INVARIANT MEASUREMENT AND THE ASSESSMENT OF AFFECTIVE VARIABLES INTERNATIONAL STUDIES OF MATHEMATICS

by

CIGDEM TOPTAS

(Under the Direction of George Engelhard, Jr.)

ABSTRACT

The goal of this dissertation is to investigate the measurement invariance related to affective variables used in international educational research. Large-scale test assessments, such as PISA, require invariance in order to compare across countries. One approach to examining invariance is based on residual analyses that examine the differences between the observed data and the expected values based on a measurement model. In addition to examining item fit, this study stresses the description of methods for evaluating person fit for affective scales including proposed substantive interpretations for person misfit. Data from four countries (6,856 from Chile, 6,351 from Japan, 4,848 from Turkey, 4,978 from United States) are used to illustrate the framework. The mathematics self-efficacy, anxiety, self-concept, and behavior scale were used, and students responded in four categories. This dissertation explored model-data fit of both items and persons evaluated from the perspective of Rasch measurement theory using data from PISA 2012.

INDEX WORDS: Rasch measurement theory, item response theory, psychometrics, item-person fit, model-data fit, differential item functioning, international research, affective variables

INVARIANT MEASUREMENT AND THE ASSESSMENT OF AFFECTIVE VARIABLES IN INTERNATIONAL STUDIES OF MATHEMATICS

by

CIGDEM TOPTAS

BA, Uludag University, 2015

M.Sc, Uludag University, 2017

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSPHY

ATHENS, GEORGIA

2023

INVARIANT MEASUREMENT AND THE ASSESSMENT OF AFFECTIVE VARIABLES INTERNATIONAL STUDIES OF MATHEMATICS

by

CIGDEM TOPTAS

Major Professor:	George Engelhard, Jr.
Committee:	Allan S. Cohen
	Matthew J. Madison
	Kristen Bub

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
August 2023

DEDICATION

I dedicate this dissertation to my family. Without their support, I could never achieve anything I chose to do. I love them without measure. Thank you to my advisor Prof. Dr. Engelhard and to my diligent professors for their guidance, patience, and endless help. To express my thanks to all those contributed in many ways to the success of this dissertation and made it an unforgettable experience for me.

ACKNOWLEDGEMENTS

I would like to acknowledge and give my warmest thanks to my supervisor Professor Engelhard who made this dissertation possible. His guidance and advice carried me through all the stages of writing my dissertation.

I would also like to give special thanks to my committee members for letting my defense be an unforgettable moment, and for their brilliant comments and suggestion, thanks to you. I am also grateful to my classmates, and my office mates for their help and moral support.

Finally, I must express my very profound gratitude to my parents and to my fiancé for providing me with unfailing support and continues encouragement through my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

This chapter introduces several key concepts related to this dissertation. First of all, an overview of international assessments is presented. Next, selected affective variables that are included in this study are described. This is followed by a brief introduction to the importance of invariant measurement in educational research. Finally, the purpose and the structure of the dissertation are summarized.

1.1 INTERNATIONAL ASSESSMENT

In today's world, there is great interest in student learning, and the investigation of educational processes in different cultures and societies. International assessments provide the opportunity to take a closer look at educational systems, and the explore the reasons that affect student success in different cultures. The academic achievement of the students is a major focus for these international studies. International assessments provide analyses regarding how students are performing in a range of disciplines at various ages and grade levels.

According to Marshall (2014), comparing educational system across countries can help researchers better understand each education system, improve curriculum awareness, strengthen schools and colleges, examine the influence of education on society, and address problems related to educational issues. To sum up, international studies provide a distinctive frame of reference for understanding performance patterns of students, schools, and education system across the countries. Moreover, it allows researchers to investigate trends in skills and knowledge of students in participating countries.

International assessments are implemented and designed by several international institutions, such as the Organization for Economic Co-operation and Development (OECD), the International Association for the Evaluation of Educational Achievement (IAE), and the United Nations Educational, Scientific and Cultural Organization (UNESCO). The Programme for International Student Assessment (PISA), The Trends in International Mathematics and Science Study (TIMSS), and Progress in International Reading Literacy Study (PIRLS) are some examples of assessments with an international focus. With these assessments, it is essential to gather consistent information about schools, students, and teachers in participating countries to enhance student learning and international comparisons. By analyzing these test results, researchers have information about students' school experience, interests, and learning environment. This information from international data has the potential to help policymakers. In my dissertation, I examine four countries: Japan, Turkey, US, and Chile. These countries are chosen based on variation in their past mathematics performance, and to reflect diversity in culture. In PISA 2012, Japan was a high mathematics performing country, while the US was close to average mathematics performance. Turkey has below-average mathematics performance, while Chile can be considered as a low mathematics performer.

International assessments can be effective as a benchmarking tool and for highlighting the extent of the achievement gap across nations for researchers. However, international assessments have a limited capacity to provide explanations for these phenomena to policymakers (Jerrim, 2015). The effectiveness of international assessments and evidence-based policy to impact educational administration is not only

dependent on their capacity to rationally describe and modify aspects of educational processes. For example, conceptions of whether students have an excellent teacher, or they are in the high-performing school or schooling system, and international best practices can be explored. International assessments have a significant role in shaping how education is conceptualized (Lewis & Lingard, 2015). Although the findings are important, they need to be examined before implementation in order to keep some factors in mind, such as different cultural, socio-economic and environmental factors.

The next section briefly describes PISA because this is the international assessments used in this dissertation.

1.2 PROGRAM OF INTERNATIONAL STUDENT ASSESSMENT (PISA)

In this dissertation, the focus is on PISA which is the Organization for Economic Cooperation and Development (OECD)'s Programme for International Student Assessment which is the one of the major international assessment programs. The first PISA assessments were administered in 2000 in 32 countries. The OECD encourages the use of data from the internationally administered PISA database (https://www.oecd.org/pisa) that includes surveys, performance data, and instructions. PISA surveys take place every three years. In PISA, students are at the same age group, which is 15 years old, and enrolled in grades 7 or above. In PISA, countries administered a student, parent, and school questionnaires. Item Response Theory (IRT)-based scaling models are used in PISA. IRT analysis is used to create the stated proficiency measures, which serve as the fundamental instrument for reporting PISA results. Data is collected on the subject areas of mathematical literacy, science literacy, and reading skills, data about students' motivations, opinions about themselves, learning styles, school

environments and family background. PISA provides information about how well a country's education system is doing, and the improvements that are needed, and the standards should be met. The PISA 2012 survey focuses on mathematics.

Mathematics should not be considered only as a subject. It plays a role in helping young people gain different perspectives on their daily problems and mistakes in their personal lives (OECD, 2013). Mathematics achievement also plays an important role in students' career path. For example, when students have a positive attitude toward mathematics, they believe that mathematics is an important subject which can help them in their career.

Four affective variables are examined in this study mathematics self-efficacy, mathematics anxiety, mathematics self-concept, and mathematics behavior. These affective variables are as follows:

- Mathematics self-efficacy refers to a student's confidence when student answering mathematics questions and overcome difficulties. The PISA mathematics self-efficacy scale was constructed based on student's perceptions of their capacity to solve a variety of pure and applied mathematics problems.
- Mathematics anxiety refers to a student's feelings about themselves in regard to mathematics. The PISA mathematics anxiety scale was constructed based on how much stress and helplessness students reported experiencing when dealing with mathematics.
- Mathematics self-concept refers to a student's beliefs in their mathematics skills. The PISA mathematics self-concept scale was constructed based on students' perceptions of their mathematics abilities.

- Mathematics behavior refers to a student's engagement in mathematics activities. The PISA mathematics behavior was constructed based on student's responses about their involvement in a variety of mathematics activities.

1.3 INVARIANT MEASUREMENT

Invariance can be seen as a key element in science. There should be an objectivity beyond opinions which are supposed to be universal. Measurement refers the process of locating an item and a person on a continuum. Measurement invariance states that the same questionnaire is used in different groups when the same item is measured in the same way. Invariant measurement provides rules and requirements that offer the opportunity to create scales with a set of desired qualities (Engelhard, 2013). Self-reported measurement instruments are often used by researchers which is useful for assessment of latent variables. It can be claimed that latent instrument can be assessed in the same approach if the measuring instrument meets the conditions. For example, a researcher may design a survey to measure population views, and, in this population, there might be various gender, ages, and nationalities. The respondents may reply to questions differently depending on their past experiences, perceptions, and other factors which would influence their response. For this reason, measurement invariance requirements are important and must be met, otherwise the assessment of the comparison would not be valid.

Invariant measurement plays a major role in international assessment research because it provides a framework for investigating the comparability of measures across countries. The disparities between countries, differences in translation of survey items, modification of survey items to a country's context may cause some problems, such as

less comparability across countries and the application of validated measures to groups (OECD, 2020).

In this study, the evaluation of invariance focuses on the fitting of data to models. The goal of item-invariant person measurement is to create a scale consisting of items that can be used to measure an individual's position on a latent variable that is not dependent on certain items in the item sets. On the other hand, the purpose of person-invariant item calibration is to create a set of calibrated items that can be used for measurement that is not dependent on specific individuals or subsets of test takers using item sets. In order to gain benefits from invariant measurement, it is crucial to emphasize the needs of measurement based on ideal-types, which places a focus on developing data structures that satisfy these criteria for invariance.

1.4 PURPOSE OF THE STUDY

The purpose of this study is to use the underlying concepts related to invariant measurement to evaluate assessments in the affective domain. Rasch measurement theory is used to explore invariant measurement. Specifically, the objective is to examine whether or not the measures are comparable across countries. Four variables in the affective domain are examined: mathematics self-efficacy, mathematics anxiety, mathematics self-concept, and mathematics behavior.

CHAPTER 2

LITERATURE REVIEW

In this chapter, related literature is presented and organized as follows. First, a description of international educational assessments is presented. Next, the basic principles of invariant measurement are described. This is followed by a brief discussion of the four affective variables included in this dissertation. An expanded discussion of each affective variable is presented in later chapters that focus on the specific affective variables: mathematics self-efficacy, mathematics anxiety, mathematics self-concept, and mathematics behavior.

2.1 INTERNATIONAL EDUCATIONAL ASSESSMENTS

Education is a crucial component for both economic development and democracy in the modern world. International educational assessment systems offer an environment for governments to compare their policy experiences, coordinate their educational policies, and look for widespread issues in education. Basically, countries that participate in international assessments have evidence to inform policy, technical capacity building, funding and aid, international relations, national politics, economic rationales, and curriculum and pedagogy (Addey & Sellar, 2018).

International educational assessment scales, developed with the use of IRT, describe a construct generally and allow us to generalize beyond the specific items in the assessment to the construct domain those items represent, allowing us to draw conclusions about populations from samples of individuals. These methodologies are

used by The National Assessment of Educational Progress (NAEP; 1969), the International Association for the Evaluation of Educational Achievement (IEA; 1974) and the Organization for Economic Co-operation and Development (OECD; 1948) (Kirsch et al., 2013). These organizations began to conduct assessment cycles. For example, the Trends in Mathematics and Science Study (TIMSS; 1995) and Progress in International Reading Literacy Study (PIRLS) are sponsored by the International Association for the Evaluation of Educational Achievement (IEA), and The Program for International Student Assessment (PISA) is developed and administered under the Organization for Economic Co-operation and Development.

The governments of the participating countries collaborate on PISA surveys because they have similar policy priorities. Each nation and economy represented in the PISA program has a voice on the PISA Governing Board, which makes the final decision on how test results will be used and shared (OECD, 2014). International assessment systems gather a wealth of supplemental data about the instructional setting. Attitudes toward learning, home resources, pedagogical approaches, and school resources are only few of the topics that may be probed using background questionaries given to participating kids, their parents, and related teachers and administrators (Rutkowski & Rutkowski, 2010).

PROGRAMME OF INTERNATINOAL STUDENT ASSESSMENT (PISA)

PISA is the one of the major international assessment programs. The first PISA assessments were administered in 2000 in 32 countries using written tasks completed in schools under carefully monitored test settings. The OECD encourages the use of data from the internationally administered PISA database (https://www.oecd.org/pisa) that

includes surveys, performance data, and instructions on how to utilize data technology approaches. PISA surveys take place every three years. PISA is an age-based survey that evaluates 15-year-olds enrolled in grades 7 or above. In PISA, countries administered a student, parent, and school questionnaires. PISA evaluates students' ability to apply their information in new contexts and expand from what they have learnt in addition to their ability to duplicate knowledge. PISA stresses the need of process proficiency, understanding and adaptability. Item Response Theory (IRT)-based scaling models are used in PISA. IRT analysis is used to create the stated proficiency measures, which serve as the fundamental instrument for reporting PISA results.

Many nations view PISA as a feasible and informative assessment of their progress to reach their educational goals, even if PISA scores sometimes do not match what is thought about the quality of the nation's education systems. PISA has produced a major change in national policy, and the result may be a remarkable increase in educational quality (Ritzen, 2013).

2.2 INVARIANT MEAUSREMENT

Wright (1984) argues that "there have been enough successful theoretical and practical work on the nature and implementation of fundamental measurement to establish its necessity as a basic tool of science and its ready accessibility for educational research" (p. 282).

Several important advances in the area of social measurement have been made in the previous years. Psychometrics has also benefited from technological developments in the field (Mislevy, 2016). Even if developments in methodology and theory have assisted measurement theory, there are still basic difficulties that have not changed much

(Engelhard, 2008). Measurement invariance is a group of methods for determining invariance that helps to understand that an item measures the same trait in all subgroups of a population or measurement conditions (Bulut et al., 2015).

According to Millsap (2011), contrary to expectations, psychological tests are not always accurate and perfect measurements do not exist in most field of psychological measurement. For example, two persons being measured may get different scores from the same test item on the same day or without matching on the qualities, systemic group variations in items scores may be experienced even if the item is fair.

2.3 AFFECTIVE VARIANBLES IN EDUCATION

The affective domain in learning (Krathwohl, Bloom, and Masia, 1973) covers how person interact with things emotionally, such as feelings, perception, attitudes, and motives. According to Grootenboer and Marshman (2016), the affective domain of mathematics represents an individual's personal beliefs, attitudes, and emotions. Therefore, it is crucial that these qualities be displayed, and that affective variables of education, especially mathematical learning, be taken into account.

Affective variables in education are at least as important as cognitive variables. Doruk et al. (2016) investigated the anxiety, attitude, and self-efficacy perceptions of 246 middle school students towards mathematics, as well as the relations between these variables. The findings demonstrated that students had low level of mathematical anxiety, but a high level of positive attitude and self-efficacy towards mathematics. There was a strong relationship between students' anxiety, attitude, and self-efficacy perceptions of mathematics. All of the correlations between mathematical anxiety, attitude toward mathematics, and self-efficacy perception are negative, except for the one between

attitude toward mathematics and self-efficacy perception. Furthermore, it was discovered that mathematical anxiety and student attitude towards mathematics might explain a significant portion of the shift in self-efficacy assessment of students towards mathematics. Ayotola and Adedeji (2009) stated that there is a significant relationship between self-belief and mathematics achievement. Lipnevich at al. (2016) states that two different samples from different cultural backgrounds both found that students' attitudes about mathematics were significant in explaining students' performance in mathematics. From studies regarding the mathematics learning process are affected by cognitive and affective factors.

According to the idea of planned behavior, certain norms and perceived behavioral control influence voluntary conduct. Azjen (1991) states that intentions impact behaviors, which are affected by three factors: attitudes, subjective norms, and perceived behavioral control. Intentions are widely acknowledged to play a significant role between the three components that influence conduct and the behavior itself. According to the idea of planned behavior, intentions also play a mediating role between perceived behavioral control and actual conduct. The model is shown in Figure 1. The three determinants are predicted to have a positive relationship with intentions and behavior, and the construct of intentions is also expected to have a positive relationship with behavior within the theory of planned behavior as a whole. Academic achievement is influenced by the behavior factor.

DIFFERENTIAL ITEM FUNCTIONING

The Rasch model represents an ideal of how a scale should function. Ideals are always violated in practice. However, most violations are inconsequential, even when

they are statistically significant (Linacre, 2012). Differential item functioning (DIF) analyses can be useful in determining when these violations require further attention. When the conditions of the Rasch model are met, it is possible to assess the differences between groups using a DIF analysis. In general, such an analysis identifies a focal group and a reference group. The focal group is the group of interest, and the group which could possibly be disadvantaged by the item. The reference group is used for the purposes of comparison (de Ayala, 2009). A DIF analysis provides insight into how an item is experienced by the reference and focal groups. When no DIF, or non-significant DIF, is detected then group membership does not matter because the items function in a comparable way for everyone. If there is a statistically significant difference, members of the disadvantaged group may feel substantive consequences, depending on the purpose of the scale or test (Linacre, 2012).

In this this dissertation, the theory of planned behavior undergirds the explanations of student responses on affective variables in mathematics in large-scale assessment data by using the PISA 2012. In the international assessment test context, students' attitudes, perceived control, and subjective norms may influence their motivation to spend time on mathematics for example homework or engage in related exercises to reduce anxiety, or feeling comfortable asking questions in the classroom, or their confidence which could improve their mathematics performance. Students' attitudes toward mathematics and their perceptions of their own mathematical abilities are two of the most important determinants of students' choices in mathematics-related areas of study (Hwang and Son, 2021). In PISA 2012, the analytic potential of the research to

explain student effort, student behavior, and, if possible, student results in mathematics are greatly enhanced by using the whole Ajzen model (OECD, 2014).

2.4 ITEM AND PERSON FIT

Rasch analysis of measurement involves approaches for assessing model-data fit. Model-data fit analyses in the framework of Rasch measurement theory (Rasch, 1960) look for evidence to model requirements of item, person and other related response patterns. This method of model-data fit allows us to understand whether there is any difference between observed and expected answers, which refers residuals. Residual analysis indicates how well the responses to the item of interest match the predictions of the model. Engelhard (2013) pointed out that the measurement of persons and the calibration of items must be independent. Based on Engelhard (2013), there are five requirements for invariant measurement: person measurement, item calibration, and Wright map. He specified these requirements as follows:

> The measurement of persons must be independent of the particular items that happen to be used for the measuring: item-invariant measurement of person. A more able person must always have a better chance of success on an item than a less able person: non-crossing person response functions. The calibration of the items must be independent of the particular persons used for calibration: person invariant calibration of test items. Any person must have a better chance of success on an easy item than on a more difficult item: non-crossing item response functions. Person and items must be simultaneously located on a single underlying latent continuum: Wright map (Engelhard, 2013, p.14).

Rasch model fit statistics are used to determine the extent to which the requirements of invariant measurement are satisfied. Infit Mean Square and Outfit Mean Square are the most common indicators used to evaluate model fit. Infit Mean Square and Outfit Mean Square have an expected value of 1.0. When compared to Infit Mean Square, Outfit Mean Square provide less of a hazard to measurement but are simpler to handle. According to Linacre (2012), if a misfit is suspected, replacing the suspect responses as a missing value can be helpful to define statistics change. Item and person fit values between 0.8 and 1.2 have been declared satisfactory. According to Engelhard and Wind (2018), broad fit classifications (A, B, C, and D) provides a framework for interpreting model-data fit. This is shown in Table 1.

2.4 RASCH APPROACH TO DIFFERENTIAL ITEM FUNCTIONING

It is very important when using an assessment with multiple subgroups to evaluate whether or not the items are invariant across subgroups. Different subgroups may respond differently to items, and there might be differences in terms of the percentage of student with comparable achievement levels correctly answering an item. The various approaches used to examine item invariance have been called differential item functioning (DIF). DIF analyses provide insight into this type of unexpected error. The performance of items across a variety of test takers can be assessed using DIF techniques (Holland & Wainer, 2012; Zumbo, 2007). DIF evidence can alert researchers to potential biases in items. It may be beneficial to treat items demonstrating DIF as separate items for various groups if it happens within the Rasch model framework.

In PISA, there are many countries with different cultures and educational systems. Therefore, it is important to check if there is any DIF across countries. This research aims

to evaluate potential DIF across countries in PISA 2012. DIF analyses provides opportunity to check different countries responses in terms of items. Most of the methods for examining DIF compare two subgroups, and there is not general agreement on how to explore DIF with multiple subgroups, such as countries.

A Rasch approach to DIF is described in this section that is suitable for analyzing two or more subgroups of students. This Rasch approach takes advantage of the use of residuals to identify DIF. The Rasch model represents an ideal perspective on how items should function, and ideals are always violated in reality. However, most violations are inconsequential, even when they are statistically significant (Linacre, 2012). DIF analyses can be useful in determining when these violations require further attention. When the requirements of the Rasch model are met, then it is possible to access the differences between groups.

The Rasch approach used in this dissertation consists of several steps:

- Fit the Rasch model with three facets (Student, Item, Country)
- Calculate expected values based on the Rasch model
- Obtain the residual matrix (Student x Item): Observed – Expected Values = Residuals
- Standardize the residuals
- Summarize the residuals by country for items using means
- Check the omnibus test for the overall significance of the interaction effects
- Create summary table of means and Excel plots
- Flag values with an absolute value greater than .30 as potential DIF items

- Describe patterns in residuals

The Rasch model creates a useful framework for comparing observed with expected outcomes. The model predicts that the expected outcome on item for a person is based on item and person locations on the Rasch scale. These residual differences between observed and expected values can be organized into a variety of indices for items. The residuals are differences between observed and expected outcomes. These residuals can be used to conduct DIF analyses. The residual-based approach is very helpful when there are more than two groups. The observed, expected, and residuals matrices are shown in Figure 2.

CHAPTER 3

MATHEMATICS SELF-EFFICAY

This chapter describes the first affective variable: mathematics self-efficacy. This chapter describes the fit statistics for persons and items based on the PISA data set.

3.1 INTRODUCTION

Self-efficacy refers to beliefs about one's performance capabilities. Bandura (1977) explained self-efficacy as a person's confidence in their abilities can determine their decision and reaching their goal. The student's self-efficacy level specifies student's effort and ability when student face with problems. Self-efficacy has been hypothesized to influence a variety of important outcomes (Bandura, 1977). For example, self-efficacy predicts important career access behavior indices such as college-major choices and academic performance (Hackett & Lent, 1992; Akter et al, 2018). Students who have poor self-efficacy may have the misconception that topics or tasks are more difficult than they actually are. Low self-efficacy or lack of confidence causes students to question their potential to succeed and makes them unwilling to engage in learning or studying. For instance, students who are confident in their abilities have a better chance of succeeding (Peters, 2013; Schöber et al., 2018). In recent research, it also shows that self-efficacy has a positive and significant effect on mathematics achievement (Ayotola and Adedeji 2009; Roslan and Maat, 2019; Muhtadi et al., 2022).

Mathematics self-efficacy refers to a student's confidence in their capacity to handle with challenges when completing mathematical tasks. Self-efficacy is a significant

affective factor for students learning mathematics and the students' self-efficacy in mathematics should be high so that they have success of learning process (Masitoh and Fitriyani, 2018). According to research, mathematics self-efficacy has been defined by researchers as the degree to which students feel competence in the subject of mathematics (Cheema & Kitsantas, 2014; Toland & Usher, 2016). Lee (2008) analyzed mathematics self-concept, mathematics self-efficacy, and mathematics anxiety and those are highly related self-constructs. These constructs are found within- and between-country levels. It can be seen that mathematics self-efficacy has positively related to the math performance both at the between- and within- country levels. Even if studies show that there is a positive relationship between performance and self-efficacy, different contribution could be occurred. It would be good to focus on students' pattern for each country.

RASCH MODEL

The Rasch measurement model was used to calibrate the items in the scale. Persons and items are ordered according to location on the construct of self-efficacy in mathematics. The log odds of a person endorsing an item for Model I can be expressed as follows:

$$\ln\left[\frac{P_{nijk}}{P_{nijk-1}}\right] = \theta_n - \delta_i - \lambda_j - \tau_k \qquad (1)$$

where:

P_{nijk} = probability of student *n* providing a rating *k* on item *i* from country j,

P_{nijk-1} = probability of student *n* providing a rating *k-1* on item *i* from country j,

θ_n = logit-scale location of student *n,*

δ_i = logit-scale location of item i*,*

λ_j = logit-scale location of country *j,* and

τ_k = difficulty of category *k* relative to category *k*-1.

The log of the odds that a student gives a rating in category *k* rather than in category *k – 1* is estimated given student locations on the self-efficacy in mathematics. The student location is based on the locations of the items, countries, and category item difficulties. The item parameter τ_k reflects the structure of the rating scale, and it is not considered as a facet in the model. Equation 2 includes the interaction term, $\delta_i\lambda_j$, that can be estimated after the model in Equation 1 is held constant:

$$\ln\left[\frac{P_{nijk}}{P_{nijk-1}}\right] = \theta_n - \delta_i - \lambda_j - \delta_i\lambda_j - \tau_k \quad (2)$$

It should be noted that the model in Equation 2 examines whether or not items function in the same way across countries.

The most common measures used to diagnosis fit to the Rasch model are Infit and Outfit Mean Squares. According to Linacre (2012), Infit Mean Square, also known as an inlier-sensitive or information-weighted fit, is more sensitive to how a person answers questions about themselves when compared to other methods. Outfit is a fit that is especially sensitive to outlying data, such as reactions to items that are difficult to measure from a distance. Both mean square error statistics have an expected value of 1.0. Values less than 1.0 suggest observations are too predictable whereas greater than 1.0 suggest observations are too unpredictable. Therefore, it can be said that small Infit values indicate over fit to the Rasch model, while values greater than 1.0 may indicate an aberrant pattern. Generally, acceptable model-data fit is defined as a fit statistic between 0.5 and 1.5. In this study, we used guidelines suggested by Engelhard and Wind (2018).

PURPOSE

The purpose of this study is to investigate invariance in a mathematics self-efficacy scale using both item and person fit for Chile, Japan, Turkey, and US.

These countries were selected because they represented a diversity of cultures. This study addresses the following questions:

1. Does the mathematics self-efficacy scale function in a comparable way across countries?
2. Are there differences in item functioning between countries in terms of mathematics self-efficacy?
3. Are there any items and persons who tend to misfit related to country?

3.2 METHODOLOGY

PARTICIPANTS

For this study, a paper-based assessment data set from PISA 2012 international database was used for this investigation. This data set includes affective and other variables about students. In this file, they offer all student responses for several variables. It should be note that if student responses are missing for all items, then these responses were removed from data set. The mathematics self-efficacy scale used in PISA 2012 consists of eight items. Students were asked "how confident do you feel having to do mathematics tasks?" and response categories included "very confident", "confident", "not very confident", "not at all confident" answers. These items are shown in Table 2.

This study examines mathematics self-efficacy of 15-year-old children from Turkey, the United States, Japan, and Chile towards mathematics in the PISA 2012. This study includes 23,033 students (11,677 boys and 11,356 girls) who participated in PISA 2012. There were 4,978 participants from United States, 4,848 participants from Turkey, 6,351 participants from Japan, and 6,856 participants from Chile. The amount of missing

data was similar across countries: Turkey (34.2%), Japan (34.4%), US (34.5%), and Chile (34%).

Rasch measurement theory was used to calibrate and examine differential item, person functioning on the mathematics self-efficacy scale. The analyze was programmed in FACETs that is specialized software for Rasch analyses.

3.3 RESULTS

The three-facet Rasch model (person, items, and country) explained 47.14% of the variance in the data, and this supports the inference that the scale is unidimensional (Bond and Fox, 2015). Rasch summary statistics for the data are presented in Table 3. Overall, Infit Mean Square and Outfit Mean Square statistics were good for items, and countries. Persons had an average self-efficacy level of 1.15 ($SD = 1.68$) logits. Items had an average difficulty of 0.00 ($SD = .74$) logits. Countries had an average self-efficacy of 0.00 ($SD = .15$) logits. The reliability of separation was high for students [$R = 0.81$; χ^2 (15135) = 60594.3, $p < .05$]. The reliability for students corresponds to the traditional coefficient alpha in classical test theory. The items [$R > .99$, χ^2 (7) = 21396.7, $p < .05$] and countries [$R > .99$, χ^2 (3) = 797.5, $p < .05$] also had a significant reliability of separation index (Table 3).

Infit Mean Square and Outfit Mean Square statistics are given, as well as classifications of the items for fit statistics. The categories are based on recommendations made by Engelhard and Wind (2018) for interpreting the fit statistics calculated in a Rasch context. Chile, Turkey, Japan, and US were in the same category which is category A (Table 4). Infit Mean Square was the lowest for the United States [Infit Mean Square =

0.89], and the highest for Turkey [Infit Mean Square =1.13]. Outfit Mean Square was close to expected value and fell within category A for all countries.

Table 5 shows that in all countries the majority of student fall into category A and that are the least number of student fall into category D. Generally, for all items, Infit Mean Square was close to the expected value, and fell within category A (Table 6). Infit Mean Square was the lowest for the Item 3 [Infit Mean Square = 0.80], and the highest for Item 7 [Infit Mean Square =1.29]. Outfit Mean Square was close to expected value and fell within category A for all items. The Wright Map (Figure 3) is a visual representation of students and item among four countries. The distribution is negatively skewed, with more students falling in the higher on the logit scale (i.e., having higher level of self-efficacy). Chile and Turkey were located in the same line, whereas the United States had the highest and Japan had the lowest score on the map.

DIF RESULTS FOR MATHEMATICS SELF-EFFICACY

Table 7 summarizes the DIF analyses for mathematics self-efficacy. This table shows the mean residual statistics in logits. Positive values indicate that the item is more difficult than expected, while negative values indicate the item is less difficult than expected. Since these units are logits, it is useful to identify interaction effects that have an absolute value greater than .30.

The first step in examining DIF is to examine the overall interaction effects. The data for self-efficacy indicate that the interaction effect explains 3.49% additional variance. There is an overall statistically significant interaction effects, $\chi 2\ (32, 15135) = 6729.3, p < .01$. Table 7 shows the individual mean residuals for each country and item. These effects are also shown graphically in Figure 4.

The data suggest that Item 5 is interpreted differently across countries. Item 5 refers to "Solving an equation like 3x+5 = 17". It is also important to note that Japan has interaction effects greater than absolute value of .30 for the most items: Items 3, 4, 5, 6, 7, and 8. These interaction effects suggest that students in Japan have a different perspective on self-efficacy as measured by these items. Model-data fit guidelines are shown in Table 8.

3.4 DISCUSSION

In this chapter, an approach for examining invariance of a mathematics self-efficacy scale Rasch based on model-data fit was conducted. Differential item functioning (DIF) was conducted based on a residual-based approach within the context of Rasch measurement theory. Most DIF methods focus on comparing two groups (Holland and Wainer, 1993), but Rasch approach can give researchers opportunity to investigate more than two groups at a time. Person fit is also important when examining model-data fit. If there is a good model-data fit for item and person, the benefit of invariant measurement can be obtained.

In the illustrative example on PISA 2012 mathematics self-efficacy, persons, items and countries are examined. Infit and Outfit statistics indicate that the scale function in a comparable way across countries. Persons tend to be in category A which implies good model-data fit. All items have relatively good fit to model-data expectation. However, the data suggest that some items have DIF across countries. For example, item related solving an equation appear to be differently interpreted across countries. Future qualitative research should explore the translation of this items across countries.

CHAPTER 4

MATHEMATICS ANXIETY

This chapter describes the second affective variable: mathematics anxiety. The structure of Chapters 3, 4, 5 and 6 are comparable. Model-data fit statistics and classifications for items and persons are based on Rasch measurement theory using the PISA data set.

4.1 INTRODUCTION

Mathematics is perceived by most students as difficult, and abstract. When student face difficult mathematical issues and challenges, they may experience mathematics anxiety. Mathematics anxiety as defined by Richardson and Suinn (1972) is characterized by "feelings of tension and anxiety that interfere with the manipulation of numbers and the solving of mathematical problems" (p. 551). Mathematics anxiety affects many students, and students may avoid learning mathematics, taking part of mathematics activities, or choosing career in mathematics fields because of this anxiety. Student fear of mathematics may lead to avoidance of mathematics, locally and globally. For example, students may take a short time for mathematics questions or may not choose mathematic related courses (Gabriel et al., 2020). Math anxiety may have negative effects on academic performance. One of the strongest predictors of poor performance in mathematics was the level of mathematics anxiety (Beilock and Maloney, 2015; Gunderson et al., 2018).

There are many factors may affect mathematics anxiety, such as gender, self-awareness, learning challenges and numerical skills (Khasawneh et al., 2021). Moreover, cultural background is the one of many factors that may affect mathematics anxiety (Brown et al., 2020, Fan et al., 2019). According to Lee (2009), students who are less skilled in mathematics tend to have greater rates of math anxiety in higher and lower mathematics performance countries. Mathematics anxiety is not exclusive to a certain group of people (e.g., female or male students) or to any one nation. Cross-cultural comparisons and discussions can help researchers, educators, teachers, parents, and students to understand and reduce mathematics anxiety (Yuan et al., 2022).

RASCH MODEL

The Rasch measurement model was used to calibrate the items in the mathematics anxiety scale. Persons and items are ordered according to location on the construct of anxiety in mathematics. The log odds of a person endorsing an item for Model I can be expressed as follows:

$$\ln\left[\frac{P_{nijk}}{P_{nijk-1}}\right] = \theta_n - \delta_i - \lambda_j - \tau_k \qquad (1)$$

where:

P_{nijk} = probability of student n providing a rating k on item i from country j,

P_{nijk-1} = probability of student n providing a rating k-1 on item i from country j,

θ_n = logit-scale location of student n,

δ_i = logit-scale location of item i,

λ_j = logit-scale location of country j, and

τ_k = difficulty of category k relative to category k-1.

The log of the odds that a student gives a rating in category *k* rather than in category *k – 1* is estimated given student locations on the anxiety in mathematics. The student location is based on the locations of the items, countries, and category item difficulties. The item parameter τ_k reflects the structure of the rating scale, and it is not considered as a facet in the model. Equation 2 includes the interaction term, $\delta_i\lambda_j$, that can be estimated after the model in Equation 1 is held constant:

$$\ln\left[\frac{P_{nijk}}{P_{nijk-1}}\right] = \theta_n - \delta_i - \lambda_j - \delta_i\lambda_j - \tau_k \quad (2)$$

It should be noted that the model in Equation 2 examines whether or not items function in the same way across countries.

The most common measures used to diagnosis fit to the Rasch model are Infit and Outfit Mean Squares. According to Linacre (2012), Infit Mean Square, also known as an inlier-sensitive or information-weighted fit, is more sensitive to how a person answers questions about themselves when compared to other methods. Outfit is a fit that is especially sensitive to outlying data, such as reactions to items that are difficult to measure from a distance. Both mean square error statistics have an expected value of 1.0. Values less than 1.0 suggest observations are too predictable whereas greater than 1.0 suggest observations are too unpredictable. Therefore, it can be said that small Infit values indicate over fit to the Rasch model, while values greater than 1.0 may indicate an aberrant pattern. Generally, acceptable model-data fit is defined as a fit statistic between 0.5 and 1.5. In this study, we used guidelines suggested by Engelhard and Wind (2018).

PURPOSE

The purpose of this chapter is to investigate invariance in a mathematics anxiety scale for both items and persons using Rasch measurement theory in Chile, Japan, Turkey, and US.

These countries were selected because they represented a range in levels of mathematics achievement and a diversity of cultures. This study addresses the following questions:

1. Does the mathematics anxiety scale function in a comparable way across countries?
2. Are there differences in item functioning between countries in terms of mathematics anxiety?
3. Are there any items and persons who tend to misfit related to country?

4.2 METHODOLOGY

PARTICIPANTS

For this chapter, a paper-based assessment data set from PISA 2012 international database was used for this investigation. This data set includes affective and other variables about students. In this file, they offer all student responses for several variables. It should be noted that if student responses are missing for all items, then these responses are removed from data set. The mathematics anxiety scale used in PISA 2012 consists of five items. Students were asked whether they agree with the mathematics anxiety statements. Response categories included “strongly agree”, “agree”, “disagree”, “strongly disagree” answers. These items are shown in Table 9.

This chapter examines mathematics anxiety of 15-year-old children from Turkey, the United States, Japan and Chile towards mathematics in the PISA 2012. This study includes 23,033 students (11,677 boys and 11,356 girls) who participated in PISA 2012. There were 4,978 participants from United States, 4,848 participants from Turkey, 6,351 participants from Japan, and 6,856 participants from Chile. The amount of missing data

was similar across countries: Turkey (34.2%), Japan (34.3%), US (34.5%), and Chile (34.3%).

Rasch measurement theory was used to calibrate and examine differential item, person functioning on the mathematics anxiety scale. The analyze was programmed in FACETs that is specialized software for Rasch analyses

4.3 RESULTS

The three-facet Rasch model (person, items, and country) explained 54.81% of the variance in the data, and this supports the inference that the scale is unidimensional (Bond and Fox, 2015). Rasch summary statistics for the data are presented in Table 10. Overall, Infit Mean Square and Outfit Mean Square statistics were good for items and countries. Persons had an average anxiety level of 0.19 ($SD = 1.99$) logits. Items had an average difficulty of 0.00 ($SD = .00$) logits. Countries had an average anxiety of 0.00 ($SD = .00$) logits. The reliability of separation was high for students [$R = 0.81$; χ^2 (15,118) = 54761.7, $p < .05$]. The reliability for students corresponds to the traditional coefficient alpha in classical test theory. The items [$R > .99$, χ^2 (4) = 17597.3, $p < .05$] and countries [$R > .99$, χ^2 (3) = 365, $p < .05$] also had a significant reliability of separation index (Table 10).

Infit Mean Square and Outfit Mean Square statistics are given, as well as classifications of the items for fit statistics. The categories are based on recommendations made by Engelhard and Wind (2018) for interpreting the fit statistics calculated in a Rasch context. Chile, Turkey, Japan, and US were in the same category which is category A (Table 11). Infit Mean Square was the lowest for the United States [Infit Mean Square

= 0.78], and the highest for Chile [Infit Mean Square =1.12]. Outfit Mean Squares were close to expected value of 1.00 and fell within category A for all countries.

Table 12 shows that in all countries the majority of students fall into category A and that the least number of students fall into category C. Generally, for all items, Infit Mean Square was close to the expected value, and fell within category A (Table13). Infit Mean Square was the highest for Item 5 [Infit Mean Square =1.45]. Outfit Mean Square was close to expected value and fell within category A for all items. The Wright Map (Figure 5) is a visual representation of students and item among four countries. The distribution is negatively skewed, with more students falling in the higher level on the logit scale (i.e., having higher level of math anxiety). Japan and Turkey were located in the same line, whereas the Chile had the highest and United States had the lowest score on the map.

DIF RESULTS FOR MATHEMATICS ANXIETY

Table 14 summarizes the DIF analyses for the mathematics anxiety scale. This table shows the mean residual statistics in logits. Positive values indicate that the item is more difficult than expected, while negative values indicate the item is less difficult than expected. Since these units are logits, it is useful to identify interaction effects that have an absolute value greater than .30.

The first step in examining DIF is to investigate the overall interaction effects. The data for anxiety indicate that the interaction effect explains 2.55% additional variance. There is an overall statistically significant interaction effects, $, \chi 2\ (20, 15119) =\ 3412.6, p < .01$. Table 14 shows the mean residuals for each country and item. These effects are also shown graphically in Figure 6.

The data suggest that Item 5 is interpreted differently for Chile and the US with student answers looking to be opposite of each other. Item 5 refers to "I worry that I will get poor <grades> in mathematics". It may happen because of variation in grading systems within each country in their mathematics education. Moreover, Japan also has a significant interaction effect for Item 5. Items 2 and 4 have significant interaction effects in Chile compared to other countries. In conclusion, Chilean students appear to have a distinct perspective on mathematics anxiety as indicated by these items. Model-data fit guidelines are shown.

4.4 DISCUSSION

This chapter examines the invariance of a mathematics anxiety scale Rasch based on model-data fit. Differential item functioning (DIF) was conducted based on a residual-based approach within the context of Rasch measurement theory. Most current DIF approaches focus on comparing two groups (Holland and Wainer, 1993), but Rasch approach can give researchers opportunity to examine more than two groups at a time. Person fit is also important when examining model-data fit. The advantage of invariant measurement is realized if the model and data match well for both the item and the person.

In the illustrative example on PISA 2012 mathematics anxiety, persons, items and countries are examined. Infit and Outfit statistics suggest that the scale function in a comparable way across countries. Persons tend to be in category A which implies good model-data fit. All items have relatively good fit to model-data expectation. However, the data suggest that some items have DIF across countries. For example, item related getting

poor grade in mathematics found different approach across countries. Future research should explore the grading system across countries.

CHAPTER 5

MATHEMATICS SELF-CONCEPT

This chapter describes another affective variable related to mathematics: mathematics self-concept. This chapter continues the analyses on model-data fit presented in previous chapters.

5.1 INTRODUCTION

Self-concept is a very important human characteristics. It has been described as "person's perception of himself" (Shavelson, et al., 1976, p.411). Even if self-efficacy and self-concept constructs seem similar conceptually, Bandura (1977) has reported that self-concept and self-efficacy are different constructs. They should not be interchanged. Pajares and Miller (1994) argued that self-concept and self-efficacy are not same, and that the self-efficacy is a person's opinion of how well they can do certain things in certain situations. Self-concept on the other hand is not measured in that way, and it includes beliefs about one's worth.

Academic self-concept is of the aspects of self-concept, and it is associated with students' academic performance and their capacity for learning. Mathematics self-concept has to do " with how sure a person is of being able to learn new topics in mathematics, perform well in mathematics class, and do well in mathematics tests" (Reyes, 1984, p. 560). When students believe that their mathematics achievement depend on their dedication on assignment or tasks, they would be willing to be positive self-concept. Students' views of their own mathematics abilities and their perceived competence in

mathematics affect their learning processes and achievement. It is found that self-concept is strongly related with academic achievement (Wu, et al., 2021; Marsh, 1992; Marsh and Craven, 1997), and mathematics achievement (Emmanuel, et al., 2014; Lee & Kung, 2018). Mathematics self-concept has not only an impact on achievement, but it also has an impact on well-being and character development (OECD, 2013). Cvencek et al. (2020) pointed out that student self-concept of mathematics are flexible, and interventions can improve students' perceptions about themselves and mathematics.

According to Ahmed et al. (2012), there is a reciprocal relationship between mathematics self-concept and mathematics anxiety. After investigating 41 countries, Lee (2009) reported that mathematics self-concept and mathematics anxiety have an inverse relationship each other. It is expected that low levels of mathematics self-concept would lead to high mathematics anxiety. Researchers have investigated the differences in self-concept among countries, there is a general tendency of lower self-concept in Asian countries as compared to other countries (Lee, 2009). Therefore, thinking about cross-cultural research in terms of self-concept would be beneficial to explore and whether there is any difference among countries.

RASCH MODEL

The Rasch measurement model was used to calibrate the items in the self-concept scale. Persons and items are ordered according to location on the construct of self-concept in mathematics. The log odds of a person endorsing an item for Model I can be expressed as follows:

$$\ln\left[\frac{P_{nijk}}{P_{nijk-1}}\right] = \theta_n - \delta_i - \lambda_j - \tau_k \qquad (1)$$

where:

P_{nijk} = probability of student *n* providing a rating *k* on item *i* from country j,

P_{nijk-1} = probability of student *n* providing a rating *k-1* on item *i* from country j,

θ_n = logit-scale location of student *n,*

δ_i = logit-scale location of item i*,*

λ_j = logit-scale location of country *j,* and

τ_k = difficulty of category *k* relative to category *k*-1.

The log of the odds that a student gives a rating in category *k* rather than in category *k – 1* is estimated given student locations on the self-concept in mathematics. The student location is based on the locations of the items, countries, and category item difficulties. The item parameter τ_k reflects the structure of the rating scale, and it is not considered as a facet in the model. Equation 2 includes the interaction term, $\delta_i\lambda_j$, that can be estimated after the model in Equation 1 is held constant:

$$\ln\left[\frac{P_{nijk}}{P_{nijk-1}}\right] = \theta_n - \delta_i - \lambda_j - \delta_i\lambda_j - \tau_k \quad (2)$$

It should be noted that the model in Equation 2 examines whether or not items function in the same way across countries.

The most common measures used to diagnosis fit to the Rasch model are Infit and Outfit Mean Squares. According to Linacre (2012), Infit Mean Square, also known as an inlier-sensitive or information-weighted fit, is more sensitive to how a person often answers questions about themselves when compared to other methods. Outfit is a fit that is especially sensitive to outlying data, such as reactions to items that are difficult to measure from a distance. Both mean square error statistics have an expected value of 1.0. Values less than 1.0 suggest observations are too predictable whereas greater than 1.0

suggest observations are too unpredictable. Therefore, it can be said that small Infit values indicate over fit to the Rasch model, while values greater than 1.0 may indicate an aberrant pattern. Generally, acceptable model-data fit is defined as a fit statistic between 0.5 and 1.5. In this chapter, we used the guidelines suggested by Engelhard and Wind (2018).

PURPOSE

The purpose of of this chapter is to investigate invariance in a mathematics self-concept scale using both item and person fit for Chile, Japan, Turkey, and US. These countries were selected because they represented a diversity of cultures. This chapter addresses the following questions:

1. Does the mathematics self-concept scale function in a comparable way across countries?
2. Are there differences in item functioning between countries in terms of student's mathematics self-concept?
3. Are there any items and persons who tend to misfit related to country?

5.2 METHODOLOGY

PARTICIPANTS

For this chapter, a paper-based assessment data set from PISA 2012 international database was used for this investigation. This data set includes affective and other variables about students. In this file, they offer all students' responses for each of variables. It should be note that if student's response missing for all items, then these responses were removed from data set. The self-concept scale used in PISA 2012 consists of five items. Students were asked whether they agree with the mathematics self-concept statements and response categories included "very likely", "likely", "slightly likely", "not at all likely" answers. These items are shown in Table 15.

This chapter examines self-concept of 15-year-old children from Turkey, the United States, Japan and Chile towards mathematics in the PISA 2012. This chapter includes 23,033 students (11,677 boys and 11,356 girls) who participated in PISA 2012. There were 4,978 participants from United States, 4,848 participants from Turkey, 6,351 participants from Japan, and 6,856 participants from Chile. The amount of missing data was similar across countries: Turkey (34.3%), Japan (34.3%), US (34.6%), and Chile (34.2%).

Rasch measurement theory was used to calibrate and examine differential item, person functioning on the mathematics self-concept scale. The analyze was programmed in FACETs that is specialized software for Rasch analyze.

5.3 RESULTS

The three-facet Rasch model (persons, items, and country) explained 62.34% of the variance in the data, and this supports the inference that the scale is unidimensional (Bond and Fox, 2015). Rasch summary statistics for the data are presented in Table 16. Overall, Infit Mean Square and Outfit Mean Square statistics were good for items, and countries. Persons had an average mathematics self-concept level of -0.89 ($SD = 2.12$) logits. Items had an average difficulty of 0.00 ($SD = 1.07$) logits. Countries had an average mathematics self-concept of 0.00 ($SD = .21$) logits. The reliability of separation was high for students [$R = 0.79;\ \chi 2\ (15118) = 65051.7, p < .05$]. The reliability for students corresponds to the traditional coefficient alpha in classical test theory. The items [$R > .99$, $\chi^2\ (4) = 23034.4$, $p < .05$] and countries [$R > .99$, $\chi^2\ (3) = 1033.1$, $p < .05$] also had a significant reliability of separation index (Table 16).

Infit Mean Square and Outfit Mean Square statistics are given, as well as classifications of the items for fit statistics. The categories are based on recommendations made by Engelhard and Wind (2018) for interpreting the fit statistics calculated in a Rasch context. Chile, Turkey, Japan, and US were in the same category which is category A (Table 17). Infit Mean Square was the highest for Turkey [Infit Mean Square = 0.99], and the lowest for the US [Infit Mean Square =0.88]. Outfit Mean Square was close to expected value and fell within category A for all countries.

Table 18 shows that in all countries the majority of student fall into category A and there are the least number of students falling into category C. Generally, for all items, Infit Mean Square was close to the expected value, but item 1 fell within category C (Table 19). Infit Mean Square was the lowest for the Item 3 [Infit Mean Square = 0.61], and the highest for Item 1 [Infit Mean Square =1.96]. Outfit Mean Square was close to expected value and fell within category A, except for item 1. The Wright Map (Figure 7) is a visual representation of students and item among four countries. The distribution is symmetric, with students falling in the middle on the logit scale. Chile and Turkey were located in the same line, whereas the United States had the highest and Japan had the lowest score on the map.

DIF RESULTS FOR MATHEMATICS SELF-CONCEPT

Table 20 summarizes the DIF analyses for the mathematics anxiety scale. This table shows the mean residual statistics in logits. Positive values indicate that the item is more difficult than expected, while negative values indicate the item is less difficult than expected. Since these units are logits, it is useful to identify interaction effects that have an absolute value greater than .30.

The first step in examining DIF is to investigate the overall interaction effects. The data for anxiety indicate that the interaction effect explains 3.36% additional variance. There is an overall statistically significant interaction effects, $\chi 2\ (20, 15119) =\ 3493.1, p < .01$. Table 20 shows the mean residuals for each country and item. These effects are also shown graphically in Figure 8.

The data suggest that Item 1 on the self-concept scale is interpreted differently for Japan and the US with student answers looking to be opposite of each other. Item 1 refers to "I am not just good at mathematics". The reason that this item shows DIF may be related to the use of the word "not". These may have led students to misunderstand items with potential variations, also related to how this item was. For Item 2, The US students and for Item 5 Japanese students have significant interaction effects. The framework for interpreting model-data fit is shown.

5.4 DISCUSSION

This chapter describes the approach an invariance of a mathematics self-concept scale Rasch based on model-data fit was used. Differential item functioning (DIF) was performed based on a residual-based approach within the context of Rasch measurement theory. Although most DIF methods focus on comparing two groups (Holland and Wainer, 1993), Rasch approach can provide researchers opportunity to investigate more than two groups at a time. Person fit is also important, when examining model-data fit. If there is a good model-data fit for item and person, the benefit of invariant measurement can be obtained.

In the illustrative example on PISA 2012 mathematics self-concept, persons, items and countries are examined. Infit and Outfit statistics emphasize that the scale

function in a comparable way across countries. Persons tend to be in category A which implies good model-data fit. Items have relatively good fit to model-data expectation, but Item 1 is underfit. However, the data suggest that some items have DIF across countries. For example, item related being not good at mathematics appear to be differently interpreted across countries. Future qualitative research should explore the translation of this items across countries.

CHAPTER 6

MATHEMATICS BEHAVIOR

This chapter describes a scale designed to measure mathematics behavior. This chapter provides analyses that are similar to those conducted in earlier chapters.

6.1 INTRODUCTION

The term "behavior" refers to all acts that living organisms exhibit toward the outside environment. Education cooperates with psychology in order to change the behavior of students in a desired direction. Previous studies have shown a positive correlation between participation in extracurricular activities and student achievement (Eccles and Barber, 1999; Espinoza, 2011). According to Wijers et al. (2010), computer games is another example of engagement in mathematical activities can help meaningful learning.

The student’s approach to a mathematical problem can be shaped by a number of experiences including some experiences that students may not be aware of experiencing. For example, the student may approach to problem using their prior knowledge and experience (Schoenfeld, 1989). The outside of school activities related mathematics, their enrollment of mathematics class, their peer interaction about mathematics subject might be important to get positive behavior. Xiao and Sun (2021) found that student who have involvement mathematics with high motivation and low anxiety activities showed higher performance when they have mathematics tasks.

The students' behavior can be better understood if socio-economical, environmental, and cultural factors are taken into account (Koyuncu, 2020). For example, outside activities might be different across countries; some schools do not have mathematics clubs or students may have to work after school. Students from different countries and cultures may engage in different out-of-school. Considering and examining the mathematics behavior of student in different countries could be helpful in understanding how it effects.

RASCH MODEL

The Rasch measurement model was used to calibrate the items in the scale. Persons and items are ordered according to location on the construct of behavior in mathematics. The log odds of a person endorsing an item for Model I can be expressed as follows:

$$\ln\left[\frac{P_{nijk}}{P_{nijk-1}}\right] = \theta_n - \delta_i - \lambda_j - \tau_k \qquad (1)$$

where:

P_{nijk} = probability of student *n* providing a rating *k* on item *i* from country j,

P_{nijk-1} = probability of student *n* providing a rating *k-1* on item *i* from country j,

θ_n = logit-scale location of student *n,*

δ_i = logit-scale location of item i*,*

λ_j = logit-scale location of country *j,* and

τ_k = difficulty of category *k* relative to category *k*-1.

The log of the odds that a student gives a rating in category *k* rather than in category *k* – 1 is estimated given student locations on the behavior in mathematics. The student location is based on the locations of the items, countries, and category item difficulties.

The item parameter τ_k reflects the structure of the rating scale, and it is not considered as a facet in the model. Equation 2 includes the interaction term, $\delta_i\lambda_j$, that can be estimated after the model in Equation 1 is held constant:

$$\ln\left[\frac{P_{nijk}}{P_{nijk-1}}\right] = \theta_n - \delta_i - \lambda_j - \delta_i\lambda_j - \tau_k \quad (2)$$

It should be noted that the model in Equation 2 examines whether or not items function in the same way across countries.

The most common measures used to diagnosis fit to the Rasch model are Infit and Outfit Mean Squares. According to Linacre (2012), Infit Mean Square, also known as an inlier-sensitive or information-weighted fit, is more sensitive to how a person often answers questions about themselves when compared to other methods. Outfit is a fit that is especially sensitive to outlying data, such as reactions to items that are difficult to measure from a distance. Both mean square error statistics have an expected value of 1.0. Values less than 1.0 suggest observations are too predictable whereas greater than 1.0 suggest observations are too unpredictable. Therefore, it can be said that small Infit values indicate over fit to the Rasch model, while values greater than 1.0 may indicate an aberrant pattern. Generally, acceptable model-data fit is defined as a fit statistic between 0.5 and 1.5. In this study, we used guidelines suggested by Engelhard and Wind (2018).

PURPOSE

The purpose of this study is to investigate invariance in a mathematics behavior scale using both item and person fit for Chile, Japan, Turkey, and US.

These countries were selected because they represented different levels of mathematics achievement and a diversity of cultures. This study addresses the following questions:

1. Does a mathematics behavior scale function in a comparable way across countries?
2. Are there differences in item functioning between countries in terms of student's math behavior?
3. Are there patterns in person misfit related to country?

6.2 METHODOLOGY

PARTICIPANTS

For this study, a paper-based assessment data set from PISA 2012 international database were used for this investigation. This data set includes affective and other variables about students. In this file, they offer all students' responses for each of variables. It should be note that if student's response missing for all items, then these responses were removed from data set. The mathematics behavior scale used in PISA 2012 consists of eight items. Students were asked "how often do you do things at school and outside of school?" and response categories included "always or almost always", "often", "sometimes", "never or rarely" answers. These items are shown in Table 21.

This study examines behavior of 15-year-old children from Turkey, the United States, Japan and Chile towards mathematics in the PISA 2012. This study includes 23,033 students (11,677 boys and 11,356 girls) who participated in PISA 2012. There were 4,978 participants from United States, 4,848 participants from Turkey, 6,351 participants from Japan, and 6,856 participants from Chile. The amount of missing data was similar across countries: Turkey (34.2%), Japan (34.6%), US (35.3%), and Chile (34%).

Rasch measurement theory was used to calibrate and examine differential item and person functioning on the mathematics behavior scale. The analysis was conducted in FACETs that is specialized software for Rasch analyzes.

6.3 RESULTS

The three-facet Rasch model explained 42.17% of the variance in the data, and this supports the inference that the scale is unidimensional (Bond and Fox, 2015). Rasch summary statistics for the data are presented in Table 22. Overall, Infit Mean Square and Outfit Mean Square statistics were good for items, and countries. Persons had an average mathematics behavior level of -1.77 ($SD = 1.37$) logits. Items had an average difficulty of 0.00 ($SD = .64$) logits. Countries had an average mathematics behavior of 0.00 ($SD = .18$) logits. The reliability of separation was moderate for students [R = 0.64; χ^2 (15,089) = 45150.7, $p < .05$]. The reliability for students corresponds to the traditional coefficient alpha in classical test theory. The items [$R > .99$, χ^2 (7) = 15937.1, $p < .05$] and countries [$R > .99$, χ^2 (3) = 1449.6, $p < .05$] also had a significant reliability of separation index (Table 22).

Infit Mean Square and Outfit Mean Square statistics are given, as well as classifications of the items for fit statistics. The categories are based on recommendations made by Engelhard and Wind (2018) for interpreting the fit statistics calculated in a Rasch context. Chile, Turkey, Japan, and US were in the same category which is category A (Table 23). Infit Mean Square was the lowest for the Japan [Infit Mean Square = 0.96], and the highest for Chile [Infit Mean Square =1.10]. Outfit Mean Square was close to expected value and fell within category A for all countries.

Table 24 shows that in all countries the majority of student fall into category A and there are the least number of student fall into category C. Generally, for all items, Infit Mean Square was close to the expected value, and fell within category A (Table 25). Infit Mean Square was the lowest for the Item 3 and Item 7 [Infit Mean Square = 0.69], and the highest for Item 8 [Infit Mean Square =1.60]. Outfit Mean Square was close to expected value and fell within category A, except Item 8. The Wright Map (Figure 9) is a visual representation of students and item among four countries. The distribution is positively skewed, with more students falling in the lower on the logit scale (i.e., having lower level of math behavior). Japan and the United States were located in the same line, whereas the Turkey had the highest score in the map.

DIF RESULTS FOR MATHEMATICS BEHAVIOR

Table 26 summarizes the DIF analyses for the mathematics anxiety scale. This table shows the mean residual statistics in logits. Positive values indicate that the item is more difficult than expected, while negative values indicate the item is less difficult than expected. Since these units are logits, it is useful to identify interaction effects that have an absolute value greater than .30.

The first step in examining DIF is to investigate the overall interaction effects. The data for anxiety indicate that the interaction effect explains 2.28% additional variance. There is an overall statistically significant interaction effects, $\chi 2\ (32, 15089) = \ 3711.1, p < .01$. Table 26 shows the mean residuals for each country and item. These effects are also shown graphically in Figure 10.

The data demonstrate that Item 1 is interpreted differently for Japan and Turkey. Item 2 has a significant interaction effect for Turkey and the US. Item 4, Japan has a

significant DIF, as well. Item 6, Chile and Turkey have a significant item interaction. The US has a DIF for Item 7, while Item 8, Japan has a DIF. Framework for interpreting model-data fit guidelines are shown in Table 7.

6.4 DISCUSSION

In this chapter, an approach for determining invariance of a mathematics behavior scale Rasch based on model-data fit was carried out. Differential item functioning (DIF) was performed based on a residual-based approach within the context of Rasch measurement theory. Most DIF methods focus on comparing two groups (Holland and Wainer, 1993), but Rasch approach can provide researchers opportunity to investigate more than two groups at a time. Person fit is also important when examining model-data fit. If there is a good model-data fit for both item and person, the advantage of invariant measurement can be obtained.

In the illustrative example on PISA 2012 mathematics behavior, persons, items and countries are examined. Infit and Outfit statistics indicate that the scale function in a comparable way across countries. Persons tend to be in category A which implies good model-data fit. Items have relatively good fit to model-data expectation, but Item 8 is in category C which is undefit. Item 8 is related participating mathematics club, so each country may not have mathematics club and it makes students to answer differently. However, the data suggest that some items have DIF across countries. For example, item related taking part in mathematic competitions appear to be differently interpreted across countries. Future qualitative research should explore the translation of this item across countries.

CHAPTER 7

USING AFFECTIVE VARIABLES TO PREDICT MATHEMATICS ACHIEVEMENT

The purpose of this chapter is to briefly examine the relationships between the four affective variables and mathematics achievement using PISA 2012 data. The analyses are conducted for the combined sample of four countries, and then separately for each country. These analyses are essentially exploratory and designed to examine the relationships among the four affective variables (mathematics self-efficacy, mathematics anxiety, mathematics self-concept, mathematics behavior) with mathematics achievement.

Pearson correlation coefficients were calculated using SPSS. Next, multiple linear regression analyses were conducted with mathematics achievement as the dependent variable, and the four affective variables used as predictors. It should be noted that the first plausible value in the PISA data set is used to define mathematics achievement. Finally, correlations and regression coefficients are reported for the combined sample, as well as separately for each country. The last section discusses patterns observed across countries.

COMBINED ANALYSES

Table 27 shows the correlations for the combined analyses, as well as results for each country. The combined correlations indicate that there are statistically significant correlations between mathematics self-efficacy, anxiety, and self-concept variables with mathematics achievement. Mathematics self-efficacy and self-concept both have positive

relationships with mathematics achievement, while math anxiety is negatively correlated. It is interesting to note that mathematics behavior does not have a statistically significant correlation with mathematics achievement. For Japan and the US, all affective variables were found to be related to mathematics achievement.

The correlational results vary somewhat across countries. The positive correlations of mathematics self-efficacy and math self-concept appear across countries, as do the negative correlations with mathematics anxiety. Mathematics performance has a statistically significant correlation with achievement in Japan, and a small correlation in the United States.

Tables 28 to 32 show the correlation between mathematics achievement and each of the affective variables. The direction of the correlations are similar across countries, although there is some difference in the strength of the correlations between countries. For example, mathematics anxiety has a negative correlation with achievement, while mathematics self-efficacy, self-concept and behavior have positive correlations with achievement.

Table 33 shows the regression analyses with mathematics achievement as the dependent variable and affective variables as predictors. Both unstandardized B coefficients (standard errors), as well as standardized Beta coefficients are reported. The adjusted r-squares range from .25 to .34 for the four countries.

The strongest predictor across countries is math self-efficacy. Mathematics self-efficacy has a positive relationship that indicates that the higher levels of self-efficacy are related to higher mathematics achievement. Mathematics anxiety have a negative direction which indicates that lower levels of anxiety is associated with higher

mathematics achievement. It can be seen that mathematics self-efficacy is significant for each country and has a positive direction. Mathematics anxiety is significant for Chile, Turkey, and the US with a negative direction for Chile, Turkey, and the US. Mathematics self-concept is statistically significant for Chile and the direction is positive. Mathematics behavior is significant for four countries; the direction is negative for Chile, Turkey, and the US whereas the direction is positive for Japan. The r-square for the data is highest for the US which is 0.34.

RESULTS FOR CHILE

For Chile, mathematics self-efficacy, anxiety, and self-concept variables have significant correlations with mathematics achievement. Mathematics achievement is related to mathematics self-efficacy, $r = .375, p < .001$, mathematics anxiety, $r = -.385, p < .001$ and mathematics self-concept, $r = .394, p < .001$. The multiple regression analyses suggest that the four affective variables are statistically significant for Chile with mathematics self-efficacy having the largest regression coefficient. The adjusted r-square for the data is .25.

RESULTS FOR JAPAN

For Japan, these four affective variables have statistically significant correlations with mathematics achievement. Mathematics achievement is related to mathematics self-efficacy, $r = .559, p < .001$, mathematics anxiety, $r = -.176, p < .001$, mathematics self-concept, $r = .281, p < .001$, and mathematics behavior, $r = .249, p < .001$. The multiple regression analyses suggest that mathematics self-efficacy and behavior are significant predictor of mathematics achievement for Japan with mathematics self-efficacy being the most influential. The adjusted r-square for the data is .31.

RESULTS FOR TURKEY

For Turkey, three variables have statistically significant correlations with mathematics achievement: mathematics self-efficacy, anxiety, and self-concept. Mathematics achievement is related to mathematics self-efficacy, $r = .430, p < .001$, mathematics anxiety, $r = -.237, p < .001$, and mathematics self-concept, $r = .211, p < .001$. The multiple regression analyses suggest that mathematics self-efficacy, anxiety and behavior are statistically significant for Turkey. Mathematics self-efficacy is the strongest predictor of mathematics achievement in Turkey. The adjusted r-square for the data is .24.

RESULTS FOR UNITED STATES

For the US, these four variables have statistically significant correlations with mathematics achievement. Mathematics achievement is related to mathematics self-efficacy, $r = .533, p < .001$, mathematics anxiety $r = -.441, p < .001$, mathematics self-concept, $r = .425, p < .001$, and mathematics behavior, $r = .089, p < .001$. The multiple regression analyses suggest that mathematics self-efficacy, anxiety, self-concept and behavior are statistically significant for the US. As was found in other countries, mathematics self-efficacy is the strongest predictor of mathematics achievement. The adjusted r-square for the data is .34.

SUMMARY

Overall, the correlations of the affective variables with mathematics achievement have similar patterns across countries. The regression analyses indicate that mathematics self-efficacy is the strongest predictor of mathematics achievement among these affective variables.

CHAPTER 8

DISCUSSION

The purpose of this dissertation was to investigate invariant measurement for assessments in the affective domain within cross-cultural research. In this chapter, the results of separate studies of four affective variables are summarized. These findings are organized around the requirements of invariant measurement. Future research areas are also highlighted in this chapter. First, the research questions are revisited, and the key results presented. Next, the limitations and suggestions for future research are discussed. Finally, the overall lessons from the dissertation are summarized.

RESEARCH QUESTIONS

The first research question is as follows:

- Are there differences in item functioning between countries in terms of mathematics affective variables?

The data suggest that the scales for measuring the affective variables do vary in their meaning across countries. Specifically, Japanese students appear to have a different perspective on mathematics self-efficacy items. For mathematics anxiety, students in Chile has a distinctive perspective as indicated by mathematics anxiety items. Mathematics self-concept also appears to vary for Japanese and the US students. Last, mathematics behavior seems to vary the most across the four countries.

Overall, it appears that the items do vary across countries. This lack of invariance places limitations on the inferences that can be drawn from international data. In the future, qualitative analyses should be conducted to explore why items seem to be

interpreted differently in some countries. For example, there are many issues related to the translation of these items across countries.

The next question is:

- Are there any persons who misfit related to each country?

The data indicate that Chile, Japan, Turkey and the US students tend to fall into Category A, based on model-data fit statistics for person fit. Category A suggest that person fit model is productive for measurement. Overall, there was some variation in person fit, but in general the student responses fit the Rasch model.

The last question is:

- Do the mathematics affective variables function in a comparable way across countries?

In order to answer this question, exploratory analyses were conducted to examine the relationship between mathematics achievement and these four affective variables. Across countries, there are consistent patterns in the correlations of the affective variables with mathematics achievement. Overall, the regression analyses indicate that mathematics self-efficacy is the best predictor of mathematical achievement.

LIMITATIONS AND SUGGESTIONS FOR FUTURE RESEARCH

First of all, there are some limitations related to PISA. PISA is the one of the world's largest empirical data sets, but there are limitations to PISA 2012 that should be taken into consideration. Even though PISA is a large assessment, it cannot assess all students, schools, teachers or parents. Sampling of respondents may cause ignorance of missing respondents and educational processes. The test questions are translated into different languages, and this may influence the results. PISA 2012 asks students about

their feelings or motivations about subject, and student honesty and care cannot be ensured.

Another limitation is that only four countries with different cultures were included. Although these four countries reflect different cultures and educational systems, it may not generalize to other countries. Future research should examine other affective variables, as well as other countries.

It should also be noted that for the purposes of this study, unweighted analyses were used, and this might cause some differences between regression coefficients. Another methodological limitation is that although PISA 2012 provides that five plausible values for each student, this study only used the first plausible value to represent student mathematics achievement.

Future research should explore strategies for improving invariant measurement. Improvements in translations, and qualitative analyses may add to our knowledge. This study was designed without taking into account qualitative information. Integrating quantitative and qualitative methodological approaches is promising area for future research.

SUMMARY

Cross cultural research provides an important lens for exploring educational process. As with other approaches to educational research, there are certain constraints that may limit the generalizability of the findings. One constraint is that the requirements of invariant measurement may not be met. If scales are not invariant across countries, then this makes quantitative comparisons challenging. This dissertation seeks to evaluate

invariant measurement using the Rasch model. Specifically, four different affective variables from four different countries were examined.

It is important to note that the affective scales differed significantly across the four countries included in this dissertation. This variation suggests that mathematics is viewed differently in these countries, and that this may be due to distinctive cultural characteristics within each country. Culture is a broad concept that includes not only the ways in which people act and organize themselves in society but also their ideas, values, arts, legal systems, conventions, talents, and patterns of behaviors. Education plays a significant role in shaping and being shaped by culture. For example, some countries may emphasize the importance of independence and individualized achievement, while other countries may encourage cooperative behaviors related to mathematics. There is also the potential for some of the items in the scales to address activities, such as chess clubs, that may not be available in every country. In order to adequately address these potential cultural differences across countries, future research should include qualitative analyses of the mathematics culture with each country.

Overall, the results from this dissertation provide a way of thinking about and evaluating affective variables in cross-cultural research. Detailed analyses of invariance are essential for understanding the strengths and weakness connected to large-scale international assessments.

REFERENCES

Addey, C., & Sellar, S. (2018). Why do countries participate in PISA? Understanding the role of international large-scale assessments in global education policy. *Global education policy and international development: New agendas, issues and policies*, *97*, 117.

Ajzen, I. (1991). The theory of planned behavior. *Organizational behavior and human decision processes*, *50*(2), 179-211.

Akter, S., Jabbar, A., & Khatun, M. T. (2018). Factors Affecting Career Choice Of The Female Secondary Students in Khulna District Of Bangladesh. *Khulna University Studies*, 91-103.

Ayotola, A., & Adedeji, T. (2009). The relationship between mathematics self-efficacy and achievement in mathematics. *Procedia-Social and Behavioral Sciences*, *1*(1), 953-957.

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

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.

Bloom, B. S., & Krathwohl, D. R. (2020). *Taxonomy of educational objectives: The classification of educational goals. Book 1, Cognitive domain*. Longman.

Bulut, O., Palma, J., Rodriguez, M. C., & Stanke, L. (2015). Evaluating measurement invariance in the measurement of developmental assets in Latino English

language groups across developmental stages. *Sage Open*, *5*(2), 2158244015586238.

Brown, J., Ortiz-Padilla, M., & Soto-Varela, R. (2020). Does mathematical anxiety differ cross-culturally?. *Journal of New Approaches in Educational Research (NAER Journal)*, *9*(1), 126-136.

Cheema, J. R., & Kitsantas, A. (2014). Influences of disciplinary classroom climate on high school student self-efficacy and mathematics achievement: A look at gender and racial–ethnic differences. *International Journal of Science and Mathematics Education*, *12*, 1261-1279.

Cvencek, D., Paz-Albo, J., Master, A., Herranz Llácer, C. V., Hervás-Escobar, A., & Meltzoff, N. (2020). Math is for me: A field intervention to strengthen math self-concepts in Spanish-speaking 3rd grade children. *Frontiers in Psychology*, *11*, 593995.

Doruk, M., Öztürk, M., & Kaplan, A. (2016). Investigation of the self-efficacy perceptions of middle school students towards mathematics: Anxiety and attitude factors. *Adıyaman University Journal of Educational Sciences*, *6*(2), 283-302.

Eccles, J. S., & Barber, B. L. (1999). Student council, volunteering, basketball, or marching band: What kind of extracurricular involvement matters? *Journal of adolescent research*, *14*(1), 10-43.

Emmanuel, A. O., Adom, E. A., Josephine, B., & Solomon, F. K. (2014). Achievement motivation, academic self-concept and academic achievement among high school students. *European Journal of Research and Reflection in Educational Sciences*, *2*(2).

Engelhard Jr, G. (2008). Historical perspectives on invariant measurement: Guttman, Rasch, and Mokken. *Measurement*, *6*(3), 155-189.

Engelhard Jr, G., & Wind, S. (2017). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. Routledge.

Engelhard Jr, G (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences.* Routledge.

Espinoza, J. A. (2011). School sponsored extracurricular activities and math achievement among Hispanic students. *Journal of Youth Development*, *6*(2), 48-57.

Fan, X., Hambleton, R. K., & Zhang, M. (2019). Profiles of mathematics anxiety among 15-year-old students: A cross-cultural study using multi-group latent profile analysis. *Frontiers in Psychology*, *10*, 1217.

Gabriel, F., Buckley, S., & Barthakur, A. (2020). The impact of mathematics anxiety on self-regulated learning and mathematical literacy. *Australian Journal of Education*, *64*(3), 227-242.

Grootenboer, P., Marshman, M., Grootenboer, P., & Marshman, M. (2016). The affective domain, mathematics, and mathematics education. *Mathematics, affect and learning: Middle school students' beliefs and attitudes about mathematics education*, 13-33.

Gunderson, E. A., Park, D., Maloney, E. A., Beilock, S. L., & Levine, S. C. (2018). Reciprocal relations among motivational frameworks, math anxiety, and math achievement in early elementary school. *Journal of Cognition and Development*, *19*(1), 21-46.

Gür, B. S., Celik, Z., & Özoğlu, M. (2012). Policy options for Turkey: A critique of the

interpretation and utilization of PISA results in Turkey. *Journal of Education Policy*, *27*(1), 1-21.

Hackett, G., & Lent, R. W. (1992). Theoretical advances and current inquiry in career psychology. *Handbook of counseling psychology*, *2*, 419-452.

Holland, P. W. (2012). Differential item functioning. *Routledge.*

Jerrim, J. (2015). Why do East Asian children perform so well in PISA? An investigation of Western-born children of East Asian descent. *Oxford Review of Education*, *41*(3), 310-333.

Khasawneh, E., Gosling, C., & Williams, B. (2021). What impact does maths anxiety have on university students?. *BMC psychology*, *9*(1), 1-9.

Kirsch, I., Lennon, M., von Davier, M., Gonzalez, E., & Yamamoto, K. (2013). On the growing importance of international large-scale assessments. *The role of international large-scale assessments: Perspectives from technology, economy, and educational research*, 1-11.

Koyuncu, M. K., & Özdemir, A. (2020). Analysis of philosophy of mathematics activities on students' attitudes and beliefs towards mathematics. *International Journal of Educational Studies in Mathematics*, *7*(2), 57-71.

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, C. Y., & Kung, H. Y. (2018). Math self-concept and mathematics achievement: Examining gender variation and reciprocal relations among junior high school

students in Taiwan. *Eurasia Journal of Mathematics, Science and Technology Education*, *14*(4), 1239-1252.

Lewis, S., & Lingard, B. (2015). The multiple effects of international large-scale assessment on education policy and research. *Discourse: Studies in the cultural politics of education*, *36*(5), 621-637.

Linacre, J. M. (2012). *Winsteps tutorial 4: Differential item functioning and dimensionality*. Beaverton, OR: Winsteps.com.

Lipnevich, A. A., Preckel, F., & Krumm, S. (2016). Mathematics attitudes and their unique contribution to achievement: Going over and above cognitive ability and personality. *Learning and Individual Differences*, *47*, 70-79.

Mahanta, D. (2012). Achievement in mathematics: effect of gender and positive/negative attitude of students. *International Journal of Theoretical & Applied Sciences*, *4*(2), 157-163.

Marsh, H. W., & Craven, R. (1996). Academic self-concept: Beyond the dustbowl. In *Handbook of classroom assessment* (pp. 131-198). Academic Press.

Marshall, J. (2019). Introduction to comparative and international education. *Introduction to Comparative and International Education*, 1-248.

Masitoh, L. F., & Fitriyani, H. (2018). Improving students' mathematics self-efficacy through problem based learning. *Malikussaleh Journal of Mathematics Learning (MJML)*, *1*(1), 26-30.

Miele, D. (2009). *Handbook of motivation at school (Vol. 704). K. R. Wentzel, & A. Wigfield (Eds.).* New York, NY: Routledge.

Millsap, R. E. (2012). *Statistical approaches to measurement invariance*. Routledge.

Muhtadi, A., Assagaf, G., & Hukom, J. (2022). Self-Efficacy and Students' Mathematics Learning Ability in Indonesia: A Meta Analysis Study. *International Journal of Instruction*, *15*(3), 1131-1146.

OECD. (2013). PISA Results.

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

OECD (2013). *Results: Ready to Learn (Volume III): Students' Engagement, Drive and Self-Beliefs.*

OECD. (2014). *PISA 2012 Technical Report.*

OECD. (2020). PISA 2018 Results (Volume VI) Are Students Ready to Thrive in an Interconnected World?. *OECD Publishing.* https://doi.org/10.1787/d5f68679-en.

Oettingen, G. (1997). Culture and future thought. *Culture & Psychology*, *3*(3), 353-381.

Pajares, F., & Miller, M. D. (1994). Role of self-efficacy and self-concept beliefs in mathematical problem solving: A path analysis. *Journal of educational psychology*, *86*(2), 193.

Penfield, R. D., & Camilli, G. (2006). 5 Differential Item Functioning and Item Bias. *Handbook of statistics*, *26*, 125-167.

Peters, M. L. (2013). Examining the relationships among classroom climate, self-efficacy, and achievement in undergraduate mathematics: A multi-level analysis. *International Journal of Science and Mathematics Education*, *11*, 459-480.

Rasch, G. (1993). *Probabilistic models for some intelligence and attainment tests*. MESA Press, 5835 S. Kimbark Ave., Chicago, IL 60637; e-mail: MESA@ uchicago.

edu; web address: www. rasch. org; tele.

Reyes, L. H. (1984). Affective variables and mathematics education. *The elementary school journal*, *84*(5), 558-581.

Richardson, F. C., & Suinn, R. M. (1972). The mathematics anxiety rating scale: psychometric data. *Journal of counseling Psychology*, *19*(6), 551.

Ritzen, J. (2013). International large-scale assessments as change agents. *The role of international large-scale assessments: Perspectives from technology, economy, and educational research*, 13-24.

Roslan, N. A., & Maat, S. M. (2019). Systematic Literature Review on Secondary School Students' Mathematics Self-Efficacy. *International Journal of Academic Research in Progressive Education and Development*, *8*, 975-987.

Rutkowski, L., & Rutkowski, D. (2010). Getting it 'better': the importance of improving background questionnaires in international large-scale assessment. *Journal of Curriculum Studies*, *42*(3), 411-430.

Schoenfeld, A. H. (1989). Explorations of students' mathematical beliefs and behavior. *Journal for research in mathematics education*, *20*(4), 338-355.

Schöber, C., Schütte, K., Köller, O., McElvany, N., & Gebauer, M. M. (2018). Reciprocal effects between self-efficacy and achievement in mathematics and reading. *Learning and Individual Differences*, *63*, 1-11.

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

Theisen, G. L., Achola, P. P., & Boakari, F. M. (1983). The underachievement of cross-national studies of achievement. *Comparative Education Review*, *27*(1), 46-68.

Thien, L. M., Darmawan, I. G. N., & Ong, M. Y. (2015). Affective characteristics and mathematics performance in Indonesia, Malaysia, and Thailand: what can PISA 2012 data tell us? *Large-scale Assessments in Education*, *3*, 1-16.

Toland, M. D., & Usher, E. L. (2016). Assessing mathematics self-efficacy: How many categories do we really need?. *The Journal of Early Adolescence*, *36*(7), 932-960.

Wijers, M., Jonker, V., & Drijvers, P. (2010). MobileMath: exploring mathematics outside the classroom. *ZDM*, *42*, 789-799.

Wu, H., Guo, Y., Yang, Y., Zhao, L., & Guo, C. (2021). A meta-analysis of the longitudinal relationship between academic self-concept and academic achievement. *Educational Psychology Review*, 1-30.

Wright, B. (1984). Despair and hope for educational measurement. *Contemporary Education Re-view*, Volume 3, 281-288.

Xiao, F., & Sun, L. (2021). Students' motivation and affection profiles and their relation to mathematics achievement, persistence, and behaviors. *Frontiers in psychology*, *11*, 533593.

Yuan, Z., Tan, J., & Ye, R. (2022). A Cross-national Study of Mathematics Anxiety. *The Asia-Pacific Education Researcher*, 1-12.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language assessment quarterly*, *4*(2), 223-233.

APPENDIX A

TABLES

Table 1 Framework for interpreting model-data fit

Category	**Mean-Square Error**	**Polyserial Correlations**	**Monotonicity and outliers**	**Value Judgment**	**Description**	**Potential Interpretations**
A	0.50 ≤ MSE < 1.50	Medium-High	Monotonic with a few outliers	Productive for measurement	Good model-data fit (Informative)	Good fit to model-based expectation
B	MSE < 0.50	High	Monotonic with very few outliers	Less productive for measurement, but not distorting of measures	Overfit (Informative)	Guttman patterns with limited category usage Intentional distortion: Acquiescence (Faking Good) Sabotaging (Faking Bad)
C	1.50 ≤ MSE < 2.00	Medium	Monotonic with some outliers	Unproductive for measurement, but not distorting of measures	Underfit (Informative)	Unmotivated Responding Person Unreliability
D	MSE ≥ 2.00	Low-Negative	Non-monotonic with many outliers	Unproductive for measurement, distorting of measures	Underfit (Non-informative)	Random Responding Multidimensionality

Table 2 Mathematics Self-efficacy items in PISA 2012

Variable Name	Item
ST37Q01	Using a <train timetable> to work out how long it would take to get from one place to another.
ST37Q02	Calculating how much cheaper a TV would be after a 30% discount.
ST37Q03	Calculating how many square meters of tiles you need to cover a floor.
ST37Q04	Understanding graphs presented in newspapers.
ST37Q05	Solving an equation like $3x+5 = 17$.
ST37Q06	Finding the actual distance between two places on a map with a 1:10 000 scale.
ST37Q07	Solving an equation like $2(x+3) = (x + 3)(x - 3)$.
ST37Q08	Calculating the petrol consumption rate of a car.

Note. Likert scale was used "very confident", "confident", "not very confident", "not at all confident"

Table 3 Rasch summary statistics for person, item and country (Self-efficacy)

	Persons	**Items**	**Countries**
Measure			
M	1.15	0.00	0.00
SD	1.68	0.74	0.15
Outfit			
M	1.00	1.00	1.01
SD	0.74	0.14	0.10
Infit			
M	1.02	1.02	1.01
SD	0.72	0.17	0.10
Reliability of separation	0.81	> 0.99	0.99
χ^2	60594.3	21396.7	797.5
df	15135	8	4
Variance explained by Rasch model	47.14%		

Table 4 Country fit statistics (Self-efficacy)

ID	Country	N	Measure	S.E.	Infit MS	Outfit MS	Fit Categories
1	Chile	6856	0.02	0.01	0.98	0.98	A
2	Japan	6351	-0.21	0.01	1.03	1.00	A
3	Turkey	4848	0.07	0.01	1.13	1.14	A
4	US	4978	0.12	0.01	0.89	0.90	A

Note. MSE is the mean square error based on the Rasch model.

Table 5 The categorical distribution of students (Self-efficacy)

	Total	**Chile**	**Japan**	**Turkey**	**US**
Fit Category	**Outfit MS**				
A	7962 (56.2%)	2535(56.2%)	2083(55%)	1673(58.2%)	1671(59.2%)
B	3572 (25.2%)	1044(24.9%)	1042(27.5%)	855(29.8%)	631(22.3%)
C	1340 (9.5%)	387(9.7%)	390(10%)	242(8.4%)	321(11.5%)
D	1287 (9.1%)	225(5%)	275(7.5%)	102(3.6%)	199(7%)

Table 6 Item fit statistics (Self-efficacy)

Items	Measure	S.E.	Infit MS	Outfit MS	Fit Category	Polyserial Correlation
8	0.96	0.01	1.01	1.02	A	0.44
6	0.89	0.01	1.03	1.03	A	0.40
3	0.37	0.01	0.81	0.80	A	0.50
4	-0.10	0.01	0.96	0.97	A	0.45
2	-0.14	0.01	0.88	0.87	A	0.46
1	-0.19	0.01	0.92	0.99	A	0.41
7	-0.50	0.01	1.29	1.27	A	0.34
5	-1.30	0.01	1.25	1.08	A	0.34

Table 7 Mathematics Self-efficacy (Item by country interaction)

Country	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8
Chile	0.24	0.34	0.14	0.19	-0.41	-0.31	-0.27	0.02
Japan	-0.09	-0.22	-0.43	-0.53	0.86	0.38	0.89	-0.51
Turkey	0.12	0.07	0.32	0.35	-0.7	0.17	-0.68	0.19
US	-0.34	-0.24	0.06	0.15	0.3	-0.22	-0.06	0.43
Mean	-0.02	-0.01	0.02	0.04	0.01	0.01	-0.03	0.03

Note. Rasch DIF statistics – values greater than .30 -- difference from average

Table 8 Fit statistics framework

Category	Mean-Square Error	Value Judgment
A	$0.50 \leq MSE < 1.50$	Productive for measurement
B	$MSE < 0.50$	Less productive for measurement, but not distorting of measures
C	$1.50 \leq MSE < 2.00$	Unproductive for measurement, but not distorting of measures
D	$MSE \geq 2.00$	Unproductive for measurement, distorting of measures

Table 9 Mathematics Anxiety items in PISA 2012

Variable Name	Item
ST42Q01	I often worry that it will be difficult for me in mathematics classes
ST42Q03	I get very tense when I have to do mathematics homework
ST42Q05	I get very nervous doing mathematics problems
ST42Q08	I feel helpless when doing a mathematics problem
ST42Q10	I worry that I will get poor <grades> in mathematics

Note. Likert scale was used “strongly agree”, “agree”, “disagree”, “strongly disagree”

Table 10 Rasch summary statistics for person, item and country (Anxiety)

	Persons	**Items**	**Countries**
Measure			
M	0.19	0.00	0.00
SD	1.99	0.00	0.12
Outfit			
M	1.01	1.01	1.00
SD	1.02	0.23	0.13
Infit			
M	1.01	1.00	0.98
SD	0.96	0.23	0.13
Reliability of separation	0.81	>.99	>.99
χ^2	54761.7	17597.3	365.0
df	15119	5	4
Variance explained by the Rasch model	54.81%		

Table 11 Country fit statistics (Anxiety)

ID	Country	N	Measure	S.E.	Infit MS	Outfit MS	Fit Categories
1	Chile	6856	0.11	0.01	1.12	1.13	A
2	Japan	6351	0.07	0.01	0.95	0.97	A
3	Turkey	4848	0.02	0.01	1.08	1.11	A
4	US	4978	-0.20	0.01	0.78	0.80	A

Table 12 The categorical distribution of students (Anxiety)

	Total	Chile	Japan	Turkey	US
Fit Category	**Outfit MS**				
A	7073(46.8%)	2240(49.7%)	1929(46.3%)	1532(48.1%)	1372(42.1%)
B	5284(35%)	1471(32.6%)	1492(35.8%)	1028(32.3%)	1293(39.7%)
C	921(6.1%)	261(5.8%)	244(5.9%)	186(5.8%)	230(7.1%)
D	1841(12.1%)	536(11.9%)	504(12.1%)	438(13.8%)	363(11.1%)

Table 13 Item fit statistics (Anxiety)

Items	Measure	S.E.	Infit MS	Outfit MS	Fit Category	Polyserial Correlation
5	-1.22	0.01	1.45	1.47	A	0.35
1	-0.72	0.01	0.90	0.97	A	0.44
2	0.18	0.01	0.86	0.86	A	0.48
3	0.69	0.01	0.88	0.89	A	0.45
4	0.97	0.01	0.88	0.88	A	0.44

Table 14 Mathematics Anxiety (Item by country interactions)

Country	Item 1	Item 2	Item 3	Item 4	Item 5
Chile	-0.13	-0.36	-0.12	-0.36	1.14
Japan	0	0.3	0	0.09	-0.4
Turkey	-0.01	0.05	-0.05	0.32	-0.3
US	0.21	0.12	0.25	0.09	-0.64
Mean	0.02	0.03	0.02	0.03	-0.05

Note. Rasch DIF statistics – values greater than .30 – difference from average

Table 15 Mathematics Self-concept items in PISA 2012

Variable Name	Item
ST42Q02	I am not just good at mathematics
ST42Q04	I get good <grades> in mathematics
ST42Q06	I learn mathematics quickly
ST42Q07	I have always believed that mathematics is one of my best subjects
ST42Q09	In my mathematics class, I understand even the most difficult work

Note. Likert scale was used “very likely”, “likely”, “slightly likely”, “not at all likely”

Table 16 Rasch summary statistics for person, item and country (Self-concept)

	Persons	**Items**	**Countries**
Measure			
M	-0.89	0.00	0.00
SD	2.12	1.07	0.21
Outfit			
M	1.04	1.04	1.04
SD	1.06	0.47	0.08
Infit			
M	0.95	0.96	0.92
SD	0.78	0.37	0.04
Reliability of separation	0.79	>.99	>.99
χ^2	65051.7	23034.4	1033.1
df	15119	5	4
Variance explained by the Rasch model	62.34%		

Table 17 Country fit statistics (Self-concept)

ID	Country	N	Measure	S.E.	Infit MS	Outfit MS	Fit Categories
1	Chile	6856	0.00	0.01	0.89	0.97	A
2	Japan	6351	-0.33	0.01	0.90	1.12	A
3	Turkey	4848	0.06	0.01	0.99	1.11	A
4	US	4978	0.27	0.01	0.88	0.95	A

Table 18 The categorical distribution of students (Self-concept)

	Total	Chile	Japan	Turkey	US
Fit Category	**Outfit MS**				
A	6461(42.73%)	1960(45.48%)	1671(40.08%)	1365(42.84%)	1567(48.11%)
B	5364(35.48%)	1757(38.98%)	1618(38.81%)	1205(37.82%)	1166(35.80%)
C	1831(12.11%)	392(8.70%)	481(11.54%)	281(8.82%)	218(6.79%)
D	1464(9.68%)	399(8.84%)	399(9.58%)	335(10.52%)	306(9.39%)

Table 19 Item fit statistics (Self-concept)

Items	Measure	S.E.	Infit MS	Outfit MS	Fit Category	Polyserial Correlation
1	-1.97	0.02	1.65	1.96	C	0.02
2	1.04	0.01	0.75	0.79	A	0.53
3	0.83	0.01	0.61	0.64	A	0.57
4	0.29	0.01	0.97	0.96	A	0.53
5	-0.19	0.01	0.97	0.96	A	0.53

Table 20 Mathematics Self-concept (Item by country interactions)

Country	Item 1	Item 2	Item 3	Item 4	Item 5
Chile	-0.22	0.06	0.27	-0.22	0.02
Japan	0.89	-0.19	-0.23	0.21	-0.47
Turkey	-0.05	-0.29	0.0	0.21	0.13
US	-0.64	0.41	-0.1	-0.13	0.27
Mean	-0.01	0.00	-0.02	0.02	-0.01

Note. Rasch DIF statistics – values greater than .30 – difference from average

Table 21 Mathematics Behavior items in PISA 2012

Variable Name	Item
ST49Q01	I talk about mathematics problems with my friends
ST49Q02	I help my friend with mathematics
ST49Q03	I do mathematics as an <extracurricular> activity
ST49Q04	I take part in mathematics competitions
ST49Q05	I do mathematics more than 2 hours a day outside of school
ST49Q06	I play chess
ST49Q07	I program computers
ST49Q09	I participate in a mathematics club

Note. Likert scale was used “never or rarely”, “sometimes”, “often”, “always or almost always”

Table 22 Rasch summary statistics for person, item and country (Behavior)

	Persons	**Items**	**Countries**
Measure			
M	-1.77	0.00	0.00
SD	1.37	0.64	0.18
Outfit			
M	0.98	0.98	0.97
SD	0.70	0.33	0.09
Infit			
M	1.00	1.06	1.03
SD	0.63	0.33	0.05
Reliability of separation	0.64	>.99	>.99
χ^2	45150.7	15937.1	1449.6
df	15090	8	4
Variance explained by the Rasch model	42.17%		

Table 23 Country fit statistics (Behavior)

ID	Country	N	Measure	S.E.	Infit MS	Outfit MS	Fit Categories
1	Chile	6856	0.04	0.01	1.10	1.06	A
2	Japan	6351	-0.16	0.01	0.96	0.86	A
3	Turkey	4848	0.28	0.01	1.06	1.06	A
4	US	4978	-0.16	0.01	1.02	0.92	A

Table 24 The categorical distribution of students (Behavior)

	Total	Chile	Japan	Turkey	US
Fit Category	**Outfit MS**				
A	8886(58.9%)	2852(63%)	2185(52.6%)	2036(63.8%)	1813(56.3%)
B	3756(24.9%)	933(20.6%)	1302(31.3%)	601(18.8%)	920(28.6%)
C	1172(7.8%)	379(8.4%)	279(6.7%)	289(9.1%)	225(7%)
D	1276(8.4%)	360(8.4%)	390(9.4%)	263(8.2%)	263(8.1%)

Table 25 Item fit statistics (Behavior)

Items	Measure	S.E.	Infit MS	Outfit MS	Fit Category	Polyserial Correlation
2	-1.02	0.01	0.90	0.95	A	0.35
1	-0.71	0.01	0.92	0.96	A	0.34
6	-0.10	0.01	1.55	1.47	A	0.30
3	0.06	0.01	0.69	0.58	A	0.53
7	0.06	0.01	0.69	0.58	A	0.53
5	0.20	0.01	1.01	1.00	A	0.39
8	0.25	0.01	1.60	1.52	C	0.27
4	1.27	0.01	1.13	0.76	A	0.42

Table 26 Mathematics Behavior (Item by country interactions)

Country	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8
Chile	-0.09	-0.03	0.15	0.16	-0.09	-0.45	0.15	0.28
Japan	0.42	-0.02	0.11	-1.26	-0.04	-0.11	0.11	-0.41
Turkey	-0.34	-0.44	0.01	0.17	0.19	0.59	0.01	-0.05
US	-0.03	0.56	-0.49	0.3	-0.09	-0.07	-0.49	0.01
Mean	-0.01	0.02	-0.06	-0.16	-0.01	-0.01	-0.06	-0.04

Note. Rasch DIF statistics -values greater than .30- difference from average

Table 27

Correlations of affective variables with mathematics achievement (Combined and by country)

	Combined	**Chile**	**Japan**	**Turkey**	**US**
Math Self-Efficacy	.365*	.375*	.559*	.430*	.533*
Math Anxiety	-.275*	-.385*	-.176*	-.237*	-.441*
Math Self-Concept	.231*	.394*	.281*	.211*	.425*
Math Behaviors	.001	.025	.249*	-.010	.089*

* $p < .05$

Table 28

Correlations of Mathematics Achievement and the Four Affective Variables (Total)

	Self-efficacy	Anxiety	Self-concept	Behavior	Achievement
Self-efficacy	1	-.317*	.514*	.337*	.365*
Anxiety	-.317*	1	-.641*	-.111*	-.275*
Self-concept	.514*	-.641*	1	.376*	.231*
Behavior	.337*	-.111*	.376*	1	.001
Achievement	.365*	-.275*	.231*	.001	1

Note. *. Correlation is significant at the 0.01 level

Table 29

Correlations of Mathematics Achievement and the Four Affective Variables (Chile)

	Self-efficacy	Anxiety	Self-concept	Behavior	Achievement
Self-efficacy	1	-.285*	.518*	.309*	.375*
Anxiety	-.285*	1	-.564*	-.087*	-.385*
Self-concept	.518*	-.564*	1	.398*	.394*
Behavior	.309*	-.087*	.398*	1	.025
Achievement	.375*	-.385*	.394*	.025	1

Note. *. Correlation is significant at the 0.01 level

Table 30

Correlations of Mathematics Achievement and the Four Affective Variables (Japan)

	Self-efficacy	Anxiety	Self-concept	Behavior	Achievement
Self-efficacy	1	-.276*	.427*	.361*	.559*
Anxiety	-.276*	1	-.726*	-.188*	-.176*
Self-concept	.427*	-.726*	1	.345*	.281*
Behavior	.361*	-.188*	.345*	1	.249*
Achievement	.559*	-.176*	.281*	.249*	1

Note. *. Correlation is significant at the 0.01 level

Table 31

Correlations of Mathematics Achievement and the Four Affective Variables (Turkey)

	Self-efficacy	Anxiety	Self-concept	Behavior	Achievement
Self-efficacy	1	-.194*	.479*	.347*	.430*
Anxiety	-.194*	1	-.533*	-.084*	-.237*
Self-concept	.479*	-.533*	1	.476*	.211*
Behavior	.347*	-.084*	.476*	1	-.010
Achievement	.430*	-.237*	.211*	-.010	1

Note. *. Correlation is significant at the 0.01 level

Table 32

Correlations of Mathematics Achievement and the Four Affective Variables (US)

	Self-efficacy	Anxiety	Self-concept	Behavior	Achievement
Self-efficacy	1	-.460*	.525*	.319*	.533*
Anxiety	-.460*	1	-.742*	-.149*	-.441*
Self-concept	.525*	-.742*	1	.341*	.425*
Behavior	.319*	-.149*	.341*	1	.089*
Achievement	.533*	-.441*	.425*	.089*	1

Note. *. Correlation is significant at the 0.01 level

Table 33 Summary of regression analyses (Mathematics achievement as dependent variable)

	Combined		**Chile**		**Japan**		**Turkey**		**US**	
	B (SE)	**Beta**	**B (SE)**	**Beta**	**B (SE)**	**Beta**	**B (SE)**	**Beta**	**B (SE)**	**Beta**
Math Self-Efficacy	35.49* (1.21)	.36	24.34* (2.02)	.26	48.19* (1.91)	.52	44.30* (2.42)	.47	37.53* (2.15)	.43
Math Anxiety	-19.97* (1.43)	-.20	-24.36* (2.60)	-.21	1.65 (2.48)	.02	-16.10* (2.52)	-.17	-16.12* (2.67)	-.19
Math Self-Concept	-2.92 (1.50)	-.03	16.76* (2.16)	.20	5.42 (2.81)	.05	-1.88 (3.09)	-.02	8.52* (3.00)	.09
Math Behaviors	-12.04* (1.08)	-.13	-13.38* (1.76)	-.15	4.29* (1.98)	.04	-15.12* (2.23)	-.18	-8.58* (1.78)	-.11
R-square	.18		.25		.32		.24		.34	
Adjusted R-square	.18		.25		.31		.24		.34	
N	7482		2262		2080		1561		1579	

* $p < .05$

APPENDIX B

FIGURES

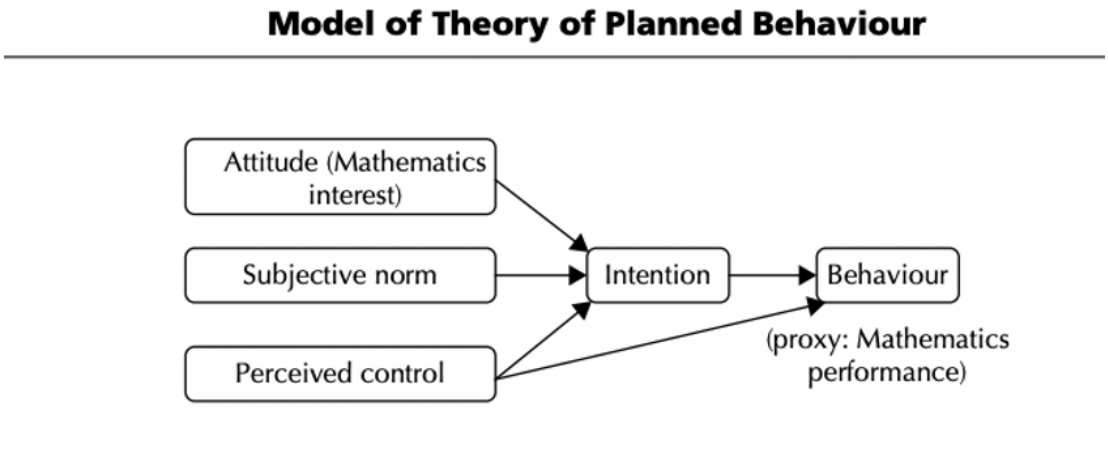

Figure 1. Model of the theory of planned behavior in PISA 2012

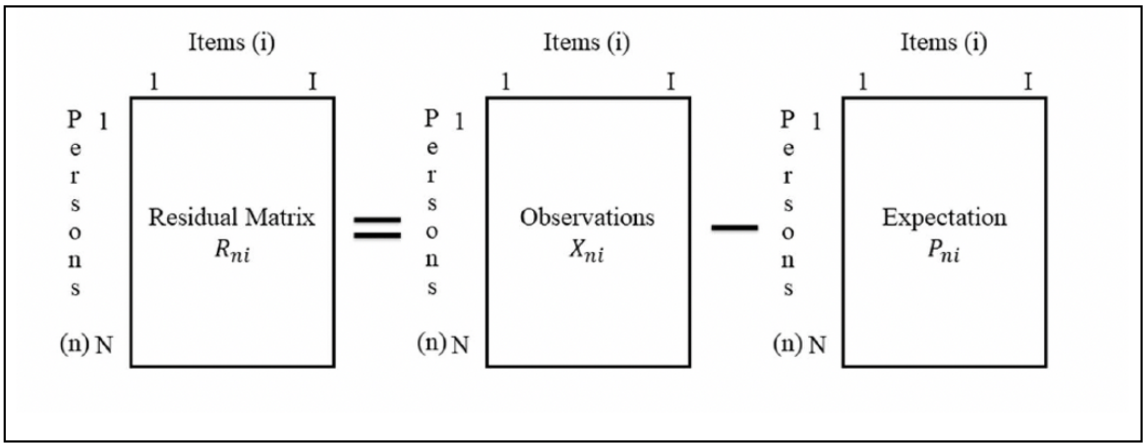

Figure 2. Definition of residuals in measurement

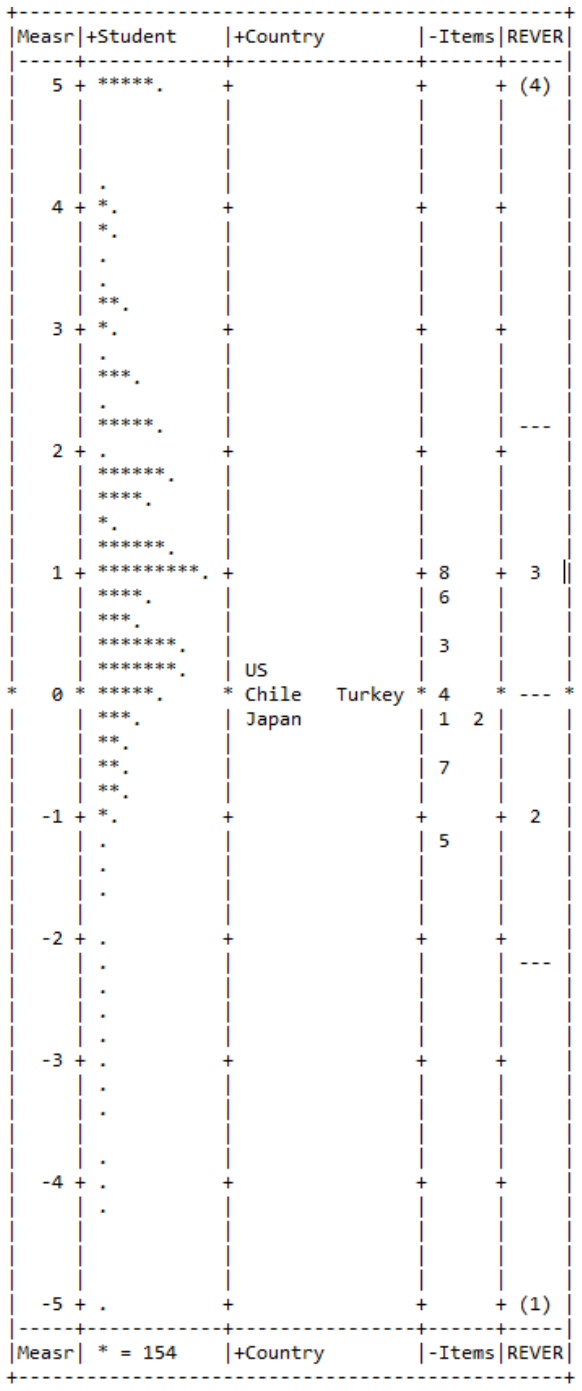

Figure 3. Wright map of mathematics self-efficacy scale

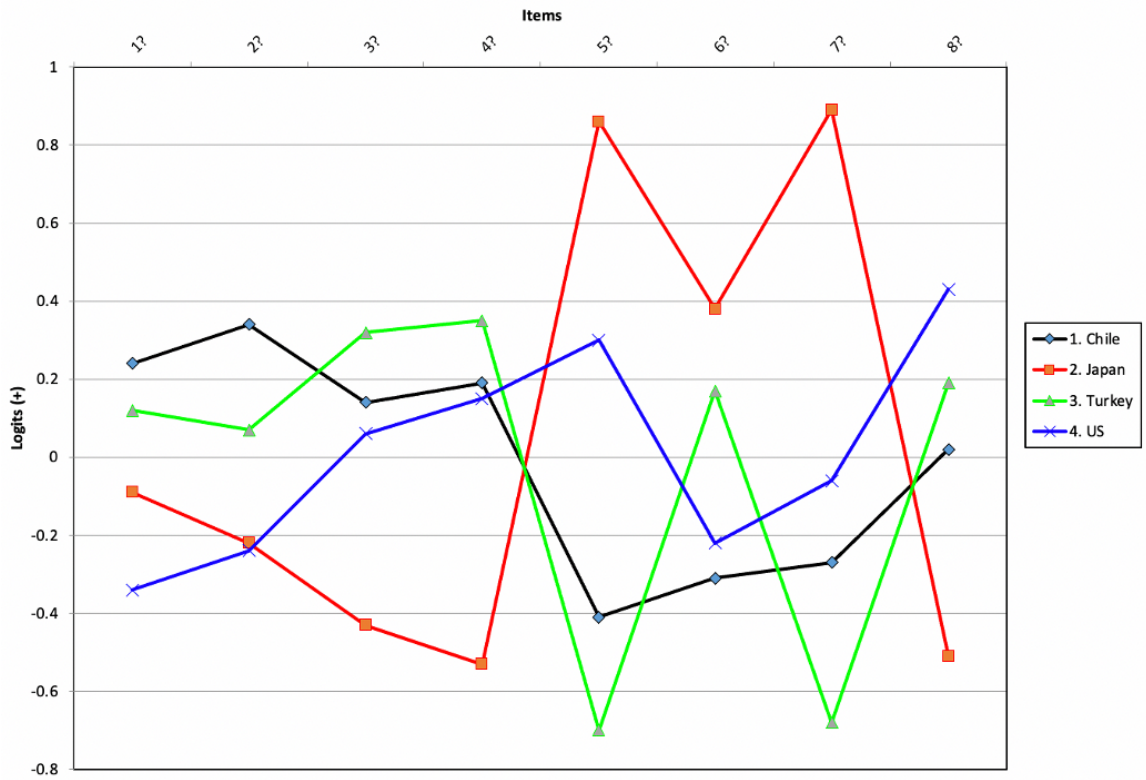

Figure 4. DIF by country (Self-efficacy)

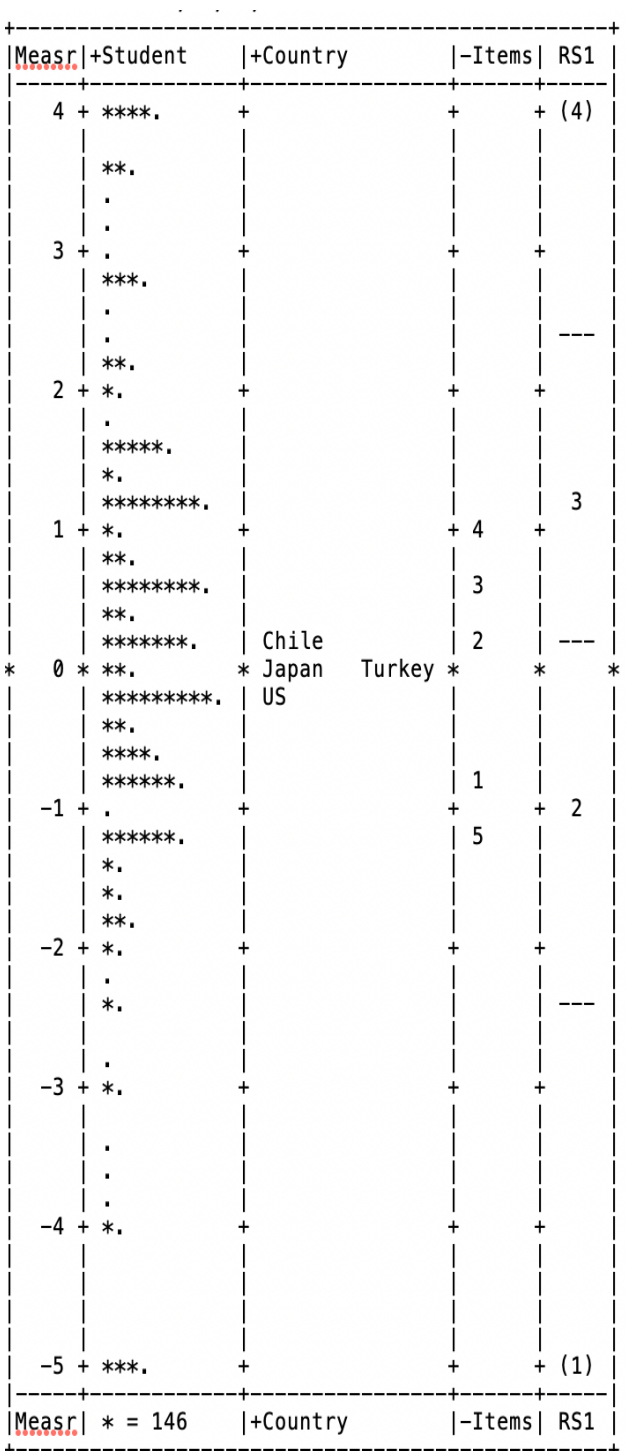

Figure 5. Wright map of mathematics anxiety scale

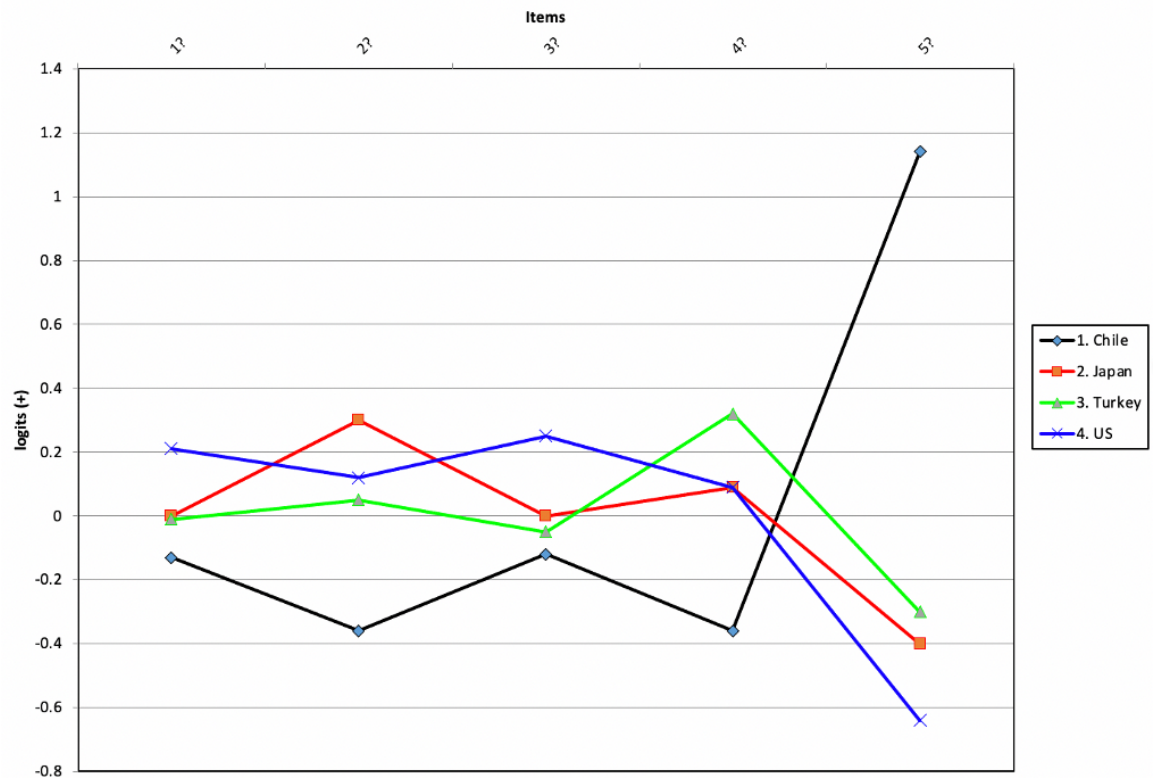

Figure 6. DIF by country (Anxiety)

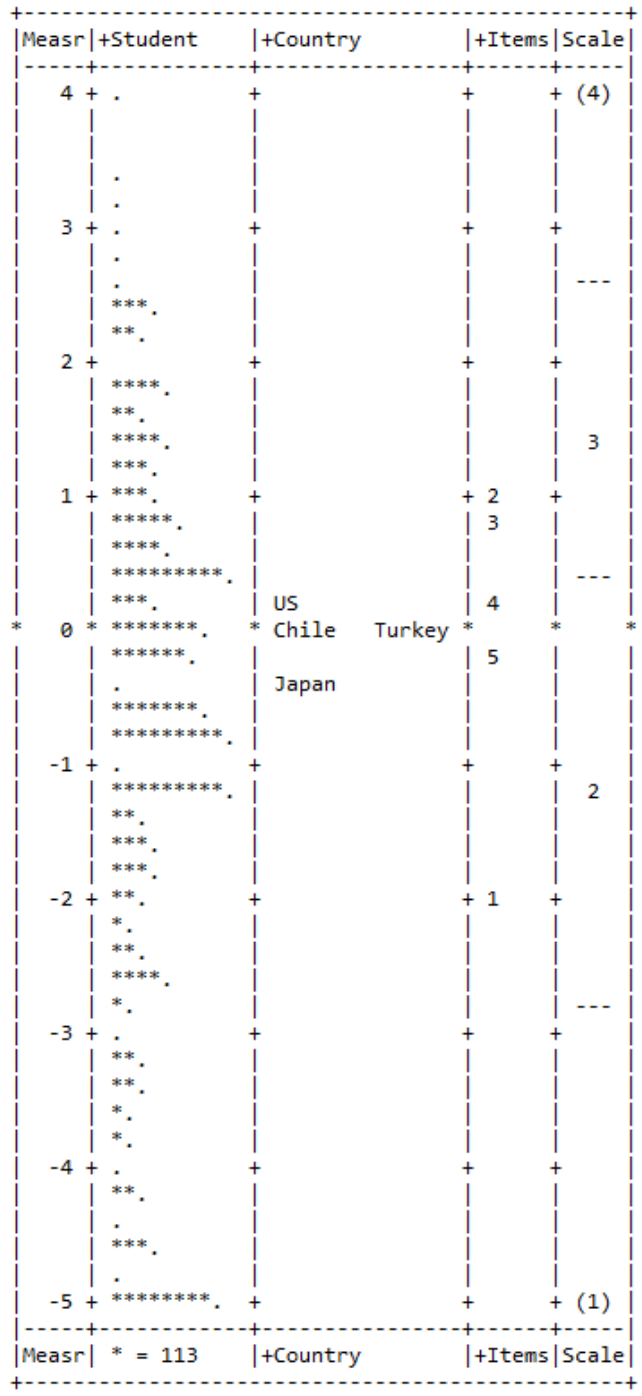

Figure 7. Wright map of mathematics self-concept scale

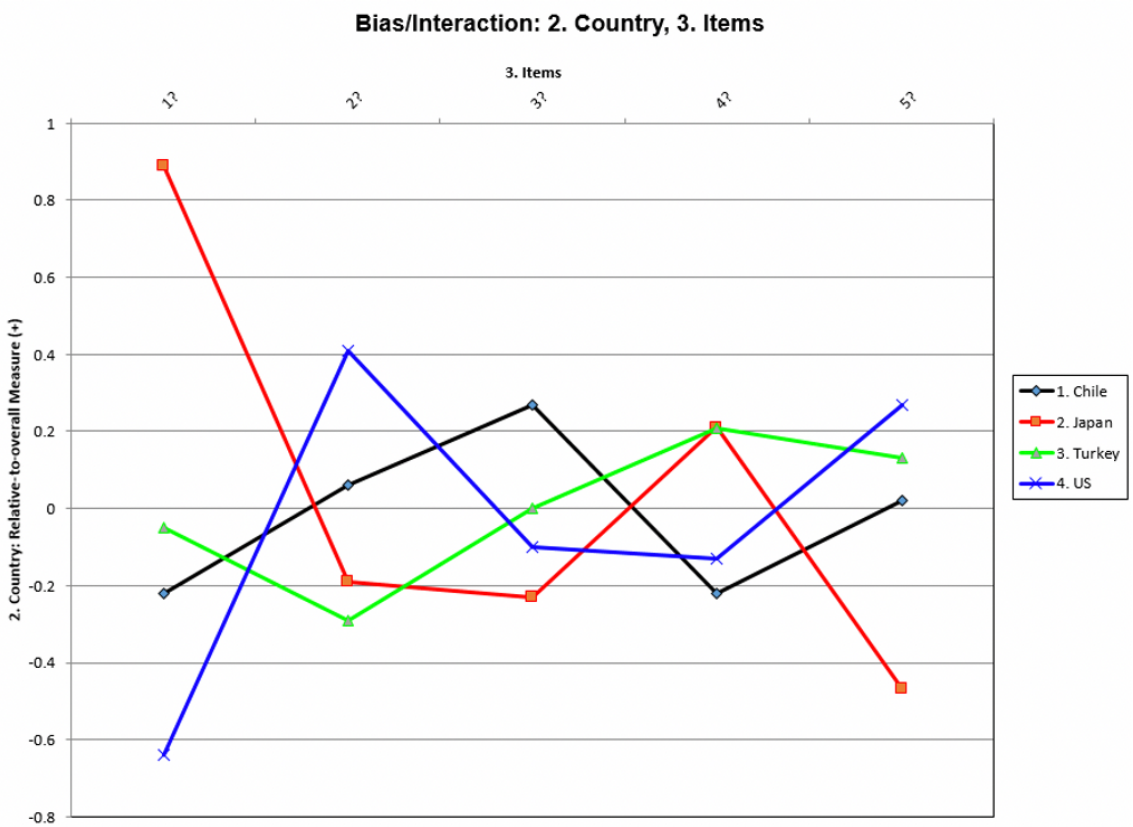

Figure 8. DIF by country (Self-concept)

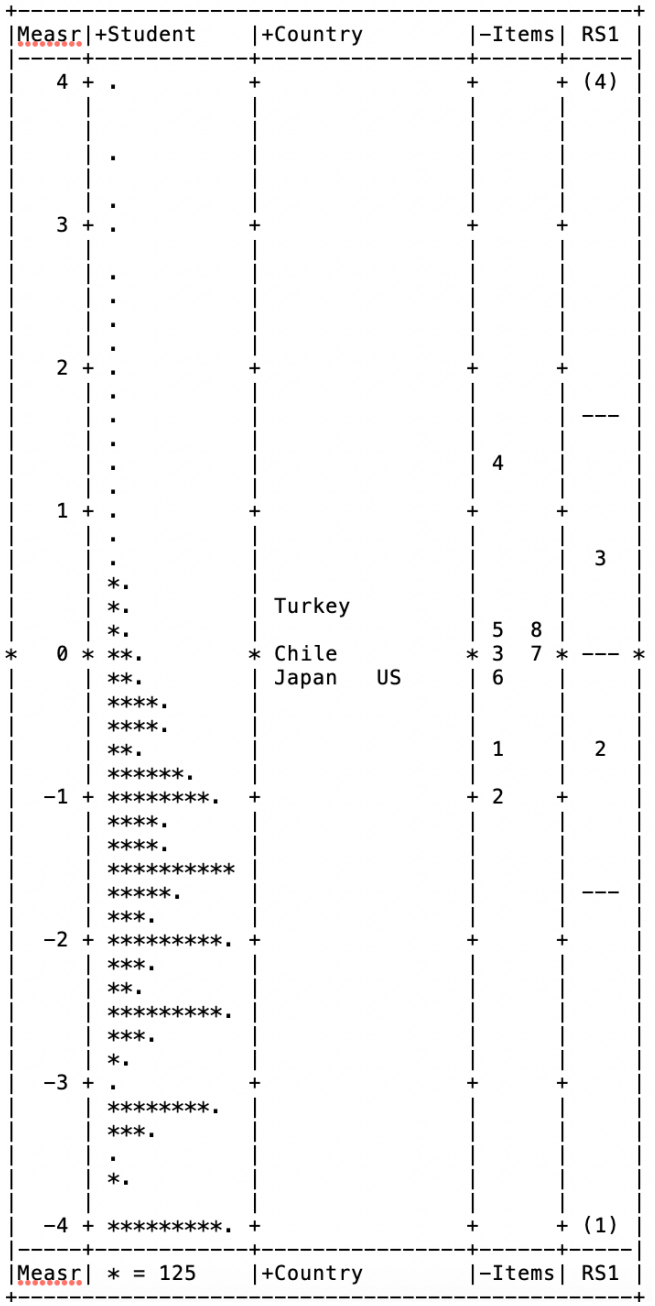

Figure 9. Wright map of mathematics behavior scale

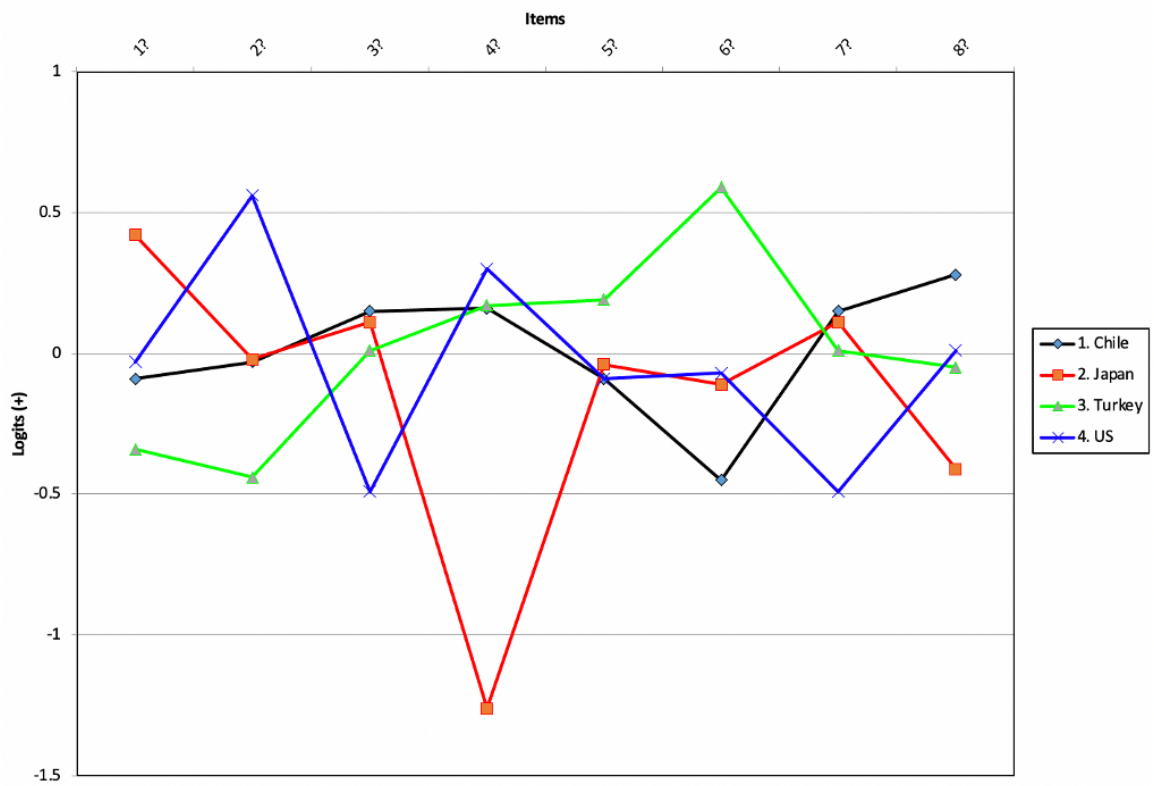

Figure 10. DIF by country (Behavior)