APPLIED MEASUREMENT IN EDUCATION, 22: 164–184, 2009

ISSN: 0895-7347 print / 1532-4818 online
DOI: 10.1080/08957340902754635

Gender Differences in Large-Scale Math Assessments: PISA Trend 2000 and 2003

Ou Lydia Liu
Educational Testing Service, Princeton

Mark Wilson
Graduate School of Education, University of California, Berkeley

Many efforts have been made to determine and explain differential gender performance on large-scale mathematics assessments. A well-agreed-on conclusion is that gender differences are contextualized and vary across math domains. This study investigated the pattern of gender differences by item domain (e.g., Space and Shape, Quantity) and item type (e.g., multiple-choice[i] items, open constructed-response items). The U.S. portion of the Programme for International Student Assessment (PISA) 2000 and 2003 mathematics assessment was analyzed. A multidimensional Rasch model was used to provide student ability estimates for each comparison. Results revealed a slight but consistent male advantage. Students showed the largest gender difference ($d = 0.19$) in favor of males on complex multiple-choice items, an unconventional item type. Males and females also showed sizable differences on Space and Shape items, a domain well documented for showing robust male superiority. Contrary to many previous findings reporting male superiority on multiple-choice items, no measurable difference has been identified on multiple-choice items for both the PISA 2000 and the 2003 math assessments. Reasons for the differential gender performance across math domains and item types were speculated, and directions of future research were discussed.

[i]In this paper, two kinds of multiple-choice items are discussed: traditional multiple-choice items and complex multiple-choice items. A sample complex multiple choice item is shown in Table 6. The terms "multiple-choice" and "traditional multiple-choice" are used interchangeably to refer to the traditional multiple choice items throughout the paper, while the term "complex multiple-choice" is used to refer to the complex multiple-choice items.

Correspondence should be addressed to Ou Lydia Liu, Educational Testing Service, 666 Rosedale Road, Mail Stop 16-R, Princeton, NJ 08541. E-mail: lliu@ets.org

INTRODUCTION

The male advantage on large-scale math assessments has been a controversial issue in the United States for more than three decades. Males have significantly outperformed females on many domestic and international assessments such as the National Assessment of Educational Progress (NAEP), SAT, Trends in Mathematics and Science Study (TIMSS), and Programme for International Student Assessment (PISA) in the United States (College Board, 2005; Organization for Economic Co-operation and Development [OECD], 2000a, 2004; Perie, Moran, & Lutkus, 2005; Mullis, Martin, & Foy, 2005; Mullis et al., 2000; Mullis, Martin, Fierros, Goldberg, & Stemler, 2000). According to the NAEP report (Perie et al., 2005), 9-year-old males obtained higher math scores than comparable females for all of the administrations from 1992 to 2004, 13-year-old males outperformed females from 1982 to 2004, and 17-year-old males surpassed females from 1973 to 2004. On the SAT-Mathematics (see Figure 1), the score difference has remained somewhere constant between 33 and 40 points from 1972 to 2005 in favor of males (College Board, 2005). It has been reported that the gender performance gap appears to decrease when males and females are matched on their math course preparation (Bridgeman & Wendler, 1991, 2005). However, there is no compelling evidence that the math achievement gap will be eliminated in the near future.

The disparities in standardized math performance are likely to be associated with the discrepancy in math-related educational and occupational opportunities. Females comprise fewer than 1/3 of the doctoral degrees in Physical and Computer Sciences, and fewer than 1/4 in Engineering and Mathematics (National Science Foundation, 2000). Females comprise only 8% of the mathematics professors in degree-granting institutions, and 10% of the engineers in the United States (U.S. Bureau of the Census, 2001).

The substantial male advantage on standardized math tests has aroused continuous attention and discussion among researchers and practitioners. Many efforts have been made to determine and explain the factors that are associated with

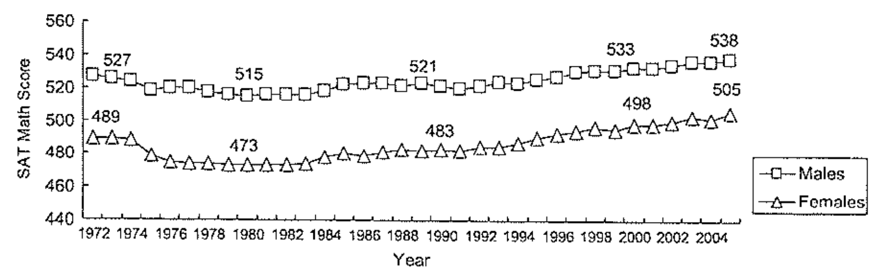

FIGURE 1 SAT mathematics score over the past three decades.

math gender differences. Most of these investigations fall into two categories: those examining the reasons from the student side and those examining the reasons from the assessment side. Research in the first category tends to offer social, cultural, psychological, or psycho-bio-social models to explore the differences in individual characteristics, and how these underlying differences have causal or correlational relationships with the observed gender differences in math performance (Buss, 1995; Gaulin, 1995; Geary, 1996; Halpern, 2000; Silverman, Phillips, & Silverman, 1996). Research from the assessment category tends to examine the item characteristics that are related to differential gender performance. Compared to the explanatory investigations from the student side, studies focused on the item characteristics are rather exploratory. In the context of gender research, as valuable and necessary as the explanatory investigations are, it is hard for the results to have a direct and immediate impact on standardized math performance. The exploratory investigations, though relatively simple and straightforward, are able to inform test developers of items that are associated with substantial gender differences, and caution decision makers about the interpretation of these results. Two main item-related factors have been identified to influence the pattern and magnitude of gender differences: (1) the different cognitive domains measured by the math tests and (2) the item types employed to elicit information from students.

Item Domain

A wealth of research has shown that gender differences are not homogenous across different math tasks. Spatial ability is the domain that has shown the largest and most robust gender differences in favor of males (Casey, Nuttall, Pezaris, & Benbow, 1995; Gierl, Bisanz, Bisanz, & Boughton, 2003; Halpern, 1997). Items assessing spatial abilities have strong ties to traditional geometry but go far beyond it in content and measures. For example, it is important for students to understand the properties of objects, their relative positions, and the relationship between actual shapes and visual representations. Males have also been reported to perform better on tasks that assess reasoning and problem-solving abilities (Burton & Lewis, 1996; Doolittle & Cleary, 1987; Gallagher, 1992, 1998; Gallagher et al., 2000; O'Neill & McPeek, 1993). In addition, males have been noted to demonstrate some advantages in understanding statistics and scored higher on a series of statistics exams when compared to females (Schram, 1996). On the other hand, females have occasionally surpassed males on items involving computational skills, items requiring an understanding of numerical patterns, and items involving a representation of quantities and quantifiable attributes of objects (Marshall & Smith, 1987; OECD, 2003; Willingham & Cole, 1997). Females also performed better on items that require memorization and recollection of practiced knowledge.

Gallagher and De Lisi (1994) noted that females tend to perform better when solving conventional items related to textbook context, whereas males are better at solving unconventional items which depart from textbook context or require an unusual use of a familiar method. These unconventional items usually require students to provide justifications for their assertions, and can be classified as reasoning problems (Armstrong, 1985; Gallagher & De Lisi, 1994). Females' propensity to rely on textbook examples may place them at a disadvantage on large-scale standardized tests, because it is unlikely that these tests will closely resemble classroom assessment (Linn & Hyde, 1989).

Algebra is a domain that has shown consistent female advantages. Females were found to perform better on algebra items when matched with males on quantitative test scores; males, when similarly matched, perform relatively better on geometry and problem-solving items (Carlton & Harris, 1989; O'Neill, Wild, & McPeek, 1989). In addition, females have been documented to exceed males on items purely presented by formulas, equation, or theory. Males are reported to perform better on word problems involving the applications of mathematical theories. Gallagher (1992) conducted think-alouds with 50 high school students on SAT-Mathematics items that exhibited differential item functioning. She found that most of the items favoring males required logical estimation strategies and most of the items favoring females required routine algorithmic strategies.

Compelling evidence has demonstrated that in order to understand the causes of gender math differences, the investigations must be conducted in a situational context, taking the characteristics of the math domain into consideration.

ITEM FORMAT

Studies have documented equal math grades for males and females in math courses (Bridgeman & Wendler, 1991; Kimball, 1989). However, gender differences on standardized tests are sizable. This discrepancy sparks the discussion about the relationship between the gender performance difference and item format. Many previous studies examining gender differences on standardized tests have reported that traditional multiple-choice items favor males and open constructed-response items favor females (Bolger & Kellaghan, 1990; DeMars, 2000, 1998; Klein et al., 1997). DeMars (1998) has noted a similar finding that among the highest ability students, males scored higher on multiple-choice items whereas females scored higher on the constructed-response section. Because multiple-choice items normally constitute a large portion of standardized tests, the better performance of males might in part stem from their advantage on multiple-choice items.

Higher risk-taking tendency of males has been offered as an explanation of their superior performance on multiple-choice items. Ben-Shakhar and Sinai

(1991) noted that male students are more likely to guess when they are uncertain about the answers. Females, on the other hand, often penalize themselves by leaving items blank. Even when permissive instructions were given, females still tended to omit more items than males (Hanna, 1986). When students were given a second chance to guess the answers to originally omitted items, females gained more score points than males (Harris, 1971).

As for why females performed better on open constructed-response items, it has been hypothesized that females, on average, have higher language abilities than males and are thus able to express themselves more effectively (Bolger & Kellaghan, 1990; Maccoby & Jacklin, 1974; Murphy, 1982). In addition, females are more likely to provide a thorough description of the mathematical procedures as required by some open-ended questions whereas males tend to simply record the results.

Although accumulating evidence has shown a male advantage on multiple-choice items, there are some contrary findings. For instance, Perie (1994) developed a test containing an equal number of multiple-choice items and open constructed-response items. Through a follow-up interview, Perie found that females were more likely than males to use guessing or memorization strategies. Furthermore, other studies found that item format did not explain any of the gender differences in performance (Beller & Gafni, 1996; Breland, Danos, Kahn, Kubota, & Bonner, 1994). When the item format was altered, there was no associated change in math performance (Wester, 1995; Wester & Henriksson, 2000). Beller and Gafni's (2000) investigation, based on the 1988 and 1991 International Assessment of Educational Progress mathematics tests, revealed that the gender effects were larger on multiple-choice items than on open constructed-response items for the 1988 assessment but smaller for the 1991 assessments. Further investigations disclosed that item format alone does not account for much of the variation in gender differences on math performance, but that differences in task difficulty could explain contradictory findings. Males were found to perform relatively better on the more difficult tasks, regardless of the item format (Beller & Gafni, 2000).

Given the discussion thus far, the role of item format remains ambiguous in the explanation of males' higher performance on standardized math tests. In addition, as standardized tests are driven to cover as much information as possible, the multidimensional nature of these tests further complicates the situation. Gender differences may be underestimated by comparing performance on the total score. If males and females are each favored by certain items, then the effect of gender differences may get cancelled out when a total score is used. To clarify the relationship between item domain, item format, and student math achievement, this study aims to serve two objectives: (1) to investigate the trend of gender differences across math content domains and (2) to investigate the trend of gender differences within math item type. The primary purpose was to shed light

on the patterns and particulars of gender differences to identify the domains and item types that are likely to introduce substantial differences. Specifically, test developers should be informed of the item domains and types that have been identified as potential sources of gender bias.

DATA SOURCES

Instruments

The data came from the U.S. portion of the PISA 2000 and 2003 mathematics assessment. PISA was launched and developed by the Organization for Economic Co-operation and Development (OECD) in 2000, to define educational goals, monitor educational progress, and provide a basis for international comparison of student academic performance (OECD, 2000a). Two content math domains (Space and Shape and Growth and Change[1]) were assessed in 2000, and two more domains (Quantity and Uncertainty) were added in the 2003 assessments. Three item types were used in the 2000 assessment (multiple-choice items, closed constructed-response items, and open constructed-response items). In 2003, two additional item types were employed (short response items and complex multiple-choice items). There were 32 math items in the 2000 PISA assessment and 84 math items in 2003. The item distribution is provided in Tables 1 and 2 for the two administrations, respectively.

When administered to the students, the math items were grouped with items from the other two domains (reading and science) assessed by PISA. Due to the large number of items, PISA designed test forms comprised of different combinations of the items from each domain. These test forms are called booklets in PISA assessments. There were nine booklets for the 2000 assessment and 13 for the 2003 assessment. Each student only responded to one booklet.

TABLE 1
Distribution of PISA 2000 Math Items by Content Domain and Item Type

	Content Domain (Number of Items)		
Item Type	*Space and Shape*	*Change and Relationships*	*Total*
Multiple-Choice	5	6	11
Closed Constructed-Response	9	9	18
Open Constructed-Response	—	3	3
Total	14	18	32

[1]The name was changed to Change and Relationships in the 2003 assessment.

TABLE 2
Distribution of PISA 2003 Math Items by Content Domain and Item Type

	Content Domain (Number of Items)				
Item Type	*Space/Shape*	*Change/Relationships*	*Quantity*	*Uncertainty*	*Total*
Short Response	2	4	13	3	22
Closed Constructed-Response	6	4	2	1	13
Complex Multiple-Choice	4	2	2	3	11
Multiple-Choice	4	1	4	8	17
Open Constructed- Response	4	11	1	5	21
Total	20	22	22	20	84

Subjects

PISA adopted a stratified sampling procedure that involved two stages. The first stage consisted of sampling individual schools that had 15-year-old students. Schools were sampled systematically with probabilities proportional to size, the measure of size being a function of the estimated number of 15-year-old students. A minimum of 150 schools were selected in each country. In the United States, students were sampled from 400 schools. As the schools were sampled, replacement schools were also selected, in case a sampled school was not able to participate for some reason.

The second stage of the sampling process involved sampling students within sampled schools. Once schools were selected, a list of each school's 15-year-old students was prepared. Then 35 students were selected with equal probability. In the United States, 2135 students (48.1% males) took the PISA math test in 2000 and 5456 students (50.2% males participated) in 2003. The test-takers were exclusively 15 year olds. Existing literature has documented that males and females showed no or small math performance differences in the elementary years (Lachance & Mazzocco, 2006). However, around age 15, moderate differences start to occur on timed standardized tests, and the differences continue to exist in high school and college when students take the SAT-Math test ($d = 0.40$) and the GRE-Quantitative test ($d = 0.70$) (Byrnes, 2005; Hyde, Fennema, &Lamon, 1990). Age 15 is an important milestone for students who choose to either go to high school moving on the college track, or to prepare for professional training. Academic preparation at this age can have a robust and profound impact on the student's future educational and professional aspirations and attainment.

METHODS AND ANALYSES

As aforementioned, the objectives of this study were to investigate the patterns of gender differences by item domain and item type. As described in the PISA

assessment framework, the categorization of item domain and item type is quite generic rather being than mutually exclusive, suggesting that items of different domains and types are expected to correlate with each other (OECD, 2000a, 2003). By using a multidimensional model, useful diagnostic information may be revealed as the pattern and magnitude of gender differences may vary across domains.

The MRCML Model

In this study, the multidimensional random coefficients multinomial logit (MRCML) model (Adams, Wilson, & Wang, 1997) was used for item calibration and ability estimation. The purpose of selecting this model among a few alternatives was three-fold. First, the MRCML model can deal with both dichotomously and polytomously scored items, which is the case with the PISA math assessments. Second, this model provides direct estimates of the correlations between dimensions, which disattenuate the measurement error and provide unbiased estimates (Briggs & Wilson, 2003). Lastly, this model is used by the PISA technical group for all the item calibration and ability estimation (OECD, 2000b, 2005a). Alignment with the standard procedure allows the connection and comparison between findings from this study and the findings reported in a series of PISA reports published by OECD. The mathematical formulation of this model is provided as follows:

$$P(X_{nik} = 1 | \boldsymbol{\theta}_n, \xi) = \frac{\exp[\mathbf{b}_{ik}' \boldsymbol{\theta}_n + \mathbf{a}_{ik}' \xi]}{\sum_{k=1}^{K_i} \exp[\mathbf{b}_{ik}' \boldsymbol{\theta}_n + \mathbf{a}_{ik}' \xi]} \tag{1}$$

X_{nik} assumes value 1 when PISA participant n's response to item i falls into category k, and 0 otherwise ($1 \leq i \leq I$, $1 \leq k \leq K_i$, $1 \leq n \leq N$). $\boldsymbol{\theta}$ is an ability parameter vector indicating student ability level on each dimension, and assumes the multivariate normal distribution. For instance, for the two math content domains tested in the PISA 2000 math assessments, each student will have an ability estimate on each of these two math domains. $\mathbf{b}_{ik}'$ is an element of a scoring vector, which specifies the performance level of category k of item i. For instance, for a dichotomously scored multiple-choice item with four options, elements of $\mathbf{b}_{ik}'$ specify which option is associated with score 1 and the rest of the options are scored as 0. ξ is an item parameter vector, which denotes the item difficulty parameter for dichotomously scored items, and both item difficulty and step difficulty parameters for polytomously scored items. $\mathbf{a}_{ik}'$ is an element of a design matrix that specifies linear combinations of the elements of ξ for each response category. For example, $\mathbf{a}_{ik}'$ would be 1 if the corresponding parameters exist (e.g., step difficulty for polytomous items) and 0 otherwise (e.g., no step difficulty for dichotomous items). The marginal maximum likelihood method was used in *ConQuest* to estimate the item parameters and the population means and variances. Monte Carlo integration was adopted when dealing with high dimensionality. A detailed discussion of the parameter estimation of the MRCML model can be found in Adams, Wilson, and Wang (1997). In addition, other applications of this model can be found in Allen and Wilson (2006), Briggs and Wilson (2003), Wang, (1999), Wang, Wilson, and Adams (1997), and Wolfe, Viger, Jarvinen, and Linksman (2007).

Plausible Values

The software program *ConQuest* (Wu, Adams, & Wilson, 1998) was used to perform the analyses. *ConQuest* produces five sets of plausible values, indicating math proficiency for each student on each dimension. The five plausible values were randomly drawn from the distribution of ability estimates that could reasonably be assigned to a student, and the mean of the plausible values should be equal to the expected a posteriori (EAP) estimator. The plausible values were used to compare the gender performance differences. Compared to using the EAP estimator, using plausible values when computing statistics takes into account the sampling error and imputation error (the latter also known as measurement error) (Wu & Adams, 2002), thus producing unbiased estimates (OECD, 2005a). Plausible values were first used to analyze the NAEP data in the mid 1980s, and have been used extensively to indicate student proficiency level in all subsequent NAEP, TIMSS, and PISA studies (U.S. Department of Education, 2001, 2003; OECD, 2000b, 2005a).

Booklet Effect

In each PISA booklet, the math items appeared in a different order. It was hypothesized that the different order might affect student math performance. Therefore, there is a need to monitor and adjust the booklet effect. The rationale is that if it is true that some booklets were more difficult than others, then the students who took the more difficult booklets should be compensated through some adjustment to their ability estimates. Similar consideration should be applied to those students who were assigned the easier booklets. The description of the booklet effect adjustment is provided in the Appendix.

Gender Performance Comparisons

Male and female math performance was compared, in terms of overall math achievement, content domain, and item type. Similar analyses were performed for both the PISA 2000 and 2003 math assessments. The five sets of plausible

values were used to indicate student math proficiency in each dimension. A SPSS *macro* program provided by the PISA 2003 data analysis manual was used to produce direct estimates of the mean gender difference and the standard error of the difference, on the basis of the plausible values (OECD, 2005b). The estimated standard error is comprised of the sampling error and the imputation error, thus unbiased. A z statistic, computed by dividing the mean difference by the estimated standard error, was used to indicate the statistical significance of the mean difference for each comparison. A significant z value means that the mean difference was significantly different from zero, suggesting a gender performance difference among PISA participants. Effect sizes indicated by Cohen's d[2] (Cohen, 1988) were used to describe the magnitude of the gender differences.

RESULTS

Overall Gender Differences on PISA Mathematics: 2000 and 2003

Table 3 shows the math performance of males and females on both 2000 and 2003 PISA assessments in the United States. There was no significant math gender difference on the 2000 assessments. However, in 2003, males significantly outperformed females. Because a large sample size is often associated with statistical significance, an effect size indicated by Cohen's d was provided for each comparison. The effect size of the gender performance difference was 0.13 for the 2000 math assessment and 0.11 for 2003, which are considered to be small effect sizes in social science research (Cohen, 1988). However, a small gender difference hardly suggests a negligible difference, because a small effect size

TABLE 3
Overall Math Performance in 2000 and 2003 PISA Assessments

	Male	Female			
Year	*Mean (SD)*	*Mean (SD)*	*Difference (SE)*	z	d
2000	.30 (.39)	.25 (.40)	.05 (.03)	1.34	.13
2003	.06 (.47)	.01 (.45)	.05 (.02)	2.17*	.11

Note. $^*p < .05$.
z statistic was calculated by dividing the difference by its standard error.
d = effect size, calculated by dividing the mean difference by the pooled standard deviation of the two groups.

[2]The effect size is calculated as $\sqrt{\dfrac{m_{boys} - m_{girls}}{(sd^2_{boys} + sd^2_{girls})/2}}$

may suggest some important practical difference. Discrepancies in performance could lead to a significant gap in the proportion of males and females in higher education, and an even larger gap in math-related occupations.

Students' mean math performance decreased from 2000 to 2003 by 0.24 logits. This may be a result of the wider range of skills that the 2003 PISA math assessment demands with the inclusion of two additional math domains.

Gender Differences Across Math Content Domains: 2000 and 2003

Table 4 presents the comparisons between males and females across math domains. In 2000, males and females did not show any significant difference on Change and Relationships items. Males scored significantly higher on Space and Shape items. In 2003, males outperformed females on Space and Shape, Change and Relationships, and Uncertainty items. No significant gender difference was observed for Quantity items. Results from administrations in both years indicated that males demonstrated some consistent advantage on Space and Shape items.

All of the effect sizes were below 0.20, suggesting small differences. The domain that showed that largest magnitude of difference was the Space and Shape domain (d = 0.14) in 2003. The smallest difference between males and females was found in the Quantity items (d = 0.04).

When examining the two-year data, there was no substantial change in the magnitude and pattern of gender differences from 2000 to 2003. The effect size increased by 0.02 points for the Space and Shape domain and decreased by 0.02 for the Change and Relationships domain, both suggesting small changes.

TABLE 4
Math Performance by Content Domain

	Males	*Females*			
	Mean (SD)	*Mean (SD)*	*Difference (SE)*	*z*	*d*
PISA 2000					
Space and Shape	−.20 (.24)	−.23 (.26)	.03 (.01)	3.20**	.12
Growth and Change	−.07 (.41)	−.12 (.42)	.05 (.03)	1.87	.12
PISA 2003					
Space and Shape	−.22 (.35)	−.27 (.34)	.05 (.01)	4.25**	.14
Change and Relationships	.08 (.85)	.00 (.78)	.08 (.02)	3.25**	.10
Quantity	.39 (.54)	.37 (.55)	.02 (.01)	1.68	.04
Uncertainty	−.01 (.16)	−.02 (.16)	.01 (.06)	1.99*	.06

Note. $^*p < .05$; $^{**}p < .01$.
z statistic was calculated by dividing the difference by its standard error.
d = effect size, calculated by dividing the mean difference by the pooled standard deviation of the two groups.

TABLE 5
Math Performance by Item Type

	Males	*Females*			
	Mean (SD)	*Mean (SD)*	*Difference (SE)*	*z*	*d*
PISA 2000					
Closed Constructed-Response	−.28 (.96)	−.38 (.95)	.10 (.05)	1.97*	.10
Multiple-Choice	.11 (.03)	.11 (.02)	.00	.00	.00
Open Constructed-Response	−.80 (.76)	−.86 (.75)	.06 (.50)	1.18	.08
PISA 2003					
Closed Constructed-Response	.42 (.61)	.37 (.61)	.05 (.03)	1.83	.10
Multiple-Choice	.10 (.11)	.10 (.11)	.00	.00	.00
Open Constructed-Response	−.95(1.00)	−1.03 (.94)	.08 (.04)	2.12*	.08
Short Response	.28 (.84)	.25 (.82)	.03 (.06)	.54	.14
Complex Multiple-Choice	.80 (.37)	.73 (.36)	.07 (.03)	2.61**	.19

Note. $^*p < .05$; $^{**}p < .01$.
z statistic was calculated by dividing the difference by its standard error.
d = Effect size, calculated by dividing the mean difference by the pooled standard deviation of the two groups.

Gender Differences within Item Type: 2000 and 2003

Table 5 presents the results of the gender comparisons in math performance on the five item types. Again, males showed consistent advantages on all of the item types for both 2000 and 2003 administrations, except for traditional multiple-choice items. No measurable difference was identified on the multiple-choice items for both 2000 and 2003 math assessments.

All the effect sizes were below 0.20, again suggesting small gender differences. Open constructed-response items showed the second smallest gender difference with an effect size of 0.08 for both administrations. Note that there are only three open constructed-response items for the year 2000, suggesting that the test may not be able to adequately measure student math proficiency within this item type. Males and females showed the largest performance gap on complex multiple-choice items (d = 0.19).

The pattern of gender differences across the 2000 and 2003 PISA administrations is fairly similar for each item type, suggesting that, with the exception of the traditional multiple-choice items, males consistently outperformed females, although the magnitude of these differences remained small.

CONCLUSIONS AND DISCUSSION

Three major findings emerged from the comparisons of PISA math performance among the 15-year-old males and females in the United States. First, the male advantage is small, but consistent. Males had higher overall performance than females on PISA math assessment for both the 2000 and 2003 administrations (see Table 3). In both years, males outperformed females within each content domain and item type (see Tables 4 and 5), except for on traditional multiple-choice items. The magnitude of the gender differences indicated by the effect sizes are fairly small (Cohen, 1988), all below 0.20. This finding resonates with a conclusion from many previous studies that the gender achievement gap has been declining over the last three decades (Cole, 1997; Feingold, 1988). For instance, there is an almost equal percentage of males and females taking advanced math courses. By the year 2000, among public high school graduates, 71% of the females had completed algebra II compared to 65% of the males; 11% of the females had completed calculus compared to 12% of the males; and 7% of the females had competed AP calculus compared to 8% of the males (U.S. Department of Education, 2005). Equal math training and experiences are likely to be contributors to the reduced gender gap. On the other hand, these results also reinforce concerns that gender differences in math achievement still exist and merit further attention. Nowell and Hedges (1998) examined the trend of gender differences in mathematics from 1960 to 1994, and found that differences in extreme scores are often substantial. Gender equity will not be achieved if there are considerable gaps in the high performing group, because most of the future mathematicians, scientists, and engineers will come from this cohort.

Second, no performance difference has been observed on traditional multiple-choice items on both the 2000 and 2003 assessments (see Table 5). As discussed earlier, many previous studies have reported a male advantage on multiple-choice items (e.g., DeMars, 2000, 1998; Klein et al., 1997; Murphy, 1982). Males are more likely to guess when they are unsure about the answers whereas females are more likely to the leave these items blank in similar situations (Ben-Shakhar & Sinai, 1991). This phenomenon has been hypothesized as one of the underlying explanations for the male advantage on multiple-choice items. However, using the same sample of U.S students examined in the present study, Liu (2006) estimated the percentage of random guessers on multiple-choice items in the PISA 2003 math assessment, using a two-class mixture model (Mislevy & Verhelst, 1990), and found that there were more guessing females (17%) than males (9%) on multiple-choice items. This finding provides a clue that there may have been some changes in females' guessing tendency, as they no longer left items unanswered when responding to multiple-choice items. This may be due to the growing popularity of test coaching and teacher instruction that females are now more aware of test-taking strategies that they have learned to maximize the opportunity to score higher. Of course, this speculation on guessing behavior requires further evidence. A first step would be to explore how males and females perform on multiple-choice items on other similar standardized tests

such as NAEP and TIMSS. Until replicated, the findings from the PISA assessments may be considered preliminary.

Third, males and females showed the largest gender differences on complex multiple-choice items and Space and Shape items. It is clear that the magnitude of gender differences varied across math domains and item types. Complex multiple-choice items displayed the largest effect size of gender difference among all item types ($d = 0.19$). An example of this item type is illustrated in Table 6. In general, complex multiple-choice items have a problem stem and several statements that surround the stem. Students are asked to make a judgment about the correctness of these statements. They obtain full credit only when all of the questions are answered correctly. This scoring scheme makes it more difficult for students to guess on complex multiple-choice items than on traditional multiple-choice items. The probability is very small that a student can guess right on four or five statements in a complex multiple-choice item. These kinds of items can be classified as "unconventional" item types in the sense that they do not usually appear in classroom assessment. Previous research has been documented that males performed better on unconventional or novel math tasks (Gallagher & De Lisi, 1994). The reasons are most likely two-fold. Females' inferior performance on novel tasks could be a byproduct of affective factors. Females tend to report lower confidence in learning math and higher anxiety in test-taking than males

TABLE 6
An Example of Complex Multiple-Choice Item

PAYMENTS BY AREA

People living in an apartment building decide to buy the building. They will put their money together in such a way that each will pay an amount that is proportional to the size of their apartment.

For example, a man living in an apartment that occupies one fifth of the floor area of all apartments will pay one fifth of the total price of the building.

Circle Correct or Incorrect for each of the following statements.

Statement	Correct / Incorrect
A person living in the largest apartment will pay more money for each square meter of his apartment than the person living in the smallest apartment.	Correct / Incorrect
If we know the areas of two apartments and the price of one of them we can calculate the price of the second.	Correct / Incorrect
If we know the price of the building and how much each owner will pay, then the total area of all apartments can be calculated.	Correct / Incorrect
If the total price of the building were reduced by 10%, each of the owners would pay 10% less.	Correct / Incorrect

Note. This item was selected from the sample math items provided by The PISA 2003 Assessment Framework-Mathematics, Reading, Science and Problem-Solving Knowledge and Skills (OECD, 2003, p. 67).

(Hyde et al., 1990; Wigfield et al., 1997). They also tend to underestimate their math ability, even when they demonstrate high proficiency in math learning (Reis & Park, 2001). Unfamiliar tasks are likely to hinder females' performance, because the unfamiliarity could become an additional psychological hurdle for females to overcome. Another reason for this might be related to the differential learning strategies males and females prefer in the learning process. Males tend to be independent in the learning process, willing to solve novel tasks, and persistent when encountering difficulties in problem solving; females, on the other hand, tend to rely on rote learning, featured by memorizing and following examples (Fennema & Peterson, 1985; Kimball, 1989).

Space and Shape items showed the largest effect size ($d = 0.14$) among the four domains in the 2003 assessment. Spatial ability has long been documented as a domain that displays large and robust gender differences. This finding resonates with previous research that reports that males show superior spatial ability on standardized tests (Casey et al., 1995; Gierl et al., 2003; Halpern, 1997; Hanna, 1986; Ryan & Chiu, 2001). In the K–12 curricula in the United States, spatial ability is not systematically taught. Outside of class, male students tend to have more spatial-related experience than female students through activities such as playing computer games, reading maps, and playing sports. There is a reliable correlation between spatial ability and spatial activity participation (Nuttall, Casey, & Pezaris, 2005). Spatial ability is also closely related to environmental influences (Baenninger & Newcombe, 1989). Therefore, insufficient exposure to spatial activities puts females at a disadvantage when being tested on spatial ability.

Although males have repeatedly demonstrated superior spatial ability, there is no evidence that these differences are genetically originated, nor does it suggest that this trend cannot be reversed. Researchers found that after adequate training, females performed equally well on items measuring spatial ability (Newcombe, Mathason, & Terlecki, 2002). Also, compared to the substantially large gender gap (e.g., $d = 0.66–0.89$) in spatial ability reported in previous studies (Linn & Petersen, 1986; Stumpf, 1995), the differences here, at least as measured by PISA, are fairly small, and could possibly be minimized through more systematic training in spatial ability (Newcombe et al., 2002). Therefore, females should be provided with more opportunities to participate in spatial related activities in and outside of class.

In future research, it would be both interesting and beneficial to replicate the analyses with data from other PISA participating countries. The male advantage in mathematics has shown some stability in many countries including Germany (Stumpf & Jackson, 1994), South Africa (Owen & Lynn, 1993), and Canada (Silverman et al., 1996). The patterns and magnitude of gender differences in math performance may vary across countries due to social, cultural, and psychological influences. For example, since 1972, there have been 144 winners in the U.S.A. Mathematical Olympiad contest for high school students, yet only two of them were females in the United States. However, there has been a female on the Chinese team for the four years China participated in the contest. These four females earned three silver medals and one gold medal (Halpern, 2000; Stanley, 1990). How the Chinese educational system has managed to produce mathematically talented females remains unclear to the U.S. researchers and educators. Through international comparisons, we may be able to find the underlying factors that will help minimize gender differences in math achievement. Such a factor could include a non-sexist educational environment (e.g., equal experience in class, parental belief), effective curriculum design and pedagogy, and practices that promote female math efficacy. Despite the robust gender differences in math achievement, there is some large cross-country variation in overall math performance. U.S. 8th graders ranked 28th in mathematics among the 45 participating countries in TIMSS (Colvin, 1996), and the 15 year olds from the United States ranked 27th in mathematics among the 40 participating countries in PISA (OECD, 2004). Future research should be directed to raise the overall standards in math performance in the United States, for both males and females.

ACKNOWLEDGMENTS

This article is based on work supported by the University of California Regents Dissertation Fellowship, 2006. We give special thanks to the two anonymous reviewers and the editor for their helpful comments, edits, and suggestions on an earlier draft of this article.

REFERENCES

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

Allen, D. D., Wilson, M. (2006). Introducing multidimensional item response modeling in health behavior and health education research. *Health Education Research, 21*(1), 73–84.

Armstrong, J. M. (1985). A national assessment of participation and achievement of women in mathematics. In S. F. Chipman, L. R. Brush, & D. M. Wilson (Eds.), *Women and mathematics: Balancing the equation* (pp. 59–94). Hillsdale, NJ: Lawrence Erlbaum.

Baenninger, M., & Newcombe, N. (1989). The role of experience in spatial test performance. A meta-analysis. *Sex Roles, 20*, 327–344.

Beller, M., & Gafni, N. (1996). *Can item format (multiple-choice vs. open-ended) account for gender differences in mathematics achievement?* Jerusalem: National Institute for Testing and Evaluation.

Beller, M., & Gafni, N. (2000). Can item format (multiple choice vs. open-ended) account for gender differences in mathematics achievement? *Sex Roles, 42*(1–2), 1–21.

Ben-shakar, G., & Sinai, Y. (1991). Gender differences in multiple-choice tests: The role of differential guessing tendencies. *Journal of Educational Measurement, 28*, 77–92.

Bolger, N., & Kellaghan, T. (1990). Method of measurement and gender differences in scholastic achievement. *Journal of Educational Measurement, 27*, 165–174.

Breland, H. M., Danos, D. O., Kahn, H. D., Kubota, M. Y., & Bonner, M. W. (1994). Performance versus objective testing and gender: An explanatory study of an advanced placement history examination. *Journal of Education Measurement, 31*, 275–293.

Bridgeman, B., & Wendler, C. (1991). Gender differences in predictors of college mathematics course grades. *Journal of Educational Psychology, 83*, 275–284.

Bridgeman, B., & Wendler, C. (2005). *Characteristics of minority students who excel on the SAT and in the classroom.* Princeton, NJ: Educational Testing Service.

Briggs, D. C., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement, 4*(1), 87–100.

Burton, N. W., & Lewis, C. (1996). Gender differences in college mathematics grades and SAT-M scores: A reanalysis of Wainer & Steinberg. *Journal of Educational Measurement, 33*, 257–270.

Buss, D. M. (1995). Evolutionary psychology: A new paradigm for psychological science. *Psychological Inquiry, 6*, 1–30.

Byrnes, J. P. (2005). Cognitive processes in an expanded framework. In A. M. Gallagher & J. C. Kaufman (Eds.), *Gender differences in mathematics* (pp. 73–98). Cambridge: Cambridge University Press.

Carlton, S. T., & Harris, A. M. (1989, March). *Characteristics of differential item performance on the Scholastic Aptitude Test—selected ethnic group comparisons.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Casey, M. B., Nuttall, R., Pezaris, E., & Benbwo, C. P. (1995). The influence of spatial ability on gender differences in mathematics college entrance test scores across diverse samples. *Developmental Psychology, 31*, 697–705.

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

Cole, N. S. (1997). *The ETS gender study: How females and males perform in educational settings.* Princeton, NJ: Educational Testing Service.

College Board (2005). *2005 College-bound seniors: Total group profile report.* Author.

Colvin, R. L. (1996, November 21st). Global study finds U.S. student weak in math. *Los Angeles Times*, A1, A24–25.

DeMars, C. E. (1998). Gender differences in mathematics and science on a high school proficiency exam: The role of response format. *Applied Measurement in Education, 11*(3), 279–299.

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education, 13*(1), 55–77.

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

Feingold, A. (1988). Cognitive gender differences are disappearing. *American Psychologist, 43*, 181–191.

Fennema, E., & Peterson, P. (1985). Autonomous learning behavior: A possible explanation of gender-related differences in mathematics. In L. C. Wilkinson & C. B. Marett (Eds.), *Gender influences in classroom interaction* (pp. 17–35). Orlando, FL: Academic Press.

Gallagher, A. M. (1992). *Sex differences in problem-solving strategies used by high-scoring examinees on the SAT-M* (College Board Rep. No. 92-2; ETS RR No. 92-93). New York: College Entrance Examination Board.

Gallagher, A. M. (1998). Gender and antecedents of performance in mathematics testing. *Teachers College Record, 100*(2), 297–314.

Gallagher, A. M., & De Lisi, R. (1994). Gender differences in scholastic aptitude test-mathematics problem solving among high ability students. *Journal of Educational Psychology, 86*(2), 204–211.

Gallagher, A. M., De Lisi, R., Holst, P. C., McGillicuddy-De Lisi, A. V., Morely, M., & Cahalan, C. (2000). Gender differences in advanced mathematical problem solving. *Journal of Experimental Child Psychology, 75*, 165–190.

Gaulin, S. J. C. (1995). Does evolutionary theory predict sex differences in the brain? In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 1211–1225). Cambridge, MA: MIT Press.

Geary, D. C. (1996). Response: A biosocial framework for studying cognitive sex differences. *Learning and Individual Differences, 8*, 55–60.

Gierl, M. J., Bisanz, J., Bisanz, G. L., & Boughton, K. A. (2003). Identifying content and cognitive skills that produce gender differences in mathematics: A demonstration of the multidimensionality-based DIF analysis paradigm. *Journal of Educational Measurement, 40*(4), 281–306.

Hanna, G. (1986). Sex differences in a casual model of mathematics achievement. *Journal for Research in Mathematics Education, 15*, 361–377.

Halpern, D. F. (1997). Sex differences in intelligence: Implications for education. *American Psychologist, 52*, 1091–1102.

Halpern, D. F. (2000). *Sex differences in cognitive abilities.* London: Lawrence Erlbaum Associates Publishers.

Harris, J. W. (1971). *Aspects of the guessing behavior of young Irish subjects on multiple-choice items.* Unpublished master's thesis, University College, Cork, Ireland.

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

Kimball, M. M. (1989). A new perspective on women's math achievement. *Psychological Bulletin, 105*, 198–214.

Klein, S. P., Jovanovic, J., Stecher, B. M., McCaffrey, D., Shavelson, R. J., Haertel, E., Solano-Flores, G., & Comfort, K. (1997). Gender and racial/ethnic differences on performance assessments in science. *Educational Evaluation & Policy Analysis, 19* (2), 83–97.

Lachance, J. A., & Mazzocco, M. M. (2006). A longitudinal analysis of sex differences in math and spatial skills in primary school age children. *Learning and Individual Differences, 16*, 195–216.

Linn, M. C., & Hyde, J. S. (1989). Gender, mathematics, and science. *Educational Researcher, 18*(8), 17–19, 22–27.

Linn, M. C., & Petersen, A. C. (1986). A meta-analysis of gender differences in spatial ability: Implications for ma and science achievement. In J. S. Hyde & M. C. Linn (Eds.), *The psychology of gender: Advances through meta-analysis* (pp. 67–101). Baltimore, MD: Johns Hopkins University Press.

Liu, O. L. (2006). *Evaluating differential gender performance on large-scale math assessments: A multidimensional Rasch modeling and mixture approach.* Unpublished doctoral dissertation, University of California, Berkeley.

Marshall, S. P., & Smith, J. D. (1987). Sex differences in learning mathematics: A longitudinal study with item and error analyses. *Journal of Educational Psychology, 79*, 372–383.

Maccoby, E. E, & Jacklin, C. N. (1974). *The psychology of sex differences.* London: Oxford University Press.

Mislevy, R., & Verhelst, N. (1990). Modeling item responses different subjects employ different solution strategies. *Psychometrika, 55*(2), 195–215.

Mullis, I. V. S., Martin, M. O., & Foy, P. (2005). *IEA's TIMSS 2003 international report on achievement in the mathematics cognitive domains.* Chest3nut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Gregory, K. D., Garden, R. A., O'Connor, K. M., Chrostowski, S. J., & Smith, T. A. (2000). *TIMSS 1999 international mathematics report.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V., Martin, M. O., Fierros, E. G., Goldberg, A. L. & Stemler, S. E. (2000) *Gender differences in achievement: TIMSS.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Murphy, R. J. L. (1982). Sex differences in objective test performance. *British Journal of Educational Psychology, 52*, 213–219.

National Science Foundation. (2000). *Women, minorities, and persons with disabilities in science and engineering: 2000* (NSF Publication No. 00-327). Arlington, VA: Author.

Newcombe, N. S., Mathason, L., & Terlecki, M. (2002). Maximization of spatial competence: More important than finding the cause of sex differences. In A. McGillicuddy-De Lisi & R. De Lisi (Eds.), *Biology, society and behavior: The development of sex differences in cognition* (pp. 183–206). Westport, CT: Ablex Publishing.

Nowell, A., & Hedges, L. V. (1998). Trends in gender differences in academic achievement from 1960 to 1994: An analysis of differences in mean, variance, and extreme scores. *Sex Roles, 39*(112), 21–43.

Nuttall, R. L., Casey, M. B., & Pezaris, E. (2005). Spatial ability as a mediator of gender differences on mathematics tests. In A. M. Gallagher & J. C. Kaufman (Eds.), *Gender differences in mathematics* (pp. 121–142). Cambridge: Cambridge University Press.

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* (pp. 255–279). Hillsdale, NJ: Erlbaum.

O'Neill, K. A., Wild, C. L., & McPeek, W. M. (1989). *Gender-related differential item performance on graduate admissions tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Organization for Economic Co-operation and Development (OECD) (2000a). *Knowledge and skills for life: First results from PISA 2000.* Paris: Author.

Organization for Economic Co-operation and Development (2000b). *PISA 2000 technical report.* Paris: Author.

Organization for Economic Co-operation and Development (2003). *The PISA 2003 assessment framework-mathematics, reading, science and problem-solving knowledge and skills.* Paris: Author.

Organization for Economic Co-operation and Development (2004). *Learning for tomorrow's world: First results from PISA 2003.* Paris: Author.

Organization for Economic Co-operation and Development (2005a). *PISA 2003 technical report.* Paris: Author.

Organization for Economic Co-operation and Development (2005b). *PISA 2003 data analysis manual: SPSS Users.* Paris: Author.

Owen, K., & Lynn, R. (1993). Sex differences in primary cognitive abilities among Blacks, Indians and Whites in South Africa. *Journal of Biosocial Science, 25*, 557–560.

Perie, M. (1994). Differential performance on mathematics achievement between the sexes explained through differential strategy selection. *Dissertation Abstracts International*, 55(05), 1250. (University Microfilms No. AAC 9425685)

Perie, M., Moran, R., & Lutkus, A. D. (2005). *NAEP 2004 trends in academic progress: Three decades of student performance in reading and mathematics* (NCES 2005-464), figures 2-4, 3-5, 3-6, and 3-7. Data from U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), various years, 1973–2004 Long-Term Trend Mathematics Assessment.

Reis, S. M., & Park, S. (2001). Gender differences in high-achieving students in math and science. *Journal for the Education of the Gifted, 25*, 52–73.

Ryan, K. E., & Chiu, S. (2001). An examination of item context effects, DIF, and gender DIF. *Applied Measurement in Education, 14*, 73–90.

Schram, C. (1996). A meta-analysis of gender differences in applied statistics achievement. *Journal of Educational and Behavioral Statistics, 21*(1), 55–70.

Silverman, I., Phillips, K., & Silverman, L. K. (1996). Homogeneity of effect sizes for sex across spatial tests and cultures: Implications for hormonal theories. *Brain and Cognition, 31*, 90–94.

Stanley, J. (1990, January). We need to know why women falter in math. *The Chronicle of Higher Education*, B4.

Stumpf, H., & Jackson, D. N. (1994). Gender-related differences in cognitive abilities: Evidence from a medical school admissions testing program. *Personality and Individual Differences, 17*, 335–344.

Stumpf, H. (1995). Gender differences in performance on tests of cognitive abilities: Experimental design issues and empirical results. *Learning and Individual Differences*, 7(4), 275–287.

U.S. Bureau of the Census. (2001). *Statistical abstract of the United States: 2001 (121st ed.).* Washington, DC: Author.

U.S. Department of Education. (2005). *The digest of education statistics: 2005.* Washington, DC: National Center for Education Statistics.

U.S. Department of Education. (2001). *The NAEP 1998 technical report* (NCES 2001-509). Washington, DC: National Center for Education Statistics.

U.S. Department of Education. (2003). *Third international mathematics and science study 1999 video study technical report volume I: Mathematics* (NCES 2003-012). Washington, DC: National Center for Education Statistics.

Wang, W. (1999). Direct estimation of correlations among latent traits within IRT framework. *Methods of Psychological Research Online*, 4(2), 47–70.

Wang, W., Wilson, M., & Adams, R. J. (1997). Rasch models for multidimensionality between and within items. In M. Wilson and G. Engelhard, (Eds.), *Objective measurement: Theory into practice.* Vol IV. Norwood, NJ: Ablex.

Wester, A. (1995). The importance of item format with respect to gender differences in test performance: A study of open-format items in the DTM test. *Scandinavian Journal of Educational Research, 39*, 335–346.

Wester, A., & Henriksson, W. (2000). The interaction between item format and gender differences in mathematics performance based on TIMSS data. *Studies in Educational Evaluation, 26*, 79–90.

Wigfield, A., Eccles, J. S., Yoon, K. S., Harold, R. D., Arbreton, A. J. A., Freedman-Doan, C., & Blumenfeld, P. C. (1997). Change in children's competence beliefs and subjective task values across the elementary school years: A 3-year study. *Developmental Psychology, 89*, 451–468.

Willingham, W.W., & Cole, N. S. (1997). *Gender and fair assessment.* Mahwah, NJ: Lawrence Erlbaum Associates.

Wolfe, E., Viger, S. G., Jarvinen, D. W., & Linksman, J. (2007). Validation of scores from a measure of teachers' efficacy toward standards-aligned classroom assessment. *Educational and Psychological Measurement, 67*(3), 460–474.

Wu, M., & Adams, R. J. (2002, April). *Plausible values—Why they are important.* Paper presented at the International Objective Measurement Workshop, New Orleans.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *ConQuest.* Melbourne, Australia: ACER.

APPENDIX

The procedure recommended by the PISA 2003 technical report (OECD, 2005a) was used for booklet correction. The procedure involves two major steps: modeling the booklet effect, and estimating the booklet parameters. In the first step, the booklet effects were included in the *ConQuest* model statement to eliminate the confounding of item difficulties and booklet effects. The calibration model is expressed as *item + item * step + booklet.*

In the second step, the booklet parameters were estimated through a regression model in *ConQuest*, which produces *n–1* dummy variables for *n* booklets, with one booklet being the reference booklet. For example, 12 dummy variables were created for the 13 booklets in the PISA 2000 assessments. The booklet parameter estimate was then added to the proficiencies of students who responded to that particular booklet. Concretely, a positive value indicates that a booklet is harder than the average and a negative value indicates that a booklet is easier than the average.

APPLIED MEASUREMENT IN EDUCATION, 22: 185–205, 2009

ISSN: 0895-7347 print / 1532-4818 online
DOI: 10.1080/08957340902754650

Routledge
Taylor & Francis Group

Correlates of Rapid-Guessing Behavior in Low-Stakes Testing: Implications for Test Development and Measurement Practice

Steven L. Wise and Dena A. Pastor
Center for Assessment and Research Studies
James Madison University

Xiaojing J. Kong
Pearson Educational Measurement

Previous research has shown that rapid-guessing behavior can degrade the validity of test scores from low-stakes proficiency tests. This study examined, using hierarchical generalized linear modeling, examinee and item characteristics for predicting rapid-guessing behavior. Several item characteristics were found significant; items with more text or those occurring later in the test were related to increased rapid guessing, while the inclusion of a graphic in a item was related to decreased rapid guessing. The sole significant examinee predictor was SAT total score. Implications of these results for measurement professionals developing low-stakes tests are discussed.

The purpose of achievement testing is to measure an examinee's level of proficiency in a specified content domain. A basic requirement in this endeavor is that the test being used has been competently constructed and is capable of providing reliable and valid test scores. However, there is an additional requirement—that the examinee will try his or her best to answer the test items correctly. Without adequate effort, performance is likely to suffer, resulting in the examinee's test

Correspondence should be addressed to Steven L. Wise, Northwest Evaluation Association, 5885 SW Meadows Rd., Suite 200, Lake Oswego, OR 97035. E-mail: steve.wise@nwea.org