A Comparative Study of Mathematics Self-beliefs between Students in Shanghai-China and the US using a Multidimensional MIMIC Model with Ordered Categorical Items from PISA 2012

A Dissertation

Presented to

The Faculty of the Curry School of Education

University of Virginia

In Partial Fulfillment

of the Requirement for the Degree

Doctoral of Philosophy

by

Shi Zhu, B.A., M.A., Ph.D.

December, 2016

DEDICATION

This work is dedicated to my sweet and beautiful daughter.

Acknowledgements

First of all, I would like to express my sincere gratitude to my advisor Prof. Meyer for his patience, kindness, and continuous support and encouragement during my Ph.D. study. The completion of this dissertation would not be possible without his motivation and guidance.

Besides Prof. Meyer, I would like to thank Prof. Whaley for her constructive feedback and suggestions. Her continued care and support has contributed to the better understanding of some important theories used in the dissertation. I am also grateful to Prof. Konold and Prof. Ryoo for their insightful comments and thought-provoking questions. Their statistics courses built a solid foundation for my future work. I am lucky to get help from many kind and considerate people at Curry.

I am also indebted to Dr. Kathan Shukla whose precious support and invaluable friendship I will cherish all my life. Gratitude also goes to my friends Bo Liu, Prof. Qiushi Wang, Dijing Shi, Benjamin Du, Prof. Christina Xu and my close friends from my high school and Nanjing University, China.

I am especially grateful to my father Prof. Chaoqun Zhu and my mother Dezheng Sun for their selfless and unwavering love as much as they can give to help me overcome the inconceivable pain and hardship in life.

Special thanks goes to my sister’s family, my wife and my dearest daughter, who teaches me how to love and helps me truly appreciate the meaning of life.

APPROVAL OF THE DISSERTATION

This dissertation, A COMPARATIVE STUDY OF MATHEMATICS SELF-BELIEFS BETWEEN STUDENTS IN SHANGHAI-CHINA AND THE US USING A MULTIDIMENSIONAL MIMIC MODEL WITH ORDERED CATEGORICAL ITEMS FROM PISA 2012, has been approved by the Graduate Faculty of the Curry School of Education in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

J. Patrick Meyer, Chair

Diane Whaley, Committee Member

Tim Konold, Committee Member

Ji Hoon Ryoo, Committee Member

_________________Date

Table of Contents

List of Tables

List of Figures

List of Appendices

Abstract

This study utilized MIMIC method to deal with multiple covariates, multiple dimensions and ordered categorical variables with threshold structures in both categorical confirmatory factor analysis (CCFA) and multidimensional graded response model (GRM) to study differential item functioning (DIF) among mathematics self-beliefs items across Shanghai-China and the Untied States from PISA 2012. The MIMIC approach with mediators was also applied in the study, which helped to detect variables that could account for meaningful partial or complete DIF effects. CFA indicated that the three-factor structure worked for the data. It also indicated that the self-belief questionnaire in PISA 2012 measured the same constructs among Shanghai-China and the US. Both robust weighted least square (WLSMV) estimator and robust maximum likelihood (MLR) estimator were used in the parameter estimation to identify items with DIF and quantify the DIF effect size. It was found in the study with MIMIC method that both in-school mathematics class periods and out-of-school study hours had partial effects on most items with meaningful DIF effects.

Key words: MIMIC; PISA; measurement invariance; DIF

Chapter 1 Introduction

According to PISA 2012, mathematics self-beliefs involve three dimensions: mathematics self-efficacy, mathematics self-concept and mathematics anxiety (OECD, 2013b). Measurement of these dimensions is assumed to be invariant among the various countries and economies participating in PISA, but this assumption may not be tenable given that some self-beliefs are susceptible to social norms and cultural contexts (Bong & Skaalvik, 2003). The primary purpose of the study is to assess the differential item functioning (DIF) of items in the mathematics self-beliefs questionnaire of the Program for International Student Assessment (PISA) 2012 across students from Shanghai－China and the United States. By comparing examinees from different countries, this study seeks to identify mathematics self-belief items that are sensitive to cultural differences. It also aims to use examinee characteristics to explain why some items exhibit DIF.

Similar studies of cross-cultural DIF have evaluated career adaptability (Savickas & Porfeli, 2012), math and science motivation (Marsh, et al., 2013), social trust (Freitag & Bauer, 2013), physical activity enjoyment (Jekauc, Voelkle, Wagner, Mewes, & Woll, 2012), bilingual students' school motivation (Ganotice, Bernardo, & King, 2012), entrepreneurial orientation (Runyan, Ge, Dong, & Swinney, 2011), and the relationship between teachers and students (Koomen, Verschueren, Schooten, Jak, & Pianta, 2012). Their studies indicated that scales measuring certain constructs are very often culturally sensitive. For example, in the study by Marsh et al. (2013), it was found that Arab students got higher scores in values, self-concept, positive affect and the willingness for further study in both mathematics and science, but they got poorer scores in mathematics and science performance than the students from Anglo countries. Therefore, context-sensitivity needs to be taken into consideration when cross-cultural DIF is assessed.

The term DIF is used throughout this study, but related terms are encountered in the literature. DIF occurs when examinees from two different groups have a different expected item response even though they have the same standing on the construct. It indicates that the item may be easier or more discriminating for members of one group than it is for comparable members of another group. That is, it suggests people from different groups are measured differently by the item even though they have the same latent trait value. The term DIF is common in the item response theory literature (Kim & Yoon, 2011), but in the factor analysis literature it is more commonly referred to as a lack of measurement invariance. Factor analytic studies often aim to establish measurement invariance before comparing groups on latent means. Measurement invariance holds when members from different groups with "the same standing on the construct" have the same probability of getting the same observed score (Schmitt & Kuljanin, 2008). That is, measurement invariance is the absence of DIF.

The multidimensional nature of mathematics self-beliefs presents a challenge in studying DIF because the dimensions are distinct but correlated (Lee, 2009) and best represented by a single model with multiple dimensions. Traditional IRT DIF detection methods assume a unidimensional construct, and would require separate analysis of each mathematics self-belief dimension. To overcome this limitation, this study uses the multiple indicators, multiple causes (MIMIC) model to study DIF in polytomous items for a multidimensional construct. The MIMIC model is a flexible approach to DIF analysis that is applicable to unidimensional or multidimensional data, multiple grouping variables, and continuous or ordinal variables (Cheng, Shao & Lathrop, 2016; Wang & Shih, 2010).

If the data is categorical, MIMIC model can be utilized to fit both an IRT model and a CFA model (Woods, 2009b), because both an ordered-categorical variable CFA (CCFA) and an IRT model can incorporate thresholds (Kim & Yoon, 2011). CCFA with threshold parameters can be compared with graded response model in IRT when they are used to identify measurement invariance or DIF due to similar techniques of analysis (Kim & Yoon, 2011). CCFA also shares a lot in common with MIRT. The methodology of nonlinear factor analysis is in essence the same as that used in MIRT (McDonald, 1967).

Thanks to the similarities between CFA and MIRT, the MIMIC model can be treated as a measurement invariance or DIF detection approach from a CFA or a MIRT perspective. That is, the MIMIC method of DIF detection of ordered-categorical items with multiple factors can be regarded as either a multidimensional graded response model or a CCFA model with covariates. The only real difference between these two methods is the parameter estimation method. Weighted least square (WLS) method is widely used in parameter estimation of CFA with categorical items (Wirth & Edwards, 2007) while full information maximum likelihood (FIML) or marginal maximum likelihood (MML) is usually used in IRT (Forero & Maydeu-Olivares, 2009). Each method has advantages and limitations. Computations in FIML are cumbersome especially in MIRT models with numerous dimensions. WLS is much more efficient (Forero & Maydeu-Olivares, 2009), but this efficiency comes with a cost of less information (Forero & Maydeu-Olivares, 2009).

The present study will use a multidimensional MIMIC modeling approach to study DIF among mathematics self-beliefs items from PISA 2012.

The research questions are as follows.

1. Do self-beliefs structure patterns of Shanghai-China and the US students follow the same pattern?

2. Are the items in the math self-beliefs questionnaire in PISA 2012 measuring the same constructs in both the US and Shanghai groups?

3. Does any DIF exist in any of these items in the dimensions of mathematics self-efficacy, mathematics self-concept and mathematics anxiety?

4. If DIF exists in these items, which covariates can be used to account for the DIF?

Significance of the Study

PISA has been the subject of numerous studies in recent years. Ferla, Valcke and Cai (2009) studied academic self-efficacy and academic self-concept among Belgian PISA examinees, and Lee (2009) studied mathematics self-efficacy, mathematics self-concept, mathematics anxiety, yet these studies focused on the understanding of structural relationship of the factors and were not related with measurement invariance studies. Even though some researchers (e.g. Grisay & Monseur, 2007; Kankaras & Moors, 2013) implemented measurement invariance studies, they focused on testing items in mathematics, science and reading sections. That is, they focused on the cognitive questions, and not the self-beliefs questions.

The current research focuses on the detection of DIF across Shanghai-China and the US in the mathematics self-belief questionnaire in PISA 2012. Unlike most measurement invariance studies in PISA, which mainly used multiple group confirmatory factor analysis (CFA) or IRT methods, this study utilizes MIMIC method to deal with multiple covariates, multiple dimensions and categorical variables with threshold structures in both CCFA and Multidimensional GRM.

Chapter 2 Literature Review

Self-Beliefs

Self-beliefs are “beliefs and perceptions about self” and they are related to an individual’s past experience and history of reinforcement (Bong & Skaalvik, 2003). Self-beliefs include individuals’ possessed attributes, their presumed roles, their beliefs regarding their capabilities, their comparison with others and their views in terms of others’ judgment of themselves (Bong& Skaalvik, 2003). Even though self-beliefs are subjective, they play an important part in an individual’s personal development (Bong & Skaalvik, 2003). Self-beliefs have been widely studied around the world especially in the United States.

Introduction to Self-efficacy

According to Bandura (1977), self-efficacy is the belief of an individual that “one can successfully execute the behavior required to produce the outcomes”, and it can have an impact on people’s feelings, thoughts, motivations and behaviors (Bandura, 1993). An individual’s judgment of his/her efficacy is not related to his/her real skills or capabilities possessed, but to “whatever skills and abilities” they think they possess (Bong & Skaalvik, 2003). In the Program for International Student Assessment (PISA), mathematics self-efficacy was defined as students’ beliefs in terms of their capabilities to effectively deal with mathematics problems and get rid of difficulties (OECD, 2013b). If students do not believe they are capable of fulfilling specific tasks, they will not put the necessary effort required to accomplish the tasks (OECD, 2013b).

Self-efficacy and Academic Performance. In recent years, some researchers have focused on the link between self-efficacy and academic achievements of college and university students. Komarraju and Nadler (2013) found that self-efficacy has a predictive power of academic achievement. Undergraduate students with high levels of self-efficacy are able to use a variety of "metacognitive strategies and resources" which are of vital importance for higher GPA. Gore (2006) also studied the predictive power of academic self-efficacy and college success as well as students' persistence. Compared with undergraduates' academic self-efficacy beliefs at the beginning of the first semester, the beliefs at the end of the first semester could predict students' college success much better. Richardson, Abraham and Bond (2012) used intelligence test results, scores of SAT and ACT, high school GPA, A level points, gender, age, SES and "42 non-intellective constructs" to analyze their relationship with students' university GPA. Their study indicated that both performance and academic efficacy, grade goal as well as effort regulation could predict university GPA, and performance efficacy had the strongest correlate with university GPA.

MacPhee, Farro and Canetto (2013) studied the relationship between academic self-efficacy and university academic performance among underrepresented students (e.g. minority students, low-SES students, female students) in STEM majors. They found that if students fall into minority and low-SES categories, they "had significantly lower (i) academic self-efficacy, (ii) test taking skills; and (iii) academic performance as indicated by GRE and critical thinking scores as wells as cumulative GPA…" (p.365). Thus, individual characteristics such as SES may affect the relationship between self-efficacy and academic performance.

Although much academic self-efficacy research focuses on American students, more and more researchers have shown interest in students from other areas in the world. Zuffianò et al. (2013) studied the influence of “self-efficacy beliefs in self-regulated learning (SESRL)” on academic achievement of students in a junior high school in Italy and found that SESRL accounted for about 2% of the variance in academic achievement. Lee, Lee and Bong (2014) analyzed the predictive power of academic self-efficacy and testing interest to academic achievement and self-regulation among 500 Korean middle school students and confirmed the association between self-efficacy and both academic achievement and grade goals. Chen and Zimmerman (2007) made a comparison between middle school students from the United States and Taiwan regarding their self-efficacy and mathematics achievement. They found that self-efficacy decreased when the mathematics items became more difficult. They also found that American students had higher self-efficacy when dealing with easy mathematics items; however, their self-efficacy declined more quickly than their peers from Taiwan when encountering moderate and difficult items. These results suggest that cultural characteristics influence self-efficacy.

Introduction to Self-concept

Self-concept is another important dimension of self-beliefs. It refers to “a composite view of oneself” (Bong & Skaalvik, 2003), and it “reflects both cognitive and affective responses” (Bong & Clark, 1999). According to Shavelson, Hubner and Stanton (1976), self-concept is how a person perceives himself/herself, and it is “organized, multifaceted, hierarchical, stable, developmental, evaluative, and differentiable" (p.411). The cognitive aspect of self-concept is composed of "awareness and understanding of the self and its attributes" (Bong & Clark, 1999).

Marsh and Craven (2006) indicated that self-concept is domain specific, while self-esteem is a global measure, which is not associated with different academic outcomes.

Self-concept and Academic Performance. Valentine, DuBois and Cooper (2004) implemented a meta-analysis to analyze the connection between self-concept and academic achievement. The findings of the study indicated that academic-domain-specific self-concept has a more powerful impact on academic achievement than global beliefs. When the academic self-concept measures and academic achievement fall into the same academic domain, the relation is more evident. Huang (2011) also did a meta-analysis to study the relationship between self-concept and academic achievement based upon 32 previous studies. The findings indicated that there are "positive effects of self-concept on academic achievement and of academic achievement on self-concept" (p.524). The author recommended domain-specific examination of academic achievement and self-concept, otherwise the strength between the two may be underestimated.

Several researchers have carried out transnational investigations regarding the association between self-concept and academic achievements in Trends in International Mathematics and Science Study (TIMSS) and PISA. Wilkins (2004) used TIMSS data to study the relationship between self-concepts of mathematics and science and academic achievements in these subjects. It was found in the study that students from higher mathematics achieving countries and economies did not necessarily have higher self-concept than those from countries and economies that performed worse in mathematics. Even though students from Japan, South Korea and Hong Kong achieved better in mathematics, their self-reported mathematics self-concept was the lowest. Yoshino (2012) used the data from TIMSS 2007 to study the association of mathematics self-concept and mathematics achievement of 8th grade students, and found a positive relationship between mathematics self-concept and mathematics achievement for both Japanese and American students. Yet American students had higher mathematics self-concept than their Japanese peers. This study supported the perspective that cultural factors may have an impact on the connection between mathematics self-concept and students' mathematics achievement.

Marsh (1990) used a reciprocal effects model (REM) to study the relationship between academic achievement and academic self-concept. His study indicated that prior self-concept has a significant impact on students' subsequent academic performance, which in turn affects subsequent self-concept. REMs of the relationship between self-concept and students' achievement have been generalized to different cultures based upon the studies of Canadian students (Guay, Marsh, & Boivin, 2003), students from mainland China (Yeung & Lee, 1999) and Hong Kong students (Marsh, Hau, & Kong, 2002). Their studies all supported the causality between prior academic self-concept and subsequent academic achievement (e.g. mathematics scores).

Introduction to Test Anxiety

Anxiety exists in all human beings and it is one of the most fundamental feelings or emotions that can be found in individuals (Huberty, 2012). It reflects "concerns about subjective, anticipatory events" (Huberty, 2012, p.29). Spielberger and Rickman (1990) defined anxiety as a state that is combined with both "experiential qualities and physiological changes" (p.69). It manifests in physiological, behavioral and cognitive symptoms (Huberty, 2012). Physiological symptoms include headache, increased heart rate, stomachache and the tension of muscles. Cognitive symptoms include memory problems, lack of attention and concentration. Behavioral symptoms can be "withdrawal or lack of participation" (Huberty, 2012).

There are a variety of anxiety forms (e.g. test anxiety, social anxiety, statistics anxiety and mathematics anxiety), which are universal across cultures. Test anxiety is one of the most studied anxiety forms. It refers to "the set of cognitive, affective, and behavioral reactions that accompany concern over possible negative consequences" on tests or exams (Zeidner, 1998, pp.25-26). There is an inverted U curve relationship between anxiety and academic performance, with some researchers indicating that a moderate level of anxiety benefits performance (Ma, 1999). Yet Students with high levels of test anxiety tend to treat evaluative situations as threats. Whether a situation is regarded as threatening is dependent on several variables that may include past experience, situational requirements, awareness of consequences, different aptitude, self-efficacy and even trait anxiety (Zeidner, 1998). Test anxiety can be caused by students' ability level, gender, the level of the grade, ethnicity, and the environment of the school (Hembree, 1988).

Test Anxiety can be divided into two categories: trait anxiety and state anxiety (Huberty, 2012; Zeidner, 1998). Trait anxiety refers to the tendency to be anxious in situations related to evaluations (Zeidner, 1998) and it is almost impossible to get rid of trait anxiety completely; however, it can be reduced (Huberty, 2012). State anxiety refers to “fluctuating states experienced in a test situation” (Zeidner, 1998, p.18) and it is “the tendency to experience high anxiety in specific situations” (Huberty, 2012, p.31). Even though students with the same academic aptitude may have different levels of test anxiety in different cultures, test anxiety can be regarded as a universal construct that exists in all cultures (Cassady & Johnson, 2002).

Test Anxiety and Academic Performance. Several researchers have found a negative relationship between test anxiety and academic performance. According to Chapell, Blanding, Silverstein, Takahashi, Newman, Gubi, and McCann (2005), both male and female undergraduate students with low levels of test anxiety had higher cumulative GPAs than those with high levels of anxiety. In terms of graduate students, it was found that female students of high levels of test anxiety got lower GPAs than female students with low levels of test anxiety, yet no significant differences could be found among male graduate students. Similar results were also obtained in the study implemented by Cassady and Johnson (2002). It was found in their study that students with various cognitive test anxiety levels performed differently in the SAT and course examinations. Students with high levels of cognitive test anxiety showed an obvious disadvantage in the SAT. Students also performed significantly differently in course examinations. According to the researchers, “cognitive test anxiety accounts for approximately 7 to 8% of the variance in student performance on actual course examinations”, which was quite significant. Cassady (2004) pointed out that high levels of cognitive anxiety had strong negative influence on test preparation, test performance and test reflection, and “cognitive test anxiety was the only measure” that had a predictive power of students’ future test performance (Cassady, 2004).

The relationship between test anxiety and academic performance was also found among students in some other countries. For example, Rana and Mahmood (2010) studied the relationship between test anxiety and students’ academic performance in a university in Pakistan. They arrived at the same conclusion as many other researchers that high levels of test anxiety are associated with low academic performance. They found that “a cognitive factor (worry) contributes more in test anxiety than affective factors (emotional).”

However, test anxiety is not always a bad thing. Cassady and Johnson (2002) pointed that if the levels of physiological reaction caused by test anxiety is moderate, then students may perform better than those with low or high levels of test anxiety.

Factors Affecting Test Anxiety. Test anxiety is affected by student characteristics such as gender and socioeconomic status. Cassady and Johnson (2002) found in their study that female undergraduate students showed higher levels of test anxiety than male “in both emotionality and cognitive test anxiety” (p. 290); however, there was no difference in terms of exam performance between male and female students. These high levels of anxiety were caused by female students’ personal doubt in terms of their own capabilities to deal with tests and exams when faced with evaluative situations.

Socioeconomic status is another factor that can have an impact on test anxiety. Zeidner and Safir (1989) studied the effect of both gender and SES on levels of test anxiety of 416 junior high school students in northern Israel. Their findings were consistent with the previous studies that “female and lower SES students generally fear negative evaluation more than do males and upper SES students and that they are more likely to devalue their cognitive performance” (Zeidner & Safir, 1989, p.183).

Very little research on test anxiety considered cultural factors that can have an impact on the relationship between anxiety and test scores, which cannot be neglected. Test anxiety is usually regarded as an individual trait and contextual factors are usually not considered in test anxiety research (Bodas & Ollendick, 2005). However, contextual variables such as “socialization practices and value systems” can have important impact on children’s test anxiety (Bodas & Ollendick, 2005).

Mathematics Anxiety. Mathematics anxiety is closely related to test anxiety. Mathematics anxiety is a “genuine fear of mathematics” (Beilock & Maloney, 2015). If people are nervous about mathematics, they may be unwilling to engage in mathematics tasks and will not take any profession that may have something to do with mathematics, which in turn constrains their occupation options (Chipman, Krantz, & Silver, 1992). According to Beilock and Willingham (2014), 80% of community college students and 25% of college students experienced mathematics anxiety of different levels. Mathematics anxiety can interfere in the procedure of cognitive activities with the working memory activity compromised (Ashcraft, 2002). People with certain levels of mathematics anxiety may have worries when facing mathematics problems, which in turn affects “working memory resources” (Ashcraft & Kirk, 2001). Since working memory responsible for storing and processing information has limited space (Engle, 2002), students who have high levels of mathematics anxiety actually have to deal with two things simultaneously when facing mathematics problems: solving mathematics problems and coping with their worries (Beilock & Maloney, 2015).

Neuroscientific evidence has also supported this finding. Young, Wu, and Menon (2012) found that when students with high levels of mathematics anxiety dealt with mathematics problems, they had more activity in the brain regions that are related to negative feelings. Yet they showed less activity in areas that are associated with working memory.

Many studies both in the United States and other countries (e.g. Aschcraft & Kirk, 2001; Jameson, 2013) have indicated that there is a relationship between students' mathematics anxiety and mathematics performance. Ma (1999) implemented a meta-analysis based upon 26 studies in terms of the relationship between mathematics anxiety and mathematics achievement. The result of the analysis supported the negative correlation between mathematics anxiety and mathematics achievement. Yet the instruments used in different studies affect the strength of the relationship. According to a study by Ho, Senturk, Lam, Zimmer, Hong, Okamota, Chiu, Nakazawa and Wang (2000) based upon 6th-grade student samples from US, mainland China, and Taiwan, affective dimension of mathematics anxiety had a strong association with mathematics achievement across the three samples while there was little connection between cognitive dimension and the achievement among students from US and mainland China while the cognitive dimension had a positive predictive power of student achievement in Taiwan. However, Krinzinger, Kaufman, and Willmes (2009) found no causal relationship between high levels of mathematics anxiety and poor calculation aptitude for children aged between grade one to grade three.

Mathematics anxiety can be caused by previous mathematics experience that was not positive or by negative parental influence, teachers' bad teaching strategies, or some negative classroom factors (Ma, 1999). Adults whose mathematics foundation is poor tend to be more mathematics anxious than their peers when dealing with more difficult mathematics problems (Beilock & Maloney, 2015). It is the same with school-aged children. If children start school with difficulty in mathematics, children will develop mathematics anxiety and may avoid further mathematics learning (Maloney & Beilock, 2012). "Exposure to negative attitudes about mathematics" can also increase people's mathematics anxiety (Beilock & Maloney, 2015). Parents whose mathematics anxiety level is high may transmit their high anxiety to their children while they help their children with their mathematics. Teachers also have impact on students mathematics anxiety level (Beilock & Maloney, 2015). According to Beilock, Gunderson, Ramirez, and Levine (2010), female elementary teachers with high levels of mathematics anxiety negatively affect their female students. As a result, their female students tend to learn less mathematics and are more willing to accept the stereotype that male students are superior to female students in terms of mathematics performance.

The Relationship between Mathematics Anxiety, Mathematics Self-efficacy and Mathematics Self-concept. According to Bong and Skaalvik (2003), academic self-concept is dependent upon social comparison and it is peer-referenced while self-efficacy is goal-referenced and does not demand peer comparison. However, Schunk and Hanson (1985) pointed out that there was evidence indicating social comparison is involved in self-efficacy judgments. Both vicarious experience (e.g. observing others) and verbal persuasions (e.g. social persuasion) are two of the four sources of information for self-efficacy beliefs; therefore, they involve social comparisons that may have powerful impact on self-efficacy beliefs (Parajes, 2003).

Even though both academic self-concept and self-efficacy are domain specific, the latter is dependent upon context. Ferla, Valcke, and Cai (2009) studied the structural relationship between academic self-concept and academic self-efficacy and found that academic self-concept and self-efficacy are two distinct constructs. Lee (2009) analyzed the mathematics self-belief constructs based upon the survey data in PISA 2003 and confirmed that mathematics self-efficacy, mathematics self-concept and mathematics anxiety are three distinct but related constructs. The study rendered support to “the universality of the self-constructs” and all the constructs were associated with math performance; however, East Asian countries and economies (Hong Kong, Japan, and South Korea) tended to score lower than their peers in other countries and economies even though they obtained higher grades than their peers in mathematics achievement.

Zeidner (1998) assumed that the relationship between mathematics anxiety and mathematics self-concept is reciprocal. Ahmed, Minnaet, Kuyper, and Werf (2012) studied the relationship between mathematics self-concept and mathematics anxiety among 7th grade students in the Netherlands and their study provided strong support for the reciprocal effects model in terms of mathematics anxiety and mathematics self-concept. They stated that “the effect of self-concept on anxiety is twice the effect of anxiety on self-concept”（Ahmed, Minnaet, Kuyper, and Werf, 2012, p.387).

There are four sources of self-efficacy beliefs: mastery experience, vicarious experience, persuasion and anxiety (Pajares, 2003). However, anxiety can also be the product of self-efficacy (Bandura, 1997).

McMullam, Jones, and Lea (2012) examined the relationship between mathematics anxiety, mathematics drug calculation self-efficacy, numerical ability, and drug calculation. They found that numerical ability had the most predictive power of drug calculation performance, followed by mathematics anxiety. Moreover, mathematics self-efficacy and drug calculation self-efficacy are also positively associated with the performance.

Jameson and Fusco (2014) tried to find out whether adult learners in the undergraduate population had the lower mathematics anxiety, mathematics self-efficacy, and mathematics self-concept than the traditional students. Their study found that adult learners had significantly higher mathematics self-efficacy than traditional students but there was no difference regarding mathematics self-concept and mathematics anxiety.

DIF Detection with MIMIC Method

Multiple-indicators, multiple-causes (MIMIC) model

The MIMIC model is a form of structural equation model (SEM; Woods, 2009b) and it is a special form of Confirmatory factor analysis (CFA). The ordinary CFA model can be represented in the following way (Kim & Yoon, 2011):

$$X_{ij} = \tau_j + \lambda_j \xi_i + \delta_{ij}$$

where X_{ij} indicates a continuous observable variable of person i on item j, τ_j is the intercept of item j, λ_j represents a factor loading on item j, ξ_i is a common factor score, and δ_{ij} is the measurement residual. It is assumed that in CFA measurement errors are not correlated with each other and are uncorrelated with latent variables (Raju, Laffitte, & Byrne, 2002); therefore, the population variance-covariance matrix can be expressed as

$$\Sigma = \Lambda_x \Phi \Lambda'_x + \phi_\delta$$

where Φ is a variance-covariance matrix of the common factor and ϕ_δ is the variance-covariance matrix of measurement residuals or unique factors.

A special form in CFA is the ordered-categorical model, in which the observed score is dichotomous or polytomous, and X'_{ij} is regarded as the latent variable of the observed one (Millsap & Yun-Tein, 2004). “The variance covariance matrix of the latent response variates is termed tetrachoric variance-covariance for dichotomous variables and polychoric variance-covariance for polytomous variables” (Kim, 2011, p.17). Thresholds are taken into account in ordered-categorical CFA and they can be shown as follows:

$$X_{ij} = c \text{ if } v_{ic} \leq X'_{ij} < v_{j(c+1)}$$

where X_{ij} is observed score; X'_{ij} is the unobserved latent variable of $X_{ij.} c = 0, 1, \ldots, C\text{-}1$; C is the largest possible number of categorical responses; v_{ic} is the threshold parameter (Wirth & Edwards, 2007). Therefore, if the total number of item thresholds is C, then the number of item categories is $C\text{-}1$ (Kim & Yoon, 2011).

In a MIMIC model, there is one or more than one observed variable that can be used as casual indicator to predict a latent variable (Jöreskog & Goldberger, 1975). MIMIC model can be used in both unidimensional and multidimensional measurement or DIF detection due to its flexibility (Wang & Shih, 2010), and the incorporation of multiple grouping variables (Cheng, Shao, & Lathrop, 2016). A MIMIC model is illustrated below in figure 1.

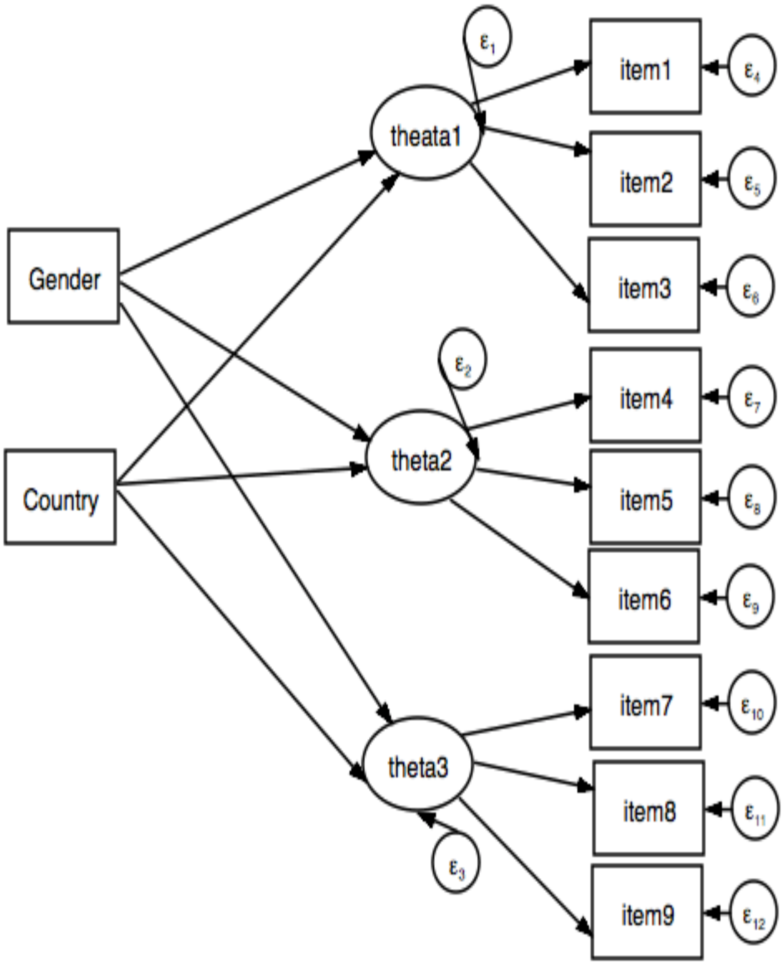


Figure 1. A MIMIC Model with Three Factors and Two Covariates

If the data is categorical, MIMIC model can be utilized to fit both an IRT model, and a CFA model using either a polychoric or a tetrachoric correlation matrix (Woods, 2009b). Many researchers have advocated the use of MIMIC method to test measurement invariance or for DIF detection due to its flexibility (e.g. Muthén, 1989; Woods, 2009b; Jin, Myers, Ahn, & Penfiled, 2012).

Measurement Invariance

Due to the increasing use of international tests like PISA and TIMSS, more and more researchers have shown interest in cross-country or cross-culture studies regarding students' academic achievement, learning motivations and self-beliefs. However, simple comparisons of mean scores is not meaningful unless there can be sufficient psychometric evidences to indicate that in each group there is the same "theoretical construct" and there is no difference in terms of the associations between items and the latent trait or latent construct (Vandenberg & Lance, 2000). Therefore, measurement invariance is the precondition of multi-group comparison.

According to Yoon and Millsap (2007), measurement invariance can be defined in the following way:

$$P(X|W,G) = P(X|W)$$

where X is an observable variable vector, W refers to a latent trait vector, G refers to the membership of a group. This equation implies that if the relationship between observable variables and latent variables is independent of the grouping variable, there is measurement invariance (Yoon & Millsap, 2007). Or put it in another way, if members from different groups with the same latent trait have the same probability of getting the same score on one construct measured, then there is measurement invariance (Schmitt & Kuljanin, 2008).

If measurement invariance does not hold, measurement bias arises (He & van de Vijver, 2012). There are three types of bias, which are respectively "construct bias, method bias and item bias"(Van de Vijver & Tanzer, 2004; He & van de Vijver, 2012). According to He and van de Vijver (2012), construct bias arises when the construct studied is not the same in different groups or cultures, and method Bias can be divided into sampling bias, instrument bias, response styles, and administrative bias.

Measurement invariance has been evaluated in a variety of cross-cultural or cross-group studies such as career interest (Savickas & Porfeli, 2012), math and science motivation (Marsh, et al., 2013), social trust (Freitag & Bauer, 2013), physical activity enjoyment (Jekauc, Voelkle, Wagner, Mewes, & Woll, 2012), bilingual students' school motivation (Ganotice, Bernardo, & King, 2012), entrepreneurial orientation (Runyan, Ge, Dong, & Swinney, 2011), and the relationship between teachers and students (Koomen, Verschueren, Schooten, Jak, & Pianta, 2012).

Even though full measurement invariance cannot be established, "a subset of the subgroup parameters" can be tested. And this test is called partial measurement invariance test (Schimitt & Kuljanin, 2008).

Recommended Practice of measurement Invariance in CFA

Vandernberg and Lance (2000), and Schmitt and Kuljanin (2008) recommended the following steps to test measurement invariance in CFA. The first step is "an omnibus test of the equality" of the variance-covariance matrix across the groups (Vandernberg & Lance, 2000). If there is a full invariance, no further invariance tests are needed (Schmitt & Kuljanin, 2008).

The second step is configural invariance test, which requires similar factor patterns across groups. The values of the parameters can vary except for a few constraints (Schmitt & Kuljanin, 2008). This step serves as the baseline so that the subsequent tests can be meaningful (Vandenberg & Lance, 2000). Configural invariance can be violated due to problems with the translation, data collection (Cheung & Rensvold, 2002, p.237), and culture-dependent understanding of the studied construct (Tayeb, 1994).

The third step is called metric equivalence, which tests whether factor loadings of items on each factor are invariant across groups (Schmitt & Kuljanin, 2008). This is also called "a test of a strong factorial invariance null hypothesis that factor loadings for like items are invariant across groups" (Vandenberg & Lance, 2000, p.12). Once metric equivalence is established, the test of factor variances and covariances can be implemented (Schmitt & Kuljannin, 2008).

The next step is called scalar invariance test, which involves the invariance of intercepts of "of the regression equations of the observed variables on the latent factors" across groups (Schmitt & Kuljanin, 2008, p.212).

Step five tests the equivalence of "items' unique variances" (Vandenberg & Lance, 2000) or residual variances. This test can be implemented "only if (at least partial) metric and scalar invariance has been established first" (Vandenberg & Lance, 2000, p.13). The test of the invariance of factor variances is in this step. In this test, the variances of the latent factors are constrained to be equal across groups (Schmitt & Kuljanin, 2008).

The last step tests the equivalence of factor means (Vandenberg & Lance, 2000), which is usually of great interest to many researchers (Schmitt & Kuljanin, 2008). "Comparing latent means between groups has merits over comparing observed group means because measurement errors are taken into account in latent factor means" (Kim, 2011, p.25).

MIMIC method mainly focuses on the invariance of intercepts and factor means, which means either Shanghai-China or US students were biased in a certain direction on a specific latent variable.

Goodness-of-fit indexes (GFI) are of vital importance to evaluate measurement invariance. Twenty GFIs were evaluated by Cheung and Rensvold (2002) in their study and they found that "Δcomparative fit index, ΔGamma hat, and ΔMcDonald's Noncentrality Index" are better than others, because these indexes are not dependent upon complexity and sample size.

Measurement invariance can also be detected in IRT. From IRT perspective, lack of measurement invariance is treated as DIF (Kim & Yoon, 2011). There are two types of DIF: uniform DIF and nonuniform DIF. If the probability of a certain response is lower for the focal group than for a reference group, then this DIF is called uniform DIF. Regarding nonuniform DIF, the probability of a certain response is higher/lower for the focal group for part of the ability continuum, but "this relationship is reversed for a different part of the continuum" (De Ayala, 2009, p.325).

The relationship between IRT and CFA

Item Response Theory. 2PL model is represented as

$$P(X_{ij} = 1 \mid \theta_i, a_j, b_j) = \frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]}$$

where a_j, b_j are the item discrimination parameter and the item difficulty parameter for item j respectively. θ_i is the latent trait for person i. The discrimination parameter in the 2PL model describes the extent to which an item can differentiate examinees with different latent traits or abilities (De Ayala, 2009).

The 2PL IRT model can be extended to a 3PL model with the addition of the guessing parameter:

$$P(X_{ij} = 1 \mid \theta_i, a_j, b_j, c_j) = c_j + (1 - c_j)\frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]}$$

where θ_i is the person i's latent construct or trait; a_j, b_j are respectively the difficulty and discrimination parameters of item j, and c_j is the guessing parameter for item j. It is difficult to give each item a guessing parameter due to estimation problems. As a result, there is one common guessing parameter for a group of items in a test (Embretson & Reise, 2000).

CFA and IRT with ordered-categorical variables. Both CFA and IRT models are widely used in measurement invariance/DIF detection. A one-factor CFA can be transformed into a 2PL IRT model with the robust maximum likelihood estimator (MLR) (Woods, Oltmanns, & Turkheimer, 2009). Regarding categorical items such as Likert-scale items, DIF detection with an ordinary CFA that deals with continuous variables is not suitable; therefore, both ordered-categorical variable CFA (CCFA) and IRT model can be taken into account to detect DIF due to their capabilities to incorporate thresholds (Kim & Yoon, 2011).

In a CFA model with one factor and ordered-categorical responses, the CCFA can be illustrated as

$$y_{ij}^{*} = \tau_j + \lambda_j \theta_i + \varepsilon_{ij}$$

where y_{ij}^{*} is the latent response variable, λ_j is the factor loading, θ_i is the latent trait, and ε_i is the random error (Kim & Yoon, 2011). The latent response variable y_i^{*} is associated with the observed ordinal response y_i with threshold parameters:

$$\mathrm{y_i} = \begin{cases} 0 \; when \; y_i^{*} \leq \tau_{i1} \\ 1 \;\; when \; \tau_{i1} < y_i^{*} \leq \tau_{i2} \\ \cdots \\ C \; when \; y_i^{*} > \tau_{iC} \end{cases}$$

where τ_{iN} is a threshold between two score categories adjacent to each other (Cheng, Shao, & Lathrop, 2016). The number of categories of an item is equivalent with the number of threshold parameters of the item plus 1. CCFA with threshold parameters can be compared with graded response model in IRT when they are used to identify measurement invariance or DIF due to similar techniques of analysis (Kim & Yoon, 2011).

Samejima (1997) proposed graded response model (GR). According to GR, the probability of a person with latent trait θ responding in category c-1 can be represented as

$$P_{(c-1)j}(\theta_i) = p_{c-1} - p_c$$

$$= \exp \alpha_j(\theta_i - b_{(c-1)j}) /[\exp \alpha_j(\theta_i - b_{(c-1)j})] - \exp \alpha_j(\theta_i - b_{cj}) /[1+\exp \alpha_j(\theta_i - b_{cj})]$$

where θ of person i is the latent trait, b_{cj} is the threshold parameter for category score c, p is the cumulative probabilities while P indicates the probability that an examinee gives response in a certain category (De Ayala, 2009, pp. 219-220), and α_j is the discrimination parameter. If there is an item with four categories (0,1,2,3), the probability of responding to category 0 is

$$P_0=\exp \alpha_j(\theta_i - b_{0j}) /[1+\exp \alpha_j(\theta_i - b_{0j})] - \exp \alpha_j(\theta_i - b_{1j}) /[1+\exp \alpha_j(\theta_i - b_{1j})]$$

$$= 1 - \exp \alpha_j(\theta_i - b_{1j}) /[1+\exp \alpha_j(\theta_i - b_{1j})]$$

The threshold structure in CCFA can be compared with parameters of item difficulty in IRT (Kim & Yoon, 2011). Therefore, transformation of parameters from a 2PL IRT model to a one-factor CCFA is straightforward.

$$\alpha_j = \left(\frac{\lambda_j}{\sqrt{1-\lambda_j^2}}\right)D$$

and

$$b_j = \frac{\tau_j}{\lambda_j}$$

where α_j is the item discrimination parameter for item j in 2PL-IRT model, b_j is the item difficulty parameter for item j in IRT model, λ_j is the factor loading of item j in CCFA, τ_j is the threshold parameter in CCFA. If D is 1.7, the IRT model is a logistic model, yet the model is normal ogive, when D=1 (Wirth & Edwards, 2007). For the transformation is between a graded response model and a CCFA model, then b_j is changed to b_{jc} and τ_j to τ_{jc} (Wirth & Edwards, 2007).

Measurement invariance is different between ordinary CFA with continuous variable and CCFA, because thresholds can be a source of measurement invariance violation in CCFA. Furthermore, there is a direct relationship between CFA model and "the observed means and covariance structure" while the relationship is indirect in CCFA (Millap & Yun-Tein, 2004). So CCFA is different from ordinary CFA while it is interchangeable with IRT when they are used to detect DIF in ordered-categorical items.

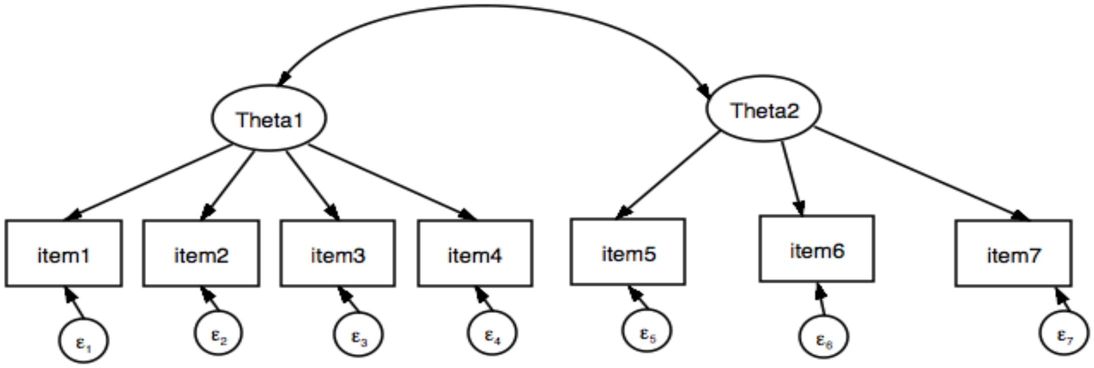
If one item measures multiple, related dimensions, then multidimensional item response theory should be used for DIF detection. Multidimensional item response theory is closely related to CFA.

Multidimensional Item Response (MIRT) models

In educational and psychological tests, examinees may need a variety of latent traits to respond correctly to an item. Therefore, MIRT models can be used to deal with multiple dimensions at the same time (Hartig & Höhler, 2009). Take GMAT and GRE writing tests for example, a non-English speaker with strong logic ability may compensate for his/her lower language proficiency to get a relatively satisfactory score in writing. This scenario reflects a compensatory multidimensional nature of MIRT.

Compensatory, partially compensatory, and conjunctive MIRT models. Compensatory MIRT models the probability of an item response given a combination of different latent traits (Reckase, 2009). In partially compensatory MIRT model, the probability of a successful response for an item is the product of the probabilities of a variety of ability dimensions that are required for that item (Hartig & Höhler, 2009; Reckase, 2009). In a conjunctive MRT model, it is assumed that test takers need multiple ability dimensions required by an item to get a correct answer to that item (Wang & Nydick, 2015).

Between- and within-item MIRT models. If the probability of getting a response correctly is influenced by only one of the dimensions in MIRT, then the model involved is a between-item MIRT model (Hartig & Höhler, 2009). In a between-item MIRT model, the items are grouped to measure only one dimension of the model. Regarding within-item MIRT, it is a model in which several dimensions are needed at the same time when examinees give correct response to items in an exam (Hartig & Höhler, 2009). Figure 2 illustrates between- and within-item MIRT models.



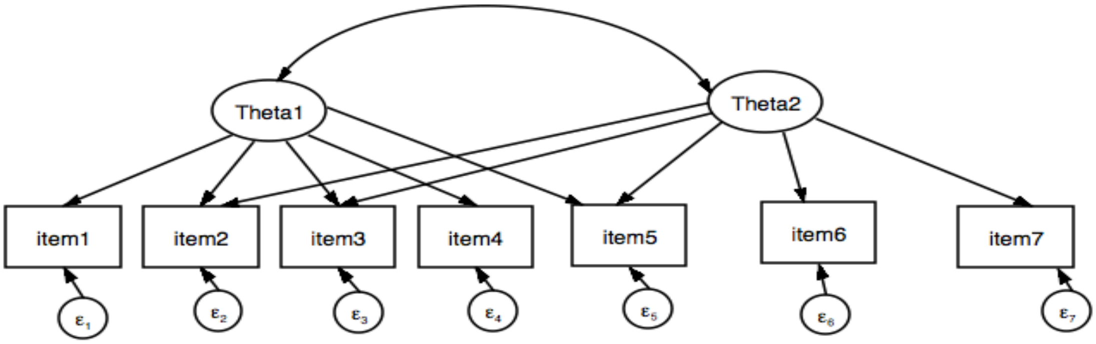


Figure 2. A Between-item MIRT Model (up) and a Within-item MIRT Model (down)

Multidimensional graded response model (MGRM). MGRM is the extension of a one dimensional IRT graded response model. And the normal ogive form of MGRM can be illustrated as:

$$P(u_{ij} = k|\theta_j) = \frac{1}{\sqrt{2\pi}} \int_{a_i^{'}\theta_j + d_{i,k+1}}^{a_i^{'}\theta_j + d_{ik}} e^{\frac{-t^2}{2}} dt$$

where *k* is the score of the item (there are *K* categories and *K-1* thresholds), a_i is a vector of item discrimination parameters, θ is the vector of constructs, and d_{ik} is the parameter that reflects the ease a test taker reaches the *k*th category (Reckase, 2009).

The relationship between MIRT and CFA. The methodology of nonlinear factor analysis is in essence the same as that used in MIRT (McDonald, 1967). Both methods have to depend on rotation of the coordinate axes to analyze the features of the data (Reckase, 2009, p.63). Given the similarities between CCFA and IRT, and CFA and MIRT, MIMIC model can be treated as a measurement invariance or DIF detection approach from both CCFA and MIRT perspectives.

If several subscales are distinct but correlated, it is better to implement a joint analysis of the model with all subscales included in order to improve the accuracy and quality of parameter estimation (Adams, Wilson, & Wang, 1997). Therefore, MIRT is an ideal alternative to separate unidimensional analyses because it utilizes the connection and the relationship of different dimensions to produce both item and person parameter with more accuracy (Cheng, Wang, & Ho, 2008).

From the relationship between CCFA and GRM as the connection between MIRT and CFA, it can be concluded that MIMIC method of DIF detection of ordinal items or ordered-categorical items with multiple factors can be regarded as a multidimensional graded response model or a CCFA model with covariates. However, different parameter estimation methods are used for IRT and CCFA.

Parameter estimation in GRM and CCFA

Weighted least square (WLS) method is widely used in parameter estimation of CFA with categorical items (Wirth & Edwards, 2007) while full information maximum likelihood (FIML) or marginal maximum likelihood (MML/ML) is usually used in GRM (Forero & Maydeu-Olivares, 2009). Maximum likelihood estimation utilizes the information of the observed variables to make their probability distribution most likely

(Myung, 2003). In MML, the unconditional or marginal probability of a response vector $\underline{x}$ can be represented as

$$p\left(\underline{x}\right) = \int_{-\infty}^{+\infty} p\left(\underline{x}\middle|\theta, \underline{\vartheta}\right) g\left(\theta\middle|\underline{\vartheta}\right) d\theta$$

where $p\left(\underline{x}\middle|\theta, \underline{\vartheta}\right)$ is the probability of a particular response pattern $\underline{x}$ that is conditioned on latent trait θ and item parameters $\underline{\vartheta}$; $g\left(\theta\middle|\underline{\vartheta}\right)$ is continuous distribution of latent traits of the population.

If f represents the number of people who get a particular response pattern $\underline{x}$, the likelihood function for the entire set of response vectors is

$$\mathrm{L}=\prod_{i=1}^{s} \left(\int_{-\infty}^{+\infty} p\left(\underline{x}\middle|\theta, \underline{\vartheta}\right) g\left(\theta\middle|\underline{\vartheta}\right) d\theta\right)^{f}$$

where s represents the total number of response patterns for the items (Hambelton & Swainathan, 1985). This function is maximized with respect to the item parameters in MML. When there exist missing values, full information maximum likelihood estimation can be used to estimate the parameters (Enders & Bandalos, 2001).

Embretson & Reise (2000) made an evaluation of the advantages of MML, which include the applicability in many IRT models including multidimensional models, justified item standard errors and no loss of information for examinees with perfect scores. However, the computations in ML is cumbersome especially in MIRT models while WLS is much more efficient (Forero & Maydeu-Olivares, 2009). MML estimation of a multidimensional requires evaluation of a multidimensional integral. As a result, computation time increases exponentially as the number of dimensions increases.

Alternative estimation methods, such as WLS estimation, improves working efficiency but sacrifices information (Forero & Maydeu-Olivares, 2009).

Both MLR and WLSMV will be used in the analysis. With MLR in Mplus, more than half a month of computational time will be needed to select anchor items; however, it takes less than 30 seconds to select each anchor item in WLSMV. Therefore, the more efficient WLSMV will be used to more quickly select anchor items. However, the use of WLSMV comes at the cost of a loss of information in the estimation. If the computational efficiency of MLR improves, it will be ideal to use it throughout the entire MIMIC model process.

MIMIC approach in DIF detection

The MIMIC model that can be used for DIF detection of ordered-categorical items has two parts: the measurement component and the structural component. The measurement model can be represented as:

$$y_i^* = \lambda_i\theta + \beta_i z + \varepsilon_i$$

where y_i^* is the latent response variable, λ_i is the factor loading, θ is the latent trait, z is the grouping variable, which can have more than two categories, β_i is the group variable effect or the relationship between the latent response variable y_i^* and the group variable, and ε_i is the random error (Wang & Shih, 2010; Chen, Shao, & Lathrop, 2016). If β_i is significant, it indicates that the group variable z has a direct effect on the latent response and it is an indication of uniform DIF in item *i* (Wang & Shih, 2010).

If the latent response variable y_i^* is associated with the observed ordinal response

y_{i} with a threshold model, then it can be illustrated as:

$$\mathrm{y_i} = \begin{cases} 0 \quad when\ y_i^* \leq \ \tau_{i1} \\ 1 \ \ when\ \tau_{i1} < y_i^* \leq \ \tau_{i2} \\ \quad \ldots \\ C \quad when\ y_i^* > \tau_{iC} \end{cases}$$

where τ_{iC} is a threshold between two score categories adjacent to each other（Cheng, Shao, & Lathrop, 2016).

Compared with non-IRT DIF detection techniques like The Mantel-Haenszel (MH) procedure (Rogers & Swaminathan, 1993), the simultaneous item bias test (SIBTEST; Shealy & Stout, 1993) and logistic regression (De Ayala, 2009), and IRT-based DIF detection methods such as Lord's (1980) chi-square, Raju's (1990) area measures, Thissen, Steinberg, and Wainer's (1988) likelihood ratio test, and Raju, van der Linden, and Fleer's (1995) procedure based upon the differential item functioning of item and tests (DFIT) framework, MIMIC method has several advantages.

MIMIC model can be used in both unidimensional and multidimensional DIF detection with one or more grouping variables due to its flexibility (Cheng, Shao, & Lathrop, 2016; Wang & Shih, 2010). Many researchers have advocated the use of MIMIC as a DIF detection method (e.g. Muthén, 1989; Woods, 2009b; Jin, Myers, Ahn, & Penfiled, 2012).

Woods, Oltmanns, and Turkheimer (2009) pointed out that in a MIMIC model, items with multiple dimensions or multiple factors are modeled without much difficulty. At least two groups can be tested for DIF at the same time and it is easy to control for additional covariates. In addition, both continuous and categorical variables can be used as covariates in MIMIC model.

According to Finch (2005), when the test is over 20 items or when there is no “pseudo-guessing” parameter in, performance of the MIMIC model is comparable to SIBTEST, IRT likelihood ratio and Mantel-Haenszel procedures. If a test has 50 or more items and is analyzed with 2PL IRT model, the MIMIC method is a competitive alternative to traditional DIF detection methods regarding type I error control regardless of size of the focal group, latent mean differences, and anchor item contamination (Finch, 2005). A basic MIMIC model to detect DIF is illustrated in Figure 3, in which item 3 and item 5 are tested for uniform DIF and item 1, item 2 and item 4 are assumed to be DIF free.

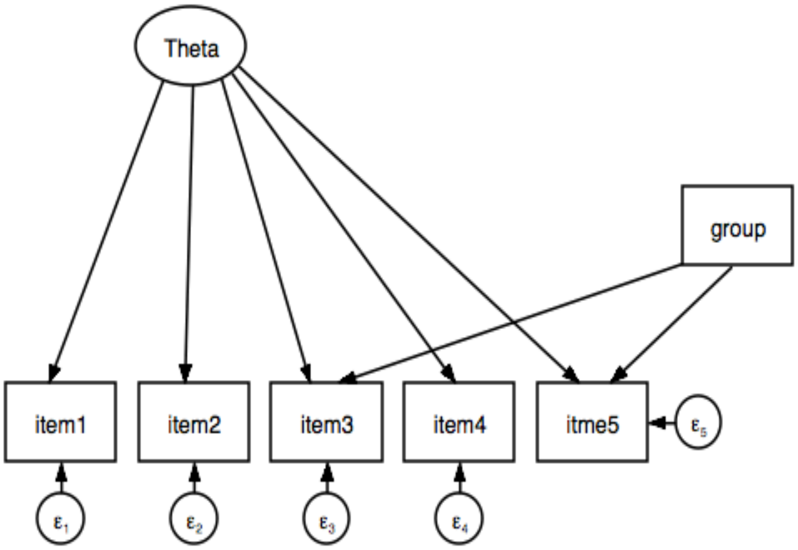


Figure 3. A Basic MIMIC Model for Uniform DIF Detection

The structural model can be represented as:

$$\theta = \gamma z + \xi$$

where γ is the regression coefficient for the grouping variable z. If γ is significant, then there is group difference in latent trait, which is often called "impact" (Cheng et al., 2016).

In a MIMIC model, item discrimination parameters or factor loadings are constrained to be equal across groups. As a result, the model can only test uniform DIF. According to Woods and Grimm (2011), if there is an interaction between grouping variable z and the latent trait θ, then the interaction "is the key to testing for nonuniform DIF" (Woods & Grimm, 2011). A MIMIC with interaction is given by:

$$y_i^* = \lambda_i\theta + \beta_i z + \omega_i\theta z + \varepsilon_i$$

If ωi is significant, it indicates nonunifrom DIF.

Cheng, Shao, and Lathrop (2016) took into account mediation effect in MIMIC model, and the new model incorporates one or more variables that have complete or partial mediation effect on DIF. If a newly added mediator that "significantly mediates the effect of X" after DIF effect is detected between the grouping variable X and the observed variable Y, then there is a mediation effect in the MIMIC model. If the DIF effect between X and Y disappears after mediator is taken into account, this is a complete mediation effect. If there is a complete mediation effect, the mediator can be used as a background or grouping variable that has a direct path to the observed variable Y (Cheng et al., 2016). If both the path between the grouping variable(s) and the latent variable (s), and the path between the grouping variable(s) and the mediator are significant, then it indicates partial mediation effect, which means the mediator can help to explain part of the DIF effect but not completely (Cheng, et al., 2016).

The procedure of DIF detection via mediator MIMIC model involves several steps (Cheng, et al., 2016). The first step is to find the DIF items in the MIMIC model without mediators. After the DIF items are identified, a mediator is incorporated into the model to account for the DIF effect. Each DIF item is put in the new model with a mediator and both direct and indirect effects can be obtained. However, "it requires substantial content knowledge and familiarity with the items to identify possible mediators" (Cheng, et al., 2016, p.57).

MIMIC testing procedure

There are different DIF detection procedures with MIMIC method. Muthén (1989) recommended modification indices (MIs) in MIMIC DIF study. Items with a large MIs of the regression coefficient will be selected as the ones with DIF and these items are put in the model for subsequent DIF analysis. The problem with this procedure is that there is no clear definition of large index; therefore, it may not be appropriate to use MIs in a study (Woods, 2009b).

Other MIMIC DIF detection procedures involve anchor item selection that should never be neglected in MIMIC DIF study, because if the matching variable that includes anchor items are contaminated, the groups under study are compared with "a biased measure" that can lead to misleading results (Shin & Wang, 2009). The standard

MIMIC method (M-ST) is to use one item tested for DIF when all other items are used as anchor items. Yet when the percentage of items with DIF is high, M-ST may not perform well (Wang & Shih, 2010). Another method is to obtain the DIF index of all other items when one item is used as the anchor item. The same step is repeated until every individual item in the test is used as the anchor item. The mean of the absolute values of the DIF index of all items across iterations are calculated and items with the smallest mean value are selected as anchor items (Wang & Shih, 2010). Woods (2009b) recommended the use of one anchor item in order to "minimize the chance of contamination", and this is of vital importance when the percentage of DIF items is large. If the sample size is large, then the one-anchor estimation is quite accurate (Wang, 2004; Woods, 2009b).

Different researchers proposed different index for anchor selection in MIMIC model. The regression coefficient β, the effect from the grouping variable to the item, is the one that can be used (Wang, Shih, & Yang, 2009). If the regression coefficient β is not significant, then the item can be selected as anchor items.

Both "the smallest *SS/SE* ratios" (Woods, 2009b) and "the β / Standard error (SE) ratios" (Woods and Grimm, 2011) are also used to select anchor items. When the ratios are obtained for all items, they are ranked by order with the absolute value. The items with the smallest absolute values are picked as anchor items. Assessing the model fit change (Oishi, 2006) and using MIMIC effect size $(-\beta_i / \lambda_i)$ (Jin, Myers, Ahn, & Penfield, 2012) are other alternatives that can be considered.

The application of MIMIC DIF test

DIF test with MIMIC approach has been used in political science, medical and health studies, and personality.

Pérez (2011) used MIMIC approach to study the Latino political attitudes. It was found that language used in the survey of the attitudes attribute to DIF.

Medical researchers have used MIMIC model to study the DIF in various forms of medical and health surveys. Fleishaman, Specton, and Altman (2002) assessed DIF in "functional disability" in terms of gender and age groups with MIMIC models. Jones (2006) applied MIMIC model to identify DIF "between English and Spanish language version of the Mini-Mental State Examination." MIMIC models were also used by Fan, Yu, and Ahn (2007) to assess DIF in a health survey among different subgroups "regarding chronic illness among items of the physical functioning (PF) and mental health (MH) of the SF-36 health survey."

Woods, Oltmanns, and Turkheimer (2008) applied MIMIC models in personality studies and they assessed uniform DIF in "the schedule for Non adaptive and Adaptive Personality" scales.

MIMIC models have been used for DIF studies in recent years, but the literature review has indicated that there are still gaps in MIMIC application. There is not much applied research regarding DIF detection in educational context with MIMIC method especially when the items are order-categorical and the factors are distinct but related. From the literature review, it can also be found that no study in MIMIC approach has been executed in large-scale international tests like PISA and TIMSS except the one in

2016 (Chen, et al., 2016). The previous research in MIMIC tended to focus more on one factor or one dimension (e.g. Woods et al., 2009; Perez, 2011)). Some researchers have taken into account multiple factors in their study (e.g. Fan, Yu, & Ahn, 2007); however, they did not specify the procedure of anchor item selection, which is of vital importance in MIMIC study.

The present study intends to utilize the flexibility of MIMIC model to implement a comparative study of PISA 2012 mathematics self-beliefs (mathematics self-efficacy, mathematics self-concept and mathematics anxiety) of students from Shanghai-China and the United States. Because the items studied will be ordered-categorical items, and the three factors are distinct but closely related (Lee, 2009), a MIMIC method with between-item multidimensional graded response model will be used in the study to detect DIF and the method will be used from both CCFA and MIRT perspectives.

Chapter 3 Methodology

The primary purpose of the study was to apply the multiple indicators multiple causes (MIMIC) approach to assess the differential item functioning of categorical items that incorporate thresholds in the mathematics self-beliefs questionnaire of the Program for International Student Assessment (PISA) 2012 across students from Shanghai-China and the United States. This study tried to answer the following research questions:

1. Do self-beliefs structure patterns of Shanghai-China and the US students follow the same pattern?

2. Are the items in the math self-beliefs questionnaire in PISA 2012 measuring the same constructs in both the US and Shanghai groups?

3. Does any DIF exist in any of these items in the dimensions of mathematics self-efficacy, mathematics self-concept and mathematics anxiety?

4. If DIF exists in these items, which covariates can be used to account for the DIF?

An Introduction to PISA

The Program for International Student Assessment (PISA) 2012 data was used for the study. These data are publically available at *www.pisa.oecd.org*. PISA, an international low-stake test, was organized by the Organization for Economic Cooperation and Development (OECD) and it was first carried out in 1997 (OECD, 2013a). PISA 2012 focused on the content of mathematics, reading and science. Assessment of financial literacy was introduced into PISA 2012 (OECD, 2013a).

In addition to the items that were used to assess the knowledge and skills related to mathematics, reading, science and financial literacy, students and school principals also answered questionnaires about their background and school information. In some countries and economies, parents were also asked to respond to questions about their attitude toward schools and their engagement in their children's mathematics-related expectations of careers (OECD, 2014).

In PISA 2012, students from Shanghai-China had the highest average score in mathematics at 613 points. The five best performing countries and economies were all from Asia, which were Shanghai, Singapore, Hong Kong, Taiwan and Korea. The average score of the United States in mathematics was 418 and it ranked 27th (OECD, 2014).

Participants

In 2012, 51,000 students whose ages ranged from 15 years 3 months to 16 years 2 months from 65 countries and economies participated in PISA including students from Shanghai-China and the United States (OECD, 2014). The number of students from Shanghai-China was 5,177, of whom 49% were male students and 51% were female students. The number of students from the United States was 4,978, of whom 51% were male and 49% were female.

Self-belief Measures

Mathematics self-efficacy. Eight items were used to measure students' mathematics self-efficacy. The items asked students regarding “their perceived ability to solve” a variety of “pure and applied mathematics problems” (OECD, 2013b). Each item had a 4-point Likert-scale (1=very confident, 2=confident, 3=not very confident, 4=not at all confident). The eight items asked students how confident they were when “using a train timetable to work out how long it would take to get from one place to another”, “calculating how much cheaper a TV would be after a 30% discount”, “calculating how many squares meters of tiles you need to cover a floor”, “understanding graphs presented in newspapers”, “solving an equation like 3x+5=17”, “finding the actual distance between two places on a map with a 1:10000 scale”, “solving an equation like 2(x+3)=(x+3)(x-3)”, and “calculating the petrol-consumption rate of a car” (OECD, 2013b).

Mathematics self-concept. The scale had 5 items. These five items also had a 4-point Likert-scale (1=strongly agree, 2=agree, 3=disagree, 4=strongly disagree). The five items were “I am just not good at mathematics”, “I get good grades in mathematics”, “I learn mathematics quickly”, “I have always believed that mathematics is one of my best subjects”, and “In my mathematics class, I understand even the most difficult work” (OECD, 2013b).

Mathematics anxiety. This measure was also a 5-item scale that was on 4-point Likert-scale (1=strongly agree, 2=agree, 3=disagree, 4=strongly disagree). The items were “I often worry that it will be difficult for me in mathematics classes”, “I get very tense when I have to do mathematics homework”, “I get very nervous doing mathematics problems”, “I feel helpless when doing a mathematics problem”, and “I worry that I will get poor grades in mathematics” (OECD, 2013b).

Table 1. Mathematics Self-belief Items

Domain	Item ID	Item Text
Self-efficacy	SE1	Using a train timetable to work out how long it would take to get from one place to another
	SE2	Calculating how much cheaper a TV would be after a 30% discount
	SE3	Calculating how many squares meters of tiles you need to cover a floor
	SE4	Understanding graphs presented in newspapers
	SE5	Solving an equation like 3x+5=17
	SE6	Finding the actual distance between two places on a map with a 1:10000 scale
	SE7	Solving an equation like 2(x+3)=(x+3)(x-3)
	SE8	Calculating the petrol-consumption rate of a car
Self-concept	SC1	I am just not good at mathematics
	SC2	I get good grades in mathematics
	SC3	I learn mathematics quickly
	SC4	I have always believed that mathematics is one of my best subjects
	SC5	In my mathematics class, I understand even the most difficult work
Mathematics Anxiety	MA1	I often worry that it will be difficult for me in mathematics classes
	MA2	I get very tense when I have to do mathematics homework
	MA3	I get very nervous doing mathematics problems
	MA4	I feel helpless when doing a mathematics problem
	MA5	I worry that I will get poor grades in mathematics

Data Analysis Procedure

Raw data from the PISA 2012 website was cleaned and recoded with Stata 13. In particular, items were recoded so that higher item scores indicated more of the construct. For mathematics self-efficacy scale, 1 was recoded as 4, 2 was recoded as 3, 3 was recoded as 2 and 4 was recoded as 1, so students with higher levels of self-efficacy got higher scores on self-efficacy items. All items in the mathematics self-concept were recoded in the same way (1 was recoded as 4, 2 was recoded as 3, 3 was recoded as 2 and 4 was recoded as 1) except the item “I am just not good at mathematics”. High scores in the scales indicated high levels of mathematics self-concept and low levels of anxiety.

Mplus 6.11 was used for the data analysis. Both robust weighted least squares (WLSMV) estimator and robust maximum likelihood (MLR) estimator were used in parameter estimation. WLSMV estimator helped to estimate the model fit of the three-factor CCFA model with a covariate. In addition, since there were18 items in the three dimensions, WLSMV estimator helped to select anchor items with high efficiency. However, part of the information would be lost during the procedure. MLR estimator was used after anchor items were selected to utilize the full information to obtain the parameter estimation of the multidimensional MIMIC Model of DIF detection of ordered categorical items from PISA 2012 within a MIRT framework.

Before the anchor selection and DIF detection stages, a standard three-factor CFA model was built, and the factors (mathematics self-efficacy, mathematics self-concept and mathematics anxiety) were regressed on the dummy variable "countries and economies" which was labeled as "CNT" in PISA 2012 with Shanghai-China recoded as 1 and the United States recoded as 0.

Anchor items

Before DIF detection with MIMIC method, anchor items without DIF were selected. The selection of anchor items is of great importance in DIF detection in MIMIC model, because if the matching items are contaminated, the accuracy of DIF detection rate will be affected (Woods, 2009b).

Anchor item selection was based upon the procedures recommended by Wang and Shih (2010), Woods (2009a, 2009b), and Woods and Grimm (2011). One item was used as the anchor item while the other 17 items from mathematics self-efficacy scale, self-concept scale and mathematics anxiety scale were tested for uniform DIF. Parameters was estimated with robust weighted least squares (WLSMV) estimator and an absolute value of beta index for each item was obtained.

The remaining items in the self-belief scales were used as the anchor item one by one as what had been done in the first step. The absolute value ratio of the beta/standard error (SE) of each item was obtained across 17 iterations. The absolute values of the ratios were ranked by order and the ones with the smallest absolute value were selected as anchor items.

After the selection of anchor items, the remaining studied items were regressed on the dummy variable country and economies (CNT) using the MLR estimator. Odds ratios were obtained and rank-ordered in the analysis to indicate the magnitude of DIF. Based upon the recommendation by Cole, Kawachi, Maller and Berkman (2000), items with the odds ratio values greater than 2 or less than than 0.5 were flagger as items with meaningful DIF.

After the detection of DIF items, other variables (e.g. out-of-school study hours per week, in-school mathematics class periods per week) were added to the model to see whether these variables could account for the uniform DIF in the MIMIC model. Evaluating whether the mediators can partially or fully account for the DIF effect of those items in the dimensions offers “valuable information on how to revise a DIF item or even to implemented targeted intervention” (Cheng, et al., 2016). If the indirect path from the grouping variable to the mediator then to the item is significant, it means the mediator can significantly mediate the grouping variable on the response of the item after the latent variable is controlled for. If there exists significant mediation effect, partial or complete mediation effect will be tested to see whether the mediator should be used as a “direct background variable” (Cheng, et al., 2016).

Chapter 4 Results

This chapter provides results based upon MIMIC method to answer the research questions in terms of the structural patterns of the mathematics self-beliefs of PISA 2012 and mathematics self-beliefs items with differential item functioning (DIF). Mediators were also added to the MIMIC model in order to account for the meaningful DIF effects.

CFA Model

The fit of the multidimensional structure of self-beliefs was examined respectively for both Shanghai-China and the US students in order to confirm that the three-factor structure was suitable in both populations. To assess the model fit, model fit indices were obtained. According to Hu and Bentler (1999), comparative fit index (CFI) greater than 0.95 and root mean square error of approximation (RMSEA) that is equal or smaller than 0.06 indicate good model fit. And the value between 0.05 and 0.08 indicates adequate model fit (Browne & Cudeck, 1993).

In this stage, since the indicators were ordered categorical variables, robust weighted least square (WLSMV) estimator was used for the analysis of the data from the mathematics self-belief scales. As far as the model results from the Shanghai-China data were concerned, the model fit the data adequately well: RMSEA=0.065, CFI=0.978 and TLI=0.975. In terms of the data analysis results of the US data in PISA 2012, the results indicated that the model also fit data well: RMSEA=0.061, CFI=0.975 and TLI=0.971.

After the baseline models for both Shanghai and the US were determined, multi-group analysis was implemented to test configural invariance. The results indicated good model fit: RMSEA=0.064, CFI=0.976, TLI=0.972. Therefore, it could be concluded that the three-factor multidimensional structure worked well for the data of mathematics self-beliefs of students in both Shanghai-China and the United States in PISA 2012.

Anchor Item Selection with MIMIC Method

The grouping variable "countries and economies" (CNT), in which Shanghai-China served as the focal group (coded as 1) and the United States was chosen as the reference group (coded as 0), was added to the CFA model. The grouping variable was treated as a dummy variable. Table 2 provides the parameter estimates of regression coefficients of all items in three factors of mathematics self-beliefs after each individual item was regressed on the grouping variable "CNT" when all other items were fixed and treated as anchor items and were assumed to be without measurement variance or DIF free. WLSMV was used in the parameter estimation.

Table 2 The Regression Coefficients of Items on the Grouping Variable in Math Self-efficacy, Self-Concept and Anxiety Domains

Item	Estimate	S.E.	Estimate/S.E.	*p*
Mathematics Self-efficacy				
SE1	-0.032	0.022	**-1.476**	0.140
SE2	0.183	0.021	8.879	0.000
SE3	0.107	0.019	5.494	0.000
SE4	-0.294	0.022	-13.286	0.000
SE5	-0.462	0.026	-17.682	0.000
SE6	0.716	0.022	32.041	0.000
SE7	0.042	0.024	1.776	0.076
SE8	-0.275	0.022	-12.425	0.000
Mathematics Self-concept				
SC1	0.131	0.019	6.764	0.000
SC2	-0.750	0.021	-36.561	0.000
SC3	0.113	0.018	6.186	0.000
SC4	0.351	0.017	20.611	0.000
SC5	0.086	0.019	**4.547**	0.000
Mathematics Anxiety				
MA1	0.219	0.019	11.500	0.000
MA2	0.192	0.018	10.836	0.000
MA3	0.089	0.019	**4.822**	0.000
MA4	-0.111	0.020	-5.628	0.000
MA5	-0.426	0.021	-20.187	0.000

One item from each factor with the smallest absolute regression coefficients divided by the standard error, β/S.E. value, was chosen as the anchor item. These three items were SE1 from the mathematics self-efficacy domain; SC5 from the self-concept domain; and MA3, which concerned nervousness about mathematics problems.

Another MIMIC model with these three anchor items fixed was estimated with the default WLSMV estimator. Based upon *p* values as well as absolute β/S.E ratios, the results shown in Table 3, indicate that SE7 (β=0.069, absolute ratio= 2.048, *p*=0.041), SC1 (β=0.016, absolute ratio=0.611, *p*=0.541) and SC3 (β= 0.017, absolute ratio=0.736, *p*=0.462) were DIF-free items. No item was assigned to the DIF-free items in the anxiety domain. Therefore, SE1 and SE7 in the mathematics self-efficacy dimension, SC1, SC3 and SC5 in the mathematics self-concept dimension, and MA3 in the mathematics anxiety dimension were used as anchor items in the DIF detection procedure of the next stage.

Table 3. Items Parameters in Math Self-efficacy, Self-concept and Anxiety Domains with Three Anchor Items Fixed

Item	Estimate	S.E.	Estimate/S.E.	*p*
Mathematics Self-efficacy				
SE2	0.190	0.028	6.888	0.000
SE3	0.123	0.029	4.259	0.000
SE4	-0.231	0.029	-8.067	0.000
SE5	-0.377	0.035	-10.649	0.000
SE6	0.664	0.030	22.193	0.000
SE7	**0.069**	**0.033**	**2.048**	**0.0041**
SE8	-0.211	0.028	-7.515	0.000
Mathematics Self-concept				
SC1	**0.016**	**0.027**	**0.611**	**0.541**
SC2	-0.698	0.025	-28.174	0.000
SC3	**0.017**	**0.024**	**0.736**	**0.462**
SC4	0.203	0.023	8.657	0.000
Mathematics Anxiety				
MA1	0.089	0.026	3.449	0.001
MA2	0.073	0.021	3.391	0.001
MA4	-0.163	0.024	-6.878	0.000
MA5	-0.423	0.026	-16.365	0.000

Note: Bolded values indicated that the items had NO DIF.

Results also showed a significant effect of the grouping variable “CNT” on two domains of mathematics self-beliefs: mathematics self-efficacy and mathematics self-concept; however, “CNT” did not have significant effect on anxiety. The results indicated that the factor means of mathematics self-efficacy and mathematics self-concept between students of Shanghai-China and those of the United States were variant. Students from Shanghai-China had higher mathematics self-efficacy (0.539, *p*=0.000) while lower mathematics self-concept (-0.300, *p*=0.000) than their American peers. However, there was no significant difference in terms of mathematics anxiety factor means (-0.034, *p*=0.239) between Shanghai-China and the United States, meaning that mathematics anxiety was a universally existing psychological state regardless of the differences in educational systems and mathematics requirements.

Detection of Items with Meaningful DIF Effects

After the six DIF-free items in the mathematics self-belief factors were identified and selected as anchor items, the next step was the procedure of the detection of items with meaningful DIF effects within a MIRT framework.

MLR estimator was used in a MIMIC Model of ordered categorical items from PISA 2012. With MLR, logistic odds ration (OR) could be obtained to estimate the effect size of DIF magnitude. Items with an odds ratio value bigger than 2 or smaller than 0.5 were flagged as having meaningful DIF (Cole, Kawachi, Maller, & Berkman, 2000). Regarding mathematics self-efficacy, SE4, SE5, SE6 and SE8 exhibited meaningful amounts of DIF (See Table 4). In terms of the mathematics self-concept domain, SC2 and SC4 could be considered as items with meaningful DIF. Only MA5 in the mathematics anxiety domain exhibited meaningful amounts of DIF. In terms of the direction of the items with meaningful DIF effects, it could be found that SE4, SE5, SE8, SC2 and MA5 favored American students while SE6 and SC4 favored Shanghai students (See Table 3 and Table 4).

Table 4. DIF Effect Size

Item	Logistic odds-ration estimates
Mathematics Self-efficacy	
SE2	1.535
SE3	1.262
SE4	**0.421**
SE5	**0.345**
SE6	**5.455**
SE8	**0.488**
Mathematics Self-concept	
SC2	**0.123**
SC4	**2.128**
Mathematics Anxiety	
MA1	1.372
MA2	1.269
MA4	0.578
MA5	**0.336**

Note: Bolded values indicated that the items had meaningful DIF.

Causes of DIF

Two mediator variables were evaluated in separate MIMIC models. One variable concerned the number of hours students spent studying outside of regular school hours. Another variable involved the mathematics class periods students had in their schools.

According to a regression model, students in Shanghai-China had 9.85（p=0.000) more hours than their American peers in terms of studying hours spent outside their regular school hours per week, and Shanghai students had 2.51 (*p*=0.000) more hours of class time in school per week than did American students.

If a mediator has full effect on DIF items, then the significant relationship between "CNT" and the mathematics self-belief items with meaningful DIF effects completely vanishes after the mediator is added to the model. However, if there is a partial effect of the mediator or the mediator can partially account for the DIF effect, then there will be a significant direct relationship between "CNT" and the mathematics self-belief items with meaningful DIF effects as well as the significant mediation effect.

When the variable of in-school mathematics class periods was utilized as the mediator, both direct and indirect effects for DIF items could be obtained as shown in Table 5. It was found that the direct effects from the grouping variable "CNT" to SE4, SE5, SE6, SE8, SC2, SC4 and MA5 were all statistically significant. However, the indirect effects from "CNT" to the mediator, in-school mathematics class periods, then to SE4, SE5, SE6, SE8, SC2 and SC4 were significant while the indirect path from "CNT" to in-school mathematics class periods then to MA5 was not significant. Since the direct paths from the grouping variable "CNT" to these six items were significant, the mediator of in-school mathematics class periods could help to partially account for the DIF effect of these items within the dimensions of mathematics self-efficacy and mathematics self-concept.

Table 5. Direct Effects and Indirect Effects for Items with Meaningful DIF with In-School Mathematics Class Periods as the Mediator

Item	Direct	*p*	Indirect	*p*
Mathematics Self-efficacy				
SE4	-0.486	0.000	0.174	0.000
SE5	-0.605	0.000	0.134	0.000
SE6	0.400	0.000	0.183	0.000
SE8	-0.465	0.000	0.173	0.000
Mathematics Self-concept				
SC2	-0.802	0.000	0.093	0.000
SC4	0.107	0.000	0.083	0.000
Mathematics Anxiety				
MA5	-0.468	0.000	0.042	0.050

When the variable of out-of-school study hours was utilized as the mediator, both direct and indirect effects for DIF items could be obtained as shown in Table 6. As far as the mediator out-of-school class study hours was concerned, there existed significant direct paths from the grouping variable "CNT" to all items in the three subscales (mathematics self-efficacy, mathematics self-concept and mathematics anxiety). The indirect paths from "CNT" to the mediator, out-of-school study hours, then to SE4, SE5, SE6, SE8, SC2, SC4 and MA5 were also significant, meaning that the mediator, out-of-school study hours, was a partial mediator for these items and it could partially explain the causes of DIF of these items in the scales of mathematics self-efficacy, mathematics self-concept and mathematics anxiety.

Table 6. Direct Effects and Indirect Effects for Items with Meaningful DIF with Out-of-school Study Hours as the Mediator

Item	Direct	*p*	Indirect	*p*
Mathematics Self-efficacy				
SE4	-0.457	0.000	0.149	0.000
SE5	-0.594	0.000	0.125	0.000
SE6	0.428	0.000	0.159	0.000
SE8	-0.413	0.000	0.122	0.000
Mathematics Self-concept				
SC2	-0.746	0.000	0.038	0.015
SC4	0.160	0.000	0.032	0.026
Mathematics Anxiety				
MA5	-0.392	0.000	-0.034	0.016

Chapter 5 Discussion

The current study used the multiple indicators multiple causes (MIMIC) method to assess measurement invariance or differential item functioning (DIF) of mathematics self-beliefs (mathematics self-efficacy, mathematics self-concept and mathematics anxiety) items between 15-year old students in Shanghai-China and the United States in PISA 2012. By comparing students from Shanghai-China and the United States, this study identified mathematics self-belief items in PISA 2012 that were sensitive to cultural and educational system differences. Two variables, out-of-school study hours and in-school mathematics class periods per week, were used in two separate MIMIC models with mediators to try to explain the reasons why some items exhibited DIF and whether the variables had partial or complete effects on the items with meaningful DIF effects.

Unlike most measurement invariance studies in PISA, which mainly used multiple group confirmatory factor analysis (CFA) or item response theory (IRT) methods, this study utilized a MIMIC method to deal with multiple covariates, multiple dimensions and ordered categorical variables with threshold structures in both categorical confirmatory factor analysis (CCFA) and multidimensional graded response models (GRM). The MIMIC approach with mediators was also applied in the study, which helped to detect variables that could partially or fully account for meaningful DIF effects. The MIMIC model in the study incorporated three factors and one grouping variable. The two MIMIC models with mediators incorporated one individual mediator respectively to help to account for the DIF effects.

A CFA indicated that the three-factor (mathematics self-efficacy, mathematics self-concept and mathematics anxiety) structure fit the data of 15-year-old students from both Shanghai-China and the United States in PISA 2012, meaning that self-beliefs structure patterns of Shanghai-China and the US students followed the same pattern. It also indicated that the self-belief questionnaire in PISA 2012 measured the same constructs in Shanghai-China and the US.

According to the latent means of the three subscales, students from Shanghai-China reported statistically higher levels of mathematics self-efficacy than their American peers. This finding was different from the findings in the study of Lee (2009) in which the researcher found that Asian students, especially those from Japan, South Korea and Thailand, had lower mathematics self-efficacy than their peers in the United States and Europe based upon the data from PISA 2003 when Shanghai did not participate in the international test. It should be noted that Lee's study involved a different sample than in the current study. Since all the items in the mathematics self-efficacy domain demanded students' specific mathematics knowledge and aptitude in a variety of mathematics subdomains, it is not unusual for Shanghai students to have higher levels of self-efficacy in mathematics that might be attributed to higher mathematics requirements and the high school entrance examination pressure encountered by many 15-year-old students in Shanghai-China.

The finding that students from the United States had statistically higher levels of mathematics self-concept than those from Shanghai-China was consistent from a variety of comparative studies that involved Asian and American students (e.g. Wilkins, 2004;

Lee, 2009; Yoshino, 2012). Students from higher achieving countries and economies in mathematics do not necessarily have higher self-concept (Wilkins, 2004).

However, as far as the mathematics anxiety was concerned, there was no significant difference between students from Shanghai-China and those from the United States. The result was consistent with the findings from Cassady and Johnson (2002) that test anxiety is universal regardless of countries and cultures.

It was found that within the domain of mathematics self-efficacy, SE4, SE5, SE6 and SE8 had a meaningful DIF effect size. In terms of the mathematics self-concept domain, SC2 and SC4 could be considered as items with meaningful DIF effect. Only one item, MA5, in the mathematics anxiety domain could be treated as an item with meaningful DIF effect. Further study found that SE4, SE5 and SE8 favored American students while SE6 favored Shanghai students. The finding was consistent with the previous study implemented by Chen and Zimmerman (2007) that American students had higher self-efficacy when dealing with easy mathematics items; however, their self-efficacy declined more quickly than their peers from Taiwan when encountering moderate and difficult items. SE6 asked students how confident they were to find the actual distance between two places on a map with a 1:10000 scale and the question asked about a more difficult mathematics problem than what were asked in SE4, SE5 and SE8; therefore, American students' mathematics self-efficacy decreased dramatically when encountering a more difficult question compared with their peers from Shanghai, who are more capable of solving challenging mathematics items owing to more practice and more experience in intensive mathematics training.

Due to cultural influence, American students are more confident while their Shanghai peers are relatively humble possibly due to cultural influences as well as higher levels of pressure and expectation experienced both at homes and in schools. Therefore, it could be found from SC2 and MA5 that students in the US thought they could get good grades in mathematics compared with their Shanghai peers while Shanghai students worried a lot that they could get poor grades in mathematics. It was also found that SC4 favored students from Shanghai. There is a possibility that the focus on mathematics, science and engineering education in China due to government's decades' efforts and priority in the development of industry and technology may make Shanghai students and parents regard mathematics as one of the most important subjects in schools and one of the subjects that are of vital importance in the high school entrance examination in the city; therefore, it was not unusual to find that students in Shanghai-China had significantly higher number of hours spent outside their regular school hours on their study and had more in-school mathematics class periods than their peers in the United States.

With MIMIC approach, it was found in the study that in-school mathematics class periods had partial effects on all items with meaningful DIF effects in students' mathematics self-efficacy and mathematics self-concept domains, and out-of-school study hours had partial effects on all items with meaningful DIF effects in the three domains of mathematics self-beliefs, meaning that both variables could partially account for the meaningful DIF effect size of these items respectively.

Limitations

Firstly, the MIMIC model used in the study can only take into account uniform DIF. MIMIC method mainly focuses on the invariance of intercepts and factor means, which means the study can only detect that either Shanghai-China or the US students were biased in a certain direction on a specific latent variable (mathematics self-efficacy, mathematics self-concept or mathematics anxiety). There may exist non-uniform DIF items; therefore, future work needs to be done to expand the analysis to non-uniform DIF, which takes into account the interaction effects of grouping variable and item responses. Since the items in the study are items with ordered categories; therefore, differential step functioning that takes into account the DIF effect in different categories (Penfield, 2010) also needs to be considered in any future study.

Second, the model was not able to incorporate some factors, which might contribute to the explanation of meaningful DIF effects of the items. These factors may include differences of general public's perceptions and attitudes towards mathematics, school requirements, mathematics teaching quality, parents' expectations, and teachers' teaching qualifications in different cultural contexts. It is difficult to quantify all these factors, some of which may demand qualitative analysis. It should also be noted that many 15-year-old students in Shanghai who participated in PISA 2012 had to face or had passed the challenging high school entrance examination. How the examination itself impacts students' mathematics self-beliefs and helps to explain the DIF effects of the items should also be considered in the future study. Selection of variables that can be used to explain the meaningful DIF effects in the MIMIC model is a tedious process, which demands more than the quantitative study or the methodology knowledge. A better understanding of differences in terms of culture and educational system may help us better locate the variables. Further study needs to be carried out to find specific and convincing reasons that can fully account for the measurement variance of the items.

Third, the findings of the current study cannot be generalized to the 15-year students in China. Students from the United States in PISA 2012 were highly diversified. However, the Shanghai sample was relatively homogeneous. Since Shanghai is the most developed metropolitan city with a booming well-educated middle class in China, it enjoys the highest standard of educational resources in a country with a widening gap between the underdeveloped west and the developed east coastal area. As a result, the students in the city have easy access to a variety of excellent educational resources compared with their peers in many other Chinese cities. Therefore, there is a high probability that students' level of mathematics self-efficacy, mathematics self-concept and mathematics anxiety in some parts of China may be very different from those in Shanghai.

References

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

Ahmed, W., Minnaert, A., Kuyper, H., & van der Werf, G. (2012). Reciprocal relationships between math self-concept and math anxiety. *Learning and Individual Differences, 22*(3), 385-389.

Ashcraft, M. H. (2002). Math anxiety: Personal, educational, and cognitive consequences. *Current Directions in Psychological Science, 11*(5), 181-185.

Ashcraft, M. H., & Kirk, E. P. (2001). The relationships among working memory, math anxiety, and performance. *Journal of Experimental Psychology: General, 130*(2), 224.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*(2), 191.

Bandura, A. (1993). Perceived self-efficacy in cognitive development and functioning. *Educational Psychologist, 28*(2), 117-148.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.

Beilock, S. L., & Maloney, E. A. (2015). Math anxiety A factor in math achievement not to be ignored. *Policy Insights from the Behavioral and Brain Sciences, 2*(1), 4-12.

Beilock, S. L., & Willingham, D. (2014). Ask the cognitive scientist–math anxiety: Can teachers help students reduce it. *American Educator, 38*(2), 28-32.

Beilock, S. L., Gunderson, E. A., Ramirez, G., & Levine, S. C. (2010). Female teachers' math anxiety affects girls' math achievement. *Proceedings of the National Academy of Sciences of the United States of America, 107*(5), 1860-1863.

Bodas, J., & Ollendick, T. H. (2005). Test anxiety: A cross-cultural perspective. *Clinical Child and Family Psychology Review*, *8*(1), 65-88.

Bong, M., & Clark, R. E. (1999). Comparison between self-concept and self-efficacy in academic motivation research. *Educational psychologist*, *34*(3), 139-153.

Bong, M., & Skaalvik, E. M. (2003). Academic self-concept and self-efficacy: How different are they really? *Educational Psychology Review, 15*(1), 1-40.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), Testing structural equation models (pp. 136- 162). Newbury Park, CA: Sage.

Brunner, M., Keller, U., Hornung, C., Reichert, M., & Martin, R. (2009). The cross-cultural generalizability of a new structural model of academic self-concepts. *Learning and Individual Differences, 19*(4), 387-403.

Cassady, J. C. (2004). The influence of cognitive test anxiety across the learning–testing cycle. *Learning and Instruction, 14*(6), 569-592.

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology, 27*(2), 270-295.

Chapell, M. S., Blanding, Z. B., Silverstein, M. E., Takahashi, M., Newman, B., Gubi, A., & McCann, N. (2005). Test anxiety and academic performance in undergraduate and graduate students. *Journal of Educational Psychology, 97*(2), 268.

Chen, P., & Zimmerman, B. (2007). A cross-national comparison study on the accuracy of self-efficacy beliefs of middle-school mathematics students. *The Journal of Experimental Education, 75*(3), 221-244.

Cheng, Y., Shao, C., & Lathrop, Q. N. (2016). The mediated MIMIC model for understanding the underlying mechanism of DIF. *Educational and Psychological Measurement, 76*(1), 43-63.

Cheng, Y., Wang, W., & Ho, Y. (2008). Multidimensional rasch analysis of a psychological test with multiple subtests: A statistical solution for the bandwidth–fidelity dilemma. *Educational and Psychological Measurement, 69*(3), 369-388.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233-255.

Chipman, S. F., Krantz, D. H., & Silver, R. (1992). Mathematics anxiety and science careers among able college women. *Psychological Science, 3*(5), 292-295.

Cole, S. R., Kawachi, I., Maller, S. J., & Berkman, L. F. (2000). Test of item-response bias in the CES-D scale: Experience from the New Haven EPESE study. *Journal of clinical epidemiology*, *53*(3), 285-289.

De Ayala, R. J. (2013). *The theory and practice of item response theory*. New York: Guilford Publications.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling*, *8*(3), 430-457.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science, 11*(1), 19-23.

Ferla, J., Valcke, M., & Cai, Y. (2009). Academic self-efficacy and academic self-concept: Reconsidering structural relationships. *Learning and Individual Differences, 19*(4), 499-505.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with mantel-haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29*(4), 278-295.

Fleer, P., Raju, N., & van der Linden, W. (1995). A Monte Carlo assessment of DFIT with dichotomously scored unidimensional tests. *Annual Meeting of the American Educational Research Association, San Francisco,*

Fleishman, J. A., Spector, W. D., & Altman, B. M. (2002). Impact of differential item functioning on age and gender differences in functional disability. *The Journals of Gerontology.Series B, Psychological Sciences and Social Sciences, 57*(5), S275-84.

Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods, 14*(3), 275.

Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling, 16*(4), 625-641.

Freitag, M., & Bauer, P. C. (2013). Testing for measurement equivalence in surveys dimensions of social trust across cultural contexts. *Public Opinion Quarterly, 77*(S1), 24-44.

Ganotice, F. A., Bernardo, A. B., & King, R. B. (2012). Testing the Factorial Invariance of the English and Filipino Versions of the Inventory of School Motivation With Bilingual Students in the Philippines. *Journal of Psychoeducational Assessment*, *30*(3), 298-303.

Gore, P. A. (2006). Academic self-efficacy as a predictor of college outcomes: Two incremental validity studies. *Journal of Career Assessment, 14*(1), 92-115.

Grisay, A., & Monseur, C. (2007). Measuring the equivalence of item difficulty in the various versions of an international test. *Studies in Educational Evaluation*, *33*(1), 69-86.

Guay, F., Marsh, H. W., & Boivin, M. (2003). Academic self-concept and academic achievement: Developmental perspectives on their causal ordering. *Journal of Educational Psychology, 95*(1), 124.

Hambelton, R., & Swaminathan, H. (1985). Item response theory: Principles and application. Boston, MA: Kluwer-Nijhoff.

Hartig, J., & Höhler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation, 35*(2), 57-63.

He, J., & van de Vijver, F. (2012). Bias and equivalence in cross-cultural research. *Online Readings in Psychology and Culture, 2*(2), 8.

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research, 58*(1), 47-77.

Ho, H., Senturk, D., Lam, A. G., Zimmer, J. M., Hong, S., Okamoto, Y., Chiu S., Nakazawa, Y., Wang, C. (2000). The Affective and Cognitive Dimensions of Math Anxiety: A Cross-National Study. *Journal for Research in Mathematics Education*, *31*(3), 362–379.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, *6*(1), 1-55.

Huang, C. (2011). Self-concept and academic achievement: A meta-analysis of longitudinal relations. *Journal of School Psychology, 49*(5), 505-528.

Huberty, T. J. (2012). Anxiety and depression in children and adolescents: Assessment, intervention, and prevention. Springer Science & Business Media.

Jameson, M. M. (2013). The Development and Validation of the Children's Anxiety in Math Scale. *Journal of Psychoeducational Assessment*, *31*(4), 391-395.

Jameson, M. M., & Fusco, B. R. (2014). Math Anxiety, Math Self-Concept, and Math Self-Efficacy in Adult Learners Compared to Traditional Undergraduate Students. *Adult Education Quarterly*, *64*(4), 306-322.

Jekauc, D., Völkle, M., Wagner, M. O., Mewes, N., & Woll, A. (2012). Reliabilität, Validität und Messinvarianz der deutschen Version der Physical Activity Enjoyment Scale. In *Sportpsychologische Kompetenz und Verantwortung* (p. 86).

Jin, Y., Myers, N. D., Ahn, S., & Penfield, R. D. (2013). A Comparison of Uniform DIF Effect Size Estimators Under the MIMIC and Rasch Models. *Educational and Psychological Measurement*, *73*(2), 339-358.

Jones, R. N. (2006). Identification of measurement differences between English and Spanish language versions of the Mini-Mental State Examination: detecting differential item functioning using MIMIC modeling. *Medical care*, *44*(11), S124-S133.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association, 70*(351a), 631-639.

Kankaraš, M., & Moors, G. (2014). Analysis of Cross-Cultural Comparability of PISA 2009 Scores. *Journal of Cross-Cultural Psychology*, *45*(3), 381-399.

Kim, E. S. (2011). Testing measurement invariance using MIMIC: Likelihood ratio test and modification indices with a critical value adjustment (Doctoral dissertation, TEXAS A&M UNIVERSITY).

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling, 18*(2), 212-228.

Komarraju, M., & Nadler, D. (2013). Self-efficacy and academic achievement: Why do implicit beliefs, goals, and effort regulation matter? *Learning and Individual Differences, 25*, 67-72.

Koomen, H. M., Verschueren, K., van Schooten, E., Jak, S., & Pianta, R. C. (2012). Validating the student-teacher relationship scale: Testing factor structure and measurement invariance across child gender and age in a Dutch sample. *Journal of School Psychology, 50*(2), 215-234.

Krinzinger, H., Kaufmann, L., & Willmes, K. (2009). Math anxiety and math ability in early primary school years. *Journal of Psychoeducational Assessment, 27*(3), 206-225.

Lee, J. (2009). Universals and specifics of math self-concept, math self-efficacy, and math anxiety across 41 PISA 2003 participating countries. *Learning and Individual Differences, 19*(3), 355-365.

Lee, W., Lee, M., & Bong, M. (2014). Testing interest and self-efficacy as predictors of academic self-regulation and achievement. *Contemporary Educational Psychology, 39*(2), 86-99.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Routledge.

Ma, X. (1999). A Meta-Analysis of the Relationship between Anxiety toward Mathematics and Achievement in Mathematics. *Journal for Research in Mathematics Education*, *30*(5), 520-40.

MacPhee, D., Farro, S., & Canetto, S. S. (2013). Academic self-efficacy and performance of underrepresented STEM majors: Gender, ethnic, and social class patterns. *Analyses of Social Issues and Public Policy, 13*(1), 347-369.

Maloney, E. A., & Beilock, S. L. (2012). Math anxiety: Who has it, why it develops, and how to guard against it. *Trends in Cognitive Sciences, 16*(8), 404-406.

Marsh, H. W. (1990). Causal ordering of academic self-concept and academic achievement: A multiwave, longitudinal panel analysis. *Journal of Educational Psychology, 82*(4), 646－656.

Marsh, H. W., Abduljabbar, A. S., Abu-Hilal, M. M., Morin, A. J., Abdelfattah, F., Leung, K. C., ... & Parker, P. (2013). Factorial, Convergent, and Discriminant Validity of TIMSS Math and Science Motivation Measures: A Comparison of Arab and Anglo-Saxon Countries. *Journal of Educational Psychology*, *105*(1), 108-128.

Marsh, H. W., Hau, K., & Kong, C. (2002). Multilevel causal ordering of academic self-concept and achievement: Influence of language of instruction (English compared with Chinese) for Hong Kong students. *American Educational Research Journal, 39*(3), 727-763.

Marsh, H. W., & Martin, A. J. (2011). Academic self-concept and academic achievement: Relations and causal ordering. *British Journal of Educational Psychology, 81*(1), 59-77.

Marsh, H. W., & Craven, R. G. (2006). Reciprocal effects of self-concept and performance from a multidimensional perspective: Beyond seductive pleasure and unidimensional perspectives. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science, 1*(2), 133-163.

McDonald, R. P. (1967). Numerical methods for polynomial models in nonlinear factor analysis. *Psychometrika, 32*(1), 77-112.

McMullan, M., Jones, R., & Lea, S. (2012). Math anxiety, self-efficacy, and ability in british undergraduate nursing students. *Research in Nursing & Health, 35*(2), 178-186.

McMullan, M., Jones, R., & Lea, S. (2012). Math anxiety, self-efficacy, and ability in british undergraduate nursing students. *Research in Nursing & Health, 35*(2), 178-186.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research, 39*(3), 479-515.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54*(4), 557-585.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of mathematical Psychology*, *47*(1), 90-100.

Oishi, S. (2006). The concept of life satisfaction across cultures: An IRT analysis. *Journal of Research in Personality, 40*(4), 411-423.

OECD. (2013a). PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy. PISA: OECD Publishing.

OECD. (2013b). PISA 2012 Results: Ready to Learn – Students' Engagement, Drive and Self-Beliefs (Volume III). PISA: OECD Publishing.

OECD. (2014). Results in focus: What 15-year-olds know and what they can do with what they know. PISA: OECD Publishing.

Pajares, F. (2003). Self-efficacy beliefs, motivation, and achievement in writing: A review of the literature. *Reading &Writing Quarterly*, *19*(2), 139-158.

Penfield, R. D. (2010). Explaining crossing DIF in polytomous items using differential step functioning effects. *Applied Psychological Measurement*, *34*(8), 563-579.

Pérez, E. O. (2011). The origins and implications of language effects in multilingual surveys: A MIMIC approach with application to Latino political attitudes. *Political Analysis, 19*(4), 434-454.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*(2), 197-207.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87*(3), 517.

Raju, N. S., Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied psychological measurement*, *19*(4), 353-368.

Rana, R. A., & Mahmood, N. (2010). The relationship between test anxiety and academic achievement. *Bulletin of Education and Research, 32*(2), 63-74.

Reckase, M. (2009). *Multidimensional item response theory*. New York: Springer.

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin, 138*(2), 353.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, *17*(2), 105-116.

Runyan, R. C., Ge, B., Dong, B., & Swinney, J. L. (2012). Entrepreneurial orientation in cross-cultural research: Assessing measurement invariance in the construct. *Entrepreneurship Theory and Practice, 36*(4), 819-836.

Runyan, R. C., Ge, B., Dong, B., & Swinney, J. L. (2012). Entrepreneurial orientation in cross-cultural research: Assessing measurement invariance in the construct. *Entrepreneurship Theory and Practice, 36*(4), 819-836.

Samejima, F. (1997). Graded response model. In W. J. v. d. Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York: Springer.

Savickas, M. L., & Porfeli, E. J. (2012). Career adapt-abilities scale: Construction, reliability, and measurement equivalence across 13 countries. *Journal of Vocational Behavior, 80*(3), 661-673.

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review, 18*(4), 210-222.

Schunk, D. H., & Hanson, A. R. (1985). Peer models: Influence on children's self-efficacy and achievement. *Journal of Educational Psychology*, *77*(3), 313.

Shavelson, R. J., Hubner, J. J., & Stanton, G. C. (1976). Self-concept: Validation of construct interpretations. *Review of Educational Research, 46*(3), 407-441.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*(2), 159-194.

Shih, C., & Wang, W. (2009). Differential item functioning detection using the multiple indicators, multiple causes method with a pure short anchor. *Applied Psychological Measurement, 33*(3), 184-199.

Tayeb, M. (1994). Organizations and national culture: Methodology considered. *Organization Studies, 15*(3), 429-445.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer H. I. Braun (Ed.), *Test Validity* (pp. 147-172). Hillsdale, NJ: Erlbaum.

Valentine, J. C., DuBois, D. L., & Cooper, H. (2004). The relation between self-beliefs and academic achievement: A meta-analytic review. *Educational Psychologist, 39*(2), 111-133.

Van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 54*(2), 119-135.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4-70.

Wang, C., & Nydick, S. W. (2015). Comparing two algorithms for calibrating the restricted non-compensatory multidimensional IRT model. *Applied Psychological Measurement*, *39*(2), 119-134.

Wang, W. C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *The Journal of Experimental Education*, *72*(3), 221-261.

Wang, W., & Shih, C. (2010). MIMIC methods for assessing differential item functioning in polytomous items. *Applied Psychological Measurement, 34*(3), 166-180.

Wang, W. C., Shih, C. L., & Yang, C. C. (2009). The MIMIC method with scale purification for detecting differential item functioning. *Educational and Psychological Measurement*.

Wang, J., & Wang, X. (2012). Structural equation modeling: Applications using Mplus. John Wiley & Sons.

Wilkins, J. L. (2004). Mathematics and science self-concept: An international investigation. *The Journal of Experimental Education, 72*(4), 331-346.

Wirth, R., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12*(1), 58.

Woods, C. M. (2009a). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, *33*(1), 42-57.

Woods, C. M. (2009b). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research, 44*(1), 1-27.

Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement, 35*(5), 339-361.

Woods, C. M., Oltmanns, T. F., & Turkheimer, E. (2009). Illustration of MIMIC-model DIF testing with the schedule for nonadaptive and adaptive personality. *Journal of Psychopathology and Behavioral Assessment, 31*(4), 320-330.

Yeung, A. S., & Lee, F. L. (1999). Self-concept of high school students in china: Confirmatory factor analysis of longitudinal data. *Educational and Psychological Measurement, 59*(3), 431-450.

Yoon, M., & Millsap, R. E. (2007). Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo study. *Structural Equation Modeling, 14*(3), 435-463.

Yoshino, A. (2012). The relationship between self-concept and achievement in TIMSS 2007: A comparison between American and Japanese students. *International Review of Education, 58*(2), 199-219.

Young, C. B., Wu, S. S., & Menon, V. (2012). The neurodevelopmental basis of math anxiety. *Psychological Science, 23*(5), 492-501.

Yu, Y. F., Yu, A. P., & Ahn, J. (2007). Investigating differential item functioning by chronic diseases in the SF-36 health survey: A latent trait analysis using MIMIC models. *Medical Care, 45*(9), 851-859.

Zeidner, M. (2010) Test anxiety. In I. B. Weiner & E. Craighead (Ed.), *Corsini's Encyclopaedia of Psychology (*4th ed., pp. 1766-1768*)*. New York: Wiley.

Zeidner, M. (1998). Test anxiety: The state of the art. New York: Plenum.

Zeidner, M., & Safir, M. P. (1989). Sex, ethnic, and social differences in test anxiety among Israeli adolescents. *The Journal of Genetic Psychology, 150*(2), 175-185.

Zuffianò, A., Alessandri, G., Gerbino, M., Kanacri, B. P. L., Di Giunta, L., Milioni, M., & Caprara, G. V. (2013). Academic achievement: The unique contribution of self-efficacy beliefs in self-regulated learning beyond intelligence, personality traits, and self-esteem. *Learning and Individual Differences, 23*, 158-162.

Appendix A. Data Management in Stata

```
******* PISA 2012 student file
use "PISA Student 2012.dta"
***QCN is Shanghai(China),and we just keep Shanghai and US for futhure comparative study
drop if CNT!="QCN" & CNT!="USA"
********  recoding self-efficacy so that strong agree is 4, agree is 3, disagree is 2 and
strongly
*******  disagree is 1. 7 N/A, 8 Invalid and 9 Missing are treated as missing data
foreach efficacy of varlist  ST37Q01-ST37Q08{
label variable `efficacy' "self efficacy"
}
local efficacy " ST37Q01-ST37Q08"
rename `efficacy'  (ef1 ef2 ef3 ef4 ef5 ef6 ef7 ef8)
recode ef1-ef8 (4=1) (3=2) (2=3) (1=4)(7=.)(8=.)(9=.), pre(new)test
*******************recoding and renaming mathematics anxiety variables
foreach anxiety of varlist  ST42Q01 ST42Q03 ST42Q05 ST42Q08 ST42Q10{
recode `anxiety' (7=.)(8=.)(9=.)
label variable `anxiety' "math anxiety"
}
rename ST42Q01 anx1
rename ST42Q03 anx2
rename ST42Q05 anx3
rename ST42Q08 anx4
rename ST42Q10 anx5
****** recoding and renaming self-concept variables
foreach concept of varlist  ST42Q02 ST42Q04 ST42Q06 ST42Q07 ST42Q09{
label variable `concept' "self concept"
}
rename ST42Q02 con1
rename ST42Q04 con2
rename ST42Q06 con3
rename ST42Q07 con4
rename ST42Q09 con5
recode "con1" (7=.)(8=.)(9=.), gen(newcon1)
recode con2 con3 con4 con5  (4=1) (3=2) (2=3) (1=4)(7=.)(8=.)(9=.), pre(new)test
**** rename gender and in-class mathematics class periods
rename ST04Q01 gender
rename ST70Q02 mathhours
***** recoding gender (female=1, male=0)
recode gender (1=1) (2=0)
******recoding country Shanghai=1, US=0
encode CNT, gen(CNT1)
codebook CNT1
recode CNT1 (1=1) (2=0)
***** recoding out-of-school study hours & in-class mathematics class periods
```

```
recode " mathhours" (9997=.) (9998=.)(9999=.)
recode "OUTHOURS" (9997=.)(9998=.)(9999=.)
save "PISA_2012_China-US_clean.dta",replace
clear
**** change Stata file into Mplus file
findit stata2mplus
stata2mplus using "PISA_2012_China-US_clean data.dta"
```

Appendix B. Mplus Code for Detection of DIF Effect Size

```
Variable:
     USEVAR=CNT1 newef1 newef2 newef3 newef4 newef5 newef6 newef7
newef8 newcon1 newcon3 newcon4 newcon5 newcon2 anx1 anx2 anx3 anx4 anx5;
     CATEGORICAL= newef1 newef2 newef3 newef4 newef5 newef6 newef7
newef8 newcon1 newcon3 newcon4 newcon5 newcon2 anx1 anx2 anx3 anx4 anx5;
     Missing are all (-9999);

ANALYSIS: ESTIMATOR= MLR;

MODEl:
    efficacy by newef1-newef8;

    anxiety by anx1;
    anxiety by anx2;
    anxiety by anx3;
    anxiety by anx4;
    anxiety by anx5;

    concept by newcon1;
    concept by newcon2;
    concept by newcon3;
    concept by newcon4;
    concept by newcon5;

    efficacy on CNT1;
    concept on CNT1;
    anxiety on CNT1;

    newef2-newef5 on CNT1;
    newef6 on CNT1;
    newef8 on CNT1;

    newcon2 on CNT1;
    newcon4 on CNT1;

    anx1 on CNT1;
    anx2 on CNT1;
    anx4 on CNT1;
    anx5 on CNT1;
```

Appendix C. MIMIC Model with Out-of-school study hours as the Mediator

```
Variable:
     Missing are all (-9999);
     USEVAR=  CNT1 newef1 newef2
     newef3 newef4 newef5 newef6 newef7 newef8 newcon1 newcon3 newcon4
     newcon5 newcon2 anx1 anx2 anx3 anx4 anx5  OUTHOURS;
     CATEGORICAL= newef1 newef2
     newef3 newef4 newef5 newef6 newef7 newef8 newcon1 newcon3 newcon4
     newcon5 newcon2 anx1 anx2 anx3 anx4 anx5;

MODEl:
    efficacy by newef1-newef8;
    anxiety by anx1;
    anxiety by anx2;
    anxiety by anx3;
    anxiety by anx4;
    anxiety by anx5;

    concept by newcon1;
    concept by newcon2;
    concept by newcon3;
    concept by newcon4;
    concept by newcon5;

    efficacy on CNT1;
    concept on CNT1;
    anxiety on CNT1;

    OUTHOURS on CNT1;

    newef4-newef5 on CNT1;
    newef4-newef5 on OUTHOURS;
    newef6 on CNT1;
    newef6 on OUTHOURS;
    newef8 on CNT1;
    newef8 on OUTHOURS;

    newcon2 on CNT1;
    newcon2 on OUTHOURS;
    newcon4 on CNT1;
    newcon4 on OUTHOURS;

    anx5 on CNT1;
    anx5 on OUTHOURS;

    MODEL INDIRECT:
```

```
newef4 ind CNT1;
newef5 ind CNT1;
newef6 ind CNT1;
newef8 ind CNT1;

anx5 ind CNT1;

newcon2 ind CNT1;
newcon4 ind CNT1;
```

Appendix D. MIMIC Model with In-school Math Class Periods as the Mediator

```
 Variable:
     Missing are all (-9999);

     USEVAR= CNT1 newef1 newef2
     newef3 newef4 newef5 newef6 newef7 newef8 newcon1 newcon3 newcon4
     newcon5 newcon2 anx1 anx2 anx3 anx4 anx5  mathhours;

     CATEGORICAL= newef1 newef2
     newef3 newef4 newef5 newef6 newef7 newef8 newcon1 newcon3 newcon4
     newcon5 newcon2 anx1 anx2 anx3 anx4 anx5;


MODEl:
    efficacy by newef1-newef8;

    anxiety by anx1;
    anxiety by anx2;
    anxiety by anx3;
    anxiety by anx4;
    anxiety by anx5;

    concept by newcon1;
    concept by newcon2;
    concept by newcon3;
    concept by newcon4;
    concept by newcon5;
    efficacy on CNT1;
    concept on CNT1;
    anxiety on CNT1;

    mathhours on CNT1;

    newef4-newef5 on CNT1;
    newef4-newef5 on mathhours;
    newef6 on CNT1;
    newef6 on mathhours;
    newef8 on CNT1;
    newef8 on mathhours;

    newcon2 on CNT1;
    newcon2 on mathhours;
    newcon4 on CNT1;
    newcon4 on mathhours;

    anx5 on CNT1;
```

```
anx5 on mathhours;

MODEL INDIRECT:

newef4 ind CNT1;
newef5 ind CNT1;
newef6 ind CNT1;
newef8 ind CNT1;

anx5 ind CNT1;

newcon2 ind CNT1;
newcon4 ind CNT1;
```