Summer 7-2015

Examining Sources of Gender DIF Using Cross-Classification Multilevel IRT Models

Liuhan Cai
University of Nebraska-Lincoln, cliuhan@gmail.com

EXAMINING SOURCES OF GENDER DIF USING CROSS-CLASSIFICATION MULTILEVEL IRT MODELS

by

Liuhan Cai

A THESIS

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfilment of Requirements

For the Degree of Master of Arts

Major: Educational Psychology

Under the Supervision of Professor Anthony D. Albano

Lincoln, Nebraska

August, 2015

EXAMINING SOURCES OF GENDER DIF USING CROSS-CLASSIFICATION MULTILEVEL IRT MODELS

Liuhan Cai, M.A.

University of Nebraska, 2015

Adviser: Anthony D. Albano

A substantial amount of research has focused on the detection of differential item functioning (DIF) in the past. However, DIF detection and the estimation of DIF effect size do not explain why it occurs. Recent studies have investigated how or why DIF may occur. Improvements in DIF analysis models have made it possible to explore additional covariates as potential sources of DIF by measuring the extent to which these covariates account for variation in performance. The current study examines variability in math performance accounted for by gender, which is referred as gender DIF. This study then investigates how the presence of gender DIF is explained by both person predictors (i.e., opportunity to learn; OTL) and item characteristics (i.e., item format). A cross-classification multilevel IRT model framework is used to demonstrate the relationship among item difficulty, gender, OTL, and item format. Data come from three countries participating in an international study of pre-service math teachers, the Teacher Education and Development Study in Mathematics (TEDS-M).

ACKNOWLEDGMENTS

I would like to thank my adviser, Dr. Anthony Albano, for his encouragement and guidance throughout all stages of this thesis. I would also like to thank Dr. Kurt Geisinger for his constructive advice and feedback. My sincere gratitude also goes to my parents, sister, and brothers, for their unconditional love and constant support. I am especially thankful for Steven Dunn, for his patience and helping me get to this point.

Contents

List of Tables

Chapter 1

Introduction

Gender differences in math performance have been widely studied. Results of large-scale assessments such as the Programme for the International Student Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMSS) indicate consistently higher average math scores for male students across countries (Else-Quest, Hyde, & Linn, 2010). Since policy reforms and teaching practices are sometimes informed from these assessments, it is imperative to ensure score comparability. Measurement invariance is the essential psychometric property for scores to be comparable (Meredith, 1993). Measurement invariance is a statistical property where item parameters do not vary across multiple groups of examinees and person parameters do not vary across time points or measurement conditions (Rupp & Zumbo, 2006). A lack of item parameter invariance, a special case of measurement invariance, will result in differential item functioning (DIF, Hambleton, Swaminathan, & Rogers, 1991). The existence of DIF items in an assessment can invalidate score interpretations and threaten test fairness.

DIF results from the influence of variables other than the construct of interest (Ackerman, 1992). An item is identified to be free of DIF if all individuals with the same underlying ability or construct have equal probability of answering the item correctly, regardless of variables such as group membership (Hambleton et al., 1991). DIF detection is a process of identifying items that are impacted by these extraneous variables. Various DIF analysis techniques and models have been developed to examine invariance in both person and item parameters, including contingency tables (e.g., Holland & Thayer, 1988), regression models (e.g., Swaminathan & Rogers, 1990), item response theory models (IRT, e.g., Thissen, Steinberg, & Wainer, 1993), multidimensional models (e.g., Roussos & Stout, 1996), structural equation models (e.g., Muthén, Kao, & Burstein, 1991), and multilevel models (e.g., Cheong, 2006; Kamata, 2001).

A substantial amount of research has focused on the detection of DIF in the past. However, DIF detection and the estimation of DIF effect size do not explain why it occurs (Kim, Cohen, Alagoz, & Kim, 2007). Recent studies have investigated how or why DIF may occur. Improvements in DIF analysis models have made it possible to explore additional covariates as potential sources of DIF by measuring the extent to which these covariates account for variation in performance.

The purpose of the present study is to examine variability in math performance accounted for by gender, which is referred as lack of measurement invariance by gender or gender DIF. This study then investigates how the presence of gender DIF is explained by both person predictors (i.e., opportunity to learn; OTL) and item characteristics (i.e., item format). A cross-classification multilevel IRT model framework is used to demonstrate the relationship among item difficulty, gender, OTL, and item format. Data come from an international study of pre-service math teachers, the Teacher Education and Development Study in Mathematics (TEDS-M).

The following section first provides a brief review on the implementation of multilevel item response models for testing item difficulty parameter invariance and how these models are extended to explore sources of DIF. Previous research on gender differences in math performance at both the test level and the item level are reviewed. Finally, previous work on OTL is reviewed.

Chapter 2

Literature Review

2.1 Modeling Parameter Variance

Many different methods have been developed to investigate item bias or DIF. DIF analyses are usually conducted through multiple statistical tests for individual items or each pair of focal and reference groups. Too many statistical tests can result in false positives (Longford, Holland, & Thayer, 1993). Compared to the traditional DIF detection procedures, the logistic mixed model is more economical as it is carried out to detect the DIF existence in an omnibus test, rather than with individual items. DIF can be interpreted by the significant interactions between item difficulty (at the item-level) and group membership (at the person-level). Moreover, sources of DIF can be explained by modeling item or person covariates through exploratory mixture model analysis (Cohen & Bolt, 2005; Van den Noortgate & De Boeck, 2005).

Researchers have demonstrated the feasibility of formulating the traditional IRT models as multilevel logistic models (e.g., Adams, Wilson, & Wu, 1997; Kamata, 2001). In the basic one-parameter IRT or Rasch model (Rasch, 1960), the log-odds of correct response to item i for person j are modeled as

$$Logit\ P(Y_{ij} = 1) = ln(\frac{P}{1-P}) = \eta_{ij} = \theta_j - b_i. \tag{2.1}$$

Here, Y_{ij} represents the scored response of person j to item i (1 = correct, 0 = incorrect). Item response scores are explained by a logistic function of the difference between person ability θ_j and item difficulty b_i. The model can also be reformulated as a cross-classification multilevel model with random person and random item effects (Van den Noortgate, De Boeck, & Meulders, 2003):

$$\begin{gathered} Logit\ P(Y_{ij} = 1) = \eta_{ij} = \beta_0 \\ \beta_0 = \gamma_0 + u_{0i} + u_{0j} \end{gathered} \tag{2.2}$$

with $u_{0i} \sim N(0, \sigma^2_{u_i}), u_{0j} \sim N(0, \sigma^2_{u_j})$. This two-level model includes item responses as the level-one unit with persons and items as level-two units. Item responses are nested within both persons and items. Both items and persons are assumed to be random samples from a population of items and a population of persons. The log-odds of correct response are modeled as a summation of the item and person parameters. In this baseline model in equation 2.2, γ_0 represents the estimated log-odds of correct response of a person with an average ability on an item with an average difficulty.

The baseline model is only a descriptive model. However, as item and/or person covariates are incorporated in the subsequent steps, the possibility of overall DIF detection and DIF source investigation opens up (Van den Noortgate & De Boeck, 2005). Variability in ability for people is estimated by the random effect u_{0j}; in the same way, variability in item difficulty is estimated by the random effect u_{0i}. DIF can be tested by allowing group effects to vary over items at level two (see examples below). When the group main effect and item-by-group interaction effects are included in the model, the random effects of group over items represent the residual DIF. The model can be further extended by adding item predictors or person characteristics predictors to explain the DIF. If the item or person covariates explain the DIF effects, one would expect that the group main effect on the additional item or person covariates would differ from zero and the variance of the random group effects over items would decrease.

Van den Noortgate and De Boeck (2005) used cross-classified multilevel models to examine item parameter invariance across gender. Gender and level of processing (retrieving, structuring, and evaluating) were both examined as potential sources of variability in item difficulty parameters. Here, students and items were the level-two units within a cross-classified structure. Item responses as the level one unit were nested within students and items. A model similar to the baseline model in equation 2.2 was extended to include gender as a grouping covariate at level two. The results demonstrated small yet significant random gender effects, indicating gender DIF. Level of processing as an item covariate at level two was then examined as a potential source of gender DIF. Results suggested that the level of processing was not a source of variability in item difficulty parameters by gender; the gender-by-level of processing interaction term was not statistically significant, and the variance of the gender random effects over items did not decrease.

Van den Noortgate and De Boeck (2005) also presented a second example demonstrating DIF detection and DIF source exploration. The second study investigated parameter invariance across schools. Responses were nested within students and items. Students and items were nested within schools. There were 33 secondary schools, and schools were regarded as a random sample of a population of secondary schools. In this case, items, students, and schools were all considered to be random effects. School as a level three covariate was the grouping variable. Group effects were modeled by means of additional random parameters. The school DIF effects were represented by the significance of random interaction term between schools and items. A third level covariate, school type, was incorporated into the model to explain the DIF effects. School type was found to partly explain school DIF. Not only did the school type effect vary over items to a statistically significant degree, but the school DIF magnitude was also reduced from 0.27 to 0.22 with school type in the model.

2.2 Gender Differences

Gender differences in math achievement have been a concern for decades as researchers have been searching for the cause of women's underrepresentation in Science, Technology, Engineering and Mathematics (STEM) (Hyde, Lindberg, Linn, Ellis, & Williams, 2008). Male students are reported to achieve higher math scores than female students in national and international large-scale assessments (e.g., Baker & Jones, 1993; Beller & Gafni, 1996; Gallagher & Kaufman, 2005; Gamer & Engelhard Jr, 1999). Meanwhile, research has consistently reported math and reading achievement parity between genders in early grades with increasing male advantages in math and female advantages in reading achievement as they move up through the grades (e.g., Willingham & Cole, 1997).

A considerable amount of research has documented the gender gap in math performance at the overall test level. Hyde, Fennema, and Lamon (1990) conducted a meta-analysis on gender effects on math performance. A weighted mean effect size of 0.15 was found over 100 studies. This small effect size indicated that, overall, males outperformed females by a small but not negligible amount. The study also reported a slight female superiority in computation, but no gender differences in understanding of math concepts. Starting from high school to college, the gender discrepancy favoring male students emerged in the area of complex problem solving, which represented the highest cognitive level studied. In terms of content domain, there was a slight male advantage in geometry, but no gender differences in arithmetic or algebra performance.

A more recent meta-analysis reported similar findings. Lindberg, Hyde, Petersen, and Linn (2010) examined 242 studies of gender math performance from 1990 to 2007 and indicated small gender variations in mean math achievement. However, the study did not find a decline in the gender gap from 1990 to 2007. The gender gap was not prominent prior to high school; performance differences favoring males peaked during high school with an effect size of 0.23, and declined among college students. With regard to item features, male students were reported to have better overall performance on multiple choice questions, whereas female students tended to do better on short answer and open ended questions. Furthermore, there were no gender variations in performance on different math content domains or the depth of knowledge. In terms of cognitive domain, male students performed slightly better on items involving problem solving skills in high school, but this difference was reversed, though small, among college students.

In the same study, Lindberg et al. (2010) conducted another meta-analysis using some large national datasets collected after 1990 in the USA. The datasets yielded an average weighted effect size of 0.07, indicating the small differences in mean performance by gender in the USA. The results also suggested that the tests with a higher proportion of algebra items favored females, and the tests with a higher proportion of measurement items favored males. The other content areas (numbers and operations, geometry, and data analysis and probability) were not found to differ by gender. On the National Assessment of Educational Progress (NAEP) datasets, male students had higher outcomes on the tests with a higher proportion of multiple choice items, whereas female students performed better on the tests with a greater proportion of short-answer and open-ended items.

Cross-national patterns of gender differences in math performance are well-documented. Else-Quest et al. (2010) examined two international datasets, the TIMSS and the PISA, and found that males and females performed similarly in most countries despite cross-national variability in the direction and magnitude of effects. An overall gender parity across content domains (algebra, number, measurement, geometry, and data) and cognitive domains (knowing, applying, and reasoning) was evident.

Wiseman (2008) investigated the phenomenon of gender segregation in national education systems using cross-nationally comparative data from 46 countries. Among the 46 countries, 34 had varying degrees of gender segregation at the education system or classroom levels, 12 had no gender segregation at all. This study provided evidence of gender parity in enrollment at both the primary and secondary levels for most countries. No consistent differences were found in achievement advantage by gender in many countries that had gender-segregated systems compared to the co-educational systems.

Gender differences at the item level have been researched on different dimensions such as item difficulty, item format, and math content domain. Penner (2003) examined the relationship between gender differences and item difficulty in math items using the 1995 TIMSS dataset. Results showed that in four out of ten countries, easier items tended to be more difficult for female students. As item difficulty increased, items tended to be even more difficult for female students. In four other countries, there were no gender differences on easy items, but as the difficulty increased, the items were more difficult for female students. In the remaining two countries, easier items were harder for female students, but gender-specific difficulty increased at the same rate for both genders. The study showed a general pattern of a male advantage on easy math items and an increasing male advantage on more difficult math items.

Other studies have also found significant gender-by-item difficulty interactions. Bielinski and Davison (1998) studied the minimum competency math test outcomes among eighth-grade and ninth-grade students and found that for both grades, easy items tended to be easier for female students than male students, while harder items tended to be harder for female students than male students. The significant negative correlations (-0.47 and -0.43) between gender differences in item difficulty, and in item difficulty estimated on the overall samples over the two studies, indicated that as item difficulty increased, the male advantage also increased. To extend their previous study, Bielinski and Davison (2001) used the 1992 NAEP, the USA cohort from the TIMSS study, and the 1988 National Educational Longitudinal Study datasets to investigate gender effects on item difficulty in math for primary and secondary students. They reported that a similar phenomenon emerged in the 1998 study where math tests with harder items generally were in favor of men and that this gender variability grew in late adolescence.

Additional research has revealed that item format is related to gender DIF. Multiple-choice (MC) and constructed-response (CR) are the two item formats that studies have typically examined. These studies have sometimes produced inconsistent results. Taylor and Lee (2012) analyzed state math tests for fourth-, seventh-, and tenth-grade students and used POLYSIBTEST and a Rasch procedure to explore gender DIF based on item format. The results indicated that even though the Rasch procedure identified many more DIF items than the SIB procedure, the directions of DIF effects were the same. Both procedures showed clear patterns that MC items favored male students and CR items favored female students, throughout all grade levels. Other studies have found similar results (e.g., DeMars, 1998; Gamer & Engelhard Jr, 1999; Becker, 1990).

In contrast, Liu and Wilson (2009) examined the USA portion of the PISA 2000 and 2003 math assessments using a multidimensional Rasch model. The results suggested no measurable gender differences on traditional MC items for both administrations. Moreover, male students showed consistent advantages on CR items on both assessments, even though the effect sizes were small. The largest gender gap was on complex MC items (an unconventional item format) where male students significantly outperformed female students with an effect size of 0.19.

Regarding math content domain, Mendes-Barnett and Ercikan (2006) used the data of 12th grade students' math exams to investigate the relationship between gender DIF and math content domain. Differential bundle functioning (DBF) analyses were utilized to identify different response patterns by gender in math achievement. They found that even though the geometry bundle did not function differentially for male students and female students, the individual geometry items exhibited high DIF, especially those ones that utilized visuals. In the content area of computation, items with no equations were found to favor female students, yet the computation items with equations displayed no DBF. Finally, there was a male advantage in algebra items. Becker (1990) reported similar results in terms of item performance on math content areas by gender. Gamer and Engelhard Jr (1999) also found geometry was not a source of gender DIF.

Conversely, in examining the math section of the SAT, Harris and Carlton (1993) revealed that after controlling mean abilities, men performed better on geometry items, while women performed better on algebra items. Among items showing DIF, eight out of 15 items came from geometry and measurement that functioned in favor of men, but none were from algebra. On the other hand, nine out of 16 items came from algebra that functioned in favor of women. Men were also found to have significant advantages in number and computation, data analysis, and proportional reasoning.

Differential course taking by gender is a potential explanation for male advantages in math performance (Meece, Parsons, Kaczala, & Goff, 1982). Beginning from high school, female students tend to take fewer advanced math and science courses in which problem solving skills are intensively trained. However, in the United States, the gender gap in course enrollment has gradually disappeared. Gender differences in patterns of interest could be a potential factor that explained the course choice variations (Su, Rounds, & Armstrong, 2009). In addition, parents' and teachers' expectation discrepancy in math ability among men and women can play an important role in their course choices (Jacobs, Davis-Kean, Bleeker, Eccles, & Malanchuk, 2005; Eccles, 1994).

In spite of the overall gender similarities in math achievement, males have demonstrated greater self-confidence and less anxiety in their math ability than females, as well as higher intrinsic and extrinsic motivations in math (Else-Quest et al., 2010). Math achievement has been correlated positively with attitudes at the student level (Shen & Tam, 2008). Gender discrepancy could be explained in part by self-perception of math ability. Implicit stereotypes of male superiority in math and science can also reinforce gender differences in math and science engagement and performance (Steffens & Jelenec, 2011). Wiseman (2008) suggested that gender parity was only achieved when there was equity in enrollment, access to resources, and opportunity to learn for both males and females. Likewise, Else-Quest et al.

(2010) concluded that cross-national variability of differential math performance by gender was associated with country-level disparity in opportunity structures for females. Gender equity in school enrollment, women's share of research jobs, and women's parliamentary representation contributed to variability in gender distinction in math performance.

2.3 Opportunity to Learn

The concept of OTL was first introduced by the International Association for the Evaluation of Educational Achievement (IEA) in the 1960s to demonstrate differential math performance across nations (McDonnell, 1995). Husen (1967) described OTL in the context of testing where "students have had the opportunity to study a particular topic or learn how to solve a particular type of problem presented by the test." He argued that the likelihood of answering test items correctly would subsequently decrease if students have not had the opportunities to learn the pertinent topics, even though they might provide solutions by utilizing knowledge of related topics. The concept of OTL has evolved since then. Highlighting the important contributions of OTL in learning and development in education, Carroll (1963) conceptualized OTL as the amount of time allowed for learning.

However, some researchers have criticized that the conceptualization of OTL as the amount of time allowed for learning only provided crude data in teacher education components (Cochran-Smith & Zeichner, 2005). Without taking content coverage into account, the qualitative similarities and differences between teacher education programs can be ignored, which may lead to inconsistent results (Blömeke & Kaiser, 2012). OTL was then further framed as the content coverage of knowledge, specifically the topics being taught, the relative emphasis on different aspects of a subject, and students' achievements on the relative important aspects of the subject (Travers & Westbury, 1989). The TEDS-M study followed the IEA tradition of connecting OTL to math achievement. In TEDS-M, OTL was construed as the occasions that pre-service teachers had to learn about particular mathematical topics during the course of teacher education. In this sense, mathematical domain specificity defined the element of an educational opportunity (Schmidt, McKnight, Valverde, Houang, & Wiley, 1997). Diversity in teacher education programs was reflected in curriculum which were established to determine what future teachers were supposed to know, educational opportunities that were provided in class, and the outlooks of how the teacher education program should be organized to offer necessary knowledge and skills for success in future teachers' professional tasks (Floden, 2002). OTL, in this case, was measured by asking the pre-service teachers what they perceived had been covered in the areas of math and math pedagogy during their teacher education.

OTL was primarily used to make cross-national comparisons. It was suggested that OTL should be considered to ensure fairness in performance comparisons (McDonnell, 1995). Through the examination of math textbooks and their use in lower secondary classrooms, Haggarty and Pepin (2002) found that learners from different countries were provided with different math knowledge and offered different levels of OTL in math. Some research has shown that in international contexts, countries with higher levels of OTL outperform those with lower levels of OTL (e.g., Mullis, Martin, & Foy, 2008). Schmidt, Cogan, and Houang (2011) examined future primary and lower secondary teachers' OTL in teacher preparation programs in the USA compared to other high-achieving countries. The results indicated that countries outperforming the USA tended to allocate more course work to math content preparation, as opposed to general and math pedagogical knowledge, especially at the lower secondary level. Even though no causal inference was supported, variation in OTL across three areas (math content, math pedagogy, general pedagogy) was related to differences in the math and math pedagogical performance. As for the primary future teachers, the difference in performance between the USA and higher-achieving countries was not as prominent. However, variation of OTL in math relative to pedagogy across the USA institutions for both the primary and lower-secondary levels was larger than its counterparts, ranging from 22% to 56% and 25% to 86%.

OTL has also been researched at the individual level. Boscardin et al. (2005) used hierarchical linear modeling to investigate the impacts of various OTL variables on student outcomes in English and algebra. The first level in the model was the student level, where each individual student was the unit of analysis. Students were nested within teachers at level two. Findings suggested that teacher expertise in these two content areas was positively correlated with student performance. Moreover, content coverage, as an indicator of OTL, was also found to have a consistently positive relationship with outcomes from the algebra and English assessments. Specifically, with one more week spent on relevant content, there was an expected increase of 0.85 in algebra test scores. On the other hand, one additional week covering English resulted in an increase of 1.59 points on the English test.

A positive association has also been found between OTL and college students' acquisitions in math and math pedagogical knowledge. Blömeke, Suhl, Kaiser, and Döhrmann (2012) found that among future primary teachers, OTL in math not only had a strongly positive direct effect on math performance, but also significantly influenced math pedagogical knowledge, presumably by mediating the effects of OTL in math pedagogy. Additionally, OTL in math pedagogy had indirect effects on both math and pedagogical knowledge, by mediating the effects of entry which was represented by students' perceived high school achievement. The higher the content coverage of math pedagogy for a program, the more attractive it was to students with high perceived high school achievement, who in turn showed higher performance in both areas.

Relatively few studies have addressed the relationship between OTL and person grouping variables at the item level performance. Albano and Rodriguez (2013) used hierarchical generalized linear modeling to investigate parameter invariance over covariates at the student level. In this study, item responses were nested within students. Gender and OTL were both examined as potential sources of variability in item difficulty parameters. A two-level model was used, where gender was the person group covariate at level two and OTL was the person covariate at the same level. Lower secondary future teachers from three countries (USA, Singapore, and Germany) were examined. For the Singapore cohort, item difficulty did not significantly differ by gender. In Germany, controlling the mean ability, items functioned in favor of men. The inclusion of OTL impact and item-by-OTL interaction effects did not reduce the number of items showing gender DIF, though a number of items did function differentially by OTL; thus OTL was not found to be a source of DIF in Germany. For the USA cohort, the best-fitting model included main effects for items, gender, and OTL, and the two-way interaction effects of item-by-gender and item-by-OTL. Difficulty estimates for eight out of 22 items were found to vary by gender when OTL was not included in the model. These items were initially identified as exhibiting gender DIF. When OTL main effect and item-by-OTL interaction effects were introduced to the model, the mean proportion corrected was expected to increase by 0.15 logits for a one unit increase in OTL. Furthermore, three out of eight items were no longer found to display gender DIF. These results indicated that person-level OTL can mediate the relationship between item difficulty and gender. Differential OTL may partly contribute to differential math performance.

As a related measure of OTL, Wu and Ercikan (2006) examined the impact of extra lesson hours after school (ELHAS) on item-level performance. Multiple-variable matching with logistic regression was used to decide whether a cultural background factor ELHAS was a source of DIF. The Taiwan and the USA cohorts from the TIMSS 1999 dataset were examined. By adding ELHAS as a covariate, the magnitude of DIF for 30% of items was reduced across four content areas, including factions and number sense, measurement, geometry, and algebra. Items found to favor students from Taiwan were mediated by the ELHAS. However, in the content areas of data representation and analysis and probability, DIF remained unchanged. Eight out of nine DIF items in this content area were detected in favor of students from the USA. These findings were attributed to differences in curriculum between Taiwan and the USA. Since the content area data and probability were not covered in Taiwan's eighth-grade curriculum, there was no ELHAS provided.

Finally, Burkes (2009) used multilevel-DIF methodology to examine item performance differences across two socioeconomic status (SES) groups with similar overall math ability. Item responses were nested within students who were nested within classrooms in the USA. In this case, SES was the person group covariate at level two, and classroom-level instructional opportunities, students' opportunity to learn the assessed math content domain and topics within their classroom, was the covariate at level three. Eight out of 71 items were detected to exhibit DIF, all of which favored students with higher SES compared to those who with lower SES.

The DIF items were from three out of five domains, including algebra, data, and number (geometry and measurement were not addressed). When item difficulties and DIF effects for SES were modeled at the classroom level as a function of OTL, only one item still exhibited DIF. For seven out of eight items, OTL was found to be the source of SES-based DIF. Under the influences of OTL, the seven items were systematically more difficult for students with lower SES.

Chapter 3

Method

3.1 Sample

Data for this study came from the lower-secondary pre-service teachers in the TEDS-M study. The target population was defined as the future teachers in their final year of teacher education program who would be eligible to teach mathematics in lower-secondary schools (Tatto et al., 2008). Future secondary teachers from 15 countries participated in the TEDS-M study. Participants were sampled following a stratified multistage probability sampling design. The analyses in this study were conducted using the data from Singapore (SGP), with 393 students (48% female, 52% male), Germany (DEU), with 768 students (61% female, 38% male), and the United States (USA), with 475 students (69% female, 31% male). These three countries were chosen because they represented distinct geographic and cultural contexts and they differed noticeably on variables of interest.

3.2 Instruments

The TEDS-M study measured future teachers' math content knowledge (MCK) and math pedagogical content knowledge (MPCK) as the outcomes at the end of secondary teacher education. The assessment was administrated in a standardized and monitored test session with a 60-minute completion time. Three test booklets were developed for the secondary level. The items were assigned to booklets following a balanced-incomplete-block design (Tatto et al., 2008). The present study used scored item responses from the MCK assessment. The MCK assessment contained a total of 76 items with four content domains including number, algebra, geometry, and data. Item formats MC (multiple-choice and complex multiple-choice) and CR (constructed-response) were used. As shown in Table 3.1, the assessment contained 58 MC and 18 CR items. Each item fell into one of the four domains: number (27 items), geometry (23), algebra (22), and data (4).

Table 3.1: MCK items by Item Format and Content Domain

	MC	CR	Total
Number	24	3	27
Geometry	17	6	23
Algebra	15	7	22
Data	2	2	4
Total	58	18	76

Measurement of opportunity to learn was conducted both at the individual and program/university level. This study focused on individual OTL. In the TEDS-M study, OTL was defined as future teachers' occasion to learn about particular topics during the course of teacher education. This study used the total OTL scores on tertiary math, as it was considered to be most relevant to secondary education. Tertiary OTL was based on the future secondary teachers' responses to whether or not they had the opportunity to learn 19 topics in four key areas: (1) geometry, e.g., axiomatic geometry or analytic geometry, (2) discrete structures and logic, e.g., linear algebra or number theory, (3) continuity and functions, e.g., multivariate or advanced calculus, (4) probability and statistics, e.g., distribution.

3.3 Preliminary Analysis

Descriptive statistics for proportion correct scores on the MCK items and OTL by gender and by country are provided in Table 3.2. Proportion correct is the probability of answering the set of items correct. Overall across the three countries, men had higher means than women. For the USA cohort, men were 0.10 higher than women on average. In SGP, the mean for men was 0.03 higher than women. In DEU, men were 0.06 higher on average than women. Similarly, for all the three countries, men had higher OTL means than women. The difference in OTL means indicated that men generally had studied one more mathematics topic than women among the three countries. By examining the correlations between proportion correct and OTL, women were found to have higher correlations than men in each country. Specifically, USA had the highest estimates (women: 0.52; men: 0.39); DEU had medium estimates (women: 0.38; men: 0.35); SGP had the lowest estimates (women: 0.11; men: 0.08). The descriptive statistics suggest that item level performance may be a function of gender, and that OTL may moderate the relationship between gender and item performance.

Table 3.3 contains descriptive statistics for average proportion correct response by gender, item format, and country. In USA, the mean proportion correct on MC and CR items were 0.08 and 0.17 higher for men than women. In SGP, the discrepancy between men and women was less; the mean proportion on MC and CR

Table 3.2: Descriptive Statistics by Gender and Country

			Prop Correct		OTL		
Country	Gender	*N*	*M*	*SD*	*M*	*SD*	*r*
USA	F	325	0.56	0.14	11.33	4.06	0.52
	M	149	0.66	0.13	13.03	3.25	0.39
SGP	F	189	0.66	0.11	9.19	4.76	0.11
	M	203	0.69	0.11	10.54	4.48	0.08
DEU	F	473	0.63	0.13	10.55	3.81	0.38
	M	294	0.69	0.14	11.90	3.89	0.35

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

items were 0.02 and 0.03 higher for men than women. In DEU, men outperformed women on both MC and CR items by 0.06 proportion correct. Across all countries, students performed better on MC than CR items. However, overall gender discrepancy was bigger on CR items than MC items. The preliminary findings indicate that item format may influence the item level performance for men and women in different ways.

Table 3.3: Descriptive Statistics by Gender and Format

			Prop Correct (MC)		Prop Correct (CR)	
Country	Gender	*N*	*M*	*SD*	*M*	*SD*
USA	F	325	0.60	0.13	0.43	0.22
	M	149	0.68	0.13	0.60	0.21
SGP	F	189	0.67	0.11	0.64	0.18
	M	203	0.69	0.12	0.67	0.19
DEU	F	473	0.67	0.14	0.51	0.22
	M	294	0.73	0.13	0.57	0.22

Note. Prop Correct is the proportion correct score across items with different formats.

3.4 Models

In this study, model fit was compared for each model, with one model considered to be a reduced form of the subsequent model. Chi-squared likelihood ratio (χ^2) tests were conducted to test the appropriateness of the more complex models. AIC (Akaike information criterion) and BIC (Bayesian information criterion) were also used to determine model fit. If the χ^2 was statistically significant and AIC reduced for a model, then the model was considered significant. BIC provided supplemental fit information. The model fit comparison approach examined the significance of the inclusion of one parameter or a set of parameters. Thus, individual effects were tested by two-sided Wald tests with an alpha level of 0.05 when necessary. This model fit comparison approach was repeated for each country.

The baseline model M0 in equation 2.2 had random effects for both items and people. Since the means of both residual terms were set to 0, the intercept represented the mean difficulty for a person with mean ability, or the estimated log-odds of a correct response of an "average" person on an "average" item. The larger the estimated value, the easier items would be for an average person.

Model M1 examines a gender main effect. $Gender_j$ equals to 0 if person j belongs to the reference group women, or 1 if person j belongs to the focal group men:

$$\eta_{ij} = \gamma_0 + \gamma_1 Gender_j + u_{0i} + u_{0j}. \tag{3.1}$$

In this model, γ_0 estimates the mean performance for women, and γ_1 estimates the difference of mean performance for men compared to women. Thus mean performance for men is $(\gamma_0 + \gamma_1)$. The residual terms u_{0i} and u_{0j} still represent the random item effects and random person effects.

Model M2 examines gender impact and item-by-gender interaction effects:

$$\eta_{ij} = \gamma_0 + \gamma_1 Gender_j + u_{1i} Gender_j + u_{0i} + u_{0j}. \tag{3.2}$$

The residual terms u_{0i} now represent the random item effects for women, and u_{1i} estimates the overall differential item effects for men compared to women. This model is used to examine gender DIF. If the variance of u_{1i} differs from 0, then there are uniform DIF effects over the gender groups. The random effects u_{0i} and u_{1i} can be correlated. A positive correlation means that by controlling the overall performance of all the people, the items with higher difficulty are harder for men. Item or person covariates would be included to explore DIF sources only if overall gender DIF is detected in Model M2.

Model M3 examines format impact and gender-by-format interaction effects. $Format_i$ is 0 if item i is MC, or 1 if item i is CR. Format is added to the model as an item covariate to determine whether it contributes to DIF:

$$\begin{aligned}\eta_{ij} = \gamma_0 + \gamma_1 Gender_j + \gamma_2 Format_i + \gamma_3 Gender_j Format_i + \\ u_{1i} Gender_j + u_{0i} + u_{0j}.\end{aligned} \tag{3.3}$$

γ_0 now estimates the mean performance for women and γ_1 estimates the difference for men in math performance controlling for item format. γ_2 estimates the difference in mean performance between the two item formats controlling for the gender effect. The interaction parameter γ_3 estimates the amount of gender differences depending on the item formats, or the magnitude of differential performance on item formats between genders. Here, one would specifically focus on how the inclusion of item format as an item covariate influences DIF effects (u_{1i}). If the gender-by-item format interaction term is significant and the variance of u_{1i} over items is reduced, one would conclude that this item covariate explains DIF.

Model M4 examines an OTL main effect, where the mean-centered OTL (OTL_j) is added to the model as a person covariate:

$$\begin{aligned} \eta_{ij} = \gamma_0 + \gamma_1 Gender_j + \gamma_2 Format_i + \gamma_3 Gender_j Format_i + \\ \gamma_4 OTL_j + u_{1i} Gender_j + u_{0i} + u_{0j}. \end{aligned} \tag{3.4}$$

γ_0 now estimates the mean performance for women and γ_1 estimates the difference between the genders, controlling for item format and OTL. γ_2 estimates the difference in mean performance between the two item formats, controlling for gender and OTL. The interaction parameter γ_3 estimates the amount of change in gender differences depending on both formats at the mean OTL score. γ_4 estimates the effect of OTL on mean performance, controlling for gender and item format.

Model M5 investigates OTL impact and item-by-OTL interaction effects:

$$\begin{aligned} \eta_{ij} = \gamma_0 + \gamma_1 Gender_j + \gamma_2 Format_i + \gamma_3 Gender_j Format_i + \\ \gamma_4 OTL_j + u_{1i} Gender_j + u_{2i} OTL_j + u_{0i} + u_{0j}. \end{aligned} \tag{3.5}$$

u_{1i} estimates the amount of gender DIF, and u_{2i} estimates OTL effect at the item level. If the item-by-OTL interaction effects are significant and the gender DIF effects $\sigma^2_{u_{0i}}$ are reduced, one can conclude that OTL contributes to the explanation of DIF.

Model M6 examines two-way interaction effects between gender and OTL:

$$\begin{aligned} \eta_{ij} = \gamma_0 + \gamma_1 Gender_j + \gamma_2 Format_i + \gamma_3 Gender_j Format_i + \\ \gamma_4 OTL_j + \gamma_5 Gender_j OTL_j + u_{1i} Gender_j + u_{2i} OTL_j + u_{0i} + u_{0j}. \end{aligned} \tag{3.6}$$

γ_5 estimates the extent to which the overall impact of OTL differs between men and women, or the extent to which the overall gender effect differs by OTL.

Model M7 adds the three-way interaction effects between items, gender and OTL:

$$\begin{aligned}\eta_{ij} = \gamma_0 + \gamma_1 Gender_j + \gamma_2 Format_i + \gamma_3 Gender_j Format_i + \\ \gamma_4 OTL_j + \gamma_5 Gender_j OTL_j + u_{1i} Gender_j + u_{2i} OTL_j + \\ u_{3i} Gender_j OTL_J + u_{0i} + u_{0j}.\end{aligned} \tag{3.7}$$

u_{3i} estimates whether or not gender DIF for all items depends on OTL, or whether the impact of OTL at the item level differs by gender.

Models were fit sequentially based on significance. First, M2 was compared to M1, providing evidence of gender DIF. Starting from M3, if the inclusion of an item covariate significantly improved model fit and reduced DIF, this covariate remained in subsequent models; if the item covariate did not explain DIF, it was omitted from subsequent models.

Chapter 4

Results

The models were fit using the “glmer” function in “lme4” package (Bates, Maechler, Bolker, & Walker, 2015) from R (R Development Core Team, 2015). All analyses were then carried out within R.

4.1 SGP: Gender

As shown in Table 4.1, in the baseline model M0, the intercept estimate was 0.92 logits. This revealed that the expected probability of a correct response for an “average” student on an “average” item was 0.72. The student variance indicated that, for a student with an ability of one standard deviation lower and a student with an ability of one standard deviation higher than the average ability, the expected probabilities of answering an item with average difficulty were 0.61 and 0.80, calculated from the antilogs of $(0.92 - \sqrt{0.22})$ and $(0.92 + \sqrt{0.22})$. As for item variance, for a student with an average ability, the probabilities of answering an item correctly with a difficulty of one standard deviation lower or one standard deviation higher than the average difficulty were 0.43 and 0.89, which were the antilogs of $(0.92 - \sqrt{1.45})$ and $(0.92 + \sqrt{1.45})$.

Table 4.1: Estimates of the Parameters for SGP

Parameter	Notation	M0	M1	M2
Fixed				
Intercept	γ_0	0.92*	0.84*	0.83*
Gender	γ_1		0.16*	0.18*
Random				
Student	$\sigma^2_{u_{0j}}$	0.22	0.22	0.22
Item	$\sigma^2_{u_{0i}}$	1.45	1.44	1.38
Gender*Item	$\sigma^2_{u_{1i}}$			0.03

Note. * of fixed coefficients denotes significance at $\alpha = 0.05$.

Model fit comparison indicated that M1 significantly fit better than M0 ($\chi^2_1 = 36.96, p < 0.001$), with decreased AIC and BIC, as shown in Table 4.2. Thus, there was a significant gender effect. M1 revealed that in SGP, the expected log-odds of correct response for women was 0.84. The overall mean performance difference for men over women was 0.16 logits. The corresponding probabilities of overall correct response for women and men were 0.70 and 0.73, respectively.

Table 4.2: Model Fit Results for SGP

Model	*df*	AIC	BIC	Log Lik	χ^2	χ^2*df*	p
M0	3	20052	20076	-10023			
M1	4	20017	20048	-10004	36.96	1	<0.001
M2	6	20017	20064	-10002	4.43	2	0.109

M2 did not significantly fit better than M1 ($\chi^2_2 = 4.43, p = 0.109$), with either similar or increased AIC and BIC. The result indicated that the gender by item interaction effects were not significant. In other words, there were no differential item effects between the gender groups. In SGP, with the same ability level, women and men had equal probability of getting an item correct. M1 was retained as the final model and no further analysis was conducted for SGP.

4.2 DEU: Gender

As shown in Table 4.3, for DEU, in the baseline model M0, the intercept estimate was 0.80 logits, indicating that the expected probability for an average student to answer correctly on an average item was 0.69. The student variance parameter reflected that for a student with an ability of one standard deviation lower and one standard deviation higher than the average ability, the expected probabilities of correctly answering an item with average difficulty were 0.52 and 0.82, which were the antilogs of $(0.80 - \sqrt{0.51})$ and $(0.80 + \sqrt{0.51})$. As for item variance, for a student with an average ability, the probabilities to correctly answer an item with a difficulty of one standard deviation lower and one standard deviation higher than the average difficulty were 0.42 and 0.87, which were the antilogs of $(0.80 - \sqrt{1.22})$ and $(0.80 + \sqrt{1.22})$.

After including gender as a predictor, M1 had significantly better model fit over M0 ($\chi_1^2 = 249.90, p < 0.001$) with decreased AIC and BIC (see Table 4.4). The person covariate equaled to 0 for women, and 1 for men. The significance of gender coefficient indicated that on average, men performed better than women by 0.42 logits. For women, the probability of correct response on an average item was 0.65, and for men, the corresponding probability was 0.74. The values were derived from the antilogs of (0.64) and (0.64 + 0.42) respectively.

After adding the gender by item interaction term, AIC favored M2 but BIC favored M1. However, the χ^2 between the two models was statistically significant ($\chi_2^2 = 11.85, p < 0.05$). Therefore, M2 was considered fit better than M1. The

Table 4.3: Estimates of the Parameters for DEU

Parameter	Notation	M0	M1	M2	M3	M4	M5
Fixed							
Intercept	γ_0	0.80*	0.64*	0.64*	0.84*	0.69*	0.70*
Gender	γ_1		0.42*	0.41*	0.43*	0.30*	0.31*
Format	γ_2				-0.82*		
Gender*Format	γ_3				-0.07		
OTL	γ_4					0.08*	0.08*
Random							
Student	$\sigma^2_{u_{0j}}$	0.51	0.47	0.47	0.47	0.38	0.39
Item	$\sigma^2_{u_{0i}}$	1.22	1.22	1.26	1.13	1.26	1.28
Gender*item	$\sigma^2_{u_{1i}}$			0.04	0.04	0.04	0.04
OTL*item	$\sigma^2_{u_{2i}}$						0.002

Note. * of fixed coefficients denotes significance at $\alpha = 0.05$.

Table 4.4: DEU: Model Fit Comparing M1 with M0, M2 with M1, M3 with M2

Model	*df*	AIC	BIC	Log Lik	χ^2	χ^2*df*	*p*
M0	3	36144	36170	-18069			
M1	4	35896	35930	-17944	249.9	1	<0.001
M2	6	35889	35939	-17938	11.85	2	<0.05
M3	8	35882	35949	-17933	10.58	2	<0.05

significance of model fit improvement of M2 revealed that there were gender DIF effects, meaning that the difference between men and women in performance varied over items ($\sigma^2_{u_{1i}} = 0.037$). In other words, with equal ability, men and women had different probabilities of correct response on some items. The negative correlation between the random effects u_{0i} and u_{1i} indicated that, conditional on the overall performance of both groups, the most difficult items were more difficult for the female group.

4.3 DEU: Gender and Item format

With an increase in BIC and a decrease in AIC, the χ^2_2 between the M3 and M2 was statistically significant ($\chi^2_2 = 10.58, p < 0.05$). Therefore, M3 fit significantly better than M2. In M3, controlling item formats, men outperformed women by 0.43 logits ($z = 6.41, p < 0.001$). The expected probabilities of correct responses for men and women were 0.70 and 0.78. Controlling for the gender effect, on average, students performed worse in CR items than MC items ($\gamma_2 = -0.82, z = -2.841, p < 0.05$). The differences between MC and CR items were the same for both men and women ($\gamma_3 = -0.07, z = -0.92, p = 0.36$), that is, the differential performance between item formats did not depend on gender. Thus, the interaction effects between gender and item format were not significant at an alpha level of 0.05. Furthermore, the magnitude of gender DIF ($\sigma^2_{u_{0i}}$) decreased only slightly (0.037 vs. 0.036). Therefore, item format was not considered a source of DIF for the DEU cohort. The analyses continued to examine OTL as a potential DIF source by excluding item format in subsequent models.

4.4 DEU: Gender and OTL

The inclusion of an OTL main effect and item by OTL interaction effects significantly improved the model fit for M4 over M2 ($\chi^2_1 = 1904.98, p < 0.001$) with decreases in AIC and BIC, and for M5 over M4 ($\chi^2_3 = 93.74, p < 0.001$) with decreases in AIC and BIC (see Table 4.5). However, M6 did not significantly improve over M5 ($\chi^2_1 = 0.91, p = 0.34$). Overall, women and men did not differ in OTL scores. Model M5 revealed that controlling for OTL, on average, men outperformed women by 0.31 logits. Also, among all students, there was a significant OTL effect on mean performance ($\gamma_4 = 0.08, z = 8.98, p < 0.001$). A one-unit increase in OTL was estimated to result in an increase of 0.08 logits in the mean performance. Furthermore, item difficulty varied by OTL. The significance of M5 indicated that the random interaction term between OTL and items was statistically significant ($\sigma^2_{u_2 i} = 0.002$), though not practically large. With different levels of OTL, people with the same ability level had different probabilities of giving correct responses on items. Nevertheless, gender DIF ($\sigma^2_{u_1 i}$) did not decrease (0.037 vs. 0.04). Therefore, OTL was not a source of gender DIF for DEU.

Table 4.5: DEU: Model Fit Comparing M4 with M2, M5 with M4, M6 with M5

Model	*df*	AIC	BIC	Log Lik	χ^2	$\chi^2 df$	*p*
M2	6	35889	35939	-17938			
M4	7	33986	34044	-16986	1904.98	1	<0.001
M5	10	33898	33981	-16939	93.74	3	<0.001
M6	11	33899	33991	-16939	0.91	1	0.34

4.5 USA: Gender

For the USA cohort, in the baseline model M0, the estimate of the intercept was 0.44 logits, thus the corresponding probability for an average student to give a correct response on an average item was 0.61. The student variance parameter reflected that, for a student with an ability of one standard deviation lower and one standard deviation higher than the average ability, the expected probabilities of giving an correct answer to an item with average difficulty were 0.45 and 0.75, as calculated from the antilogs of $(0.44 - \sqrt{0.43})$ and $(0.44 + \sqrt{0.43})$. The size of the item variance indicated that for a student with an average ability, the probabilities to answer an item correctly with a difficulty of one standard deviation lower or one standard deviation higher than the average difficulty were 0.36 and 0.81, which were the antilogs of $(0.44 - \sqrt{1.01})$ and $(0.44 + \sqrt{1.01})$.

Table 4.6: Estimates of the Parameters for USA

Parameter	Notation	M0	M1	M2	M3	M4	M5	M6
Fixed								
Intercept	γ_0	0.44*	0.28*	0.27*	0.50*	0.55*	0.55*	0.55*
Gender	γ_1		0.52*	0.52*	0.40*	0.25*	0.30*	0.30*
Format	γ_2				-0.98*	-0.98*	-1.00*	-1.00*
Gender*Format	γ_3				0.50*	0.09*	0.34*	0.34*
OTL	γ_4					0.52*	0.09*	0.09*
Gender*OTL	γ_5							-0.003
Random								
Student	$\sigma^2_{u_{0j}}$	0.43	0.38	0.38	0.38	0.27	0.28	0.28
Item	$\sigma^2_{u_{0i}}$	1.01	1.01	1.14	0.97	0.97	1.02	1.02
Gender*Item	$\sigma^2_{u_{1i}}$			0.17	0.12	0.13	0.09	0.09
OTL*Item	$\sigma^2_{u_{2i}}$						0.005	0.005

Note. * of fixed coefficients denotes significance at $\alpha = 0.05$.

Both M1 and M2 were found to have significantly better model fit over the previous models. As shown in Table 4.6 for M2, the gender effect, which represented the difference between women and men in mean performance, was 0.52 logits $(z = 6.11, p < 0.001)$. The log-odds of correct response for women was 0.27 and for men was 0.79, corresponding to the probabilities of correct response of 0.57 and 0.69. Additionally, the probability of correct response varied over students and especially over items ($\sigma^2_{u_{0j}} = 0.38$ and $\sigma^2_{u_{0i}} = 1.14$). The difference between women and men varied over items as well ($\sigma^2_{u_{1i}} = 0.168$). The improvement of M2 over M1 in model fit $(\chi^2_2 = 68.92, p < 0.001)$ (see Table 4.7) with reduced AIC and BIC indicated the presence of statistically significant item by gender interaction effects, where items tended to show gender DIF. The negative correlation of the random effects u_{0i} and u_{1i} indicated that, controlling the overall performance of all people, more difficult items were harder for the women.

Table 4.7: Model Fit Results for USA

Model	*df*	AIC	BIC	Log Lik	χ^2	$\chi^2 df$	*p*
M0	3	26627	26651	-13310			
M1	4	26528	26560	-13260	100.62	1	<0.001
M2	6	26463	26512	-13226	68.92	2	<0.001
M3	8	26447	26511	-13216	20.13	2	<0.001
M4	9	25640	25712	-12811	808.96	1	<0.001
M5	12	25458	25554	-12717	188.49	3	<0.001
M6	13	25460	25564	-12717	0.0218	1	0.88

4.6 USA: Gender and Item Format

The next step was to examine whether item format was a source of DIF. By incorporating item format as a covariate, the model fit was significantly improved for M3 over M2 ($\chi^2_2 = 20.13, p < 0.001$) with decreased AIC and BIC. As shown in Table 4.6 for M3, controlling for item format, the predicted mean performance for women was 0.50 logits. Men's expected mean performance was 0.40 logits higher. The corresponding probabilities were 0.62 and 0.71 for women and men. Meanwhile, students performed worse in CR items than MC items by 0.98 logits, controlling for gender effects. Moreover, the significance of the gender by item format interaction term ($\gamma_3 = 0.50, z = 3.97, p < 0.001$) indicated that the differential performance in item formats differed by gender, where the disparity between MC and CR items was larger for women than men regardless of the fact that men outperformed women on both item formats. More importantly, regarding the reduction of DIF effects, not only were the interaction effects between gender and item format significant, but the variance for gender DIF ($\sigma^2_{u_{1i}}$) was also reduced after the item covariate was introduced. The proportion of Gender DIF effects explained by item format was

0.29 (= 0.17 – 0.12/0.17), suggesting that item format contributed to DIF for the USA cohort. Because the magnitude of DIF effects was still relatively large, analysis continued with item format remaining in subsequent models.

4.7 USA: Gender, Item Format, and OTL

M4 and M5 additionally examined OTL impact and item by OTL interaction effects. Both models fit significantly better than the previous ones (See Table 4.7). However, M6 was not found to improve model fit over M5 ($\chi^2_1 = 0.02, p = 0.88$); AIC and BIC values both increased. The interaction effects between gender and OTL were not significant ($\gamma_5 = -0.003, z = -0.15, p = 0.88$). Thus, M5 was retained as the final model. As indicated in the last column (M5) in Table 4.6, the main effect of OTL was significant ($\gamma_4 = 0.09, z = 8.32, p < 0.001$), indicating that a one-point increase in OTL (one additional math topic being studied), corresponded to an increase of 0.09 logits, holding other variables constant. The improvement of M5 over M4 ($\chi^2_3 = 188.49, p < 0.001$), with reduced AIC and BIC, revealed that item by OTL interaction effects were significant. The random interaction term ($\sigma^2_{u_{2i}} = 0.005$) showed that item difficulty varied over different levels of OTL. DIF was explained partly by OTL; in addition to a statistically significant main effect for OTL, the effect of OTL varied in a statistically significant way over items; most importantly, the variance over items between the gender groups ($\sigma^2_{u_{1i}}$) was reduced from 0.12 to 0.09. The proportion of gender DIF that was explained by OTL was 0.25 (= 0.12 – 0.09/0.12). Even though the interaction term between gender and OTL was found not to be significant, OTL mediated the relationship between item difficulty and gender, since it reduced statistically significant amount of the gender DIF. Therefore, for USA, the conclusion was that both item format and OTL contributed to gender DIF.

Chapter 5

Discussion

This study was primarily designed to describe the relationships between math item difficulty, gender, item characteristics (specifically item format) and person characteristics (specifically OTL) with TEDS-M data from three countries. The study demonstrated how cross-classified models can be used to examine both item and person covariates as potential sources of uniform DIF. Results achieved in this study also provide indications on how pre-service teachers' math performance is influenced by gender, item format, and OTL.

Results from the final models of all three countries indicated that, overall, men tended to have higher mean math performance than women. The final SGP model M1 revealed that men were 0.16 logits higher than women in mean math performance. In DEU, the final model M5 indicated that men were 0.31 logits higher on average than women. Results of the final USA model M6 revealed the same pattern; men outperformed women by 0.30 logits. The predicted mean proportion correct for women and men were 0.63 and 0.70 correspondingly. Thus, gender discrepancies in math performance existed in this study. The gender effect in this study is consistent with the findings from other research, where male advantages in standardized math test were reported (e.g., Langenfeld, 1997; Liu & Wilson, 2009).

Overall math performance tended to be better on MC items than CR items for DEU and USA. In DEU, students had lower mean performance estimate by 0.82 logits in CR items than MC items. However, there was no interaction between gender and item format. The discrepancy between MC and CR items was not different for men and women. In USA, students performed better on MC items than CR items by 1.00 logit. In addition, the interaction term indicated that, even though women performed worse than men on both formats, the discrepancy between genders was larger on CR items. The results are consistent with the finding that men had an advantage in MC items, but contradicted the finding of female advantages on CR items (e.g., Bolger & Kellaghan, 1990; Beller & Gafni, 2000).

When OTL increased, math performance tended to improve in DEU and USA. A one-unit increase in OTL would result in an increase of 0.08 logits in DEU. Similarly, a one-unit increase in OTL resulted in an increase of 0.09 logits in the USA. In other words, with one more topic studied among the four topics which OTL measures, there were estimated increases in performance of 0.08 and 0.09 logits for DEU and USA. The relationship between OTL and mean performance is in tune with the positive correlations reported in Table 3.2 and in previous studies (e.g., Wang & Goldschmidt, 1999).

A DIF effect for gender was defined as the differential item effects of belonging to a specific group. Results showed no gender DIF effects for SGP. There was evidence of measurement invariance in the SGP test of MCK. Gender DIF was found in DEU and USA. DIF was examined in an omnibus test where no specific DIF items were identified. The results from DEU and USA both indicated that throughout the 76 items, there were some items that functioned differently between the gender groups. Conditionally on the overall performance of both groups, more difficult items favored men more than women. The results provided evidence of item bias due to gender in DEU and USA.

Item format was tested to determine whether it was associated with gender DIF. In DEU, the interaction term between gender and item format was not significant and no reduction of any gender DIF magnitude was found. As in USA, students' overall math performance on the two item formats depended on what gender group they belonged to. Even though men performed better on both formats, the gender differences on CR items were larger than on MC items. More importantly, the magnitude of gender DIF was significantly reduced with a proportion of variance explained of 0.29. Results from the literature (e.g., Taylor & Lee, 2012) have confirmed the finding that item format was associated with gender DIF.

This study also revealed that with different levels of OTL, items functioned differentially in DEU and USA. In DEU, however, OTL was not found to mediate the relationship between item difficulty and gender, provided by the evidence that the magnitude of gender DIF was not decreased after the inclusion of OTL. On the other hand, in USA, the inclusion of OTL resulted in a significant reduction of random gender effects over items. The conclusion was that OTL mediated the relationship between item difficulty and gender for some DIF items. Nevertheless, the interaction between gender and OTL did not improve model fit. The relationship between OTL and overall performance did not differ significantly by gender. The results from the USA cohort supported the findings of Albano and Rodriguez (2013) and Cheong (2006), where OTL was related to DIF.

This study is an extension of the original study of Albano and Rodriguez (2013), who examined differential math performance due to gender and OTL using hierarchical generalized linear modeling where person effects were viewed as random and item effects as fixed. To investigate DIF sources, the present study used cross-classified multilevel models, in which both item and person effects were treated as random. Item-level and person-level covariates were then estimated simultaneously. Besides OTL as the person-level covariate, this study further examined how item format could potentially explain variability in item difficulty and moderate the relationship between gender and item difficulty.

This study demonstrates the application of cross-classified multilevel models in educational research. The cross-classified multilevel model is a flexible tool to explain potential DIF sources related to item and person characteristics. This approach results in more economical models where DIF can be detected in an omnibus test. This approach can be helpful in creating and adapting appropriate measurement tools when constructing or translating items. Moreover, in terms of person characteristics, researchers can take construct-irrelevant variances such as OTL into account, and thus improve DIF detection and estimation. By doing so, item biases can be reduced and the validity of group comparisons then can be improved.

This study has some limitations. Inadequate sample sizes may have resulted in the lack of power in finding the significance of gender by OTL interaction effects. This problem also limits the possibility of incorporating more covariates that can potentially explain variability in item performance by gender. Also, the measure of OTL only accounted for self-reported exposure to certain math content. Other important factors in measuring OTL include hours in class, quality of teachers' feedback, and level of cognitive demand are not included in the instrument, which can be problematic. Future studies should seek larger sample sizes and consider more comprehensive measures of OTL.

Additionally, important item features such as item content domain and cognitive subdomain can be explored as potential sources to explain gender DIF. Past research has shown that men and women tend to adopt different strategies when responding to certain problem characteristics (e.g., Bolger & Kellaghan, 1990; DeMars, 2000). Studies have also indicated that content and cognitive skills required in items are related to gender DIF in math (e.g., Gierl, Bisanz, Bisanz, & Boughton, 2003; Harris & Carlton, 1993). Furthermore, a third level such as school level could be incorporated into the models to examine individual's social and/or psychological context effects (e.g., Entwisle, Alexander, & Olson, 1994; Van den Noortgate & De Boeck, 2005). Future work could examine other important covariates at the item and person levels while also incorporating additional levels of nesting.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of educational measurement*, *29*(1), 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*(1), 47–76.

Albano, A. D., & Rodriguez, M. C. (2013). Examining differential math performance by gender and opportunity to learn. *Educational and Psychological Measurement*, *73*(5), 836–856.

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

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). lme4: Linear mixed-effects models using Eigen and S4 [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=lme4` (R package version 1.1-8)

Becker, B. J. (1990). Item characteristics and gender differences on the sat-m for mathematically able youths. *American Educational Research Journal*, *27*(1), 65–87.

Beller, M., & Gafni, N. (1996). 1991 international assessment of educational progress in mathematics and sciences: The gender differences perspective. *Journal of Educational Psychology*, *88*(2), 365–377.

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

Bielinski, J., & Davison, M. L. (1998). Gender differences by item difficulty interactions in multiple-choice mathematics items. *American Educational Research Journal*, *35*(3), 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*(1), 51–77.

Blömeke, S., & Kaiser, G. (2012). Homogeneity or heterogeneity? Profiles of opportunities to learn in primary teacher education and their relationship to cultural context and outcomes. *ZDM*, *44*(3), 249–264.

Blömeke, S., Suhl, U., Kaiser, G., & Döhrmann, M. (2012). Family background, entry selectivity and opportunities to learn: What matters in primary teacher education? An international comparison of fifteen countries. *Teaching and Teacher Education*, *28*(1), 44–55.

Bolger, N., & Kellaghan, T. (1990). Method of measurement and gender differences in scholastic achievement. *Journal of Educational Measurement*, *27*(2), 165–174.

Boscardin, C. K., Aguirre-Munoz, Z., Stoker, G., Kim, J., Kim, M., & Lee, J. (2005). Relationship between opportunity to learn and student performance on English and algebra assessments. *Educational Assessment*, *10*(4), 307–332.

Burkes, L. L. (2009). *Identifying differential item functioning related to student socioeconomic status and investigating sources related to classroom opportunities to learn.* ERIC.

Carroll, J. (1963). A model of school learning. *The Teachers College Record*, *64*(8), 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*(1), 57–79.

Cochran-Smith, M., & Zeichner, K. M. (Eds.). (2005). *Studying teacher education: The report of the AERA panel on research and teacher education.* Mahwah, N.J.: Lawrence Erlbaum Associates.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, *42*(2), 133–148.

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

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

Eccles, J. S. (1994). Understanding women's educational and occupational choices. *Psychology of women quarterly*, *18*(4), 585–609.

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*(1), 103–127.

Entwisle, D. R., Alexander, K. L., & Olson, L. S. (1994). The gender gap in math: its possible origins in neighborhood effects. *American Sociological Review*, *59*(6), 822–838.

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

Gallagher, A., & Kaufman, J. (2005). *Gender differences in mathematics.* New York: Cambridge University Press.

Gamer, M., & Engelhard Jr, G. (1999). Gender differences in performance on multiple-choice and constructed response mathematics items. *Applied Measurement in Education*, *12*(1), 29–51.

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

Haggarty, L., & Pepin, B. (2002). An investigation of mathematics textbooks and their use in English, French and German classrooms: Who gets an opportunity to learn what? *British Educational Research Journal*, *28*(4), 567–590.

Hambleton, R. K., Swaminathan, H., & Rogers, H. (1991). *Fundamentals of item response theory (measurement methods for the social science).* Sage Publications, Inc.

Harris, A. M., & Carlton, S. T. (1993). Patterns of gender differences on mathematics items on the scholastic aptitude test. *Applied Measurement in Education*, *6*(2), 137–151.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. *Test validity*, 129–145.

Husen, 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*(2), 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*(5888), 494–495.

Jacobs, J. E., Davis-Kean, P., Bleeker, M., Eccles, J. S., & Malanchuk, O. (2005). "I can, but I don't want to": The impact of parents, interests, and activities on gender differences in math. In A. Gallagher & J. Kaufman (Eds.), *Gender differences in mathematics: An integrative psychological approach* (pp. 246–263). New York, NY, US: Cambridge University Press.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*(1), 79–93.

Kim, S.-H., Cohen, A. S., Alagoz, C., & Kim, S. (2007). DIF detection and effect size measures for polytomously scored items. *Journal of Educational Measurement*, *44*(2), 93–116.

Langenfeld, T. E. (1997). Test fairness: Internal and external investigations of gender bias in mathematics testing. *Educational Measurement: Issues and Practice*, *16*(1), 20–26.

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: a meta-analysis. *Psychological bulletin*, *136*(6), 1123–1135.

Liu, O. L., & Wilson, M. (2009). Gender differences in large-scale math assessments: PISA trend 2000 and 2003. *Applied Measurement in Education*, *22*(2), 164–184.

Longford, N. T., Holland, P. W., & Thayer, D. T. (1993). Stability of the MH D-DIF statistics across populations. In H. Wainer (Ed.), *Differential item functioning* (pp. 171–196). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

McDonnell, L. M. (1995). Opportunity to learn as a research concept and a policy instrument. *Educational evaluation and policy analysis*, *17*(3), 305–322.

Meece, J. L., Parsons, J. E., Kaczala, C. M., & Goff, S. B. (1982). Sex differences in math achievement: Toward a model of academic choice. *Psychological Bulletin*, *91*(2), 324–348.

Mendes-Barnett, S., & Ercikan, K. (2006). Examining sources of gender DIF in mathematics assessments using a confirmatory multidimensional model approach. *Applied Measurement in Education*, *19*(4), 289–304.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543.

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

Muthén, B. O., Kao, C.-F., & 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), 1–22.

Penner, A. M. (2003). International gender× item difficulty interactions in mathematics and science achievement tests. *Journal of Educational Psychology*, *95*(3), 650.

R Development Core Team. (2015). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org` (Version 3.2.1)

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen:The Danish Institute of Educational Research.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement*, *20*(4), 355–371.

Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement*, *66*(1), 63–84.

Schmidt, W. H., Cogan, L., & Houang, R. (2011). The role of opportunity to learn in teacher preparation: An international context. *Journal of Teacher Education*, *62*(2), 138–153.

Schmidt, W. H., McKnight, C. C., Valverde, G. A., Houang, R. T., & Wiley, D. E. (1997). *Many visions, many aims: A cross-national investigation of curricular intentions in school mathematics.* Dordrecht: Kluwer.

Shen, C., & Tam, H. (2008). The paradoxical relationship between student achievement and self-perception: A cross-national analysis based on three waves of timss data. *Educational Research and Evaluation*, *14*(1), 87–100.

Steffens, M. C., & Jelenec, P. (2011). Separating implicit gender stereotypes regarding math and language: Implicit ability stereotypes are self-serving for boys and men, but not for girls and women. *Sex Roles*, *64*(5-6), 324–335.

Su, R., Rounds, J., & Armstrong, P. I. (2009). Men and things, women and people: a meta-analysis of sex differences in interests. *Psychological bulletin*, *135*(6), 859–884.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement*, *27*(4), 361–370.

Tatto, M. T., Schwille, J., Senk, S., Ingvarson, L., Peck, R., & Rowley, G. (2008). Teacher education and development study in mathematics (TEDS-M): conceptual framework: policy, practice, and readiness to teach primary and secondary mathematics.

Taylor, C. S., & Lee, Y. (2012). Gender DIF in reading and mathematics tests with mixed item formats. *Applied Measurement in Education*, *25*(3), 246–280.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Travers, K. J., & Westbury, I. (1989). *The IEA study of mathematics I: Analysis of mathematics curricula, Vol.1.* Oxford, UK: Pergamon Press.

Van den Noortgate, W., & De Boeck, P. (2005). Assessing and explaining differential item functioning using logistic mixed models. *Journal of Educational and Behavioral Statistics*, *30*(4), 443–464.

Van den Noortgate, W., De Boeck, P., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics*, *28*(4), 369–386.

Wang, J., & Goldschmidt, P. (1999). Opportunity to learn, language proficiency, and immigrant status effects on mathematics achievement. *The Journal of Educational Research*, *93*(2), 101–111.

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

Wiseman, A. W. (2008). A culture of (in) equality?: A cross-national study of gender parity and gender segregation in national school systems. *Research in Comparative and International Education*, *3*(2), 179–201.

Wu, A. D., & Ercikan, K. (2006). Using multiple-variable matching to identify cultural sources of differential item functioning. *International Journal of Testing*, *6*(3), 287–300.