Educational and Psychological
Measurement
73(5) 836–856
© The Author(s) 2013
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0013164413487375
epm.sagepub.com
SAGE

Examining Differential Math Performance by Gender and Opportunity to Learn

Anthony D. Albano[1] and Michael C. Rodriguez[2]

Abstract

Although a substantial amount of research has been conducted on differential item functioning in testing, studies have focused on detecting differential item functioning rather than on explaining how or why it may occur. Some recent work has explored sources of differential functioning using explanatory and multilevel item response models. This study uses hierarchical generalized linear modeling to examine differential performance due to gender and opportunity to learn, two variables that have been examined in the literature primarily in isolation, or in terms of mean performance as opposed to item performance. The relationships between item difficulty, gender, and opportunity to learn are explored using data for three countries from an international survey of preservice mathematics teachers.

Keywords

hierarchical generalized linear modeling, differential item functioning, gender, opportunity to learn, multilevel modeling, mathematics assessments

The property of measurement invariance is considered to be one of the major strengths of item response theory (IRT). Measurement invariance, or parameter invariance, refers to the theoretical lack of variability in item parameters over examinee populations and in person parameters over measurement conditions or sets of items (Rupp & Zumbo, 2006). Testing programs rely on this property when creating and maintaining an IRT measurement scale across cohorts of examinees and

[1] University of Nebraska–Lincoln, Lincoln, NE, USA
[2] University of Minnesota, Twin Cities, MN, USA

Corresponding Author:
Anthony D. Albano, University of Nebraska–Lincoln, 235 Teachers College Hall, Lincoln, NE 68588-0345, USA.
Email: albano@unl.edu

alternate forms of a test; parameter estimates obtained with one form and one group of examinees are expected to represent values that would not change, within a linear transformation, if a different test form were used or if the test were administered to a different examinee group. A lack of invariance would invalidate score interpretations and comparisons in these contexts.

A variety of item response models have been developed to examine invariance in item and person parameters. Modeling frameworks include the many-facet Rasch model (Linacre, 1994), the hierarchical generalized linear model (HGLM; e.g., Kamata, 2001), and other explanatory item response models (De Boeck & Wilson, 2004) and structural equation models (e.g., Muthén, Kao, & Burstein, 1991). These frameworks extend the IRT model beyond the conventional parameters of item difficulty, discrimination, lower asymptote, and person ability, making it possible to explore additional covariates and the extent to which these covariates explain variability in performance.

The present study focuses first on variability that is explained by gender, which provides evidence of differential item functioning (DIF) or a lack of item parameter invariance across persons. Gender DIF is examined for math items within an HGLM framework. The main purpose of this study is to build on previous work by using HGLM to investigate opportunity to learn (OTL) as a potential source of gender DIF. The relationships between item difficulty, gender, and OTL are explored using data from an international survey of preservice mathematics teachers.

Modeling Parameter Variance

Many methods exist for detecting and testing item parameter invariance, item bias, or DIF (see Holland & Wainer, 1993). Zumbo and Hubley (2003) described these methods as falling into three general frameworks. The first includes the use of contingency tables (e.g., the Mantel–Haenszel approach) and regression models (e.g., logistic regression; Swaminathan & Rogers, 1990), the second is based on IRT techniques (e.g., item characteristic curve comparisons), and the third involves modeling multiple ability or trait dimensions (Ackerman, 1992).

As noted by Cheong (2006), multilevel modeling represents an extension of the regression approach to examining parameter variance. In a two-level unconditional HGLM, the log-odds of correct response η_{ij} are modeled as

$$\begin{aligned} \eta_{ij} &= \beta_{0j} + \sum_{q=1}^{N-1} \beta_{qj} X_{qij} \\ \beta_{0j} &= \gamma_{00} + u_{0j} \\ \beta_{qj} &= \gamma_{q0}. \end{aligned} \tag{1}$$

This base model is a multilevel representation of the one-parameter unidimensional Rasch model. Item responses i at level one are considered to be nested within persons j at level two. The difficulty of item $q = i$ is estimated by the fixed effect γ_{q0} using the item indicator variable X_{qij}, described further below, and ability for person j is estimated by the random effect u_{0j}. Without a random term u_{qj} or other terms in the model for β_{qj} at level two, individual item difficulty parameters are assumed to be invariant across people, as is typically the case in IRT. Main effects, also referred to as impact effects, can be estimated by including person covariates in the model for β_{0j} in Equation (1). Item parameter invariance or DIF can then be tested by including one or more person covariates at level two, as demonstrated below, and/or by allowing the item parameter β_{qj} to vary randomly across people at level two. For additional details on using HGLM for DIF and item analysis, see Beretvas, Cawthon, Lockhart, and Kaye (2012) and Kamata (2001).

Although a substantial amount of research has been conducted on DIF over the past 30 years, studies have focused on detecting DIF rather than on explaining how or why it may occur. Recent work has begun to explore sources of DIF using explanatory item response models such as HGLM. These models make it possible to estimate effects for item parameters and all covariates simultaneously while controlling for person ability, and to thereby avoid a two-step procedure where parameters are estimated in one model and then modeled as outcomes in another. In a two-step approach, estimates of effects such as DIF can be attenuated by measurement error in the item effects, whereas the one-step, simultaneous estimation approach has been shown to produce disattenuated estimates (e.g., Adams, Wilson, & Wu, 1997). HGLM can also accommodate complex data structures, including covariates at multiple levels of nested data. As a result, these models facilitate the study of multiple sources of DIF (Beretvas et al., 2012), including sources at the person level, such as demographic and other grouping variables, and sources at additional levels of nesting. Two examples that used HGLM in the study of DIF are reviewed here.

Randall, Cheong, and Engelhard (2011) investigated parameter invariance across testing accommodations for students with disabilities. Student disability status (with disabilities versus without) and accommodation condition (resource guide, calculator, and standard or no accommodation) were both examined as potential sources of variability in item difficulty parameters. Item responses were nested within students who were nested within schools and a model similar to Equation (1) was extended to include disability status as a person group covariate at level two and an accommodation covariate at level three, where all students in a given school had been assigned to the same accommodation condition. The best-fitting model included main effects for items, disability status, and accommodation, and all two-way and three-way interactions between them. The results indicated that difficulty for two out of ten items varied significantly by both disability status and accommodation condition, as indexed by a cross-level interaction between the two and the item indicator. Thus, parameters for these items were not found to be invariant across students and schools. Further examination of the content and other features of these items could provide an explanation for this lack of invariance.

Cheong (2006) also used a three-level HGLM to examine item parameter invariance over covariates at the student and school levels. As in Randall et al. (2011), item responses were nested within students who were nested within schools. However, in this case, ethnicity was the person group covariate at level two, and instructional opportunities, or OTL, the covariate at level three. Difficulty estimates for 3 out of 13 items were found to vary by student ethnicity when OTL was excluded from the model. These items were initially identified as displaying ethnic/racial DIF. When both item difficulties and DIF effects for ethnicity were modeled at the school level as a function of OTL only one item still exhibited ethnic DIF; difficulties for the other two items were no longer found to vary by ethnicity. These results indicate that school-level OTL may moderate the relationship between item difficulty and person level covariates such as ethnicity; differential performance by ethnic/racial groups may in part be attributed to differential OTL.

Opportunity to Learn

The concept of OTL originated in studies conducted by the International Association for the Evaluation of Educational Achievement (IEA), where it was used to explain, in part, math and science performance differences found between countries (McDonnell, 1995). In the first of these international studies, Husén (1967) described OTL as the opportunity to study a particular topic or to learn a problem solving technique as required by a given test; he argued that lower OTL would be associated with lower chances of responding correctly to relevant test items. Since this early work, OTL has become an important consideration in the development and use of achievement measures, and a considerable amount of research has documented a consistent positive association between student achievement and OTL (Floden, 2002).

The concept of OTL has evolved since its origins in IEA. Carroll (1963) presented a model of student learning where OTL was conceptualized as the amount of class time allowed for learning. Later research built on this conceptualization by considering the degree of separation between the student and the opportunity to learn. As noted by Floden (2002), this separation varies from the opportunities a student is intended to have, based on curriculum established at the national, state, or district levels (the intended curriculum), to the opportunities a student actually has, based on the instruction actually given by the teacher and attended to by the student (the implemented curriculum). These distinctions reveal that OTL can be conceptualized and measured in a variety of ways.

As a research tool, OTL has been used to ensure valid comparisons in terms of achievement across different subgroups of individuals. Within IEA, comparisons were primarily made across countries (McDonnell, 1995). In this case, differences in mean achievement by country would be more meaningful if the countries being compared were similar in terms of average OTL. Results from IEA have shown that lower performing countries tend to have lower OTL (Mullis, Martin, & Foy, 2008; Mullis, Martin, Foy, & Arora, 2012). OTL has also been discussed as a contributor to ethnic/racial achievement gaps in the United States (e.g., Kim & Hocevar, 1998), suggesting that comparisons by race or ethnicity may not be valid unless OTL is taken into account.

Few studies have examined the relationship between OTL and person grouping variables in terms of performance on individual items, as opposed to mean performance. Linn and Harnisch (1981) first referred to OTL in relation to item performance or DIF. They suggested that DIF based on ethnicity or race may more appropriately be attributed to the differing instructional opportunities associated with certain ethnic groups. The findings of Cheong (2006) provided evidence supporting this hypothesis. Clauser, Nungester, and Swaminathan (1996) examined the extent to which accounting for OTL could improve the detection of gender DIF using logistic regression. Results of their study showed that the number of items flagged for DIF on a high-stakes licensure exam was reduced when educational background was accounted for in the regression models. Thus, item difficulty parameters varied not only by gender but also by the different educational backgrounds of women and men.

Gender

Although research on OTL is limited, especially in terms of item performance, many studies have documented differential performance by gender in academic achievement (for a review, see Willingham & Cole, 1997). Gender effects in standardized testing have been described at both the test level, or averaged across items, and at the item level, in terms of gender DIF.

At the test level, results indicate that academic achievement often differs noticeably by gender. The U.S. National Assessment of Educational Progress (NAEP) has reported a trend of higher mean reading scores for female students, and higher mean math scores for male students, for multiple grade levels and across multiple years (Rampey, Dion, & Donahue, 2009). Recent NAEP results show that this gender gap in reading is somewhat consistent across grades (National Center for Education Statistics, 2011c); however, the gender gap in math appears to widen from elementary school to high school (National Center for Education Statistics, 2011b).

Other studies report that the gender gap is narrowing. Hyde, Fennema, and Lamon (1990) conducted a meta-analysis of 100 studies of math gender effects. For certain grade ranges and content areas, significant effects emerged; performance differences favoring men were found in the content area of problem solving, most notably in calculus and geometry, but only for high school and college students. However, the mean gender effect was found to be negligible when averaged across all studies and math content areas. More recently, Hyde, Lindberg, Linn, Ellis, and Williams (2008) analyzed results from 10 state math tests, representing more than seven million students, and found only small differences in performance by gender; standardized effect sizes, with positive effects favoring men and negative effects favoring women, were all less than 0.10, with a weighted mean effect of 0.0065.

Differences in test performance by gender have been explained, in part, by gender differences in course choices, with women traditionally taking fewer advanced math and science courses than men (e.g., Meece & Parsons, 1982). However, recent reports indicate that gender differences in course taking have diminished in the United States, with similar percentages of women and men completing high school math and science courses (e.g., National Center for Education Statistics, 2011a). These similarities in course taking may help explain the similarities recently reported in mean achievement. Cultural influences have also been identified as a source of gender effects (e.g., Baker & Jones, 1993). Else-Quest, Hyde, and Linn (2010) examined two international datasets, the Third International Mathematics and Science Study (TIMSS) and the Programme for International Student Assessment, and found that gender differences in math achievement and attitude toward math were associated with country-level indicators of gender equity.

Gender effects have also been explored at the item level. Bielinski and Davison (1998) found evidence of gender by item difficulty interactions, where easier math items tended to be easier for female students and more difficult items tended to be easier for male students. Correlations between gender effects and item difficulty estimates on two state math tests were -0.47 and -0.43; as items became more difficult, with lower IRT logit values, the advantage for male students in these samples increased. Bielinski and Davison (2001) later replicated this finding using three national data sets: the 1992 NAEP, the U.S. cohort from the TIMSS, and the National Educational Longitudinal Study of 1988. Correlations between gender effects and IRT item difficulty in elementary, middle, and high school were found to be statistically significant and negative. Penner (2003), also using the TIMSS, reported similar gender by item interactions in 8 out of 10 countries studied.

Additional research has revealed that certain features of test items are related to gender DIF. Some studies have examined item type, typically by comparing performance on multiple-choice (MC) and constructed-response (CR) items by gender. Results of these studies are mixed, with some reporting that males tend to do better on MC items whereas females tend to do better on CR items (e.g., Bolger & Kellaghan, 1990; DeMars, 2000), and others finding that female and male students perform similarly on both MC and CR items (e.g., Liu & Wilson, 2009). Item content has also been explored as a possible source of gender DIF. Harris and Carlton (1993) investigated what was at the time a persisting gender effect favoring men on the math section of the SAT. Controlling for mean ability, algebra items were found to be slightly easier for women than men, whereas geometry items were found to be easier for men than women. Men also performed better on questions involving real-world problems, whereas women performed better on questions involving abstract problems. Garner and Engelhard (1999) also explored the relationship between various item features and gender DIF. Using data from a state high school graduation test, they too found that algebra items tended to exhibit DIF in favor of women, whereas items in the remaining content areas of number and computation, data analysis, and geometry and measurement tended to exhibit DIF in favor of men.

Summary

Although the effects of OTL and gender have been studied in isolation, research on the relationship between them at the item level is limited. Course taking has been linked to gender effects (e.g., Meece & Parsons, 1982), as have other educational and economic opportunities (e.g., Baker & Jones, 1993; Else-Quest et al., 2010), and item-level variables (e.g., Bielinski & Davison, 2001; Garner & Engelhard, 1999; Liu & Wilson, 2009); however, the impact of OTL and the relationships between gender and OTL in terms of item performance are not well understood.

Previous research has demonstrated the usefulness of HGLM and other multilevel item response models for testing the invariance of item difficulty parameters and for examining sources of DIF. These models can shed light on the relative importance of gender, OTL, and other covariates in explaining variability in item difficulty and mean test performance. The present study extends previous work by using HGLM to examine the relative importance of gender, OTL, and their interaction in explaining variability in item difficulty in an international data set.

Method

Data

Item-level data were obtained from the Teacher Education and Development Study in Mathematics (TEDS-M), an international study of preservice mathematics teachers conducted in 2008 (Brese & Tatto, 2012). The study examined teacher preparation programs across multiple organizational units and using multiple measurement tools. The present study used scores from the math content knowledge (MCK) assessment for individuals training to teach math at the secondary level. Initial analyses focused on the U.S. cohort, which included responses from 475 students (69% female, 31% male). The analyses described below were then conducted using data from Singapore (SGP), with 393 students (48% female, 52% male), and Germany (DEU), with 768 students (61% female, 38% male). These three countries had relatively large sample sizes and high response rates, and differed noticeably on variables of interest, as described below. They were also selected to represent three distinct geographic and cultural contexts.

The full MCK assessment contained a total of 103 MC (which includes traditional MC and what are referred to as complex MC) and CR items. A subset of this full test was used in the present study for two reasons. First, the within-country sample sizes were found to be insufficient for estimating the number of parameters contained in the fullest model, as many as four fixed effects per item. Second, for test security reasons, only a portion of the actual MCK instrument was made publicly available. Item content was considered to be important in this study because it may aid in the examination of items flagged for bias as well as inform development of future items in this context.

Full content was made available for a subset of items that represented the full item set in terms of difficulty, content domain, and item type (Brese & Tatto, 2012). Of the 32 items in this subset, 23 addressed MCK, whereas the remaining 9 addressed mathematics pedagogy content knowledge. This study used responses for 22 of the MCK items. The remaining MCK item was removed because it stood out as the only item in the data content domain. Table 1 contains information on these items, including the original TEDS-M item ID, content domain, subdomain, label (a summary of the item content), item format, and the proportion correct (*p* value) for all participating countries (Avg) and for the USA, SGP, and DEU. For the full stem and response options for each item, see supplement 4 of the TEDS-M user guide.

As shown in Table 1, this subset of items contained 16 MC and 6 CR items. Each item fell into one of three domains: algebra (7 items), geometry (7 items), and number (8 items). Items were also categorized into subdomains, including applying (9 items), knowing (6 items) and reasoning (7 items). Proportion correct tended to be highest for SGP, then DEU, followed by USA, and all three countries tended to be higher than the international average.

Opportunity to learn was measured at both the individual and program/institution levels, providing measures of what students actually had the opportunity to learn and what their programs or institutions intended for them to learn. Individual OTL, as a measure of implemented curriculum, was used in the present study. At the individual level, topics were presented in university-/tertiary-level mathematics, school-level mathematics, and mathematics education/pedagogy. Students responded to each by indicating whether or not they had studied the topic as part of their current teacher preparation program. Responses were coded as 1 for studied and 0 for not studied and total scores across each topic set were examined. Because of its relevance to teaching at the secondary level, the total score on tertiary math was used to represent OTL in the present study. Tertiary math covered 19 topics, including axiomatic, analytic, non-Euclidian, and differential geometry; topology; linear and abstract algebra; set theory; number theory; beginning, univariate, multivariate, and advanced calculus; differential equations; functions; discrete mathematics; probability; statistics; and mathematical logic.

Table 2 contains descriptive statistics for proportion correct scores and OTL by country and gender. Proportion correct represents the proportion of items answered correctly by each student. The mean for men was 0.13 higher than for women in the USA. In SGP, the means for women and men were the same. In DEU, the mean for men was 0.07 higher than for women. OTL means were slightly higher for men; on average, men indicated having studied roughly one additional math topic in each country. Table 2 also contains the correlation between proportion correct and OTL by country and gender. Estimates were highest in the USA (0.48 for women, 0.40 for men), lower in DEU (0.30 for women, 0.25 for men), and lowest in SGP (0.11 for women, 0.06 for men). The correlations for women were all higher than those for men. Together, these descriptive statistics suggest that performance at the item level may depend on gender, and that OTL may moderate the relationship between the

Table 1. Subset of Released Math Content Knowledge Items.

					p			
Item ID	Domain	Subdomain	Format	Label	Avg	USA	SGP	DEU
MFC604A1	Algebra	Apply	CR	Linear relation	.72	.85	.97	.84
MFC604A2	Algebra	Apply	CR	Linear relation	.50	.65	.91	.66
MFC610A	Number	Know	MC	Irrational number	.44	.61	.85	.66
MFC610C	Number	Know	MC	Irrational number	.54	.73	.88	.85
MFC610D	Number	Know	MC	Irrational number	.37	.38	.67	.57
MFC703	Geometry	Reason	CR	Length of segment	.33	.44	.67	.45
MFC704	Geometry	Apply	CR	Length of segment	.32	.33	.48	.29
MFC705A	Geometry	Know	MC	Solve equation in plane	.53	.63	.70	.68
MFC705B	Geometry	Know	MC	Solve equation in space	.51	.64	.67	.66
MFC710A	Algebra	Apply	MC	Exponential function	.41	.53	.49	.68
MFC710B	Algebra	Apply	MC	Exponential function	.39	.36	.69	.68
MFC710C	Algebra	Apply	MC	Exponential function	.60	.84	.72	.83
MFC711	Algebra	Reason	CR	Sum of functions	.11	.07	.15	.26
MFC802A	Number	Reason	MC	Identify a proof	.46	.67	.60	.86
MFC802B	Number	Reason	MC	Identify a proof	.63	.66	.76	.74
MFC802C	Number	Reason	MC	Identify a proof	.58	.78	.87	.82
MFC802D	Number	Reason	MC	Identify a proof	.54	.72	.67	.68
MFC804	Number	Know	MC	Combinations	.35	.33	.49	.42
MFC808A	Geometry	Apply	MC	Lines of symmetry	.70	.83	.81	.83
MFC808B	Geometry	Apply	MC	Lines of symmetry	.61	.70	.83	.67
MFC808C	Geometry	Apply	MC	Lines of symmetry	.53	.64	.68	.75
MFC814	Algebra	Reason	CR	Matrix operations	.19	.37	.45	.59

Note. The item ID, domain, subdomain, format, label, and average *p* value across all countries participating in the study (Avg) were adapted from the TEDS-M user guide (Brese & Tatto, 2012). CR = constructed-response; MC = multiple-choice; SGP = Singapore; DEU = Germany; TEDS-M = Teacher Education and Development Study in Mathematics.

Table 2. Descriptive Statistics by Country and Gender.

			Prop Correct		OTL		
Country	Gender	*N*	*M*	*SD*	*M*	*SD*	*r*
USA	F	317	0.55	0.20	11.33	4.06	0.48
	M	144	0.68	0.19	13.03	3.25	0.40
SGP	F	179	0.70	0.16	9.19	4.76	0.11
	M	199	0.70	0.17	10.54	4.48	0.06
DEU	F	441	0.63	0.19	10.55	3.81	0.30
	M	282	0.70	0.19	11.90	3.89	0.25

Note. Prop Correct = the proportion correct score across the set of items administered to a student. *r* = correlation between Prop Correct and OTL; OTL = opportunity to learn; SGP = Singapore; DEU = Germany.

two. Multilevel models were used to explore the relationships between these variables for each country.

Models

A model comparison approach was used to test the statistical significance of sequentially more complex models. Three model fit indices described improvement in fit from one model to the next: likelihood ratio χ^2, AIC (Akaike information criterion), and BIC (Bayesian information criterion). If the χ^2 was statistically significant and the AIC reduced for a model, it was considered significant. The BIC provided supplemental fit information. Model comparisons served as omnibus tests for the set of parameters that entered into each subsequent model. Individual effects were then examined as necessary, as described below. This modeling approach was repeated for each country.

The base model M0 included fixed effects for items and a random effect for people, as in Equation 1. Subsequent models differed from M0 only at the person level, level two. The indicator variable for item $q = N$ was omitted from the model to make it identifiable, resulting in item indicators X_{qij} only for the first 21 items. Effect coding was used with the item indicators, where $X_{qij} = 1$ for $q = i$, $X_{qij} = -1$ for $q = N$, and $X_{qij} = 0$ otherwise. As a result, γ_{00} estimated the mean item difficulty, and γ_{q0}, the deviation for item q from the mean, with larger values indicating higher mean predicted log-odds of correct response, that is, easier items. The intercept and item effect would be combined to obtain the traditional Rasch difficulty estimate, as $-(\gamma_{00} + \gamma_{q0})$. Effects and standard errors for the excluded item were obtained by substitution.

Model M1 examines gender impact and item by gender interaction effects. Gender ($\text{Gender}_j = 0$ for women and 1 for men) is included in the level-two models for the overall math performance and item difficulty:

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + \gamma_{01}\text{Gender}_j + u_{0j} \\ \beta_{qj} &= \gamma_{q0} + \gamma_{q1}\text{Gender}_j.\end{aligned} \tag{2}$$

Here, γ_{00} estimates the mean performance for women, and γ_{01}, the overall performance difference for men. The terms γ_{q0} and γ_{q1} estimate the difficulty of item q for women and the differential effect on item difficulty for men.

Model M1 is used to examine gender DIF. The approach taken in this study was to examine individual γ_{q1} for statistical significance only if the omnibus test for M1 was statistically significant. Significance of γ_{q1} was evaluated based on results from a t-test with $\alpha = .05$, and on the effect size in comparison to the standard deviation (*SD*) of student ability u_{0j}. Standardized effects larger in absolute value than half a standard deviation, having $p<.05$, were considered significant (for examples of similar approaches, see Albano, in press; Cheong, 2006).

Model M2 additionally examines OTL impact and item by OTL interaction effects. The mean-centered OTL (OTL_j) is added to the level-two models for β_{0j} and β_{qj}:

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + \gamma_{01}\text{Gender}_j + \gamma_{02}\text{OTL}_j + u_{0j} \\ \beta_{qj} &= \gamma_{q0} + \gamma_{q1}\text{Gender}_j + \gamma_{q2}\text{OTL}_j.\end{aligned} \tag{3}$$

In M2, γ_{00} now estimates the mean performance for women at the mean OTL score and γ_{01} estimates the difference for men, controlling for OTL. γ_{02} estimates the effect of OTL on mean performance, or the change in mean performance for a 1-point change in OTL. In the same way, the item-specific γ_{q0} and γ_{q1} are now estimated while controlling for OTL. The item by OTL interaction γ_{q2} estimates the effect of OTL on the difficulty of item q as a deviation from γ_{02}.

Model M3 examines all two-way and three-way interactions between items, gender, and OTL. The gender by OTL interaction is included in both the intercept and item effect models:

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + \gamma_{01}\text{Gender}_j + \gamma_{02}\text{OTL}_j + \gamma_{03}\text{Gender}_j\text{OTL}_j + u_{0j} \\ \beta_{qj} &= \gamma_{q0} + \gamma_{q1}\text{Gender}_j + \gamma_{q2}\text{OTL}_j + \gamma_{q3}\text{Gender}_j\text{OTL}_j.\end{aligned} \tag{4}$$

In M3, γ_{03} estimates the extent to which the overall impact of OTL differs for women and men, or the extent to which the overall gender effect differs by OTL. Finally, the item by gender by OTL interaction γ_{q3} estimates whether or not gender DIF for item q depends on OTL.

Below are the mixed forms of M1, M2, and M3, with levels one and two combined. For model M1,

$$\eta_{ij} = \gamma_{00} + \gamma_{01}\text{Gender}_j + u_{0j} + \sum_{q=1}^{N-1}(\gamma_{q0} + \gamma_{q1}\text{Gender}_j)X_{qij}; \tag{5}$$

for model M2,

$$\eta_{ij} = \gamma_{00} + \gamma_{01}\text{Gender}_j + \gamma_{02}\text{OTL}_j + u_{0j} + \sum_{q=1}^{N-1}(\gamma_{q0} + \gamma_{q1}\text{Gender}_j + \gamma_{q2}\text{OTL}_j)X_{qij}; \tag{6}$$

and for model M3,

$$\begin{aligned}\eta_{ij} = \gamma_{00} + \gamma_{01}\text{Gender}_j + \gamma_{02}\text{OTL}_j + \gamma_{03}\text{Gender}_j\text{OTL}_j + u_{0j}\\ + \sum_{q=1}^{N-1}(\gamma_{q0} + \gamma_{q1}\text{Gender}_j + \gamma_{q2}\text{OTL}_j + \gamma_{q3}\text{Gender}_j\text{OTL}_j)X_{qij}.\end{aligned} \tag{7}$$

Models M1, M2, and M3 each contains 22 parameters more than the next less complex model: M1 contains 22 more parameters than M0, as does M2 over M1, and M3 over M2. These models were selected because of their focus on interactions with item difficulty, specifically item difficulty and gender in M1, item difficulty and OTL in M2, and item difficulty, gender, and OTL in M3. In the case that M2 or M3 did not significantly improve model fit, intermediate models could also be examined, ones which estimate effects on mean performance but not item performance, that is, including (a) OTL in the model for β_{0j} but not for β_{qj}, representing an extension of M1 and (b) an interaction effect for gender and OTL in β_{0j} but not β_{qj}, representing an extension of M2. In the case where M1 or M2 were selected as the final model, these intermediate models would also be examined.

Results

Results are presented first for SGP and DEU. More detailed results on the effects of gender and OTL are then presented for the USA. Results are limited for SGP and are less detailed for DEU, as gender and OTL were found to have less of an effect in these countries than in the USA.

SGP: Gender

Model comparison results are included in Table 3. In SGP, M1 did not result in a significant improvement in model fit over M0 ($\chi^2_{22} = 32$, $p = .083$). Thus, models M2 and M3 were not ued. This indicated that, overall, item difficulty in SGP did not differ significantly by gender. Model M0 was retained as the final model and no other analyses were conducted with SGP.

DEU: Gender

In DEU, both M1 and M2 were found to have significantly better model fit over the previous models. M1 revealed a significant overall mean performance difference for men over women of 0.42 logits, a standardized effect of 0.55 (the DEU M1 *SD* of

Table 3. Model Fit Results for USA, SGP, and DEU.

Country	Model	df	AIC	BIC	log Lik	χ^2	$\chi^2 df$	p
USA	M0	23	7,347	7,503	−3,651			
	M1	45	7,284	7,589	−3,597	107	22	<.001
	M2	67	7,132	7,586	−3,499	196	22	<.001
	M3	89	7,146	7,749	−3,484	30	22	.119
SGP	M0	23	5,671	5,823	−2,813			
	M1	45	5,684	5,980	−2,797	32	22	.083
DEU	M0	23	10,557	10,721	−5,256			
	M1	45	10,527	10,848	−5,218	74	22	<.001
	M2	67	10,455	10,934	−5,161	116	22	<.001
	M3	89	10,473	11,109	−5,147	26	22	.245

Note. M0 is the base model and it includes item difficulty and person ability parameters, M1 additionally includes item by gender interactions, M2 additionally includes item by OTL interactions, and M3 additionally includes all two-way and three-way interactions. Models M2 and M3 were not fit with SGP because M1 was not statistically significant. SGP = Singapore; DEU = Germany; AIC = Akaike information criterion; BIC = Bayesian information criterion; log Lik = log likelihood; OTL = opportunity to learn.

person ability u_{0j} was 0.77). On further inspection, gender effects were found to vary from this impact effect for three items in M1. These were items MFC804, MFC808A, and MFC808B (two of which were also significant under USA M1), with logits of 0.62, −0.69, and −0.55, and standardized estimates of 0.82, −0.89, and −0.71, respectively.

DEU: Gender and Opportunity to Learn

The inclusion of impact and item by OTL interaction effects in M2 did not alter the item by gender relationships; the same three items still showed significant gender DIF. In M2, the *SD* decreased to 0.72, and the overall effect of gender decreased to 0.32 with the inclusion of OTL in the model. The impact effect of OTL in M2 was 0.08, representing an estimated increase of 0.08 logits, or 0.11 *SD*, for an increase in OTL of 1 (the M2 ability *SD* was 0.72). An interaction effect for gender and OTL at model β_{0j} was not found to significantly improve model fit over M2 ($\chi^2_1 = 0.237$, $p = .626$).

USA: Gender

In the USA, M1 was found to have significantly better model fit over M0 ($\chi^2_{22} = 107$, $p<.001$), suggesting the presence of item by gender interaction effects. The AIC was reduced, whereas the BIC increased with M1. Model M1 revealed a significant difference in mean performance for women and men. The mean item difficulty for women was $\gamma_{00} = 0.17$ and the difference from this estimate for men was $\gamma_{01} = 0.74$. Thus,

Table 4. Subset of USA Item by Gender Interactions for M1 and M2.

	M1				M2			
Item ID	Est	*SE*	*p*	Est/*SD*	Est	*SE*	*p*	Est/*SD*
MFC604A2	0.743	0.336	.027	0.835	0.727	0.342	.034	0.983
MFC705A	−0.702	0.273	.010	−0.789	−0.793	0.285	.005	−1.071
MFC705B	−0.673	0.275	.014	−0.756	−0.816	0.290	.005	−1.103
MFC710A	−0.585	0.269	.030	−0.658	−0.451	0.273	.099	−0.609
MFC802D	−0.669	0.289	.021	−0.752	−0.618	0.292	.034	−0.835
MFC804	1.168	0.280	.001	1.313	1.174	0.280	.001	1.587
MFC808A	−0.847	0.335	.012	−0.952	−0.631	0.335	.060	−0.853
MFC808C	−0.574	0.271	.034	−0.645	−0.444	0.270	.100	−0.600

Note. This table contains the estimates (Est), standard errors (*SE*), *p* (*t*-test significance), and standardized estimates (Est/*SD*) from M1 and M2 for the item by gender interaction effects identified as significant in M1. *SD* for M1 was 0.89; *SD* for M2 was 0.74.

mean performance for men was 0.74 logits higher than for women, an increase of 0.83 *SD* (the USA M1 standard deviation of u_{0j} was 0.89).

Inspection of the M1 item by gender interaction effects revealed significant gender DIF on eight items. These items are listed in Table 4. The second through fifth items identified had negative gender interaction effects. After controlling for an overall increase in mean performance for men, women were estimated to perform differentially better than men on these items, on average. For example, on the item with the largest negative interaction effect, item MFC808A, the predicted mean log-odds for women was $\gamma_{00} + \gamma_{q0} = 0.17 + 1.76 = 1.93$. On this same item, the predicted mean log-odds for men was $\gamma_{00} + \gamma_{q0} + \gamma_{01} + \gamma_{q1} = 0.17 + 1.76 + 0.74 + -0.85 = 1.82$. Thus, despite performing better overall, the mean log-odds for men was estimated to be slightly lower on this item. Gender effects for two items were positive: MFC604A2 and MFC804 were estimated to be 0.743 and 1.17 logits higher, respectively, for men than the overall mean increase for men of 0.74.

USA: Gender and Opportunity to Learn

The next step was to examine the improvement in model fit for M2, which included item by OTL interactions, and M3, which included all two-way and three-way interactions. Fit results are shown in Table 3. M2 was found to improve fit over M1 ($\chi^2_{22} = 196$, p<.001), with a slight reduction in both AIC and BIC. M3 was not found to improve fit over M2 ($\chi^2_{22} = 30, p = .119$); the AIC and BIC both increased for M3. An interaction effect for gender and OTL at model β_{0j} was also not found to significantly improve model it over M2 ($\chi^2_1 = 0.187$, $p = .665$). Thus, M2 was retained as the final model.

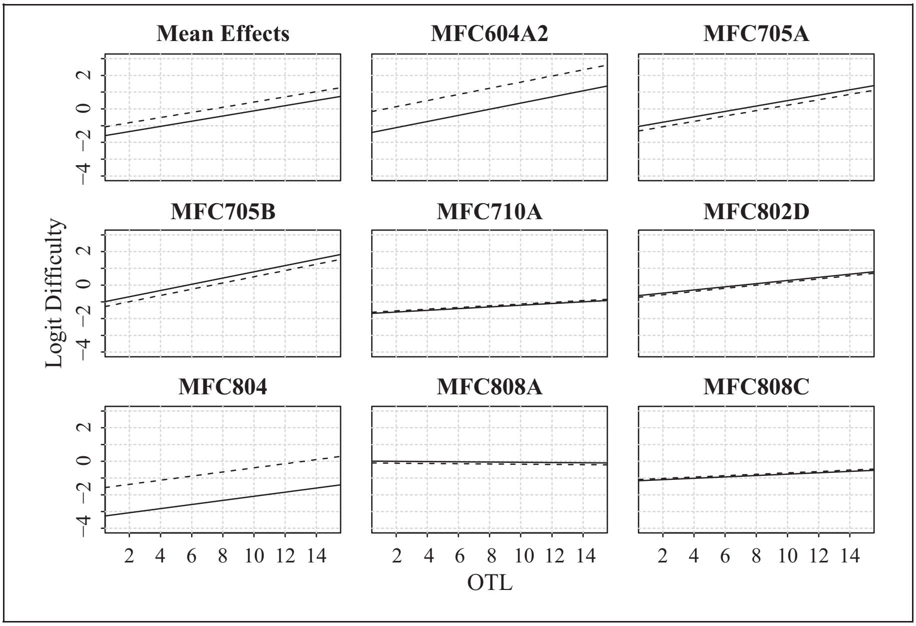

Figure 1. M2 regression lines for the mean opportunity to learn (OTL) and gender effects and for the eight items with significant M1 gender effects. Each plot shows predicted logit item difficulty on the *y*-axis by OTL on the *x*-axis, with separate lines for gender. Solid lines represent difficulty by OTL for women. Dashed lines represent difficulty by OTL for men. Axes are equal across plots.

In the USA, OTL mediated the relationship between item difficulty and gender. Comparison of the M1 and M2 item by gender interaction effects revealed that 3 of the 8 significant interactions from M1 were no longer significant in M2. In Table 4, these items are MFC710A, MFC808A, and MFC808C, all of which have *p* greater than .05 in M2. Standardized estimates (Est/*SD*) are also closer to zero for these items in M2, though they are still all greater in absolute value than 0.50 *SD* (the M2 person ability *SD* was 0.74). In M2, the overall mean gender effect was reduced to 0.52, which was noticeably lower than in M1, even considering the reduction in *SD* from 0.89 to 0.74.

Figure 1 includes the M2 regression lines for the mean OTL and mean gender effects in the first plot, and for the eight items with significant M1 gender effects in the remaining plots. Each plot shows predicted logits by OTL and gender. For all items besides MFC604A2 and MFC804, the lines for men and women are vertically close to one another, revealing that the overall performance difference between women and men, as shown in the first plot, was reduced for the these items.

In M2, the mean impact of OTL was estimated to be 0.15 logits. This slope represents the average change in log-odds of correct response for an increase of one in the number of math topics studied. The slope is shown for women and men in the first plot of Figure 1. For each additional math topic studied, overall performance was predicted to increase by 0.15 logits; for an OTL score change of 3.90, the raw OTL *SD* in the USA, performance was predicted to increase by 0.58 logits.

OTL effects for eight items were found to differ significantly from the mean effect of 0.15. Three of these items had OTL slope estimates that were larger than the mean OTL effect. The remaining five had slope estimates smaller than 0.15. Three of these items are included in Figure 1: MFC710A, MFC808A, and MFC808C, the same three items found to have significant gender effects in M1 but not in M2. Thus, of the eight items with gender DIF in M1, the three with significant OTL effects were all found to no longer have gender DIF in M2. Furthermore, the relationship between difficulty and OTL for these three items was weak, as shown by their small slopes in Figure 1.

USA: Item Content

The final step with the USA data was to examine the content of the items flagged for bias due to gender. There did not appear to be a trend in terms of item content. Of the eight items with gender effects in Table 4, four focused on geometry, two on algebra, and two on number manipulation.

Of the four geometry items, two addressed lines of symmetry in a hexagon and in a rhombus (MFC808A, MFC808C). OTL in M2 was estimated to have little to no impact on the difficulty of these items, as shown in Figure 1. The gender effects for these items were both reduced by the inclusion of OTL in M2.

One of the algebra items (MFC710A) asked whether or not the following situation could be modeled by an exponential function: ‘‘The height *h* of a ball *t* seconds after it is thrown into the air.’’ The correct answer is no. OTL in M2 had only a slight positive effect on difficulty for this item, and the gender effect was reduced from M1 to M2.

The other algebra item (MFC604A2) was the only CR item displaying DIF in M1. This item required students to identify and solve for three unknowns described in a word problem. The gender effect for this item was not reduced in M2.

The remaining two geometry items required students to describe the solution to the equation $3x = 6$ in a two-dimensional plane (item MFC705A) and in three-dimensional space (item MFC705B). The gender effects for these two items were also not reduced by OTL in M2.

Gender effects for both number items also persisted from M1 to M2. Item MFC802D began with the statement ‘‘If the square of any natural number is divided by 3, then the remainder is only 0 or 1.’’ Students then identified whether or not the following procedure represented a correct mathematical proof for the statement: ‘‘Check the statement for the first several prime numbers and then draw a conclusion based on the Fundamental Theorem of Arithmetic.’’ The correct answer is no.

Item MFC804, which was estimated to be substantially more difficult for women than men, required students to calculate the number of combinations when selecting 2 from 10 and 8 from 10. The item stem was worded as follows: ‘‘A class has 10 students. If at one time, 2 students are to be chosen, and another time 8 students are to be chosen from the same class, which of the following statements is true?'' The options compared the number of ways to choose 2 versus 8 students, with the correct response being that the number of ways is the same. With OTL in the model, the substantial gender effect for this item remained.

Discussion

The main purpose of this study was to describe the relationships between math item difficulty, gender, and OTL with data from three different countries. Results of the study provide insight into how math performance is affected by gender and OTL. This discussion focuses on the following key findings: impact effects for gender and OTL, DIF effects for gender and OTL, and differences in results by country.

Results from the final USA model, M2, indicated that mean math performance tended to be higher for men than for women. The M2 mean math performance estimate for women was 0.17 logits, which corresponds to a predicted mean proportion correct of 0.54, close to the observed value of 0.55; the predicted mean proportion correct for men was 0.67, also close to the observed value (see Table 2). This gender effect is consistent with some findings reported in the literature, where effects for gender in standardized math tests tend to favor men over women, especially in high school and college (e.g., Harris & Carlton, 1993; Liu & Wilson, 2009; National Center for Education Statistics, 2011b; Rampey et al., 2009).

Overall performance in the USA tended to increase as OTL increased. The OTL logit slope estimate of 0.15 can be used to predict changes in the mean proportion correct based on changes in the number of math topics studied. For example, reducing the mean OTL by four topics reduces the predicted mean proportion correct by 0.15 for both men and women. Conversely, increasing the mean OTL by four topics increases the predicted mean proportion correct by 0.15 for women and 0.12 for men. This positive relationship between OTL and mean performance is consistent with the positive correlations reported in Table 2 and with the positive relationships described in the literature (e.g., McDonnell, 1995). However, it also provides a direct estimate of the impact of OTL in terms of math performance, one that is not well described in the literature.

DIF effects for gender and OTL were identified as deviations from the corresponding impact effects for individual items. Gender DIF was found in M1 for eight items, six of which were estimated to be differentially easier for women than men and two of which were estimated to be differentially easier for men. Gender differences in item performance were thus found to differ from the mean gender difference, providing evidence of item bias due to gender. Trends were not evident in the content or other features of these items; however, these results suggest that math performance comparisons based on these items, at least using the USA data, should be made with caution.

Three of the eight gender DIF items were also found to function differentially by OTL, with performance on one item remaining essentially unchanged as OTL increased and performance for the other two items increasing only slightly as OTL increased. Furthermore, for these three items, OTL was found to mediate the relationship between item difficulty and gender. Thus, when OTL was found to have little or no relationship with item difficulty, gender effects were no longer significant. On the other hand, for items with OTL effects similar to the mean OTL effect, gender effects persisted.

This study also revealed a lack of significant effects in a number of contexts. In the USA and DEU, the relationship between OTL and performance overall and between OTL and item performance did not differ significantly by gender; gender by OTL interaction effects were not found to improve model fit. Additionally, for DEU, OTL was not found to mediate the relationship between gender and item difficulty; although a number of items functioned differentially by OTL, the inclusion of OTL did not lead to a reduction in the number of significant gender effects in DEU. Finally, for SGP, overall performance and performance at the item level did not differ significantly for women versus men or for different levels of OTL.

Lack of significance in the more complex models examined in this study may have been caused by a lack of statistical power resulting from inadequate sample sizes. Furthermore, a lack of significance for OTL effects may have also been because of the quality of the OTL measure itself (alpha coefficients for the set of 19 tertiary OTL topic statements were .84, .88, and .80 for USA, SGP, and DEU). The OTL scale was based on a self-report of a variety of mathematical topics studied and is only useful to the extent that it measures opportunity to learn the math content represented by the MCK assessment. Future work should consider datasets with larger numbers of individuals and alternative measures of OTL. The TEDS-M study included a measure of OTL at the program/institution level, which may be useful in this context. Results from other countries may also be informative.

Future work should also examine other covariates, both for items and persons, which may explain variability in item difficulty and which may moderate or mediate the relationship between item difficulty and gender. A substantial amount of research has examined item features and their relationship with gender effects, including item content and item type (e.g., Bolger & Kellaghan, 1990; DeMars, 2000; Garner & Engelhard, 1999; Harris & Carlton, 1993; Hyde et al., 1990; Liu & Wilson, 2009). Cross-classified models could be used to simultaneously explore the effects of both item-level and person-level covariates and interactions between them (e.g., Beretvas et al., 2012).

Overall, this study provides evidence of item parameter invariance in the SGP test of MCK, and a lack of invariance, to a certain extent, in the USA and DEU tests. Results indicate that gender effects persist into post-secondary education. Results from the USA also support the findings of Cheong (2006) and Clauser et al., (1996), showing that differential performance and OTL are related and that accounting for OTL may improve DIF detection and impact estimation, thereby reducing bias and increasing the validity of performance comparisons across person groups.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Ackerman, T. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*, 67-91.

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics, 22*, 47-76.

Albano, A. D. (in press). Multilevel modeling of item position effects. *Journal of Educational Measurement*.

Baker, D. P., & Jones, D. P. (1993). Creating gender equality: Cross-national gender stratification and mathematical performance. *Sociology of Education, 66*, 91-103.

Beretvas, S. N., Cawthon, S. W., Lockhart, L. L., & Kaye, A. D. (2012). Assessing impact, DIF, and DFF in accommodated item scores: A comparison of multilevel measurement model parameterizations. *Educational and Psychological Measurement, 72*, 754-773.

Bielinski, J., & Davison, M. L. (1998). Gender differences by item difficulty interactions in multiple-choice mathematics items. *American Educational Research Journal, 35*, 455-476.

Bielinski, J., & Davison, M. L. (2001). A sex difference by item difficulty interaction in multiple-choice mathematics items administered to national probability samples. *Journal of Educational Measurement, 38*, 51-77.

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

Brese, F., & Tatto, M. T. (Eds.). (2012). *TEDS-M 2008 user guide for the international database*. Amsterdam, Netherlands: IEA.

Carroll, J. (1963). A model for school learning. *Teachers College Record, 64*, 723-733.

Cheong, Y. F. (2006). Analysis of school context effects on differential item functioning using hierarchical generalized linear models. *International Journal of Testing, 6*, 57-79.

Clauser, B. E., Nungester, R. J., & Swaminathan, H. (1996). Improving the matching for DIF analysis by conditioning on both test score and an educational background variable. *Journal of Educational Measurement, 33*, 453-464.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer.

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

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin, 136*, 103-127.

Floden, R. (2002). The measurement of opportunity to learn. In A. C. Porter & A. Gamoran (Eds.), *Methodological advances in cross-national surveys of educational achievement* (pp. 231-266). Washington, DC: National Academies Press.

Garner, M., & Engelhard, G., Jr. (1999). Gender differences in performance on multiple-choice and constructed-response mathematics items. *Applied Measurement in Education, 12*, 29-51.

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

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Husén, T. (1967). *International study of achievement in mathematics: A comparison of twelve countries*. New York, NY: Wiley.

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

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., & Williams, C. C. (2008). Gender similarities characterize math performance. *Science, 321*, 494-495.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38*, 79-93.

Kim, S., & Hocevar, D. (1998). Racial differences in eighth-grade mathematics: Achievement and opportunity to learn. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas, 71*, 175-178.

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

Linn, R. L., & Harnisch, D. L. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement, 18*, 109-118.

Liu, O. L., & Wilson, M. (2009). Gender differences in large-scale math assessments: PISA trend 2000 and 2003. *Applied Measurement in Education, 22*, 164-184.

McDonnell, L. M. (1995). Opportunity to learn as a research concept and a policy instrument. *Educational Evaluation and Policy Analysis, 17*, 305-322.

Meece, J. L., & Parsons, J. E. (1982). Sex differences in math achievement: Toward a model of academic choice. *Psychological Bulletin, 91*, 324-348.

Mullis, I. V. S., Martin, M. O., & Foy, P. (2008). *TIMSS 2007 international mathematics report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Muthén, B. O., Kao, C., & Burstein, L. (1991). Instructionally sensitive psychometrics: Application of a new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement, 28*, 1-22.

National Center for Education Statistics. (2011a). *The nation's report card: America's high school graduates* (NCES 2011-462). Washington, DC: Institute of Education Sciences.

National Center for Education Statistics. (2011b). *The nation's report card: Mathematics 2011* (NCES 2012-458). Washington, DC: Institute of Education Sciences.

National Center for Education Statistics. (2011c). *The nation's report card: Reading 2011* (NCES 2012-457). Washington, DC: Institute of Education Sciences.

Penner, A. M. (2003). International gender by item difficulty interactions in mathematics and science achievement tests. *Journal of Educational Psychology, 95*, 650-655.

Rampey, B. D., Dion, G. S., & Donahue, P. L. (2009). *NAEP 2008 trends in academic progress* (NCES 2009-479). Washington, DC: Institute of Education Sciences.

Randall, J., Cheong, Y. F., & Engelhard, G., Jr. (2011). Using explanatory item response theory modeling to investigate context effects of differential item functioning for students with disabilities. *Educational and Psychological Measurement, 71*, 129-147.

Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement, 66*, 63-84.

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

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

Zumbo, B. D., & Hubley, A. M. (2003). Item bias. In R. Fernandez-Ballesteros (Ed.), *Encyclopedia of psychological assessment* (pp. 505-509). Thousand Oaks, CA: Sage.