JOURNAL OF APPLIED MEASUREMENT, *9*(1), 18-35

A Multidimensional Rasch Analysis of Gender Differences in PISA Mathematics

Ou Lydia Liu
Educational Testing Service, Princeton

Mark Wilson
University of California, Berkeley

Insu Paek
Educational Testing Service, Princeton

Since the 1970s, much attention has been devoted to the male advantage in standardized mathematics tests in the United States. Although girls are found to perform equally well as boys in math classes, they are consistently outperformed on standardized math tests. This study compared the males and females in the United States, all 15-year-olds, by their performance on the PISA 2003 mathematics assessment. A multidimensional Rasch model was used for item calibration and ability estimation on the basis of four math domains: Space and Shape, Change and Relationships, Quantity, and Uncertainty. Results showed that the effect sizes of performance differences are small, all below .20, but consistent, in favor of boys. Space and Shape displayed the largest gender gap, which supports the findings from many previous studies. Quantity showed the least amount of gender difference, which may be explained by the hypothesis that girls perform better on tasks that they are familiar with through classroom practice.

Requests for reprints should be sent to Ou Lydia Liu, Educational Testing Service, 666 Rosedale Road, Mail Stop 16-R, Princeton, NJ 08541, USA, e-mail: lliu@ets.org.

Introduction

Since the 1970s, much attention has been devoted to the male advantage in standardized mathematics tests in the United States. In their seminal volume on gender differences, Maccoby and Jacklin (1974) concluded that males generally score higher on mathematics tests than females and the average difference is about .5 (in standard deviation units) for high school students. Male advantage on the SAT-mathematics has remained more or less constant throughout the past three decades (Ben-Shakhar and Sinai, 1991; Feingold, 1988; Gallagher and Kaufman, 2005). Males' superior performance in mathematics also appears in NAEP, the largest national educational assessment. According to the 2005 NAEP report (NCES, 2005), 4th grade males outperformed females in the mathematics assessment from 1990 to 2005, and the same finding holds for 8th grade, except in years 1992 and 1996.

Gender differences in mathematics in favor of males tend to enlarge among highly selective groups (Benbow and Stanley, 1980, 1983; Hedges and Nowell, 1995). Mathematically talented students remained disproportionately males at the high end of the ability distribution (Becker and Hedges, 1984; Feingold, 1988; Rosenthal and Rubin, 1982). Based on 16 years of research on intellectually talented 12- and 13-year-old students, Benbow (1988) concluded that there are noticeable differences on the SAT-M in favor of males.

A glimpse at some of the large-scale standardized international assessments also indicates a male advantage in mathematics assessment. Boys outperformed girls significantly in both 2000 and 2003 PISA mathematics in the United States (OECD, 2000, 2004). Also, the TIMSS results revealed that 8th grade boys performed better than girls on mathematics assessment consecutively from 1995 to 2003 (TIMMS, 2000 a and b, 2003).

However, there have been some controversies about the existence and magnitude of gender differences in math. The conclusion of established gender differences in mathematics by Maccoby and Jacklin (1974) was challenged by some later studies which reported that gender differences in mathematics have declined over the years. For the SAT-Mathematics test, the difference has shrunk from the usual 40 points (d = .39)[1] to 33 points in 2005 (College Board, 2005). The ACT-Mathematics test has shown a similar pattern with a difference of 3.1 points in 1970 decreasing to 1.2 points in 2001 (Langenfield, 1997; U.S. Office of Education, 2001). In a review paper titled "Cognitive Gender Differences Are Disappearing", Feingold (1988) concluded that from 1960 to 1983 the gender differences on the PSAT-Mathematical declined from .34 to .17 (measured in standardized effect size d), a value he considered of trivial practical significance. Linn and Hyde (1989) conducted a meta-analysis study on gender differences in cognitive abilities and found that since 1974, gender differences in math ability have declined from .31 to .14 (measured in standardized effect size *d)*. Furthermore, many studies reported that boys and girls gained equal grades in schools and sometimes girls performed better than boys (Kimball, 1989; Xie and Shauman, 2003), and men and women get equal grades in college math classes that are matched for the difficulty of the classes (Bridgeman and Lewis, 1996).

The declining gender differences in mathematics performance have triggered some heated debate on whether the research on gender differences should continue and if so, which direction to follow and what emphasis to take. Some researchers argue that further research on gender differences will reinforce the stereotype impression and bring detrimental impact to females. Halpern and Ikier (2002) refute this allegation by raising the point that those who are against the research of gender differences implicitly believe that the research results will show that females are inferior to males. Gender gaps need to be studied before they can be closed. Especially so far as there is no solid evidence that gender differences

[1] Effect size d is computed as the difference between the female mean and the male mean, divided by the pooled within-group standard deviation. Cohen (1988) describes an effect size of .2 as small, .5 as medium and .8 as high.

have been eliminated, nor can we justify that the existing difference is indeed negligible.

Small effect sizes in performance can possibly suggest large practical differences (Cole, 1997). The gender gap in math test scores is likely to have some impact on females' career aspirations and occupational choice (Wigfield, Battle, Keller, and Eccles, 2002). Females comprise less than a third of all bachelor's degree recipients in the physical and computer sciences, and less than a quarter of all graduate degree recipients in engineering and mathematics, although they account for over half of the student body (National Science Foundation, 2000). Females only comprise 8% of the mathematics professors in the United States. Many of the male-dominant professions that require a higher level of mathematical ability belong to the prestigious and high-paying job categories. This brings some unwarranted divide in male/female socioeconomic status. There are a few underlying factors that might contribute to females' decision to stay away from math related careers (1) the widely circulated belief, stereotype thinking that girls do not perform well in math (Steele, 1997), and the lower parental expectation of females' math achievement (Eccles and Jacobs, 1986) (2) the fact that boys get better scores on high stakes standardized math tests, and (3) in general girls' lower confidence in learning math (Wigfield, Battle, Keller, and Eccles, 2002). It might be a chicken and egg issue as which factor comes first and causes the others. Halpern and Ikier (2002) advocate for a psychobiosocial model of cognitive gender differences, which recognizes the joint impact of psychological, biological, and social factors on the progress and Bandura's (1977) development of cognitive abilities. According to Bandura (1977) self-efficacy theory, one's past success or failure in executing specific tasks largely determines their efficacy in carrying out these tasks. If this is the case, improved math performance will be accompanied by enhanced efficacy for females to master mathematical understanding. Research also suggests that as girls get older, they tend to be less constrained by the self-limiting gender notion that females can only undertake certain jobs (Jackson and Tein, 1998; Sanberg, Ehrhardt, Melins, Ince, and Meyer-Bahlburg, 1987).

On the basis of the above discussion, there is a need to further clarify the relationship between math and gender on standardized tests in order to achieve gender equality in math learning. Ample evidence suggests that girls still underperform on standardized tests, and, standardized tests present critical thresholds for student future success. Very often math test scores are used in important admission decisions and award selections. In the era of the No Child Left Behind Act, both school performance and student learning are largely determined by large-scale standardized tests. To carry out the investigations, it is important to keep in mind that mathematical ability is multidimensional (De Lisi and McGillicuddy-De Lisi, 2002), as are gender differences in mathematics. Previous research has not devoted enough effort to examining various types of mathematical domains and processes.

To examine gender differences on large-scale standardized tests, the 2003 Programme for International Student Assessment (PISA) mathematics data were analyzed in this study. A multidimensional item response modeling was used to examine the four domains of the PISA mathematics assessment. Many of the previous efforts were thwarted by the fact that many studies do not go beyond reporting a mean gender difference in math test scores. The total score approach ignores the variability in the magnitude of gender differences across dimensions if multiple subdomains are measured in the math test. A potential consequence of this is that we fail to provide effective diagnostic information for classroom teachers about student strengths and weaknesses in a particular math field.

This paper aims to serve two primary objectives (1) investigate the empirical evidence supporting the four sub-content domains using a four-dimensional model, and (2) investigate gender performance within and across the four math domains.

If differential gender performance is identified within math domains, we want to know how the differences occur across domains. Gender differences are not homogenous across math content areas. Studies found that males tend to outperform females on tasks involving reasoning and problem solving ability, tasks that measure spatial skills and tasks that require multiple solution strategies (Burton and Lewis, 1996; Doolittle and Cleary, 1987; Gallagher, 1998; O'Neil and McPeek, 1993). Females have been found to score higher on items measuring computational skills, items that involve retrieving information from working memory, and items that involve an extensive amount of reading or explanation (Casey, Nuttall, Pezaris, and Benbow, 1995; Gierl, Bisanz, Bisanz, and Boughton, 2003; Halpern, 1997).

Method

Instrument

Data from the PISA 2003 mathematics assessment were analyzed. There are 84 math items in the PISA mathematics test, measuring student ability in four math domains: Space and Shape, Change and Relationships, Quantity, and Uncertainty. All of the following analyses were based on these four math domains. The math items are represented by five item types: short response, closed constructed-response, open constructed-response, multiple-choice and complex multiple-choice items. The item distribution is provided in Table 1.

Compared to other curricula-dependent mathematics surveys, PISA aims at measuring the educational "yield" of 15-year-old students. The primary goal of PISA is to monitor educational progress and provide a basis for international comparison (OECD, 2004). The PISA math assessment emphasizes student ability to make sound judgments, demonstrate evidence, and apply knowledge and skills in real-life situations.

Participants

Altogether 5456 students in the United States participated in the PISA 2003 math assessment, including 2740 boys (50.2%), 2715 girls (49.8%), and one student with missing gender value. The students are exclusively 15-year olds. PISA seeks to measure how well students at this age are prepared to meet the challenges of the knowledge societies. Studies have found that boys and girls showed no measurable difference in math performance during the elementary years (Byrnes, 2005; Hyde, Fennema, and Lamon, 1990). Larger differences are more likely to be found in the later-developing stages of mathematical cognitive abilities (Geary, 1996). Around age 15, a small to moderate gender difference appears in math performance in favor in boys (Byrnes, 2005). Therefore, 15-year-olds are a suitable target population for this study.

The accuracy of the PISA survey depends on the quality of its sampling procedure. Two-stage stratified samples were used in the PISA study. The first stage consisted of sampling individual schools where 15-year-old students were present. 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 from each country.

Table 1

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

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

Replacement schools were also selected, in case a sampled school could not participate for some reason (OECD, 2004). The second stage of the sampling involved sampling students within selected schools. Once schools were selected, a list of each school's 15-year-old students was documented. Then 35 students were selected with equal probability. The strict stratified probability sampling procedure ensures the generalizibility of the findings. The inferences apply to not only the students who participated in this survey, but also to the 15-year-old population who could have taken the test (OECD, 2004).

MRCMLM

To examine the dimensionality of the PISA math assessment and make comparisons of gender differences across dimensions, a multi-dimensional measurement model is needed. The multidimensional random coefficient multinomial logit model (MRCMLM; Adams, Wilson, and Wang, 1997) was selected for the item calibration and ability estimation in this study. The software program ConQuest (Wu, Adams, and Wilson, 1998) was used.

MRCMLM is a generalized Rasch type item response model, which is a multidimensional extension of the random coefficient multidimensional logit model (Adams and Wilson, 1996). Within the MRCMLM framework, many existing IRT models are shown to be its special cases, such as the simple logistic model (Rasch, 1960), the rating scale model (Andrich, 1978), the partial credit model (Masters, 1982), the ordered partition model (Wilson, 1992), the linear logistic test model (Fisher, 1983), the multi-facet model (Linacre, 1994), and the multidimensional versions of these models. Its flexibility comes from the use of a scoring function and a design matrix. The former allows users to specify individual item weights (not empirically estimated) in each dimension and the latter allows users to specify item parameters in a linear form. For example, assigning equal weight to all items in the scoring function and specifying one item location parameter for the item property parameter in the design matrix leads to the simple Rasch model for dichotomous response data. The capability to accommodate specified item weights in its item response modeling introduces flexibility when priori weights are decided based on some theoretical or practical reasons.

The MRCMLM is formulated as:

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

where X_{nik} = 1 if person n's response to item i is in category k, or 0 otherwise ($1\leq i \leq \mathrm{I}$, $1\leq k \leq \mathrm{K}_i$, $1\leq n \leq \mathrm{N}$, and X_{ni1} is fixed to zero as a reference category for model identification); $\boldsymbol{\theta}_n$ is a d x 1 proficiency (or ability) parameter vector of a person n ($1 \leq d$ (dimension) $\leq \mathrm{D}$); $\mathbf{b}'_{ik}$ is a 1 x d scoring vector for category k of item i; ξ is a p x 1 item parameter vector; and $\mathbf{a}'_{ik}$ is a 1 x p vector to specify linear combination of p elements of ξ for each response category.

ξ is a fixed unknown parameter vector while $\boldsymbol{\theta}_n$ is a random parameter vector. The elements of $\boldsymbol{\theta}_n$ are assumed to follow a multivariate normal (MVN) distribution:

$$\boldsymbol{\theta}_n \sim \mathrm{MVN}(\boldsymbol{\mu}, \textstyle\sum), \quad (2)$$

where $\boldsymbol{\mu}$ is a 1 x d mean vector and $\sum$ is a d x d variance-covariance matrix. The mean vector of $\boldsymbol{\mu}$ and the variance-covariance matrix of $\sum$ are fixed unknown parameters. The parameter estimates of ξ , $\boldsymbol{\mu}$, and $\sum$ are obtained by maximizing the following marginal maximum likelihood (MML):

$$L=\prod_n\int\prod_i P\left(X\,|\,\boldsymbol{\theta}_n\right)dG\left(\boldsymbol{\theta}_n\,|\,\mu,\Sigma\right), \quad (3)$$

where $\mathrm{P}(\mathrm{X}|\boldsymbol{\theta}_n)$ is Equation (1) and $\mathrm{G}(\boldsymbol{\theta}_n|\boldsymbol{\mu}, \sum)$ is the cumulative distribution function of Equation (2). For the person ability parameter $\boldsymbol{\theta}_n$, ConQuest produces expected a posterior (EAP) estimates, maximum likelihood estimate (MLE) and weighted likelihood estimates (WLE). This study used EAP estimates for student ability estimation.

A four-dimensional model was analyzed based on the four mathematics content domains (see Table 1). Because the test consisted of dichotomously and polytomously scored items, the MRCMLM is adjusted to a four-dimensional Rasch model for the dichotomous items, and to a four-dimensional partial credit model for the polytomous items. In this four-dimensional model, each item only loads on one dimension, which is referred to as a between-item multidimensional model[2] (Wang, Wilson, and Adams, 1997).

Results

ConQuest uses the MML method to estimate the item parameters and EAP methods were used to produce the student ability estimates. The joint prior distribution of student abilities was obtained during the MML item parameter estimation process (Adams, Wilson, and Wang, 1997; DeMars, 2004).

First, we present the evidence that supports the application of the four dimensional model, which in return supports the four-domain structure of the PISA math assessment. Second, we present the results from the gender difference analysis. Since multiple booklets were used in PISA 2003 math assessment, there is a need to adjust the potential booklet impact on student performance. The booklet effects were analyzed and used to adjust to the analyses results (see Appendices A and B).

Comparing to the unidimensional model

Here the unidimensional Rasch model is nested in the four-dimensional model, meaning that by applying some constraints (i.e., constraining all the inter-dimensional correlation to 1.0) to the four-dimensional model the unidimensional model is obtained. The difference in deviance between the multidimensional model and the unidimensional model approximately follows a chi-square distribution and can be used to provide index of model fit. The difference in deviance statistics of these two models is 930.3 with 9 degrees of freedom, where the degrees of freedom are the difference in the number of parameters estimated in the unidimensional and multidimensional model. The difference in deviance is statistically significant at the $\alpha = .001$ level. This provides statistical support for the use of a four-dimensional model, along with the theoretical support on the basis of the assessment design for the four content topics.

Wright Map

In the MRCMLM application, item parameters and student estimates were calibrated to be on the same logit metric, so that within a single dimension all model parameter estimates can be compared on the same scale. The Wright map is a visual representation of the relative relations between item and person estimates. It is ideal that the item difficulty distribution will cover the span of the student ability distribution, thus providing accurate measures of student proficiency over the whole scale. The information elicited from students will be maximized when the item difficulty level is close to the student ability level. A lack of items in a difficulty range will lead to large errors in ability estimation. In Figure 1, the math items cover the student ability distributions of the four dimensions quite well, except for the Quantity dimension. The ability distribution is more dispersed for the Change and Relationships, and it is more peaked for the Space and Shape and Uncertainty dimension. In either case, there are sufficient items along the continuum to provide accurate ability estimates across the whole range of students. For the Quantity dimension, even the more difficult items are relatively easy for the top students. That is related to the nature of the Quantity domain, most of the items involving basic calculation or symbolic representation. No higher-order thinking is required for this kind of items.

Both male and female mean ability estimates vary across the four content topics (Table 2). Both genders obtained the highest estimate on Quantity items, followed by Change and Relationships, Uncertainty, and Space and Shape. The variance of student performance also varies across dimensions. The Change and Relationships dimension displayed the largest variance, followed by Quantity, Space and Shape, and Uncertainty dimensions.

[2] The other multidimensional model is within-item multidimensional model, where one item loads on more than one dimension (Wang, Wilson, and Adams, 1997).

```
Logit | Space and Shape  | Change and     | Quantity        | Uncertainty       | Booklet
      |                  | Relationships  |                 |                   |
      |       |          |     |          |      |          |           | 40    | 
    3 |       |          |     |          |      |          |           |       | 
      |       |          |     |          |    X |          |           |       | 
    2 |       | 17       |   X | 26       |      |          |           |       | 
      |       | 16 18    |   X |          |    X |          |           | 78    | 
      |       | 20       |   X | 23 31 37 |   XX |          |           |       | 
      |       |          |  XX |          |   XX | 63       |           |       | 
      |       | 3 8      |   X |          |   XX |          |           | 83    | 
      |       |          |   X | 25       |  XXX | 57       |           |       | 
      |     X |          |  XX | 29       |  XXX | 56 62    |         X | 66    | 
    1 |     X | 1        |  XX |          | XXXX | 46       |        XX |       | 
      |    XX | 5        |  XX | 21       | XXXX | 53       |     XXXXX | 74    | 
      |   XXX |          |  XX | 42       | XXXX | 54       |    XXXXXX |       | 
      |  XXXX |          |  XX |          | XXXX |          | XXXXXXXXX | 80 82 | 
      |  XXXX |          | XXX | 24       | XXXX |          | XXXXXXXXX |       | 
      | XXXXX |          |  XX |          |  XXX | 44 45 47 |  XXXXXXXX | 84    | 
      | XXXXX | 7        | XXX |          |  XXX | 59       |     XXXXX | 77    | 10 12
    0 | XXXXX | 2        | XXX |          |   XX | 43 49 52 |       XXX | 71    | 2 3 4 5
      | XXXXX | 14       | XXX |          |   XX |          |         X | 73 81 | 6 8 9
      |       |          |     | 22 32 34 |      |          |           |       | 
      | XXXXX | 9        |  XX | 39       |    X |          |           |       | 1 7 11
      |       |          |     |          |      | 51 55 58 |           |       | 
      |   XXX |          |  XX |          |    X | 61       |           | 67    | 13
      |    XX | 6        |  XX |          |    X |          |           | 70 72 | 
      |    XX | 13       |  XX | 33       |    X |          |           |       | 
      |       |          |     |          |      |          |           | 65 69 | 
      |     X | 12       |  XX | 36       |      |          |           | 75    | 
   -1 |       |          |  XX | 35       |      |          |           | 76 79 | 
      |       |          |  XX |          |      |          |           |       | 
      |       |          |   X |          |      |          |           | 68    | 
      |       | 10       |   X |          |      |          |           | 10    | 
      |       |          |   X |          |      |          |           |       | 
      |       | 4        |   X | 28 30    |      |          |           |       | 
      |       |          |     |          |      | 48 50 60 |           |       | 
   -2 |       |          |     |          |      |          |           |       | 
      |       | 15       |     | 38       |      | 64       |           |       | 
      |       |          |     | 41       |      |          |           |       | 
      |       | 11 19    |     |          |      |          |           |       | 
   -3 |       |          |     |          |      |          |           |       | 
   -4 |       |          |     | 27       |      |          |           |       | 
```

Figure 1. Wright map for the four mathematics dimensions
Note: Each "X" represents about 107 cases.

Table 2
Boys and Girls' Performance within Each Math Content Domain

	Males		Females			
Content Domain	Mean	*SD*	Mean	*SD*	*t*	Effect Size
Space and Shape	–.22	.35	–.27	.34	5.35**	.14
Change and Relationships	.08	.85	.00	.78	3.62**	.10
Quantity	.39	.54	.37	.37	1.60	.04
Uncertainty	.01	.16	–.02	.16	.69	.12
Overall	.06	.47	.01	.45	4.32**	.11

Note: ** *p* < .01.

Effect size is indicated by Cohen's *d*, as calculated by $d = \frac{mean_1 - mean_2}{\sqrt{SD_1^2 + SD_2^2}}$ (Cohen, 1988).

Correlation between dimensions

The correlation between dimensions reported from ConQuest ranges from 0.77 to 0.88 (see Table 3). Note that the correlation produced in the ConQuest analysis is not the raw correlation between student ability estimates. These correlations are disattenuated or corrected for error so they are relatively free of measurement noise stemming from various sources (Briggs and Wilson, 2003, Wang, et al., 1997; Wang, 1994). The Quantity dimension is highly correlated with the other three dimensions (all correlation coefficients were the same at .88). Quantity items deal with the recognition of numerical patterns, and the processing and understanding of numbers in a real-life situation. These competencies are fundamental skills to carry out and complete other math activities. A correct answer to items in the other three domains depends on quantity knowledge to some extent. The two dimensions that showed the lowest correlation were the Space and Shape dimension with the Uncertainty dimension. Except for the fact that they both fall under the math umbrella, these two domains are loosely connected. For example, calculation of probability items does not usually involve mental rotation of objects or vice versa. On average, the correlations among the four dimensions are reasonably high, given the fact that they all measure student "math ability".

Table 3

Correlations/Covariance between Dimensions

Covariance/Correlation Matrix				
Dimension	1	2	3	4
1 Space and Shape		.49	.33	.13
2 Change and Relationships	.85		.71	.31
3 Quantity	.88	.88		.20
4 Uncertainty	.77	.87	.88	

Note: Values below the diagonal are correlations and values above are covariances.

Fit statistics

For each item parameter, ConQuest provides a weighted fit mean square (WFMS) statistic to indicate the fit between the items and the measurement model used to calibrate the items. Fit statistics can detect the discrepancies between the modeled assumptions and the empirical data (Wright and Masters, 1982; Wu, Adams, and Wilson, 1997). WFMS can be expressed as

$$\sum_{n=1}^{N} y_{ni}^{2} / \sum_{n=1}^{N} w_{ni}$$

(see Appendix C for details), with y_{ni} being the score residual between observed score and expected score for person n on item i, and w_{ni} being the variance of the observed score for person n on item i. This index is weighted so that responses made by respondents whose ability is far above or below that item will have less influence on the item fit statistics (Wright and Masters, 1982). WFMS is expected to have a value of 1. The WFMS values between .75 and 1.33 are considered acceptable by convention (Adams and Khoo, 1996). Items with fit index larger than 1 suggest more variability in the data than the model explained, and items with fit index less than 1 indicate less variability than the model expected.

Figure 2 illustrates the WFMS value for each item by the four content topics. All of the items showed reasonably good fit with the model. A Space and Shape item and an Uncertainty item displayed relatively large WFMS values, 1.10 and 1.13, respectively. Both items are among the most difficult items in the PISA math assessment. The large fit statistic could result from the fact that the items are so difficult that they failed to distinguish among students of different ability levels. There is more variability in the item responses than the model predicts.

Gender comparisons: overall and by each dimension

Table 2 also shows the descriptive statistics for the two genders on overall performance as well as within each content domain. On average, boys scored higher than girls on the PISA 2003 assessment with statistical significance. Boys also performed better across all of the four math domains. The gender performance difference was statistically significant on the Space and Shape dimension and the Change and Relation-

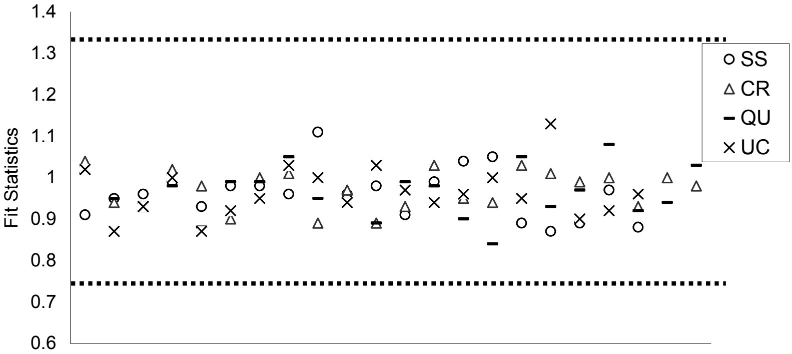

Figure 2. Fit statistics by content domain
Note: The legends stand for Space and Shape, Change and Relationships, Quantity, and Uncertainty, respectively

ships dimension. An effect size of the difference was calculated for each domain using Cohen's *d* (Cohen, 1969, 1988). All the effect sizes were below .20. However, a small effect size could be of important significance if it reflects a systematic difference (Cole, 1997a). The domain that displayed that largest effect size was the Space and Shape dimension (*d* = .14). It has been well documented that boys demonstrate superior spatial ability (Casey, Nuttall, Pezaris, and Benbow, 1995; Gierl, Bisanz, Bisanz, and Boughton, 2003; Halpern, 1997; Linn and Hyde, 1989). There is a great deal of controversy about the possible causes of these differences. Some researchers argue that the differences are the result of biological causes (Bock and Kolakowski, 1971; Kimura, 1992). Many others challenge the proposal of innate differences by providing evidence of noticed relationships between spatial ability and gender-role beliefs. In fact, gender gap in spatial ability can be decreased through specific training (Baenninger and Newcombe, 1989; Peters, Chisholm, and Laeng, 1995; Quaiser-Pohl and Lehmann, 2002). The gender differences in spatial ability are not homogeneous across students in different majors. Larger gender differences have been identified among students majoring in arts, humanities and social sciences than those majoring in computational visualistics[3] (Quaiser-Pohl and Lehmann, 2002). In current K-12 classrooms, no specific curriculum has been designed to promote student spatial ability. Based on the authors' general observation (in eastern and western countries), boys tend to spend more time on space-related activities than girls in daily life. Boys are more likely to engage themselves in sports, car toys, and playing with spatial video games (Baenninger and Newcombe, 1989; Halpern and Ikier, 2002). The amount of exposure to space-related activities has a substantial impact on the quantity and quality of student spatial ability. Indirectly, greater experiences may enhance math performance through increased familiarity with mathematical thinking and greater confidence (Kimball, 1989). In fact, females' spatial ability has been proven to increase through an appropriate amount of training (Baenninger and Newcombe, 1989; Newcombe, Mathason and Terlecki, 2002; Vasta, Knott, and Gaze, 1996). Baenninger and Newcombe (1989) designed a few spatial training studies along two dimensions: content (specific, general, and indirect) and duration (long, medium, and short). Their finding is that the training is most effective when the instructions are specific and when the duration is long. With lots of practice, females are able to perform as well as males on spatial tests (Baenninger and Newcombe, 1989). This finding refutes the explanation of biological sex differences in spatial ability.

Quantity is the dimension that showed the least amount of gender differences in terms of effect size (*d* = 0.04), and the *t*-test is insignificant

[3] Computational visualistics is a new, trans-disciplinary area of scientific endeavor that explores how pictures are created, stored, transmitted, and analyzed by computers as well as perceived, understood and processed by computer users.

for this sample. Quantity items involve relatively lower levels of mathematical complexity. Both males and females obtained highest ability estimate on Quantity items (Table 2). Quantity items mainly consist of computational items. Girls have been found to demonstrate advantages on computational items since they are more cautious than boys in dealing with arithmetic operations (De Lisi and McGillicuddy-De Lisi, 2002; Linn and Hyde, 1989). The concept of numerical manipulation is often introduced and practiced in math classes. According to the *familiarity* hypothesis raised by Kimball (1989), girls are motivated and confident to perform well on tasks that they feel familiar with. Thus, they get the same or even better grades than boys in math class but perform less well on standardized tests (Kimball, 1989). In addition, girls are found to rely on a rote approach and memorization when learning math (Fennema and Peterson, 1985; Grant, 1985), which in turn helps girls in solving computational items.

Gender frequency at each ability estimate level

To examine gender differences in a thorough manner, we compared the number of males and females at each ability estimate level for overall performance and within each math domain (see Figures 3-7). Each figure includes a pair of distributions representing each gender. From Figure 3 we can see that the males and females showed very little difference at the extreme ends, contrary to the previous findings that there are significantly more males at the top performing level (Becker and Hedges, 1984; Rosenthal and Rubin, 1982). However, there are consistently more females at the left side of the distribution, and more males at the right side, except at one estimate point (0.1). Overall, boys showed advantages in the PISA math performance.

The distribution of ability estimates was quite unimodal for the Space and Shape dimension (see Figure 4). Clearly, there are more females at the left side of the distribution and more males at the right side, suggesting that there are more males belonging to the high performing groups. Among the four pairs of distributions representing four content domains, Space and Shape dimension is the one that showed the most obvious and consistent male advantage in mean estimates. This explains the largest effect size of gender differences on this dimension among the four sub domains.

The distribution of the Change and Relationships dimension is less unimodal (see Figure 5), with many peaks in the middle, which coincides with the results shown in the Wright map (Figure 1). Males and females showed no difference in terms of numbers at the lower end of the distribution. There were clearly more females than males in the medium ability range. Males outnumbered females at the higher end of the distribution, the difference in frequency ranging from a few students to about 20. The difference is more obvious in the middle than at the upper tail.

Similar to the distribution of the Change and Relationships dimension, the Quantity dimension (see Figure 6) followed a less unimodal distribution. There are several peaks in the middle range of the ability continuum. The Quantity dimension showed small frequency differences between the two genders at several estimate points in the middle range of the distribution. The difference does not consistently favor one gender over the other for this dimension. There is no obvious difference at the two tails of the distribution, suggesting that males and females performed equally at the low and high ability level.

The distribution of the Uncertainty dimension followed a quite unimodal distribution (see Figure 7). There were very small frequency discrepancies at a few estimate points. There were more females belonging to some of the low performing groups and there were more males belonging to the middle performing groups. There was no difference at the high ability level. In general, the number of males and females was very close at all ability estimate levels on this dimension.

Generally speaking, there is no striking difference between males and females at each ability estimate level, except that males consistently outperformed females on the Space and Shape dimension, and there are more males at the top of the Change and Relationships dimension.

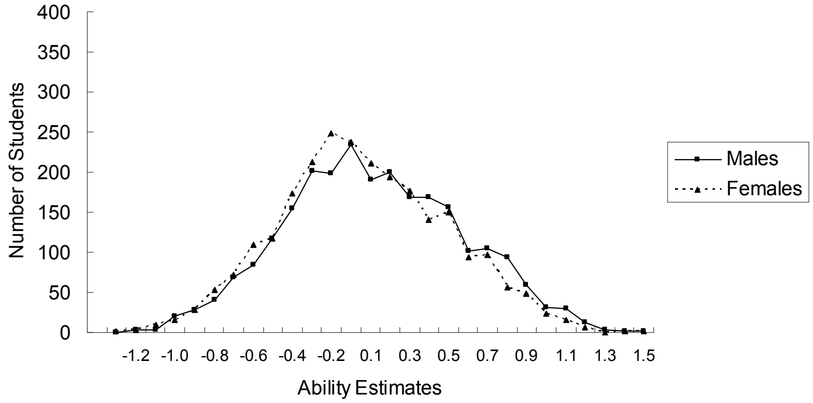

Figure 3. 2003 PISA Math: Number of Boys and Girls at Each Ability Estimate

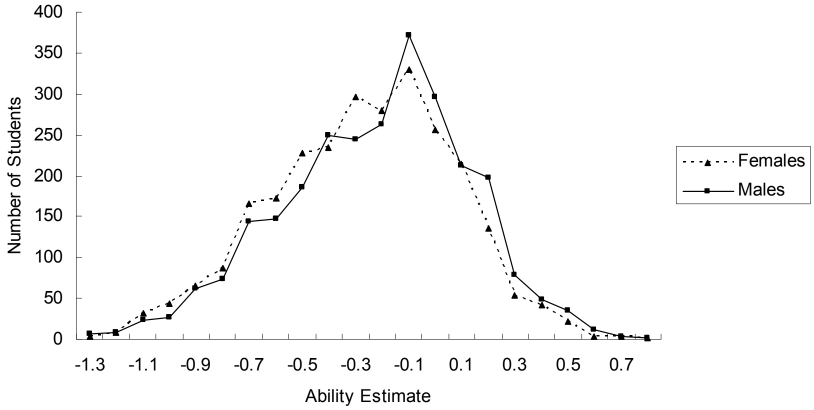

Figure 4. 2003 Space and Shape: Number of Boys and Girls at Each Ability Estimate

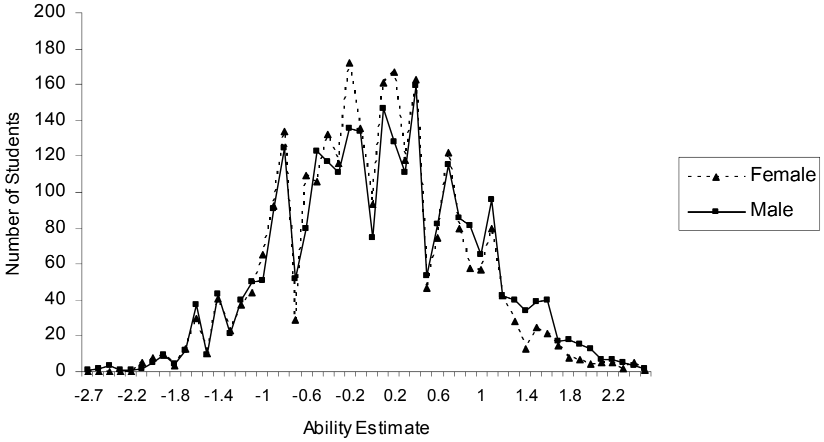

Figure 5. Change and Relationships: Number of Boys and Girls at Each Ability Estimate

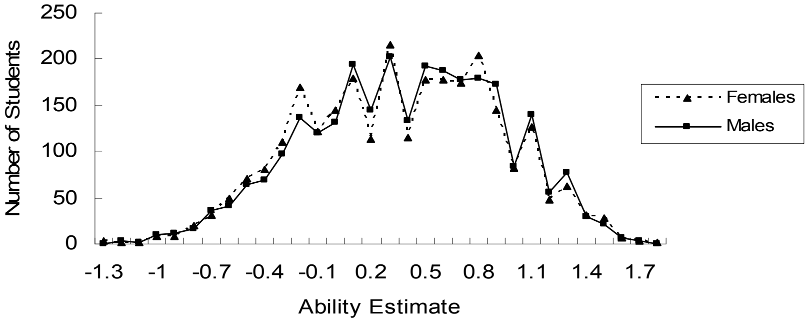

Figure 6. 2003 Quantity: Number of Boys and Girls at Each Ability Estimate

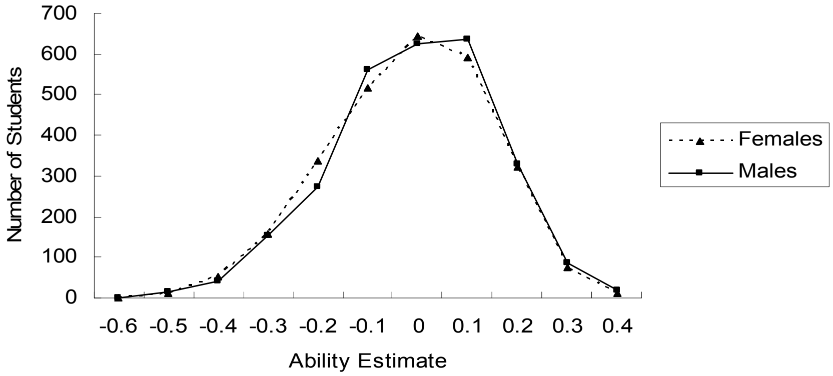

Figure 7. 2003 Uncertainty: Number of Boys and Girls at Each Ability Estimate

Conclusions and Discussion

Gender differences on large-scale standardized mathematics assessments have been a concern to educators and policy makers in the United States for over three decades. On the SAT mathematics test, males outperformed females from 1972 to 2005 and the difference has remained somewhat constant at around 40 points (College Board, 2005). The gender gap in math performance in K-12 settings is likely to amplify the gender gap in advanced math education and math related careers in the United States. Many of the math related majors are disproportionately males and also the faculties in mathematics departments are dominantly males in the United States. Since math related professions happen to be relatively better paid jobs, females' social economic status is disadvantaged by their lack of representation in these professions (Liu and Wilson, 2007).

It has long been identified that gender differences in mathematics performance are not homogeneous across tasks requiring different mathematical skills. There are certain math items that display larger differences than others. Gallagher (1998) found that males tend to perform better on items requiring spatial ability and girls tend to score higher on items involving calculations. It is crucial to identify the domains that display substantial gender differences in favor of each gender so that actions can be taken to narrow the gender gap.

This study takes a multidimensional Rasch approach to evaluate the gender differences across four math domains of the PISA 2003 mathematics assessment. The multidimensional Rasch model was shown to have better model fit than the undimensional Rasch model. The four domains are reasonably correlated to support a four-dimensional model. On average, males outperformed females across all four content domains. All the effect sizes were below .20, small but consistent. This supports the findings from the general literature that gender differences in math performance have been declined, but still exist (Cole, 1997b; Feingold, 1988).

The magnitude of the difference varied across dimensions, with the Space and Shape domain showing the largest difference and the Quantity dimension showing the smallest. It has long been documented that males showed superior spatial ability (Casey, Nuttall, Pezaris, and Benbow, 1995; Gierl, Bisanz, Bisanz, and Boughton, 2003; Halpern, 1997). The examination of the number of males and females at each ability estimate points revealed that there were consistently more females than males at the lower end of the ability distribution and there were more males at the upper end of the distribution for the Space and Shape dimension. The gender differences on the other three dimensions were less obvious or consistent than those on the Space and Shape dimension, except for Change and Relationships where there were more males at the top end. The results emphasized the importance of contextualized investigations of the math gender issue. Males and females not only obtained different scores, they also performed differently on various components of math tests (Harris and Carlton, 1993).

A larger significance of this investigation lies in that it offers an example of examining gender differences by content area, to provide diagnostic information of student strengths and weaknesses within each domain. It is hoped that teachers can utilize the findings to revamp instruction and cater to the differential learning needs of boys and girls.

For future research, it would be important to conduct a trend analysis with the PISA 2000 data to see whether the patterns of gender differences remain unchanged between the years 2000 and 2003 in the United States. If findings are consistent, if for example, the Space and Shape domain also showed the largest difference on the 2000 mathematics assessment, it will deliver a clear message to policy makers that serious actions need to be taken to narrow the gender gap in spatial ability. Probably instructional materials should be embedded in curriculum design to provide effective scaffoldings for girls to develop spatial ability. Similarly, actions need to be taken to improve boys' performance in areas such as calculation where they showed weaknesses (Gallagher, 1998; Linn and Hyde, 1989). Another interesting direction would be to examine the interaction between the math content measured and the effect of item format, possibly to identify the combination of domain and item type which may enlarge or reduce the gender gap.

Furthermore, international comparisons are needed to explore the generalizability of patterns of gender differences. The ranking of the math performance of the U.S. students is about 27th among 42 participating countries in PISA 2003. Countries of higher and lower performance than the United States could be selected for the comparisons. It would be important to investigate how specific curriculum design, pedagogy and classroom environment contribute to smaller differences in some countries. The purpose is to reduce gender gap in math performance and foster the mathematical learning of both males and females.

Acknowledgement

The authors cordially thank Dr. John Young at Educational Testing Service and the two anonymous reviewers for their valuable comments, suggestions, and edits.

References

Adams, R. J., and Wilson, M. (1996). Formulating the Rasch model as a mixed coefficients multinomial logit. In G. Engelhard and M. Wilson (Eds.), *Objective measurement: Theory into practice* (Vol III; pp. 143-166). Norwood, NJ: Ablex.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-573.

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

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

Becker, B. J., and Hedges, L. V. (1984). Meta-analysis of cognitive gender differences: A comment on an analysis by Rosenthal and Rubin. *Journal of Educational Psychology*, *76*, 583-587.

Benbow, C. P., and Stanley, J. C. (1983). Sex differences in mathematical reasoning ability: More facts. *Science*, *222*, 1029-1030.

Benbow, C. P. (1988). Sex differences in mathematical reasoning ability in intellectually talented preadolescents: Their nature, effects, and possible causes. *Behavioral and Brain Sciences*, *11*, 169-232.

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

Bock, R. D., and Kolakowski, D. (1973). Further evidence of sex- linked major- gene influence on human spatial visualizing ability. *American Journal of Human Genetics*, *25*, 1-14.

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

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

Casey, M. B., Nuttall, R., Pezaris, E., and Benbow, 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. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

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

Cole, N. A. (1997a). Understanding gender differences and fair assessment in context. In W. Willingham and N. A. Cole (Eds.), *Gender and fair assessment* (pp. 157-183). Mahwah, NJ: Lawrence Earlbaum.

Cole, N. S. (1997b). *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*. New York: Author.

De Lisi, R., and McGillicuddy-De Lisi, A. (2002). Sex differences in mathematical abilities and achievement. In A. McGillicuddy-De Lisi and R. De Lisi (Eds.), *Biology, society, and behavior: The development of sex differences in cognition* (pp. 155-181). Westport, CT: Ablex.

DeMars, C. (2004). Measuring higher education outcomes with a multidimensional Rasch model. *Journal of Applied Measurement*, *5*, 350-361.

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.

Eccles, J. S., and Jacobs, J. E. (1986). Social forces shape math attitudes and performance. *Signs*, *11*, 367-380.

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

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

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

Gallagher, A. M., and Kaufman, J. C. (Eds). (2005). *Gender differences in mathematics*. Cambridge: Cambridge University Press.

Geary, D. (1996). Sexual selection and sex differences in mathematical abilities. *Behavioral and Brain Sciences*, *19*, 229-284.

Gierl, M. J., Bisanz, J., Bisanz, G. L., and 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.

Grant, L. (1985). Race-gender status, classroom interaction, and children's socialization in elementary school. In L. C. Wilkinson and C. B. Marett (Eds.), *Gender influences in classroom interaction* (pp.57-77). Orlando, FL: Academic Press.

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

Halpern, D. F., and Ikier, S. (2002). Causes, correlates, and Caveats: Understanding the development of sex differences in cognition. In A. McGillicuddy-De Lisi and R. De Lisi (Eds.), *Biology, society, and behavior: The development of sex differences in cognition* (pp. 3-19). Westport, CT: Ablex.

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

Hedges, L. V., and Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science*, *269*, 41-45.

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

Jackson, D. W., and Tein, J. (1998). Adolescents' conceptualization of adult roles. Relationships with age, gender, work goal, and maternal employment. *Sex Roles*, *38*, 987-1008.

Kimball, M. M. (1989). A new perspective on women's math achievement. *Psychological Bulletin*, *105*(2), 198-214.

Kimura, D. (1992). Sex differences in the brain. *Scientific American*, *267*, 118-125.

Langenfield, T. E. (1997). Test fairness: Internal and external investigations of gender bias in mathematical testing. *Educational Measurement: Issues and Practice*, *16*, 20-26.

Linacre, J. M. (1994). *Many-facet Rasch measurement*. Chicago: MESA Press.

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

Liu, O. L., and Wilson, M. (2007). *Gender differences and similarities in PISA 2003 mathematics: A comparison between U.S. and Hong Kong*. Manuscript submitted for publication.

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 49-174.

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., and Terlecki, M. (2002). Maximization of spatial competence: more important than finding the cause of sex differences. In A. McGillicuddy-De Lisi and R. De Lisi (Eds.), *Biology, society and behavior: the development of sex differences in cognition* (pp.183-206). Westport, CT: Ablex.

O'Neill, K. A., and McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P.W. Holland and H. Wainer (Eds.), *Differential item functioning* (pp. 255-279). Hillsdale: NJ: Lawrence Erlbaum.

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

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

Peters, M., Chisholm, P., and Laeng, B. (1995). Spatial ability, student gender, and academic performance. *Journal of Engineering Education*, *84*, 69-73.

Quasier-Pohl, C., and Lehmann, W. (2002). Girls' spatial abilities: Charting the contributions of experiences and attitudes in different academic groups. *British Journal of Educational Psychology*, *72*, 245-260.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded edition, 1980. Chicago: University of Chicago Press.)

Rosenthal, R., and Rubin, D. B. (1982). A simple general purpose display of magnitude of experimental effects. *Journal of Educational Psychology*, *74*, 166-169.

Sanberg, D., Ehrhardt, A., Mellins, C., Ince, S., and Meyer-Bahlburg, F. (1987). The influence of individual and family characteristics upon career aspirations of girls during childhood and adolescence. *Sex Roles*, *16*(11/12), 649-668.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, *52*, 613-629.

TIMSS and PIRLS International Study Center. (2003). *IEA's TIMSS 2003 international report on achievement in the mathematics cognitive domains*. http://timss.bc.edu

TIMSS and PIRLS International Study Center. (2000a). *TIMSS 1999 international mathematics report*. http://timss.bc.edu

TIMSS and PIRLS *International Study Center. (2000b). Gender differences in achievement*. http://timss.bc.edu

Vasta, R., Knott, J. A., and Gaze, C. E. (1996). Can spatial training erase the gender differences on the water-level task? *Psychology of Women Quarterly*, *20*, 549-567.

Wang, W. C. (1994). *Implementation and application of the multidimensional random coefficients multinomial logit*. Unpublished doctoral dissertation, University of California, Berkeley, CA.

Wang, W., Wilson, M., and 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.

Wigfield, A., Battle, A., Keller, L., and Eccles, J. S. (2002). Sex differences in motivation, self-concept, career aspiration, and career choice: Implications for cognitive development. In A. McGillicuddy-De Lisi and R. De Lisi (Eds.), *Biology, society and behavior: The development of sex differences in cognition* (pp. 94-124). Westport, CT: Ablex.

Wilson, M. (1992). The ordered partition model: an extension of the partial credit model. *Applied Psychological Measurement*, *16*, 309-325.

Wu, M. L., Adams, R. J., and Wilson, M. R. (1998). ConQuest: Generalized item response modeling software [Computer program]. Camberwell, VIC, Australia: ACER.

Xie, Y., and Shauman, K. (2003). *Women in science: Career processes and outcomes*. Cambridge, MA: Harvard University.

Appendix A

Booklet Effect Correction

The 2003 PISA assessment items were allocated to 13 item clusters (seven mathematics clusters, and two clusters in each of the science, reading and problem solving skills [1], see Table 4), with each cluster representing 30 minutes of test time. The 13 clusters were organized into 13 test booklets, with four clusters for each booklet. The frequency of each booklet, namely the number of students responding to each booklet, is also included in Table 4. We can see that the thirteen booklets were evenly distributed among students. Since mathematics is the primary focus of the 2003 PISA assessments, there are seven math clusters, which would take about three-and-a-half hours to complete.

Due to the different location of domains and items within each booklet, it is expected that there might be booklet influence on student math ability estimate. That is, the particular order of items can affect student responses. The booklet effect was modeled at the booklet level. When estimating the four-dimensional model, the booklet effect parameter was modeled and estimated together with 4 different dimension abilities (EAP) and item parameters to eliminate the confounding of item difficulties and booklet effects. This booklet effect adjustment was modeled in ConQuest by using the "regression" command statement, which conducts a latent regression of ability onto a dummy variable to indicate which booklet was taken by each student. The use of the latent regression for the booklet effect estimation was specified with ease by the following ConQuest command:

Regression blooklet,

where the "Regression" is a ConQuest command statement for latent regression. For more details of the use of the regression statement in the program ConQuest, see Wu, Adams, and Wilson (1998). Because there were 13 booklets, ConQuest produced 13 booklet estimates which represent the booklet difficulty. The booklet effect adjustment for student ability was made by adding the booklet parameter estimates from the latent regression to student ability estimate for students taking the corresponding booklet. This booklet effect adjusted EAP estimates were used for gender similarity and difference analyses.

[1] PISA 2003 measures four domains: reading, science, math and problem solving.

Table 4

Cluster rotation design used to form test booklets for PISA 2003

Booklet	Frequency	Cluster[a]			
1	418	M1	M2	M4	R1
2	425	M2	M3	M5	R2
3	420	M3	M4	M6	PS1
4	408	M4	M5	M7	PS2
5	431	M5	M6	S1	M1
6	427	M6	M7	S2	M2
7	407	M7	S1	R1	M3
8	418	S1	S2	R2	M4
9	421	S2	R1	PS1	M5
10	437	R1	R2	PS2	M6
11	424	R2	PS1	M1	M7
12	404	PS1	PS2	M2	S1
13	416	PS2	M1	M3	S2

Note: [a]M stands for mathematics, R stands for reading, S stands for science, and PS stands for problem-solving.

Appendix B

Booklet Estimates

The Wright map (Figure 1) also shows the booklet difficulty estimates. The booklet estimates can be interpreted in a similar way as the item difficulty estimates, with booklet 10 being the most difficult one and booklet 13 being the easiest. The span of the booklet estimates is about 0.5 logit. The importance of monitoring the booklet effect can be illustrated by comparing the booklet estimate span to the student ability span within each dimension. For example, the range of the student ability is about 1.5 logits for the Uncertainty dimension. The booklet effect could take up to 1/3 of the true ability range. This will produce substantial difference in ability estimates between students who responded to, say, booklet 10 and booklet 13. To make corrections for this booklet effect, the booklet estimate was added to the student estimate when responding to that booklet. The correlations between ability estimates with and without booklet correction are provided in Table 5. The Quantity dimension was the one least affected by the booklet effect and the Uncertainty dimension was the one most affected. To ensure accurate student ability estimate, we can see that there is a need to address the booklet effect.

Table 5

Correlation between ability estimates with and without booklet effect correction

	Space/ Shape	Change/Relationship	Quantity	Uncertainty
Correlation	.97	.98	.99	.86

Appendix C

Score residual $$y_{ni} = x_{ni} - E_{ni}$$

Variance of x_{ni} $$W_{ni} = \sum_{k=0}^{m} (k - E_{ni})^2 \pi_{nik}$$

Standardized residual z_{ni} $$z_{ni} = y_{ni} / W_{ni}^{1/2}$$

Weighted Mean Square $$v_i = \frac{z_{1i}^2 w_{1i} + z_{2i}^2 w_{2i} + z_{3i}^2 w_{3i} + \ldots + z_{Ni}^2 w_{Ni}}{w_{1i} + w_{2i} + w_{3i} + \ldots + w_{Ni}}$$

$$= \sum_{n=1}^{N} z_{ni}^2 w_{ni} / \sum_{n=1}^{N} w_{ni}$$

$$= \sum_{n=1}^{N} y_{ni}^2 / \sum_{n=1}^{N} w_{ni}$$