Gender: Advances through Meta-Analysis* (pp. 67-101). The John Hopkins University Press, Baltimore.

Linn, M. C. and Hyde, J. S. (1989). Gender, mathematics, and science. *Educational Researcher*. November 1989.

Linn, R. L. & Baker, E. L. and Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, Vol. 20, No. 8, pp. 15-21

Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis*, Spring 1993, Vol. 15, No. 1, pp. 1-10.

Linn, R. L. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher*, Vol. 23, No. 9, pp. 4-14

Maccoby, E. & Jacklin, C. (1974). The psychology of sex differences. Stanford, CA: *Stanford University Press*.

Mathematics Curriculum Framework -- Content Standard (1995), Volume One. State of Delaware.

O'Neil, H. F., Jr. and Brown, R. S. (1997). Differential effects of question formats in math assessment on metacognition and affect. CSE Technical Report 449, National Center for Research on Evaluation, Standards, and Student Testing (CRESST) Center for the Study of Evaluation (CSE), University of California, Las Angeles.

Ryan, K. E. & Fan, M. (1996). Examining gender DIF on a multiple-choice test of mathematics: A confirmatory approach. *Educational Measurement: Issues and Practice*, 15, 4.

Ryan, K. E. (1999). Revisiting gender and achievement. *Educational Researcher*, June-July.

Shavelson, R. J., Baster, G. P., and Gao, X. (1993). Sampling variability of performance assessment. *Journal of Educational Measurement*, 30(3), 215-232.

Sowder, J. T. (1998). Perspectives from mathematical education. Educational Research 27(5). 12-13.

Stanley, J. C. (1993). Boys and girls who reason well mathematically. Paper presented in London at the Ciba Foundation Symposium on the Origins and Development of High Ability.

Tate, W. (1997). Race-ethnicity, SES, gender, and language proficiency trends in mathematics achievement: An update. *Journal for Research in Mathematics Education*, 28(6), 652-679.

Wainer H. and Steinberg, L. S. (1991). Sex differences in performance on the mathematics section of the Scholastic Aptitude Test: A bi-directional validity study. Technical Report No. 91-12, Educational Testing Service, Princeton, NJ.

Wang, N. and Lane S. (1996). Detection of gender-related differential item functioning in a mathematics performance assessment. *Applied Measurement in Education*, 9(2), 175-199.

Wang, X. (1993). Gender differences in spatial and verbal abilities and mathematics performance for Chinese students: A cross-cultural comparison. Paper presented at the 1993 NCME Annual Meeting, Atlanta, GA.

Wang, X. and Maxey J. (1995). Gender differences in the ACT mathematics tests: A cross cultural comparison. Paper presented at the 1995 Annual NCME Meeting, San Francisco, CA.

Wilson, L. & Zhang L. R. (1998). A cognitive analysis of gender differences on constructed-response and multiple-choice assessments in mathematics. Paper presented at the 1998 AREA Annual Conference, San Diego, CA.

Willingham, W. W. and Cole, N. S. (1997). *Gender and Fair Assessment*. Lawrence Erlbaum Associates, Publishers.

Zhang, L. R., Wilson, L., & Manon, J. (1999). An analysis of gender differences on performance assessment in mathematics -- A follow-up study. Paper presented at the 1999 AERA Annual Conference, Montreal, Canada.

Table 1a

Number of Items (Maximum Points) by Grdae, Content Standard, and Year

Content Standards	**Grade 3**		**Grade 5**		**Grade 8**		**Grade 10**	
	1998	*1999*	*1998*	*1999*	*1998*	*1999*	*1998*	*1999*
Estimation, Measurement & Computation	15 (17)	15 (17)	15 (17)	15 (17)	14 (16)	12 (13)	18 (19)	17 (18)
Number Sense	13 (18)	13 (18)	12 (18)	12 (18)	11 (14)	12 (16)	4 (7)	4 (7)
Algebra	5 (6)	5 (6)	5 (6)	5 (6)	12 (14)	13 (18)	14 (15)	14 (15)
Spatial Sense & Geometry	10 (14)	10 (14)	9 (11)	9 (11)	6 (8)	7 (9)	7 (8)	8 (9)
Statistics & Probability	8 (11)	8 (9)	10 (14)	10 (14)	11 (15)	11 (15)	12 (17)	12 (17)
Patterns, Relationship & Functions	9 (11)	9 (13)	9 (11)	9 (11)	7 (11)	6 (7)	6 (12)	6 (12)
Total	***60 (77)***	***60 (77)***	***60 (77)***	***60 (77)***	***61 (78)***	***61 (78)***	***61 (78)***	***61 (78)***

Table 1b

Number of Items (Maximum Points) by Grdae, Cognitive Process, and Year

	Grade 3		**Grade 5**		**Grade 8**		**Grade 10**	
	1998	*1999*	*1998*	*1999*	*1998*	*1999*	*1998*	*1999*
Conceptual Knowledge	24 (30)	24 (28)	24 (29)	23 (26)	25 (29)	26 (30)	25 (31)	26 (31)
Procedual Knowledge	28 (32)	28 (35)	28 (32)	28 (31)	26 (28)	26 928)	28 929)	28 (31)
Mathematical Process	8 (15)	8 (14)	8 (16)	9 (20)	10 (21)	9 (20)	8 (18)	7 (16)
Total	***60 (77)***	***60 (77)***	***60 (77)***	***60 (77)***	***61 (78)***	***61 (78)***	***61 (78)***	***61 (78)***

Table 2a
Test Performance on the DSTP Mathematics Assessment
by Gender, Achievement, Grade, and Year

Gender		*1998*						*1999*					
		N.	**Mean**	**S.D.**	**D**	**F/M Ratio**	**SDR**	**N.**	**Mean**	**S.D.**	**D**	**F/M Ratio**	**SDR**
Grade 3	*MP = 77*												
Total		7971	45.85	13.90	0.03	0.94	0.97	8097	48.95	14.06	-0.06	0.94	0.99
female		3871	46.03	13.71				3915	48.54	13.95			
male		4100	45.68	14.07				4182	49.32	14.16			
Bottom 10%													
female		345	21.37	4.12	0.16	0.82	0.94	363	22.38	4.36	-0.04	0.92	0.95
male		422	20.71	4.36				393	22.55	4.59			
Top 10%													
female		414	67.34	2.84	-0.05	0.90	0.98	441	69.30	2.72	0.10	0.82	1.02
male		462	67.49	2.91				540	69.04	2.67			
Grade 5	*MP = 77*												
Total		7919	41.73	14.36	0.01	1.01	0.96	8243	40.74	14.46	-0.02	0.94	0.98
female		3971	41.81	14.04				3985	40.74	14.35			
male		3948	41.65	14.68				4258	41.00	14.57			
Bottom 10%													
female		348	16.56	3.57	0.03	0.86	0.98	390	16.39	3.23	0.14	0.99	1.00
male		403	16.45	3.64				392	15.94	3.22			
Top 10%													
female		377	64.43	2.89	-0.07	0.90	0.89	433	64.72	3.56	0.08	0.87	0.95
male		420	64.65	3.24				497	64.43	3.76			

Table 2b
Test Performance on the DSTP Mathematics Assessment by Gender, Achievement, Grade, and Year

Gender		*1998*						*1999*				
	N.	Mean	S.D.	D	F/M Ratio	SDR	N.	Mean	S.D.	D	F/M Ratio	SDR
Grade 8 *MP = 78*												
Total	8235	32.88	15.51	-0.03	0.93	0.96	8226	34.54	14.96	-0.04	0.98	0.93
female	3973	32.67	15.15				4078	34.21	14.40			
male	4262	33.07	15.83				4148	34.87	15.49			
Bottom 10%												
female	331	11.60	2.15	-0.04	0.80	0.99	345	13.26	2.49	0.14	0.77	0.92
male	412	11.68	2.17				448	12.90	2.72			
Top 10%												
female	391	63.19	5.43	-0.06	0.82	1.00	374	63.87	5.15	-0.01	0.82	0.94
male	477	63.51	5.45				455	63.91	5.45			
Grade 10 *MP = 78*												
Total	7140	26.38	13.61	0.00	1.00	0.92	7469	29.49	13.63	-0.04	0.97	0.87
female	3570	26.36	13.02				3679	29.18	12.62			
male	3570	26.39	14.18				3790	29.78	14.54			
Bottom 10%												
female	253	9.48	1.59	0.03	0.74	0.96	273	10.95	2.07	0.09	0.77	0.95
male	341	9.43	1.65				356	10.75	2.17			
Top 10%												
female	358	54.33	6.89	-0.16	0.85	0.93	321	56.61	5.96	-0.26	0.68	0.88
male	422	55.49	7.39				469	58.27	6.80			

Table 3a
Test Performance on the SAT9 Mathematics Test

Gender	1998						1999					
	N.	Mean	S.D.	D	F/M Ratio	SDR	N.	Mean	S.D.	D	F/M Ratio	SDR
Grade 3 *MP = 30*												
Total	7971	19.04	5.67	0.01	0.94	0.97	8097	20.00	5.56	-0.05	0.94	0.98
female	3871	19.08	5.58				3915	19.85	5.49			
male	4100	19.00	5.75				4182	20.14	5.62			
Bottom 10%												
female	359	8.26	1.92	0.14	0.79	1.01	282	8.24	1.89	0.10	0.84	1.00
male	457	7.99	1.91				336	8.06	1.89			
Top 10%												
female	516	27.09	1.11	-0.09	0.89	0.99	639	27.00	1.14	-0.20	0.79	1.03
male	580	27.19	1.12				814	27.22	1.11			
Grade 5 *MP = 30*												
Total	7919	19.26	6.16	0.00	1.01	0.94	8243	19.67	6.00	-0.03	0.94	0.96
female	3971	19.25	5.96				3985	19.56	5.86			
male	3948	19.27	6.35				4258	19.76	6.12			
Bottom 10%												
female	538	8.50	2.13	0.09	0.84	0.98	479	8.73	2.11	0.17	0.78	0.95
male	639	8.31	2.17				613	8.35	2.22			
Top 10%												
female	654	27.36	1.21	-0.14	0.85	0.96	711	27.44	1.24	-0.05	0.81	0.96
male	767	27.53	1.26				879	27.50	1.29			

Table 3b
Test Performance on the SAT9 Mathematics Test
by Gender, Achievement, Grade, and Year

Gender	1998						1999					
	N.	Mean	S.D.	D	F/M Ratio	SDR	N.	Mean	S.D.	D	F/M Ratio	SDR
Grade 8 *MP = 30*												
Total	8235	15.10	6.15	-0.03	0.93	0.94	8226	15.11	6.04	-0.09	0.98	0.94
female	3973	15.01	5.97				4078	14.83	5.84			
male	4262	15.19	6.32				4148	15.38	6.22			
Bottom 10%												
female	403	5.63	1.47	-0.06	0.78	1.08	422	5.62	1.38	0.00	0.84	0.95
male	514	5.72	1.36				503	5.62	1.46			
Top 10%												
female	405	26.03	1.62	-0.02	0.75	0.92	394	25.93	1.61	-0.13	0.76	0.87
male	537	26.07	1.76				517	26.15	1.86			
Grade 10 *MP = 30*												
Total	7140	13.31	5.92	-0.02	1.00	0.92	7469	12.86	5.84	-0.11	0.97	0.88
female	3570	13.24	5.67				3679	12.53	5.44			
male	3570	13.38	6.16				3790	13.18	6.19			
Bottom 10%												
female	341	4.99	1.20	0.02	0.78	0.95	442	4.87	1.25	-0.01	0.88	1.01
male	437	4.97	1.26				500	4.88	1.24			
female												
male	277	25.40	1.99	-0.16	0.78	0.95	224	25.42	2.11	-0.17	0.57	0.98
Male	356	25.72	2.10				396	25.79	2.15			

Table 4a
Test Performance on the Delaware Portion and SAT9
by Gender, Grade, and Year

Gender	1998						1999					
	N.	Mean	S.D.	D	PR	SDR	N.	Mean	S.D.	D	PR	SDR
Grade 3	*DSTP MP = 48*											
Total	7971	27.59	9.05	0.03	0.57	0.98	8097	29.77	9.27	-0.05	0.62	0.99
female	3871	27.72	8.94				3915	29.52	9.24			
male	4100	27.46	9.15				4182	30.00	9.30			
Grade 3	*SAT9 MP = 30*											
Total	7971	19.04	5.67	0.01	0.63	0.97	8097	20.00	5.56	-0.05	0.67	0.98
female	3871	19.08	5.58				3915	19.85	5.49			
male	4100	19.00	5.75				4182	20.14	5.62			
Grade 5	*DSTP MP =48*											
Total	7919	23.15	9.06	0.02	0.48	0.97	8243	21.80	9.42	0.03	0.45	1.00
female	3971	23.26	8.93				3985	21.93	9.43			
male	3948	23.04	9.18				4258	21.68	9.42			
Grade 5	*SAT9 MP = 30*											
Total	7919	19.26	6.16	0.00	0.64	0.94	8243	19.67	6.00	-0.03	0.66	0.96
female	3971	19.25	5.96				3985	19.56	5.86			
male	3948	19.27	6.35				4258	19.76	6.12			

Table 4b
Test Performance on the Delaware Portion and SAT9
by Gender, Grade, and Year

Gender	1998						1999					
	N.	Mean	S.D.	D	PR	SDR	N.	Mean	S.D.	D	PR	SDR
Grade 8												
DSTP MP = 48												
Total	8235	17.77	10.16	-0.03	0.37	0.96	8226	19.43	9.79	-0.01	0.40	0.93
female	3973	17.60	9.97				4078	19.37	9.45			
male	4262	17.88	10.34				4148	19.49	10.21			
Grade 8 *SAT9 MP = 30*												
Total	8235	15.10	6.15	-0.03	0.50	0.94	8226	15.11	6.04	-0.09	0.50	0.94
female	3973	15.01	5.97				4078	14.83	5.84			
male	4262	15.19	6.32				4148	15.38	6.22			
Grade 10												
DSTP MP = 48												
Total	7140	13.07	8.58	0.01	0.27	0.93	7469	16.62	8.70	0.01	0.35	0.89
female	3570	13.13	8.26				3679	16.65	8.16			
male	3570	13.01	8.86				3790	16.60	9.20			
Grade 10 *SAT9 MP = 30*												
Total	7140	13.31	5.92	-0.02	0.44	0.92	7469	12.86	5.84	-0.11	0.43	0.88
female	3570	13.24	5.67				3679	12.53	5.44			
male	3570	13.38	6.16				3790	13.18	6.19			

Table 5a
Test Performance on the DSTP Mathematics Assessment
by Gender, Item Format, Grade, and Year

Item Format		*1998*						*1999*					
		N.	**Mean**	**S.D.**	**D**	**PR**	**SDR**	**N.**	**Mean**	**S.D.**	**D**	**PR**	**SDR**
Grade 3													
MC	*MP =50*		D_{ECR-MC} =	0.11					D_{ECR-MC} =	0.04			
	female	3463	32.04	8.77	-0.01	0.64	0.98	3674	33.45	8.70	-0.07	0.67	0.99
	male	3626	32.13	8.96		0.64		3861	34.10	8.78		0.68	
SA	*MP = 16*												
	female	3871	7.96	4.10	0.03	0.50	1.01	3915	9.71	4.05	0.00	0.61	1.00
	male	4100	7.84	4.07		0.49		4182	9.73	4.05		0.61	
ECR	*MP = 12*												
	female	3871	7.29	2.52	0.10	0.61	0.94	3915	6.48	2.80	-0.04	0.54	0.97
	male	4100	7.04	2.67		0.59		4182	6.59	2.88		0.55	
Grade 5													
MC	*MP = 50*		D_{ECR-MC} =	0.12					D_{ECR-MC} =	0.23			
	female	3777	32.18	9.67	-0.02	0.64	0.96	3790	31.55	9.35	-0.05	0.63	0.96
	male	3739	32.41	10.09		0.65		4038	32.02	9.79		0.64	
SA	*MP = 16*												
	female	3971	6.45	3.45	0.00	0.40	0.97	3985	6.85	3.68	0.00	0.43	1.02
	male	3948	6.44	3.54		0.40		4258	6.86	3.61		0.43	
ECR	*MP = 12*												
	female	3971	4.13	2.47	0.09	0.34	1.00	3985	3.37	2.93	0.18	0.28	1.04
	male	3948	3.90	2.48		0.33		4258	2.86	2.81		0.24	

Table 5b
Test Performance on the DSTP Mathematics Assessment by Gender, Achievement, Item Format, and Grade

Item Format		**1998**						**1999**					
		N.	**Mean**	**S.D.**	**D**	**PR**	**SDR**	**N.**	**Mean**	**S.D.**	**D**	**PR**	**SDR**
Grade 8													
MC	*MP = 50*		D_{ECR-MC} =	0.04					D_{ECR-MC} =	0.10			
	female	3681	25.41	9.52	-0.05	0.51	0.96	3719	25.42	9.10	-0.08	0.51	0.94
	male	3954	25.91	9.94		0.52		3844	26.16	9.71		0.52	
SA	*MP = 16*												
	female	3973	4.60	3.89	-0.01	0.29	0.95	4078	5.29	3.64	-0.01	0.33	0.94
	male	4262	4.63	4.10		0.29		4148	5.34	3.88		0.33	
ECR	*MP = 12*												
	female	3973	2.91	2.88	-0.01	0.24	1.01	4078	3.80	2.98	0.02	0.32	0.96
	male	4262	2.93	2.85		0.24		4148	3.75	3.11		0.31	
Grade 10													
MC	*MP = 50*		D_{ECR-MC} =	0.03					D_{ECR-MC} =	0.18			
	female	3276	22.08	8.55	-0.03	0.44	0.90	3407	21.47	8.15	-0.11	0.43	0.87
	male	3275	22.39	9.50		0.45		3528	22.40	9.42		0.45	
SA	*MP = 16*												
	female	3570	2.89	3.19	0.04	0.18	0.98	3679	4.65	3.30	0.02	0.29	0.92
	male	3570	2.76	3.25		0.17		3790	4.59	3.58		0.29	
ECR	*MP = 12*												
	female	3570	1.56	2.34	0.00	0.13	0.97	3679	3.28	2.33	0.07	0.27	0.91
	male	3570	1.57	2.41		0.13		3790	3.11	2.55		0.26	

Table 6a
Test Performance on the DSTP Mathematics Assessment by Gender, Content Standard, Grade, and Year

Content Standards		1998 N.	1998 Mean	1998 S.D.	1998 D	1998 SDR	1999 N.	1999 Mean	1999 S.D.	1999 D	1999 SDR
Grade 3											
Estimation, Measurement	*MP = 17*						*MP = 17*				
& Computation	*female*	3871	9.64	3.65	-0.01	0.99	3915	10.64	3.52	-0.09	0.99
	male	4100	9.68	3.68			4182	10.96	3.55		
Number Sense	*MP = 18*						*MP = 18*				
	female	3871	10.39	3.72	0.01	0.98	3915	12.31	3.80	-0.01	0.99
	male	4100	10.34	3.78			4182	12.33	3.83		
Algebra	*MP = 6*						*MP = 6*				
	female	3871	2.51	1.48	-0.03	1.01	3915	2.93	1.55	-0.03	0.98
	male	4100	2.55	1.46			4182	2.97	1.58		
Spatial & Geometry	*MP = 14*						*MP = 14*				
	female	3871	10.01	2.42	0.00	0.92	3915	8.73	2.87	-0.15	0.94
	male	4100	10.02	2.63			4182	9.16	3.04		
Statistics & Probability	*MP = 11*						*MP = 9*				
	female	3871	5.17	2.00	0.00	0.98	3915	7.57	2.24	0.03	0.99
	male	4100	5.16	2.04			4182	7.51	2.27		
Patterns, Relationship	*MP = 11*						*MP =13*				
& Functions	*female*	3871	6.13	2.64	0.04	0.99	3915	6.61	2.65	0.01	1.02
	male	4100	6.02	2.66			4182	6.59	2.61		

Table 6b
Test Performance on the DSTP Mathematics Assessment
by Gender, Content Standard, Grade, and Year

Content Standards		1998 N.	Mean	S.D.	D	SDR	1999 N.	Mean	S.D.	D	SDR
Grade 5											
Estimation, Measurement	*MP = 17*						*MP = 17*				
& Computation	*female*	3971	9.73	3.80	-0.02	0.94	3985	9.72	3.85	-0.06	0.96
	male	3948	9.79	4.03			4258	9.96	4.03		
Number Sense	*MP = 18*						*MP = 18*				
	female	3971	9.56	3.24	0.07	0.96	3985	8.44	3.69	0.04	0.98
	male	3948	9.32	3.38			4258	8.30	3.77		
Algebra	*MP = 6*						*MP = 6*				
	female	3971	3.17	1.80	0.03	0.99	3985	3.23	1.77	0.02	1.00
	male	3948	3.12	1.82			4258	3.20	1.77		
Spatial & Geometry	*MP = 11*						*MP = 11*				
	female	3971	4.85	2.04	0.02	0.93	3985	5.00	2.08	-0.01	0.96
	male	3948	4.80	2.19			4258	5.02	2.17		
Statistics & Probability	*MP = 14*						*MP = 14*				
	female	3971	8.34	3.21	-0.02	1.00	3985	7.88	3.07	0.04	1.05
	male	3948	8.40	3.21			4258	7.75	2.93		
Patterns, Relationship	*MP =11*						*MP = 11*				
& Functions	*female*	3971	6.17	2.73	-0.02	1.02	3985	6.46	2.73	-0.01	1.02
	male	3948	6.23	2.67			4258	6.50	2.67		

Table 6c
Test Performance on the DSTP Mathematics Assessment
by Gender, Content Standard, Grade, and Year

Content Standards		1998 N.	1998 Mean	1998 S.D.	1998 D	1998 SDR	1999 N.	1999 Mean	1999 S.D.	1999 D	1999 SDR
Grade 8											
Estimation, Measurement	*MP = 16*						*MP = 13*				
& Computation	*female*	3973	6.22	3.14	-0.05	0.95	4078	5.41	2.69	-0.07	0.96
	male	4262	6.39	3.30			4148	5.60	2.81		
Number Sense	*MP = 14*						*MP = 16*				
	female	3973	5.48	3.37	-0.12	0.96	4078	5.80	3.77	-0.15	0.96
	male	4262	5.89	3.51			4148	6.38	3.94		
Algebra	*MP = 14*						*MP = 18*				
	female	3973	7.08	3.71	0.09	1.01	4078	8.48	4.03	0.02	0.96
	male	4262	6.75	3.69			4148	8.39	4.21		
Spatial & Geometry	*MP = 8*						*MP = 9*				
	female	3973	3.08	2.22	-0.11	0.95	4078	2.80	2.19	-0.09	0.95
	male	4262	3.32	2.33			4148	3.00	2.31		
Statistics & Probability	*MP = 15*						*MP =15*				
	female	3973	5.55	2.85	0.04	0.96	4078	7.59	3.18	0.07	0.95
	male	4262	5.42	2.96			4148	7.37	3.36		
Patterns, Relationship	*MP =11*						*MP = 7*				
& Functions	*female*	3973	5.26	2.79	-0.02	0.99	4078	4.12	1.90	0.01	0.96
	male	4262	5.31	2.81			4148	4.10	1.98		

Table 6d
Test Performance on the DSTP Mathematics Assessment
by Gender, Content Standard, Grade, and Year

Content Standards		1998 N.	1998 Mean	1998 S.D.	1998 D	1998 SDR	1999 N.	1999 Mean	1999 S.D.	1999 D	1999 SDR
Grade 10											
Estimation, Measurement	*MP = 19*						*MP =18*				
& Computation	*female*	3570	6.71	3.24	-0.05	0.91	3679	7.49	3.20	-0.11	0.87
	male	3570	6.89	3.55			3790	7.87	3.67		
Number Sense	*MP = 7*						*MP = 7*				
	female	3570	1.53	1.69	-0.01	0.92	3679	1.52	1.60	-0.05	0.90
	male	3570	1.54	1.84			3790	1.61	1.77		
Algebra	*MP =15*						*MP =15*				
	female	3570	6.40	3.11	0.03	0.95	3679	6.29	3.15	-0.02	0.94
	male	3570	6.30	3.29			3790	6.34	3.36		
Spatial & Geometry	*MP =8*						*MP = 9*				
	female	3570	3.07	1.91	0.04	0.97	3679	2.51	1.57	-0.04	0.92
	male	3570	3.00	1.96			3790	2.58	1.71		
Statistics & Probability	*MP = 17*						*MP = 17*				
	female	3570	6.09	3.45	0.04	0.98	3679	6.65	3.21	0.06	0.94
	male	3570	5.94	3.53			3790	6.46	3.42		
Patterns, Relationship	*MP = 12*						*MP = 12*				
& Functions	*female*	3570	2.57	2.27	-0.06	0.94	3679	3.89	2.43	-0.05	0.89
	male	3570	2.72	2.41			3790	4.01	2.73		

Table 7a
Test Performance on the DSTP Mathematics Assessment
by Gender, Cognitive Process, Grade, and Year

Cognitive Process		1998 N.	Mean	S.D.	D	SDR	1999 N.	Mean	S.D.	D	SDR
Grade 3											
Conceptual	*MP = 30*						*MP =28*				
Knowledge	*female*	3871	17.63	5.30	-0.01	0.95	3915	17.44	5.51	-0.06	1.00
	male	4100	17.68	5.57			4182	17.79	5.52		
Procedural	*MP = 32*						*MP = 35*				
Knowledge	*female*	3871	21.30	6.36	0.02	0.98	3915	23.06	6.80	-0.07	0.98
	male	4100	21.17	6.46			4182	23.55	6.94		
Mathematical	*MP = 15*						*MP = 14*				
Processes	*female*	3871	7.10	3.33	0.08	0.99	3915	8.11	2.85	0.03	0.98
	male	4100	6.83	3.37			4182	8.03	2.91		
Grade 5											
Conceptual	*MP = 29*						*MP = 26*				
Knowledge	*female*	3971	15.01	5.44	0.01	0.96	3985	13.08	5.27	-0.05	0.97
	male	3948	14.97	5.64			4258	13.37	5.43		
Procedural	*MP = 32*						*MP = 31*				
Knowledge	*female*	3971	19.71	6.34	-0.01	0.94	3985	20.06	5.97	-0.03	0.97
	male	3948	19.77	6.71			4258	20.24	6.16		
Mathematical	*MP = 16*						*MP = 20*				
Processes	*female*	3971	7.09	3.44	0.05	0.99	3985	7.60	4.38	0.11	1.03
	male	3948	6.91	3.47			4258	7.12	4.26		

Table 7b

Test Performance on the DSTP Mathematics Assessment
by Gender, Cognitive Process, Grade, and Year

Cognitive Process		1998					1999				
		N.	Mean	S.D.	D	SDR	N.	Mean	S.D.	D	SDR
Grade 8											
Conceptual	*MP = 29*						*MP =30*				
Knowledge	*female*	3973	12.86	6.11	-0.07	0.95	4078	13.10	5.64	-0.10	0.91
	male	4262	13.33	6.45			4148	13.67	6.19		
Procedural	*MP = 28*						*MC = 28*				
Knowledge	*female*	3973	14.11	5.87	0.02	0.96	4078	14.92	5.77	-0.01	0.95
	male	4262	14.01	6.10			4148	15.00	6.06		
Mathematical	*MP = 21*						*MP = 20*				
Processes	*female*	3973	5.70	4.32	-0.01	0.98	4078	6.19	4.28	0.00	0.96
	male	4262	5.73	4.42			4148	6.20	4.47		
Grade 10											
Conceptual	*MP = 31*						*MP = 31*				
Knowledge	*female*	3570	10.30	5.70	0.01	0.93	679	10.93	5.57	-0.04	0.89
	male	3570	10.24	6.16			3790	11.20	6.29		
Procedural	*MP = 29*						*MP = 31*				
Knowledge	*female*	3570	13.22	5.48	0.02	0.93	3679	13.89	5.50	-0.05	0.89
	male	3570	13.09	5.87			3790	14.18	6.20		
Mathematical	*MP = 18*						*MP = 16*				
Processes	*female*	3570	2.90	3.08	-0.07	0.92	3679	4.37	2.79	-0.01	0.89
	male	3570	3.11	3.33			3790	4.40	3.15		

Appendix A

Delaware Mathematics Content Standards

Standard #1
Students will develop their ability to *Solve Problems* by engaging in developmentally appropriate problem-solving opportunity in which there is a need to use various approaches to investigate and understand mathematical concepts; to formulate their own problems, to find solutions to problems from everyday situations; to develop and apply strategies to solve a wide variety of problems; and to integrate mathematical reasoning, communication and connection.

Standard #2
Students will develop their ability to *Communicate Mathematically* by solving problems in which there is a need to obtain information from the real world through reading, listening, and observing; to translate this information into mathematical language and symbols; to process this information mathematically; and to present results in written, oral, and visual formats.

Standard #3
Students will develop their ability to *Reason Mathematically* by solving problems in which there is a need to investigate significant mathematical ideas in all content areas; to justify their thinking; to reinforce and extent their logical reasoning abilities; to reflect on and clarify their own thinking; to ask questions to extent their thinking; and to construct their own learning.

Standard #4
Students will develop their ability to make *Mathematical Connections* by solving problems in which there is a need to view mathematics as an integrated whole and to integrate mathematics with other disciplines, while allowing the flexibility to approach problems, from within and outside mathematics, in a variety ways.

Standard #5
Students will develop an understanding of *Estimation, Measurement, and Computation* by solving problems in which there is a need to measure to a required degree of accuracy by selecting appropriate tools and units; to develop computing strategies and select appropriate methods of calculation from among mental math, paper and pencil, calculators or computers; to use estimating skills to approximate an answer and to determine the reasonableness of results.

Standard #6
Students will develop *Number Sense* by solving problems in which there is a need to represent and model real numbers verbally, physically and symbolically; to use operations with understanding; to explain the relationships between numbers; to apply the concept of a unit; and to determine the relative magnitude of real numbers.

Standard #7
Students will develop an understanding of *Algebra* by solving problems in which there is a need to progress from the concrete to the abstract using physical model, equations and graphs; to generalize number patterns; and to describe, represent and analyze relationships among variable quantities.

Standard #8
Students will develop *Spatial Sense* and an understanding of *Geometry* by solving problems in which there is a need to recognize, transform, analyze properties of, and discover relationships between geometric figures.

Standard #9
Students will develop an understanding of *Statistics and Probability* by solving problems in which there is a need to collect, appropriately represent, and interpret data; to make inferences or predictions; to present convincing arguments; and to model mathematical situations to determine the probability.

Standard #10
Students will develop an understanding of *Patterns, Relationship and Functions* by solving problems in which there is a need to recognize and extend a variety of patterns; and to analyze, represent, model and describe real-world functional relationships.

Appendix B

Definitions of Cognitive Processes of DSTP Mathematics Assessment

Conceptual Knowledge involves the construction of fundamental mathematical ideas including the notions of a unit, counting, ordering, part vs. whole, mathematical operations, geometric figures, pattern, measurement, and chance. Conceptual knowledge deepens as concepts are connected one to another and applied more widely.

Procedural Knowledge involves the skills performance of a standardized routine. It tends to be most firmly held if based upon strong conceptual foundations. Not all-conceptual knowledge, however, is transformed into procedural knowledge. Only certain fundamental procedures need to be practiced to the point of fluency.

While conceptual and procedural knowledge can be assessed independently, the application of these concepts and skills is described under the broad category of **Mathematical Processes (or called Problem Solving).** Both concepts and procedures may be called upon in non-routine situations that require problem solving. Connections between diverse mathematical concepts or between mathematics and another discipline can be assessed.

Appendix C

Item Format

Multiple-Choice
This type of items consists of a stem and four response options. They are scored dichotomously.

Constructed-Response Items
This type of items requires students to create their own responses. Two constructed-response items are used in the DSTP mathematics assessment. In scoring, constructed-response items allow for the awarding of credit for more than one response or approach. They also allow for crediting growth in student knowledge. They enrich the analysis of the assessment by providing information about students' approaches to problems, misconceptions, abilities, and understanding.

- *Short Answer* requires students to reflect on a situation and communicate briefly the reasoning behind their solution. Short answer items are scored using the 0-2 scoring rubrics.

- *Extended-Response* requires students to consider a situation within or across the content strand areas, understand what is required to solve the situation, choose an approach to the problem, carry out the 'attack', and interpret the solution in terms of the original situation. The response mode requires that student provide evidence of their reasoning and communicate their decision-making steps in the context of the problem. Extended-response items are scored using the 0-4 scoring rubrics.

Appendix D

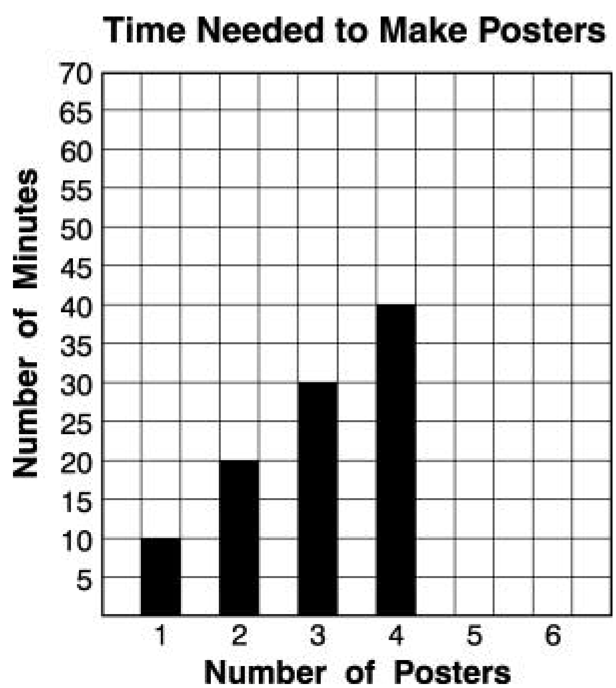
Sample Items with Scoring Rubrics

Standard Measured:

Standard 10: Students will develop an understanding of PATTERNS, RELATIONSHIPS, AND FUNCTIONS by solving problems in which there is a need to recognize and extend a variety of patterns; and to analyze, represent, model, and describe real-world functional relationships.

Item:

The graph below shows how many minutes it takes to make different numbers of posters.



If this pattern continues, how many minutes will it take to make 6 posters?

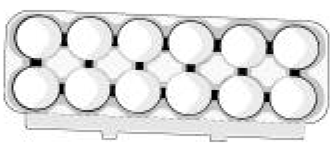
a. 50
b. 55
c. 60
d. 65

Answer: c

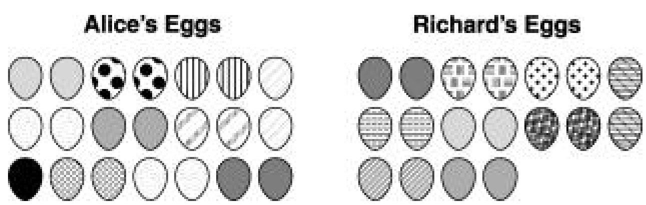
Standard Measured:

Standard 6: Students will develop NUMBER SENSE by solving problems in which there is a need to represent and model real numbers verbally, physically, and symbolically; to use operations with understanding; to explain the relationships between numbers; to apply the concept of a unit; and to determine the relative magnitude of real numbers.

Item:



Alice and Richard are coloring eggs and putting them back into the egg cartons. Alice colored 21 eggs and Richard colored 18 eggs. (There are 12 eggs in a carton.)



How many cartons will they need to hold all the eggs? Explain how you got your answer.

Scoring Rubric:

4 Correct answer (4 cartons) with clear verbal or pictorial explanation. (This may or may not involve finding the total number of eggs. For example, the problem might be successfully solved by grouping the egg images into groups of 12.)

3 Correct answer (4 cartons) but sketchy explanation or incorrect answer of 3 cartons with explanation that describes the number of cartons *filled*.

2 There is an (unsuccessful) attempt to divide eggs into cartons. Perhaps total number of eggs (39) is correct but attempt to divide into cartons is flawed.

1 An attempt is made to count eggs (perhaps even successfully), but there is no evidence of an attempt to divide eggs into cartons.

0 Trace evidence of work but without clear connection to problem situation.

Commentary:

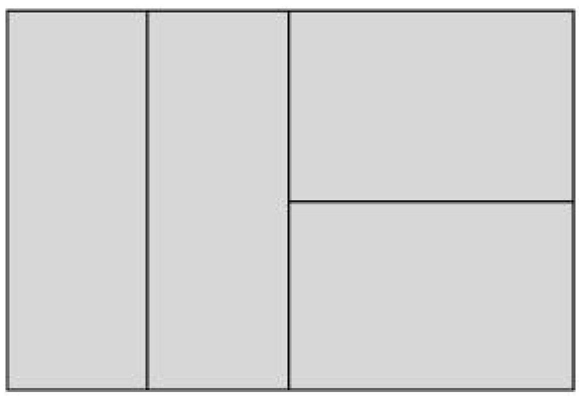
This item addresses several components of number sense. The student must use several operations with understanding including addition and division but is able to do this in a context involving a physical representation of number. A great variety of solution strategies have been observed. For example, some students actually identified the first, second, and third dozen eggs in the diagram and found that three full cartons were needed and a fourth with only three eggs. Other students found the sum directly and compared this to the number of eggs in three dozen.

Standard Measured:

Standard 6: Students will develop NUMBER SENSE by solving problems in which there is a need to represent and model real numbers verbally, physically, and symbolically; to use operations with understanding; to explain the relationships between numbers; to apply the concept of a unit; and to determine the relative magnitude of real numbers.

Item:

John split a cake into four pieces. The pieces are *not* all the same shape. Do you believe the four pieces are the same size? Please explain.



Scoring Rubric:

4 Explains that the four pieces are the same size because they are each half of a half (or one-quarter) of the full cake. Or the student might recreate a picture of the rectangular cake in the response booklet and use an area or measurement model, i.e., counting the number of answer grid squares in each piece of the cake.

3 Agrees that the four pieces are all the same size but the explanation is weak or incomplete. For example, a response in this category might describe the process of cutting the cake but fail to describe the ultimate equality of half of a half: "I cut the cake into two pieces of the same size and then cut one of these in half the long way and the other in half the other way."

2 Agrees that the four pieces are the same size but without coherent (or perhaps any) explanation.

1 Asserts that the resulting pieces are not the same size, with or without rationale.

0 Attempts a response but does not clearly address the task.

Commentary:

Operations with fractions must be built upon a firm foundation of conceptual understanding. Using an area model of fractions as in this item, students are asked to compare fractional pieces of different shapes but the same area. A strong somewhat abstract response may involve the recognition that half-of-a-half represents one-quarter. Other students may have cut each shape into smaller units in order to compare the shapes' areas.

<u>Standard Measured</u>:

Standard 5: Students will develop an understanding of ESTIMATION, MEASUREMENT, and COMPUTATION by solving problems in which there is a need to measure to a required degree of accuracy by selecting appropriate tools and units; to develop computing strategies and select appropriate methods of calculation from among mental math, paper and pencil, calculators or computers; to use estimating skills to approximate an answer and to determine the reasonableness of results.

Item:

Use your ruler to draw a line segment CD half the length of line segment AB shown below. What is the length of the segment you drew?

Scoring Rubric:

2 3.5 cm (or $1\frac{3}{8}$ inches) segment drawn and length so labeled (need not be labeled “CD”)
Range of 3.3 cm to 3.7 cm OR $\pm\frac{1}{8}$ inch

1 Segment drawn to correct length but not labeled with length.

0 Length of line incorrect (3.2 cm or shorter or 3.8 cm or longer).

<u>Commentary</u>:

This item assesses both measurement *and* computation. The segment is 7 centimeters long (or, alternatively, $2\frac{3}{4}$ inches) so a segment half as long will be 3.5 cm (or $1\frac{3}{8}$ inches). It is a performance task in that a ruler must be used to measure the given line and then construct one half as long.

Standard Measured:

Standard 6: Students will develop NUMBER SENSE by solving problems in which there is a need to represent and model real numbers verbally, physically, and symbolically; to use operations with understanding; to explain the relationships between numbers; to apply the concept of a unit; and to determine the relative magnitude of real numbers.

Item:

Mr. Garcia was assembling a set of shelves. He tried a $\frac{1}{2}$-inch wrench, but it was too small. Then he tried a $\frac{3}{4}$-inch wrench, but it was too large. Which of the following wrenches might work?

a. $\frac{1}{4}$ inch

b. $\frac{3}{8}$ inch

c. $\frac{5}{8}$ inch

d. $\frac{7}{8}$ inch

Answer: c

Commentary:

This item assesses a student's ability to order simple fractions. It is set in a context which may provide additional access for students. Solution strategies can range from the formal —rewriting both fractions in terms of eighths — to the informal — for example, using common benchmarks or a familiarity with a conventional ruler.

Standard Measured:

Standard 8: Students will develop SPATIAL SENSE and an understanding of GEOMETRY by solving problems in which there is a need to recognize, construct, transform, analyze properties of, and discover relationships between, geometric figures.

Item:

The size of a TV set is measured diagonally. Calculate the approximate area of the screen if it is considered to be a 21-inch TV with a width of 16 inches.

a. 109 sq in.
b. 211 sq in.
c. 218 sq in.
d. 222 sq in.

Answer: c

Commentary:

This item requires that students apply one of the most fundamental geometric relationships, the Pythagorean theorem, within a common context. Because students had access to calculators for the section of the test in which this item appeared, the calculations (involving the square root of 185) are easily managed. The item assesses whether or not students will be able to bring their knowledge about the Pythagorean relationship to bear in the solution of a real-world problem.

Standard Measured:

Standard 7: Students will develop an understanding of ALGEBRA by solving problems in which there is a need to progress from the concrete to the abstract using physical models, equations, and graphs; to generalize number patterns; and to describe, represent, and analyze relationships among variable quantities.

Item:

Rewrite $\sqrt{y + 2} = x - 3$ so that you can enter it into your calculator to determine its graph.

Scoring Rubric:

2 Response may be $y = x^2 - 6x + 7$ (terms on right side may be in different order) or $y = (x - 3)^2 - 2$.

1 Response involves minor algebraic error, e.g., adding 2 to both sides, hence, $y = (x - 3)^2 + 2$ or error in squaring binomial, e.g., $y = x^2 - 11$ (or $y = x^2 - 7$, etc.).

0 Any other response including those in which both sides not squared or not of the form $y = \ldots$

Commentary:

The graphing calculator aids in understanding of functional relationships by quickly and easily presenting tabular and graphical representations of functions represented algebraically. However, in order to graph functions, they may need to be manipulated algebraically to be put in the standard $y = f(x)$ form. This item features symbolic manipulation in the context of entering a function in the graphing calculator, a context that should be familiar to students who are studying mathematics in a standards-based classroom.

Standard Measured:

Standard 10: Students will develop an understanding of PATTERNS, RELATIONSHIPS, AND FUNCTIONS by solving problems in which there is a need to recognize and extend a variety of patterns; and to analyze, represent, model, and describe real-world functional relationships.

Item:

The table below shows *thinking*, *braking*, and *stopping* distances at different highway speeds.

Speed (mph)	Thinking Distance (ft)	Braking Distance (ft)	Stopping Distance (ft)
v		$\frac{v^2}{20}$	
10	10	5	15
20	20	20	40
30	30	45	75
40	40	80	120
50	50	125	175
60	60	180	240

For the values in the table, if speed is represented by v, then a formula for the braking distance would be $\frac{v^2}{20}$.

a. What formulas would represent *thinking distance* and *stopping distance*?

b. According to the formula, how many feet would it take to stop if a car is traveling 55 mph?

c. The usual rule is to allow one car length (approximately 20 feet) of space between your car and a car in front for every 10 mph of speed. How good is the rule when compared to the data above? Explain your reasoning.

Scoring Rubric:

4 Correct answers to all parts.

a. Thinking distance is v;
Stopping distance is $v + \frac{v^2}{20}$ or $\frac{20v + v^2}{20}$

b. 206.25 (feet)

c. Explanation which indicates that the customary rule works for low speeds (through 20 mph) but fails to provide enough stopping distance at higher speeds. This is because the rule is linear but stopping distance is quadratic. (Student needn't use the terms linear and quadratic but should note that the rule fails for speeds above 20 mph).

3 Parts a. and b. answered correctly, but explanation in c. does not contain enough detail, i.e., doesn't identify 20 mph as the maximum speed for which the conventional rule works.

2 Parts a. and b. answered correctly, with, perhaps, minor computational errors in b. Response to c. inadequate or even missing.

1 Unable to write formula for stopping distance though perhaps able to approximate stopping distance at 55 mph through linear interpolation from tabular values.

0 Some work but without much evidence that this work addresses the question.

Commentary:

A primary goal of high school mathematics is to promote the development of a variety of ways of modeling the world. This item features a quadratic model in a context that is important for young adults, that of the stopping distance of an automobile. The item involves scaffolding to promote access to the algebraic representation and then proceeds to require interpretation of the mathematical results.

Appendix E

Mathematical Reference Sheets

GRADE 8 MATHEMATICS
DSTP REFERENCE SHEET

AREAS OF POLYGONS

Triangle $A = \frac{1}{2}bh$

Rectangle $A = bh$

Square $A = s^2$

Parallelogram $A = bh$

Trapezoid $A = \frac{1}{2}h(b_1 + b_2)$

VOLUMES

Cube $V = e^3$

Pyramid $V = \frac{1}{3}Bh$
where B = area of the base

Cylinder $V = \pi r^2 h$

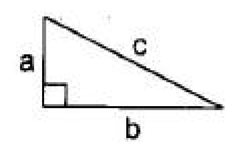
Rectangular Prism $V = lwh$

CIRCLES

$C = 2\pi r = \pi d$

$A = \pi r^2$

RIGHT TRIANGLES



$a^2 + b^2 = c^2$

SURFACE AREAS

Cube $SA = 6e^2$

Cylinder $SA = 2\pi rh + 2\pi r^2$

GRADE 10 MATHEMATICS
DSTP REFERENCE SHEET

AREAS OF POLYGONS

Triangle	$A = \frac{1}{2}bh$
Rectangle	$A = bh$
Square	$A = s^2$
Parallelogram	$A = bh$
Trapezoid	$A = \frac{1}{2}h(b_1 + b_2)$

CIRCLES

$C = 2\pi r = \pi d$

$A = \pi r^2$

SURFACE AREAS

Cube	$SA = 6e^2$
Cylinder	$SA = 2\pi rh + 2\pi r^2$
Sphere	$SA = 4\pi r^2$

DISTANCE BETWEEN POINTS (x_1, y_1) & (x_2, y_2)

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

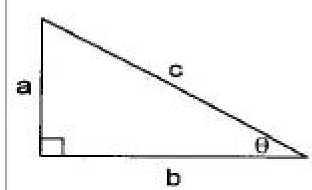
VOLUMES

Cube	$V = e^3$
Pyramid	$V = \frac{1}{3}Bh$

where B = area of the base

Cylinder	$V = \pi r^2 h$
Cone	$V = \frac{1}{3}\pi r^2 h$
Sphere	$V = \frac{4}{3}\pi r^3$

RIGHT TRIANGLES



$a^2 + b^2 = c^2$

$\sin\theta = \frac{a}{c}$

$\cos\theta = \frac{b}{c}$

$\tan\theta = \frac{a}{b}$