# ASSESSING MEASUREMENT INVARIANCE: MULTIPLE GROUP CONFIRMATORY FACTOR ANALYSIS FOR DIFFERENTIAL ITEM FUNCTIONING DETECTION IN POLYTOMOUS MEASURES OF TURKISH AND AMERICAN STUDENTS

By

DERYA EVRAN

A THESIS PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF ARTS IN EDUCATION

UNIVERSITY OF FLORIDA

2012

To my Mom, who always inspired me and believed in me

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. James Algina, for his professional patience and illimitable academic guidance during my research and study. I would also like to thank my committee member, Dr. Walter Leite, for sharing his ideas and corrections. I was pleased to take the opportunity of Turkish Republic, Ministry of Education and relate with the faculty and students in Research and Evaluation Methodology, University of Florida.

Lastly, I would like to thank my father, Sevki Evran, for his extraordinary encouragement and belief, to my mother, Gursel Evran, for her ineffable love and understanding, to my sisters, Gunsu Evran and Goksu Evran, for tolerating my problems and impressing me with wise opinions and endless patience, despite their age. I am blessed to have a great family and hoping to achieve the goals that they dreamed for me.

TABLE OF CONTENTS

page

LIST OF TABLES

Abstract of Thesis Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Master of Arts in Education

# ASSESSING MEASUREMENT INVARIANCE: MULTIPLE GROUP CONFIRMATORY FACTOR ANALYSIS FOR DIFFERENTIAL ITEM FUNCTIONING DETECTION IN POLYTOMOUS MEASURES OF TURKISH AND AMERICAN STUDENTS

By

Derya Evran

August 2012

Chair: James Algina
Major: Research and Evaluation Methodology

International assessments are often developed in one country and applied in other countries. Assessing the measurement invariance across countries is an important step in determining if valid conclusions can be drawn in these other countries and if valid comparisons can be made across countries under these circumstances. This study investigated measurement invariance, across two countries, of selected questions from the Programme for International Student Assessment 2009 student questionnaire. Turkey and United States were compared by using multiple group confirmatory factor analysis for scores on polytomous items. Specifically, the questions concerned enjoyment of reading, learning strategies, teacher student relations, and disciplinary climate, teachers' stimulation of reading engagement, and teachers' use of strategies. The results indicated that based on the chi-square goodness of fit test and root mean squared error of approximation, none of the models proposed in OECD (2012) adequately fit the data and all questions exhibited differential item functioning (DIF). However, based on the comparative fit index and the Tucker-Lewis index, models for questions, other than except learning strategies, fit the data. For these questions,

results supported the fit of the strict factorial invariance model and indicated that the questions functioned similarly for students in Turkey and the United states. Thus the evidence for DIF was ambiguous except in regard to learning strategies. Under the assumption that the questions functioned similarly in Turkey and the United States, factor mean were compared across the two countries on questions other than learning strategies. According to the comparison, Turkish and American students significantly differed on all questions. The difference was large for enjoyment of reading, moderate for teachers' use of strategies, and small for the remaining questions.

# CHAPTER 1
# INTRODUCTION

Large-scale standardized international assessments have become more important in recent decades as a consequence of globalization. In order to compare and evaluate the quality of education and future workforces across countries, many cross-national student assessments have been conducted to compare learning outcomes across countries. Three prominent examples are the Programme for International Student Assessment sponsored by the Organisation for Economic Co-operation and Development, the Trends in International Mathematics and Science Study sponsored by the Institute of Educational Sciences, and the Progress in International Reading Literacy Study, coordinated by the International Association for the Evaluation of Educational Achievement.

The reasons for translating or adapting tests in cross-cultural assessments were discussed by Hambleton and Kanjee (1995). Three main advantages of adapting tests across cultures are to improve fairness in assessment by allowing persons to be assessed in the language of their choice, to allow comparison studies of groups at an international and national level, and to reduce costs and save time in developing new tests. However, Hambleton and Kanjee (1995) emphasized that if measurements are not equivalent across different groups, then valid comparisons across these groups cannot be made. Investigation of the differences across cultures, genders, ethnicities, and nationalities is valid only to the degree that assessments provided to the various groups meet the requirements of the measurement invariance (Hambleton & Kanjee, 1995; Wolf, 1998). Measurement invariance or equivalence is defined as the degree to which test scores can be used to make comparable inferences for different examinees

(Standards for Educational and Psychological Testing, 1999, p.92). Furthermore, four levels of the measurement invariance were defined: configural invariance, metric invariance, scalar invariance and strict factorial invariance (Meredith, 1993; van de Vijver and Tanzer, 1997).

Hambleton (2005) identified four cultural/language differences that can affect tests scores: construct equivalence of the test for different cultures, test administration, item formats, and speed effect and emphasized the importance of investigating test and item equivalence. Additionally, Hambleton, Sireci and Patsula (2005) identified three sources of bias in test adaptations: construct bias, method bias, and item bias. Construct bias occurs when a construct is not relevant in all cultures in which the test will be used and when the operational definition varies across cultures. Method bias refers to variation in test administration across cultures, differences in familiarity with the items formats, and differential response styles such as social desirability (Hambleton, 1994). Lastly, item bias refers to faulty translation of items and differential relevance of items across cultures.

### 1.1 The Programme for International Student Assessment

One of the large-scale standardized international assessments is the Programme for International Student Assessment (PISA). PISA was developed under the auspices of the Organization for Economic Co-operation and Development's (OECD) and was first conducted in 2000. The main aim of the study is to evaluate education systems worldwide by testing the skills and knowledge of 15-year-old students in participating countries. The survey began with 32 countries in 2000, continued with 41 countries in 2003, 57 countries in 2006, and 67 countries in 2009, with the total of over 70 countries since 2000 (OECD, 2009; OECD, 2012).

PISA is administered to approximately 4500-10000 students in each country every three years. The content of PISA has three domains: reading, mathematics, and science with a focus on basics of subject specific knowledge and the application of the knowledge in real-life situations. The emphasis in the assessment is on the understanding of concepts and the ability to function in various situations within each domain. Test items are a mixture of multiple-choice items and open-ended questions. The items are organized in groups based on a passage setting out a real-life situation with different students taking different combinations of test items. Paper and pencil tests are used and assessments take two hours for each student. However in some countries additional time is allowed for additional digital passages that are provided via computer. Also a student questionnaire is administered to students in order to assess factors such as student background, educational history and goals, and their access to and use of computers. There is also school questionnaire for principals to acquire information about their schools and a parents' questionnaire to acquire information about students' past and present domain engagement at homes. OECD aims to report a basic profile of knowledge and skills of students among countries, contextual indicators of student and school characteristics, trend indicators of change over time, and to provide a valuable knowledge base for policy analysis and research (OECD, 2009; OECD, 2012).

### 1.2 Differential Item Functioning

Lord (1980) used the term item bias to refer to different item response functions for several groups, where an item response function plots the probability of item response as a function of the construct assessed by the items. Later, Angoff (1993) used the term differential item functioning (DIF) instead of item bias. Differential item functioning (DIF) occurs when individuals who are in different groups, for example males and females or

monolingual speakers of two different languages, have the same ability but different probabilities of selecting or constructing answers to items (Holland & Wainer, 1993). DIF analysis often concerns two different subgroups, which are called focal and reference groups. The group of interest is the focal (e.g., racial groups, minorities, women), and the comparison group for focal is the reference (e.g., whites, males) (Holland & Wainer, 1993). DIF analysis can be investigated for both dichotomous and polytomous items and is a major technique for investigating whether scores can be compared meaningfully.

### 1.3 DIF for PISA 2009 Student Questionnaire

OECD has published technical reports for every cycle of PISA. According to these reports, DIF analyses were conducted for some cycles and the focus of the analyses varied across cycles. According to the technical report for PISA 2000 (OECD 2002), DIF was examined for gender, reading competence, and socioeconomic status by using Rasch models. No information on DIF analysis was provided in the reports for PISA 2003 (OECD 2005). DIF analysis was conducted by country and gender for the PISA 2006 (OECD 2009). Item response theory (IRT), specifically Rasch models, was used. According to PISA 2009 technical report (OECD 2012), in the field trials of cognitive items, gender × item interactions and item × country interactions were investigated. After the main survey, DIF for items in French testing countries was examined with results used to modify reading passages for PISA 2009 (OECD, 2012).

Measurement invariance of PISA student questionnaire items is important. As noted in the technical report for the 2009 PISA (OECD 2012), the development of comparable measures of student background, attitudes, and perceptions is a major goal of PISA. Cross-country validity of these constructs is of particular importance as

measures derived from questionnaires are often used to explain differences in student performance within and across countries and are, thus, potential sources of policy relevant information about ways of improving educational systems. (OECD, 2012, p. 286)

Despite the obvious importance of investigating measurement invariance for questionnaire items, the approaches to cross-country validation employed in PISA 2009 were fairly limited and included monitoring translation into different languages and cross-country comparison of reliability coefficients and correlations between a limited number of latent variables. Based on the technical report, cross-country DIF analyses were not conducted for questionnaire items.

Because the instruments used in international assessments are often originated in developed western countries but applied in developed, newly industrialized, and developing countries and across countries with very different cultures, the constructs assessed may not be equally relevant in all countries and the meaning of items used to assess the constructs may vary across countries. Two such countries are the United States a Western developed country and Turkey, a newly industrialized country at the crossroads of Europe and Asia with increasing ties with the west and close cultural, political, economic and industrial relations with the Middle East and other Turkic states of Central Asia.

### 1.4 Research Questions

In this study, the measurement invariance of student questionnaires was investigated in order to examine DIF across two countries. Turkey and United States are compared by using the scores on specific items of the student questionnaire. In the PISA 2009 student questionnaire, a question (e.g., Question 24. “How much do you

agree or disagree with these statements about reading?") consists of several items. For a number of these questions, OECD (2012) reported the number of latent variables measured by the items. But this information was not provided for all of the questions. The questions for which OECD provided information about the number of latent variables measured by the items were the focus of the present study. Specifically, the questions were about enjoyment of reading, learning strategies, teacher student relations, and disciplinary climate, teachers' stimulation of reading engagement, and teachers' use of strategies.

For each such question on the PISA 2009 student questionnaire, the following research questions were addressed:

1. Is there any level of measurement invariance between the United States and Turkey?
2. Is there complete measurement invariance between the United States and Turkey?

# CHAPTER 2
# LITERATURE REVIEW

The aim of the literature review chapter is to introduce differential item functioning (DIF), describe methods for investigating DIF for polytomous items, present the concepts of measurement invariance, and relate these concepts to DIF. In the first section DIF is presented. In the second section, an overview of DIF detection methods for polytomous items is described. Measurement invariance and factor analytic methods to model to test invariance across groups are presented in the third section.

## 2.1 Differential Item Functioning

Holland and Wainer (1993) defined DIF as different probabilities of answering an item correctly by people who are in different groups but are at the same ability level. In the standards of the American Educational Research Association (1999), DIF was defined as a statistical property of a test item in which different groups of test takers who have the same total test score have different average item scores. Also, it is emphasized that fairness, bias, and differential item functioning are not synonymous terms even though they are used similarly in the literature (American Educational Research Association, 1999). In DIF analysis, examinees are in two or more different subgroups. When the number of subgroups is two, the groups are frequently referred to as the focal and reference groups. The focal group is the group of interest, such as minorities or women, and the reference group is the comparison group of the focal, such as whites or males (Holland & Wainer, 1993).

There are a wide variety of methods available for investigating DIF. Two broad classes of methods are those that define ability as an observed variable and those that define ability as a latent variable. In addition, methods can be classified by the type of

item score, dichotomous, nominal, or ordered-categorical. Ordered-categorical items are also called polytomous items. The focus of this review is on methods for detecting DIF on polytomous items. Nevertheless for the sake of simplicity, I will first introduce DIF for dichotomous items. First consider DIF when ability is defined as an observed variable. Let *Y* denote the score for a dichotomous item and *X* the observed total score for the test. Let $E_R$ be the expected value operator for the reference group and $E_F$ the expected value operator for the focal group. DIF is present if there is at least one value of *X* such that

$$E_R[Y \mid X] \neq E_F[Y \mid X] \quad (2\text{-}1)$$

(Chang, Mazzeo & Roussos, 1996). Next consider DIF when ability is defined as a latent variable. Let $\theta$ denote the latent ability underlies item scores on the test. DIF is present if there is at least one value of $\theta$ such that

$$E_R[Y \mid \theta] \neq E_F[Y \mid \theta] \quad (2\text{-}2)$$

(Chang, Mazzeo & Roussos, 1996).

According to Mellenberg (1982), DIF can be of two types: uniform and non-uniform. When DIF is uniform, the between-group difference in item performance is equal over ability levels. When DIF is non-uniform there is an interaction between group membership and ability level *($\theta$)* and the difference in item performance is not same at all ability levels.

The definition of DIF for polytomous items is similar to the definition for dichotomous items but the patterns of DIF are more complex because the number of possible response categories is greater than two. Assume there are *J* possible

response categories on a polytomous item. The focal and reference groups may differ in one or more of the categories and some or all of the categories could show DIF. Moreover, DIF might be in favor of the focal group on some response categories or the reference group on others. Therefore, each response category must be evaluated separately to detect DIF (Vaughn, 2006).

Penfield, Alvarez, and Lee (2009) explained the DIF patterns and dimensions in polytomous items. Pervasiveness is the number of category scores that are affected by DIF. If all *J* categories have DIF, DIF on the item is pervasive. If some of the *J* categories have DIF, then DIF on the item is non-pervasive. Another dimension is consistency, which is the direction and magnitude of DIF. If the category scores differ by the same amounts and always favor the same group then DIF on the item is constant. If the category scores differ by different amounts and always favor the same group then DIF on the item is convergent. If the category scores differ by different amounts and the differences can favor the reference or the focal group then DIF on the item is divergent.

## 2.2 Polytomous DIF Detection Methods

### 2.2.1 Mantel-Haenszel Procedures

Mantel-Haenszel procedures use the total score on a test or a subtest to stratify the sample and formulate a model for association between group membership (focal or reference) and an item score conditional on the stratifying variable. An association between group membership and an item score is evidence of DIF. The first Mantel-Haenszel (MH) procedure was proposed by Holland and Thayer (1998) for DIF detection with dichotomous items.

To help understand MH procedures it is useful to consider responses by the focal and reference group to one item. For the examinees in the $m^{th}$ score group *(m = 1, …,*

*M)* the data can be presented in a *2 × 2* (Group membership × Item score) contingency table, where group membership is in either the focal or reference group and item score is either right or wrong (see Table 2-1).

The estimated odds of a correct response for a member of the reference group is

$$\frac{n_{r1m}}{n_{rm}} \Big/ \frac{n_{r2m}}{n_{rm}} \tag{2-3}$$

and for the focal group is

$$\frac{n_{f1m}}{n_{fm}} \Big/ \frac{n_{f2m}}{n_{fm}} \tag{2-4}$$

If the odds are equal there is no DIF in score group *m*. The odds ratio for the reference group to the focal group is

$$\alpha_m = \frac{\frac{n_{r1m}}{n_{rm}} \Big/ \frac{n_{r2m}}{n_{rm}}}{\frac{n_{f1m}}{n_{fm}} \Big/ \frac{n_{f2m}}{n_{fm}}} = \frac{n_{r1m}\left(n_{f2m}\right)}{n_{f1m}(n_{r2m})} \tag{2-5}$$

and is equal to 1 if there is no DIF in score group *m*. The MH test assumes that the odds ratios are equal across score groups and the common odds ratio is denoted by $\alpha_{MH}$. The null hypothesis is $H_0: \alpha_{MH} = 1$ and the alternative is $H_1: \alpha_{MH} \neq 1$. When the null is rejected there is evidence of DIF.

To apply the MH test to dichotomous item scores, the data are cast in a *2 × 2 × M* (Group membership × Item score × Total score) contingency table (Holland and Wainer, 1993). The estimate of the common odds ratio is

$$\alpha_{MH} = \frac{\sum_{m=1}^{M} \frac{n_{r1m}\left(n_{f2m}\right)}{n_m}}{\sum_{m=1}^{M} \frac{n_{f1m}(n_{r2m})}{n_m}} \tag{2-6}$$

where $n_m = (n_{rm} + n_{fm})$ is the total number of people at the ability level of *m*. The specification of a common odds ratio across total scores is an assumption of the MH procedure.

In order to test the null hypothesis of MH approach, a chi-square test is used:

$$\chi^2_{MH} = [|\sum_m n_{r1m} - \sum_m E(n_{r1m})| - .5]^2 \,/\, \sum_m var(n_{r1m}) \quad (2\text{-}7)$$

where,

$$E(n_{r1m}) = n_{rm}\left(n_{r1m} + n_{f1m}\right)/(n_{rm} + n_{fm}) \quad (2\text{-}8)$$

And

$$var(n_{r1m}) = \frac{n_{rm}(n_m)n_{fm}(n_{r2m} + n_{f2m})}{(n_m)^2((n_m) - 1)} \quad (2\text{-}9)$$

#### 2.2.1.1 The generalized Mantel-Haenszel statistic

The generalized Mantel-Haenszel (GMH) statistic is a Mantel Haenzel statistic for nominal response data (Mantel & Haenszel, 1959). The data for the $m^{th}$ total score group can be presented in a *2* × *J* (Group membership × Item score) contingency table, where *J* is the number of response categories, as illustrated in Table 2-2. The comparison of the focal and reference groups is made on the group differences in the entire response distribution (Zwick, Donoghue & Grima, 1993). Comparing the differences across the entire response scale allows the GMH detect both uniform and non-uniform DIF.

According to Zwick, Donoghue and Grima (1993) the *GMH* $\chi^2$ statistic is

$$\chi^2_{GMH} = [\sum A_m - \sum E(A_m)]'[\sum V(A_m)]^{-1}[\sum A_m - \sum E(A_m)] \quad (2\text{-}10)$$

where $A_m = (n_{R1m}, n_{R2m}, \ldots, n_{R(J-1)m},)$, and $E(A_m)$ are vectors of length $J-1$, and $V(A_m)$ is the covariance matrix of $A_m$. The chi-square statistics in Equation 2-10 provides a comparison of the groups in terms of the entire response distributions. The odds ratios of focal group and reference group for a particular score could be compared (Zwick, Donoghue & Grima, 1993).

According to the formula, the comparison does not account for the order of the response categories. It calculates a chi-square distribution for the entire response distributions for the two groups. According to the Spray and Miller (1994) and Zwick, Donoghue and Grima (1993) the GMH can detect uniform DIF well when sample sizes are moderately large and DIF magnitude is large. GMH has less power for uniform DIF than does the Mantel, which is intended for use with ordered categorical responses and is described in the Section 2.2.1.2. However, Tian (1999) found that Type I error for GMH is highly inflated under conditions of high item discrimination and large group ability differences. As with other techniques, power for GMH to detect uniform DIF increases as studied item discrimination increases (Kristjansson, Aylesworth, Mcdowell & Zumbo, 2005).

**2.2.1.2 The Mantel**

The Mantel is a polytomous extension of the Mantel-Haenszel method (Mantel, 1963) and is used for ordered response categories. The method is applied by assigning index numbers *(1, 2, …)* to the response categories and then comparing the item means of the focal and reference groups who have been matched on a measure of the ability levels (Zwick, Donoghue & Grima, 1993). The data are organized into a $2 \times J \times M$ contingency table. At each of the $M$ levels the data are organized into a $2 \times J$

contingency table, such as Table 2-2. The values of *j (1, 2, 3, .., J)* represent the scores that can be obtained.

According to Zwick, Donoghue and Grima (1993) the Mantel chi-square is,

$$\text{Mantel } \chi^2 = \frac{\left(\sum_m F_m - \sum_m E(F_m)\right)^2}{\sum_m var(F_m)} \tag{2-11}$$

where $F_m$ is the sum of scores for the focal group at the $m^{th}$ level of the matching variable and defined as

$$F_m = \sum_j j n_{Fjm} \tag{2-12}$$

and the expectation of $F_m$ under the null hypothesis $H_0$ no difference is,

$$E(F_m) = \frac{n_{Fm}}{n_m} \sum_j J n_{jm} \tag{2-13}$$

Rejection of the null hypothesis indicates the examinees in the reference and focal groups who are at the same ability level differ in their mean on the specific item. The Mantel can detect uniform DIF well when item discrimination is moderate or high, sample sizes are large, and when DIF magnitude is high (Chang, Mazzeo & Roussos, 1996; Zwick, Donoghue & Grima, 1993); the Mantel cannot detect non-uniform DIF because of the mean comparison on item scores (Spray & Miller, 1994).

#### 2.2.1.3 Liu-Agresti statistic

Liu and Agresti (1996) proposed a GMH method based on the cumulative common odds ratio for several *2* × *J* tables. Penfield and Algina (2003) proposed using this statistic to detect DIF in polytomous items. To apply the Liu-Agresti the data for each of the *M* total score groups are arranged in a *2* × *J* as illustrated in Table 2-2. Then *J-1* dichotomizations are created by collapsing the frequencies at or below response level *j*

into one outcome, and the frequencies above response level *j* into a second outcome, where *j = (1, .., J-1)*. Table 2-3 presents notation for a typical item response and total score. The cumulative odds ratio for the table is

$$\psi_{jm} = \frac{n_{r(\leq j)m}\left(n_{f(>j)m}\right)}{n_{f(\leq j)m}\left(n_{r(>j)m}\right)} \tag{2-14}$$

The cumulative odds ratio is assumed to be equal across the dichotomizations and the levels of the matching variable. The common cumulative odds ratio is

$$\psi_{LA} = \frac{\sum_{k=1}^{K}\sum_{j=1}^{j-1}\frac{n_{r(\leq j)m}\left(n_{f(>j)m}\right)}{n_m}}{\sum_{k=1}^{K}\sum_{j=1}^{j-1}\frac{n_{f(\leq j)m}\left(n_{r(>j)m}\right)}{n_m}} \tag{2-15}$$

Given the dichotomizations defined by Liu and Agresti, $\psi_{LA}$ concerns the odds of a lower score on the polytomous item. The common cumulative odds ratio relevant the odds of a higher score on the polytomous item is

$$\alpha_{LA} = \frac{1}{\psi_{LA}} \tag{2-16}$$

An odds ratio of 1 indicates that the odds of a higher response in the focal group is equal to the odds of a higher response in the reference group and thus indicates absence of DIF. An odds ratio less than 1 suggests DIF in favor of focal group; an odds ratio greater than 1 suggests DIF in favor of reference group (Penfield & Algina, 2003). The advantage of the $\alpha_{LA}$ in DIF investigations is its similarity to $\alpha_{MH}$, thus allowing similar methods to be used with polytomous and dichotomous (Penfield & Algina, 2003).

#### 2.2.1.4 Cox's $\beta$

Cox (1958) proposed a procedure for investigating the relationship between a dichotomous variable and an ordered variable, while controlling for a stratifying variable.

Treating the dichotomous and ordered variable as fixed in a *2* × *J* table such as Table 2-2, Cox modeled the cell probabilities with the multivariate hypergeometric distribution. Cox's $\beta$ is an estimate of the non- centrality parameter *($\beta$)* of the distribution (Camilli & Congdon, 1999; Penfield & Algina, 2003). A value of $\beta$ that is significantly different from zero indicates a relationship between the dichotomous and ordered variable. In the context of a DIF study, the levels of the dichotomous variable are the focal and reference groups, the ordered variable is the polytomous variable, and the stratifying variable is the total score. Cox's $\beta$ was proposed as a DIF statistic and an effect size for DIF detection by Camilli and Congdon (1999). The formula for $\beta$ is

$$\beta = \frac{\sum_{m=1}^{M}\sum_{j=1}^{J} j[n_{Rjm} - E(n_{Rjm})]}{\sum_{m=1}^{M}\zeta_m^2} \tag{2-17}$$

where $n_{Rjm}$ is the frequency of outcome *j* at *m* matching variable for the reference group as in Table 2-1, $E(n_{Rjm})$ is

$$E(n_{Rjm}) = \frac{n_{Rm}n_{jm}}{n_m} \tag{2-18}$$

and $\zeta^2{}_m$ is

$$\zeta_m^2 = \frac{n_{Rm}n_{Fm}}{n_m(n_m - 1)} = \left[\left(\sum_{j=1}^{J} j^2 n_{Jm}\right) - \frac{1}{n_m}\left(\sum_{j=1}^{J} j n_{jm}\right)^2\right] \tag{2-19}$$

(Camilli & Congdon, 1999). The variance of $\beta$ is

$$var(\beta) = (\sum_{m=1}^{M} \zeta_m^2)^{-1} \tag{2-20}$$

Equation 2-17 was proposed by Camilli and Congdon (1999) for the parameter estimate and the Equation 2-21 was proposed to test the null hypothesis that $\beta = 0$ (the studied

item contains no DIF) versus the alternative that $\beta \neq 0$ (the studied item contains DIF) (Penfield and Algina, 2003):

$$vz = \frac{\beta}{\sqrt{V(\beta)}} = \frac{\sum_{m=1}^{M}\sum_{j=1}^{J} j[n_{Rjm} - E(n_{Rjm})]}{\sqrt{\sum_{m=1}^{M}\zeta_m^2}} \quad (2\text{-}21)$$

Squaring the *z* statistic in Equation 2-21 yields the *Mantel* $\chi^2$ in Equation 2-11.

### 2.2.2 Logistic Regression Procedure

Logistic regression consists of models for studying the relationship of dichotomous, nominal or ordered-categorical dependent variables and independent variables. In DIF applications the dependent variables are item scores and the independent variables are group membership and total score on the test. Swaminathan and Rogers (1990) proposed using logistic regression for DIF detection in dichotomous items:

$$P(Y = 1) = \frac{e^z}{1+e^z} \quad (2\text{-}22)$$

Where

$$z = \tau_0 + \tau_1\theta + \tau_2 g + \tau_3 X g \quad (2\text{-}23)$$

*X* is the total test score for an individual and *g* is the group membership defined as *0* for the reference group and *1* for the focal group. The coefficient $\tau_2$ is the group difference in performance on the item for examines with a score of $X = 0$ and $\tau_3$ is the coefficient for the Group × Test Score interaction. If $\tau_2 \neq 0$ and $\tau_3 = 0$ the items shows uniform DIF, if $\tau_3 \neq 0$ the items shows non-uniform DIF. Logistic regression is an effective technique for detecting DIF in dichotomous items (French & Miller, 1996; Zumbo, 1999).

### 2.2.2.1 Polytomous logistic regression (PLR)

Miller and Spray (1993) and French and Miller (1996) proposed three extensions of logistic regression for DIF detection in ordinal items: the cumulative logits model, the adjacent categories model, and the continuation ratio logits model. Each of the models has the following form:

$$logit = \tau_0 + \tau_1 X + \tau_2 X + \tau_3 Xg \tag{2-24}$$

The models differ in terms of how the logit is defined. In the cumulative logits model, the logit of the probability that *Y* is less than or equal to *j* is used:

$$logit[P(Y \leq j|x)] = log \frac{P(Y \leq j|x)}{1 - P(Y \leq j|x)} = log \frac{\pi_1(x) + \cdots + \pi_j(x)}{\pi_{j+1}(x) + \cdots + \pi_j(x)} \tag{2-25}$$

*j = 1, …, J-1* where $\pi_j$ denotes the probability of a response in the *j* $^{th}$ response category of the item. According to Agresti, 2002 no information is lost due to dichotomization of the score categories. In the adjacent categories model, adjacent category logits are used (French and Miller, 1996). The adjacent category logit for the *j* $^{th}$ response category is

$$logit[P(Y = j|Y = j \; or \; j + 1)] = log \frac{\pi_j}{\pi_{j+1}} \tag{2-26}$$

*j = 1, …, J*- 1 (Agresti, 2002). In the continuation ratio logit model, continuation ratio logits are used (French & Miller, 1996). According to Agresti (2002) a continuation ratio logit is

$$logit[P(Y = j)] = log \frac{\pi_j}{\pi_{j+1} + \cdots + \pi_j}, \qquad j = 1, \ldots, J - 1 \tag{2-27}$$

### 2.2.2.2 Logistic discriminant function analysis (LDFA)

Logistic discriminant function analysis (LDFA) is another method proposed by Miller and Spray (1993) for detecting DIF in polytomous items. The group membership

variable, *g*, is used as a dependent variable. The total score *(X)* and the score on the polytomous item *(Y)* are used as independent variables in the model. The model is

$$prob(g|X,Y) = \frac{e^{(1-G)(-a_0-a_1X-a_2Y-a_3XY)}}{1+e^{(-a_0-a_1X-a_2Y-a_3XY)}} \quad (2\text{-}28)$$

The model predicts group membership from total score and item response, and the difference between model deviances is used as a test statistic (Miller & Spray, 1993).

According to Miller and Spray (1993) LDFA detects uniform DIF as well as, or better, than other procedures when sample sizes are large, and studied item discrimination is at least *1.0*. Non-uniform DIF is detected with better power than power for the GMH.

### 2.2.3 Item Response Theory Procedures

An item response theory (IRT) model specifies the relationship between item scores and a latent ability. For example a commonly used IRT model for dichotomous items is the two-parameter model:

$$Prob\left(Y_g = 1 \middle| \theta\right) = \frac{1}{1+\exp[-1.7a(\theta-b)]} \quad (2\text{-}29)$$

where $\alpha$ is the item discrimination parameter and measures how strongly performance on the item depends on the latent ability $\theta$ and *b* is the difficulty parameter and measures how large $\theta$ be in order for examinees to have a 50% chance of answering the item correctly. The right hand side of the Equation 2-29 is called the item response function (IRF). Two approaches have been proposed to using IRT to detect DIF group differences in IRT item parameters and differences in item response function. When the first approach is used either Wald test or the likelihood ratio test can be used to test equality of the item parameters. The Wald test uses

$$PZ = \frac{b_f - b_r}{\sqrt{var\ (b_f - b_r)}} \tag{2-30}$$

And

$$Z = \frac{a_f - a_r}{\sqrt{var\ (a_f - a_r)}} \tag{2-31}$$

as test statistics. The likelihood ratio compares two model, the models in Equation 2-30, Equation 2-31, with Equation 2-32

$$Prob\big(Y_g = 1\big|\theta\big) = \frac{1}{1 + \exp\big[-1.7a_g(\theta - b_g)\big]} \tag{2-32}$$

where the *g* subscript indicates that the item parameters are allowed to vary across groups. Referring to Equation 2-29 and Equation 2-32 as models *A* and, *B* respectively the likelihood ratio test statistic is

$$G^2(df) = 2\log[\frac{likelihood(A)}{likelihood(B)}] \tag{2-33}$$

where *df* is the difference in the number of item parameters in models *B* and *A*. According to the likelihood ration model comparison, DIF is indicated if model *B* fits the data better than *A* does (Thissen, Steinberg & Wainer, 1988).

IRFs can be similar for the groups even when item parameters are quite different. Therefore, comparison of IRFs has been proposed for DIF detection by Kim and Cohen, (1998). Oshima and Morris (2008) provide a review of measures comparing IRFs.

#### 2.2.3.1 Graded response model

Samejima (1969) developed the graded response model for the relationship of scores on polytomous items to the latent ability measured by the items. In the graded response model a response to an item *i* can be classified into *J* categories with scores *1*

to *J* (Samejima, 1999). According to the Samejima (1969) the cumulative response function for category *j* is,

$$P_j = Prob(Y \geq j|\theta) = \frac{1}{1 + \exp\left[-a(\theta - b_j)\right]} \quad (2\text{-}34)$$

where $P_j$ is the probability that a person with a given level of $\theta$ has a score at or above category *j*. The category response function shows the probability of a response in category *j* for a person with a given level of $\theta$. The category response functions are found by taking differences of cumulative response functions (Samejima, 1999):

$$P(Y = j|\theta) = P(Y \geq j|\theta) - P(Y \geq j + 1|\theta) \quad (2\text{-}35)$$

As noted by Muthén and Asparouhov (2002) a more general expression for the cumulative response function is

$$P = Prob(Y \geq j|\theta) \; = \; F\left[a\left(\theta - b_j\right)\right] \quad (2\text{-}36)$$

where *F* can be either a logistic distribution function, as in Equation 2.36 or a normal distribution function. Equation 2-34 uses the logistic response function.

The category probability curve can also be developed from a factor analytic perspective (Muthén & Christoffersson, 1981; Muthén & Asparouhov, 2002). Let *Y* * denote an item-specific latent variable. According to the factor analysis approach to the graded response model

$$Y^* = \; \nu + \; \lambda\eta + \; \epsilon \quad (2\text{-}37)$$

where *Y* * represents a continuous latent variable, $\nu$ is an intercept, $\lambda$ is factor loading, $\eta$ is a factor variable and $\epsilon$ is a residual. The relationship between observed scores *(Y)* and the item-specific latent variable is

$$Y = j \text{ if } \; \tau_{j-1} \leq \; Y^* < \; \tau_j, \qquad j = 1, \cdots, J \quad (2\text{-}38)$$

where the $\tau$ parameters are referred to as thresholds and $\tau_0$ = - ∞ and $\tau_J$ = ∞

According Muthén and Asparouhov (2002), the mean and the variance of latent variable *Y* * are

$$\mu^* = \nu + \lambda\kappa \tag{2-39}$$

And

$$\sigma^* = \lambda^2\varphi + \psi \tag{2-40}$$

where $\kappa$ is the mean of $\eta$, $\varphi$ is the variance of $\eta$ and $\psi$ is the variance of the residual $\epsilon$. According to the factor analytic formulation, the cumulative response function is

$$P(Y \geq j|\eta) = F[-(\tau_j - \nu - \lambda\eta)\psi^{-1/2}] \tag{2-41}$$

where *F* can be a logistic or normal cumulative distribution function. When *F* is a logistic cumulative distribution function, the factor analytic formulation is equivalent to the graded response model in Equation 2-34.

#### 2.2.3.2 Generalized partial credit model

The generalized partial credit model (Muraki, 1992) is an IRT model that is an alternative to the graded response model. The generalized partial credit model is based on the assumption that for a person with a given level of $\theta$ the probability of a score in category *j* rather than category *j-1* is

$$P_j = Prob(Y = j|\theta) = \frac{1}{1 + \exp[-a(\theta - b_j)]} \tag{2-42}$$

Although the parameter $\alpha$ in Equation 2-42 is not subscripted, it can vary across items. The partial credit model (Masters, 1982) is a specialization of the generalized partial credit model in which $\alpha$ = *1* for all items. The generalized partial credit model and the

partial credit model can be used to study DIF by allowing the *a* and $b_j$ parameters to vary across groups such as males and females.

### 2.3 Measurement Invariance

When tests or instruments are used in different countries, measurement invariance of the tests should be assessed in order to demonstrate that the test can be used to make valid comparison of the countries (Steenkamp & Baumgartner, 1998). In international comparisons, differences between nations on a test might be due to the differences on the construct measured by the test or to systematic biases related to the way people from different nations interpret the items on the test.

Hui and Triandis (1985) proposed a conceptual model that for valid cross-cultural comparisons. The first requirement of the model is psychic unity of all mankind. All people in all cultures can be characterized by the same general properties or characteristics, such as emotions, attitudes, and intellect. The first equivalence requirement is conceptual-functional equivalence: a construct that should be meaningfully discussed across cultures and have a universal meaning across cultural groups. The second equivalence requirement is equivalence in construct operationalization: measuring the constructs with the same procedures across the different cultures. The third requirement is item equivalence: each item represents the same thing to the respondents across different cultures, that is, each test item has the same meaning across cultures. The last equivalence requirement is scalar equivalence: the construct is measured on the same measurement scale across cultures, and implies that one examinee's test performance is equivalent to another's only when the test scores are placed on a common ability scale.

### 2.3.1 Methodological Measurement Invariance

The factor analysis model is used to analyze data that are collected on several observed variables. The model for one of the observed variables is

$$Y = \nu + \lambda_1\eta_1 + \cdots + \lambda_s\eta_s + \epsilon \quad (2\text{-}43)$$

where *Y* is an observed variable, which is expressed on a interval or ratio scale, the variables $\eta_1$ to $\eta_s$ are *s* factors measured by the observed variables and $\varepsilon$ is the measurement residual for the variable. The parameters of the model are $\nu$, an intercept for the observed variable, and $\lambda_1$ to $\lambda_s$, the factor loadings which express the relationship between the observed variable and the factors.

The factor analysis model can be either an exploratory or confirmatory model. In the confirmatory factor analysis (CFA) model, a hypothesis about the absence of relationship between observed variables and factors is implemented by setting factor loadings to 0. The factor loadings that are estimated are referred to as free and the factor loadings that are set to 0 are referred to fixed. Using Equation 2-43 and adding the assumptions that (a) the factors and measurement residuals are not correlated and (b) the means for the measurement residuals are 0, Equation 2-43 implies the structural models

$$\Sigma = \Lambda\Phi\Lambda + \Psi \quad (2\text{-}44)$$

And

$$\underline{\mu} = \underline{\nu} + \Lambda\underline{\kappa} \quad (2\text{-}45)$$

In Equation 2-44 $\Sigma$ is the *p* × *p* covariance matrix for the *p* observed variables, $\Lambda$ is a *p* × *s* matrix that contains the factor loadings, $\phi$ is the *s* × *s* covariance matrix for the factors and $\Psi$ is the *p* × *p* covariance matrix for the measurement residual variances. Unless

otherwise noted $\Psi$ is assumed to be diagonal implying that the measurement residuals are not correlated. In Equation 2-45 $\underline{\mu}$ is the *p* × *1* mean vector for the observed variables, $\underline{\nu}$ is the *p* × *1* vector of intercepts, and $\underline{\kappa}$ is the *s* × *1* mean vector for the factors.

When there are several groups of participants in a study, such as the participants from several countries, the model can be extended to the multiple group factor analysis model

$$Y = \nu_g + \lambda_{1g}\eta_1 + \cdots + \lambda_{sg}\eta_s + \epsilon \tag{2-46}$$

where the *g* subscript indicated the *g* $^{th}$ population (group) *g = 1, 2* and the structural model becomes

$$\Sigma_g = \Lambda_g \Phi_g \Lambda_g + \Psi_g \tag{2-47}$$

And

$$\underline{\mu}_g = \nu_g + \Lambda_g \underline{\kappa}_g \tag{2-48}$$

Measurement invariance under the multiple group confirmatory factor analysis model is defined in terms of mathematical equality of the measurement parameters contained in $\Lambda_g$, $\nu_g$, and $\Psi_g$ (Jöreskog, 1971; Meredith, 1993; Steenkamp & Baumgartner, 1998). According to Jöreskog (1971) multiple group factor analysis is the most useful, valid and powerful analysis of testing measurement invariance. Steenkamp and Baumgartner (1998) proposed using multisample confirmatory factor analysis to assess measurement invariance across countries.

### 2.3.2 Levels of Measurement Invariance

#### 2.3.2.1 Configural invariance (Pattern invariance)

According to the factor analysis approach, configural invariance is defined as the same pattern of free and fixed factor loadings across groups. Thus the number of factors is the same in all groups and the fixed loadings are in the same positions in the $\Lambda_g$ matrices. According to the Steenkamp and Baumgartner (1998) configural invariance is supported if the specified model of CFA model fits the data well in all countries.

#### 2.3.2.2 Metric invariance (Weak factorial invariance)

Even if the data meet the requirement of configural invariance, observed scores cannot be compared meaningfully across countries. In the metric invariance model, factor loadings are constrained to be equal across groups $(\Lambda_1 = \Lambda_2)$ and it implies that the scale intervals are equal across groups. Under metric invariance, comparison across groups on difference scores on observed variables are meaningful (Steenkamp & Baumgartner, 1998).

#### 2.3.2.3 Scalar invariance (Strong factorial invariance)

According to the Meredith (1993), in order to conduct valid mean comparisons across countries, differences in means of the observed scores must be due to differences in the means of the underlying construct. The strong factorial invariance model (Meredith, 1993) specifies invariance of both factor loadings $(\Lambda_1 = \Lambda_2)$ and item intercepts ($\nu_1 = \nu_2$) and implies that differences between groups on the observed means are due to the differences between groups on the latent means.

#### 2.3.2.4 Strict factorial invariance

With strict factorial invariance, in addition to scalar invariance measurement residual variances are constrained to be equal ($\Psi_1 = \Psi_2$) across groups (Meredith, 1993).

### 2.3.3 Measurement Invariance in Ordered-Categorical Outcomes

Muthén and Christofferson (1981) extended the multiple group factor analysis model to dichotomous observed variables. The model they used is equivalent to the normal ogive model from IRT. Muthén and Asparouhov (2002) discussed multiple group factor analysis for ordered-categorical variables, using a model that is equivalent to the graded response model from IRT. Millsap & Yun-Tein (2004) showed how to use M*plus* and LISREL to conduct multiple group factor analysis model for ordered-categorical outcomes with several (i.e., *s ≥1*) factors.

To use the approach described by Millsap and Yun-Tein (2004), Equation 2-37 is extended to have multiple factors:

$$Y^* = \nu_g + \lambda_{1g}\eta_1 + \cdots + \lambda_{sg}\eta_s + \varepsilon \quad (2\text{-}49)$$

where the residual $\varepsilon$ is assumed to be uncorrelated with the factors $\eta_1 \ldots \eta_s$. Following Equation 2-38 the observed variable *(Y)* is a discretized version of an item specific latent variable:

$$Y = j \text{ if } \tau_{g(j-1)} \leq Y^* < \tau_{gj} \quad (2\text{-}50)$$

where $\tau_{g0} = -\infty$, and $\tau_{gJ} = \infty$ .The cumulative response function is

$$P(Y \geq j|\eta) = F[-(\tau_{gj} - \nu_g - \lambda_{g1}\eta_1 - \cdots - \lambda_{gs}\eta_s)\psi^{-1/2}] \quad (2\text{-}51)$$

where *F* can be a logistic or normal cumulative distribution function. Then it can be shown that

$$\Sigma_g^* = \Lambda_g \Phi_g \Lambda_g + \Psi_g \tag{2-52}$$

and

$$\underline{\mu}_g^* = \underline{\nu}_g + \Lambda_g \underline{\kappa}_g \tag{2-53}$$

where $\Sigma^*_g$ and $\underline{\mu}_g^*$ are the covariance matrix and mean vector, respectively, for the item-specific latent variables. The other parameters in Equation 2-52 and Equation 2-53 are defined as in the multiple group model for interval and ratio observed variables. With Equation 2-52 and Equation 2-53 established, the configural, metric, strong, and strict factorial invariance models can be defined as in the multiple group model for interval and ratio observed variables. If the strict factorial invariance model describes the data then the cumulative response functions are the same in the two groups and there is no evidence of DIF. If one of the less restricted models fits the data better than there is evidence of DIF.

Table 2-1. Contingency table for group membership × dichotomous item score

| Group | Item score | | Total |
|---|---|---|---|
| | Right | Wrong | |
| Reference | $n_{r1m}$ | $n_{r2m}$ | $n_{rm}$ |
| Focal | $n_{f1m}$ | $n_{f2m}$ | $n_{fm}$ |

Table 2-2. Contingency table for group membership × nominal item score.

| Group | Response level | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | … | $J$ | Total |
| Reference | $n_{R1m}$ | $n_{R2m}$ | $n_{R3m}$ | … | $n_{RJm}$ | $n_{Rm}$ |
| Focal | $n_{F1m}$ | $n_{F2m}$ | $n_{F3m}$ | … | $n_{FJm}$ | $n_{Fm}$ |
| Total | $n_{1m}$ | $n_{2m}$ | $n_{3m}$ | … | $n_{Jm}$ | $n_{m}$ |

Table 2-3. Contingency table for group membership × cumulative item score

| Group | Cumulative item score | |
|---|---|---|
| | $\leq j$ | $> j$ |
| Reference | $n_{r(\leq j)m}$ | $n_{f(>j)m}$ |
| Focal | $n_{f(\leq j)m}$ | $n_{r(>j)m}$ |

# CHAPTER 3
# METHODS

## 3.1 The Programme for International Student Assessment

The Programme for International Assessments (PISA) is an international study. Its purpose is to evaluate educational systems by testing skills and knowledge of 15-year old student in participating countries. PISA was first administered in 2000 and is administered approximately to 4500-10000 students in each participating country every three years. PISA assesses performance in three achievement domains: reading, mathematics and science. The focus of these assessments is not only on the school curriculum but also on the knowledge and skills need for adult life. PISA also assesses problem solving, the application of the knowledge in real-life situations. Although all three domains are assessed in each cycle, the emphasis was on reading in 2000 and 2009, mathematics on 2003 and science in 2006. In addition to achievement and problem solving, context questionnaires are administered to schools, students, and, optionally, to parents in order to assess a variety of constructs. In the 2009 PISA, constructs such as school features, policies and practices (e.g., grade range, student teacher ratio and ability grouping), parental background and support for learning (e.g., parent's education and support of child's reading literacy), and students' self-reports on engagement, learning and teaching (e.g., enjoyment of reading, use of metacognitive strategies, and teachers' stimulation of reading engagement) were assessed (OECD, 2009; OECD 2012). The focus in the present study is on constructs assessed by the student questionnaire. The specific constructs on which the study focuses are listed in section 3.3.

## 3.2 Participants

Participants for the present study came from the PISA 2009 dataset. PISA is approximately administered 4500-10000 students in each participating country. In 2009 475,460 students from 65 countries participated in PISA. According to PISA 2009 Technical Report, the sampling design used for the PISA assessment was a two-stage stratified sample design. The first-stage sampling units were schools having15-year-old students. Schools were sampled from a national list of all PISA-eligible schools, with systematic probability proportional to size sampling in which the measure of size was a function of the estimated number of PISA-eligible 15-year-old students enrolled in the school. Stratification variables used in PISA 2009 varied by country. The second-stage sampling units were students within sampled schools. From a complete list of 15-year-old students in sampled schools in each country, typically 35 students (the fixed target cluster size) were selected with equal probability (OECD 2009, p. 58). For lists of fewer students than the target cluster size all students on the list were selected.

The current study focused on students from the USA and Turkey. In the USA, two stratification variables were used: four regions and schools being public or private, 5233 students from 165 schools in 50 states and 1 district were selected. In Turkey, two stratification variables were used as well: regions and three programs, 4996 students from 170 schools in 12 geographical regions and 55 provinces were selected; all students provided data on the questionnaires.

## 3.3 Measures

As noted previously, the focus of the present study is on constructs assessed in the student questionnaire. The following lists the specific constructs that were the focus of the study:

**Enjoyment of reading.** The assessment of enjoyment of reading included 11 items with responses recorded on a 4-point Likert scale (*strongly disagree*, *disagree*, *agree*, *strongly disagree*). Examples items are “I read only if I have to”, “Reading is one of my favourite hobbies”.

**Learning strategies.** The assessment of learning strategies included 13 items with responses recorded on a 4-point scale (*almost never*, *sometimes*, *often*, *almost always*). According to the PISA 2009 Technical Report (OECD 2012) the 13 items measure three constructs. The first construct is memorization strategies and is measured by four items. An example item is “When I study, I try to memorize everything that is covered in the text.” The second construct, also measured by four items, is Elaboration. An example item is “When I study, I try to understand the material better by relating it to my own experiences.” The third construct is Control Strategies, which is measured by five items. An example item is “When I study, I start by figuring out what exactly I need to learn.”

**Teacher-student relations.** The assessment of teacher-student relations included five items with responses recorded on a 4-point Likert scale (*strongly disagree*, *disagree*, *agree*, *strongly disagree*). Example items are “I get along well with most of my teachers”, “Most of my teachers are interested in my well-being”.

**Disciplinary climate.** The assessment of disciplinary climate included five items with responses recorded on a 4-point scale (*never or hardly ever*, *in some lessons*, *in most lessons*, *in all lessons*). Example items are “Students don’t listen to what the teacher says”, “There is noise and disorder”.

**Teachers' stimulation of reading engagement.** The assessment of teachers' stimulation of reading engagement included seven items with responses recorded on a 4-point scale (*never or hardly ever*, *in some lessons*, *in most lessons*, *in all lessons*). Example items are "The teacher asks students to explain the meaning of a text", "The teacher asks questions that challenge students to get a better understanding of a text".

**Teachers' use of strategies.** The assessment of teachers' use of structuring and scaffolding strategies included nine items with responses recorded on a 4-point scale (*never or hardly ever*, *in some lessons*, *in most lessons*, *in all lessons*). Example items are "The teacher explains beforehand what is expected of the students", "The teacher checks that students are concentrating while working on the reading assignment".

### 3.4 Data Analysis

In order to investigate DIF between countries, multiple group confirmatory factor analysis with latent variables and categorical outcomes under was used (Millsap, Yun-Tein, 2004; Muthén, Asparouhov, 2002). The M*plus* software was used to model the factor analysis of enjoyment of reading, learning strategies of reading, and teacher relations, disciplinary climate under the teachers' stimulation of reading and teaching strategies in learning for USA and Turkey. The models implied by the 2009 PISA report were used. The conceptualizations of these item sets were set forth in the PISA 2009 Technical Report (OECD, 2012). That is, for all item sets except the set used to assess learning strategies a single-factor model was used. For learning strategies a 3-factor model was used.

For each variable, four models were estimated: the configural invariance, metric invariance, strong factorial invariance, and strict factorial invariance models. The CFA model used to investigate configural invariance is

$$Y^* = \nu_g + \lambda_{1g}\eta_1 + \cdots + \lambda_{sg}\eta_s + \varepsilon \quad (3\text{-}1)$$

and the observed variable *(Y)* is a discretized version of the item-specific latent variable *Y **:

$$Y = j \text{ if } \tau_{g(j-1)} \leq Y^* < \tau_{gj} \quad (3\text{-}2)$$

where $\tau_{g0}$ = - ∞ and $\tau_{gJ}$ = ∞ . In this model the variances of *Y ** and of $\varepsilon$ cannot be determined independently. The models were estimated by using M*plus,* which provides two alternatives parameterizations of the model, the delta and theta parameterizations. In the delta parameterization the variance of *Y ** is set equal to *1* for each variable in one group and can be estimated for the second group. In the theta parameterization the variance of $\varepsilon$ is set equal to *1* in one group and can be estimated for the second. The theta parameterization was used in this thesis.

The model in Equation 3-1 is not identified without further constraints. In this thesis the identification constraints proposed in Millsap and Yun-Tein (2004) and reiterated in Millsap (2011) were used. For the configural invariance model these are

1. For all *Y ** set $\nu_g$ = *0*.
2. Among the factor loadings for each factor, set one loading equal to 1.0. The variable with its factor loading set equal to 1.0 is referred to as the reference variable for the factor. The loading for the first variable that loaded on a factor was selected arbitrarily as the reference variable, the default in M*plus*.
3. From among the thresholds $\tau_{g1}$ to $\tau_{g(j-1)}$ set one equal across the models for the USA and Turkey. The threshold $\tau_{g3}$ was selected, arbitrarily.
4. For each reference variable, select one additional threshold to be equal across the models for the USA and Turkey. The threshold $\tau_{g2}$ was selected arbitrarily.
5. For all variables, set the residual variance to one for one group and estimate the variances in the second group. The variances were set equal to one for the USA. The selection of the USA is arbitrary.
6. For one group set the factor means to zero. The USA was arbitrarily selected.

As noted by Millsap and Yun-Tein (2004) the restrictions in the list are sufficient rather than necessary and sufficient to identify the model. That is, other sets of restrictions can also identify the model. For example, in an alternate set of restrictions, restriction 3 to 6 are replaced by

3'. Set the residual variances to one in each group.

and

4'. Set the factor means to zero in each group.

Due to the facts that (a) there are alternate sets of identification constraints and (b) arbitrariness in selecting reference variables and invariant thresholds, when there is DIF the parameter estimates models for the configural, metric, and strong factorial invariance models are not necessarily useful for pinpointing the source of DIF. For example if the fit of the configural invariance model is adequate but the fit of the metric invariance is not, then according to Millsap (2011), “At present, no respecification procedure exists that can be relied on to locate which loadings should vary across groups in all cases.” (p. 140). In this thesis evidence of DIF was be reported but the source of DIF was not investigated.

A model can be judged adequately fitting either by using the chi-square model fit test or by using model fit indices. The chi-square model fit statistic is often discounted because in large samples it can detect small degrees of lack of fit. To supplements the model comparison test the following goodness of fit indices were reported: root mean square error of approximation (RMSEA), Comparative Fit Index (CFI) and the Tucker-Lewis Fit Index (TLI). Values at or below .06 for RMSEA, and at or above .95 for CFI and TLI were considered clear evidence of good fit. These criteria are based on Hu and

Bentler (1999). Values at or below .09 for RMSEA, and at or above .90 for CFI and TLI were considered suggestive evidence of good fit. The less strict criteria for CFI and TLI were in common use before Hu and Bentler's work was published. The less strict criterion for RMSEA is just below the .10 value recommended by Browne and Cudek (1993) as indicating poor fit. It should be noted that these criteria were set forth in the context of factor analysis of continuous variables and it is not clear how well they apply to analysis of ordinal data. A model was considered adequately fitting if the model comparison test was non-significant or at least two of the three indices meet the less strict criteria for goodness of fit.

The configural invariance model is the least restricted model and the strict factorial invariance model is the most restricted. Adequate fit of the strict factorial invariance model is evidence against claims of DIF. Inadequate fit of the strict factorial invariance model is evidence of DIF. An adequately fitting model can be selected by comparing goodness of fit indices or by conducting model comparison tests. Unfortunately there is not agreement in the literature about which approach is the best and the two approaches need not agree. In this thesis the results of both approaches are presented. For each model the goodness of fit chi-square and the RMSEA, CFI, and TLI goodness of fit indices are reported.

Regarding the model comparison approach: for any model pair (e.g., the configural invariance model and the metric invariance model), the model comparison test provides a test of the hypothesis that a less restricted model (i.e., the configural invariance model) fits better than the more restricted model (i.e., the metric invariance model). The main critique of this approach is that when sample sizes are large, as they are in this

study, relatively small degrees of difference in fit, differences that may be relatively inconsequential, can lead to rejection of the more restricted model. Thus, according to this critique, claims of DIF could be based on relatively small differences between the models for the data from USA and Turkish students. Comparison of goodness of fit indices is proposed to address this critique, but there are no agreed-upon guidelines for comparing indices. In this thesis, to compare goodness of fit indices the following criteria were used. For RMSEA the change statistic was calculated by subtracting the fit index for the more restricted model (the metric invariance model) from the index for the less restricted model (the configural invariance model). For CFI and TLI, the change statistic was calculated by subtracting the fit index for the less restricted model (the configural invariance model) from the index for the more restricted model (the metric invariance model). Therefore a negative change in a goodness of fit index indicates a decline in fit.

For CFI and TLI, a change in absolute value that was (a) equal to or smaller than .025 was considered small, (b) larger than .025 and equal to or smaller than .05 was considered moderate, and (c) a change larger than .05 was considered large. For RMSEA a change in absolute value that was (a) equal to or smaller than .01 was considered small, (b) larger than .01 and equal to or smaller than .03 was considered moderate, and (c) larger than .03 was considered large. For CFI and TLI, the change statistic was calculated by subtracting the fit index for the less restricted model (the configural invariance model) from the index for the more restricted model (the metric invariance model). Therefore a negative change in a goodness of fit index indicates a decline in fit.

In addition to reporting goodness of fit information for each model, the following parameter estimates are reported for the configural invariance model: the factor loadings and the thresholds for each group, and the residual variances for Turkey. For questions for which the for the strict factorial invariance model shows adequate fit either by the model comparison test or the goodness of fit indices the factor loadings and the means and variances for both groups are reported. To be considered to show adequate fit by the goodness of fit indices, at least two of RMSEA, TLI and CFI must have meet at least the relaxed criteria for fit indices and the change in fit from the configural invariance model to the strict factorial invariance model must have been no larger than moderate in size for two of the three indices.

The M*plus* software was used to model the factor analysis of enjoyment of reading, learning strategies of reading, and teacher relations, disciplinary climate under the teachers' stimulation of reading and teaching strategies in learning for USA and Turkey. The models implied by the 2009 PISA report were used. The conceptualizations of these item sets were set forth in the PISA 2009 Technical Report (OECD, 2012). That is, for all item sets except the set used to assess learning strategies a single-factor model was used. For learning strategies a 3-factor model was used.

# CHAPTER 4
# RESULTS

Multiple group confirmatory factor analysis was conducted for the following variables: Enjoyment of reading (Question 24), Learning strategies (Question 27), Teacher student relations (Question 34), and Disciplinary climate (Question 36), Teachers' stimulation of reading engagement (Question 37), and Teachings' use of strategies (Question 38). Each question consists of several items. Following the conceptualizations of these item sets that were set forth in the PISA 2009 Technical Report (OECD, 2012), with the exception of the model used for Learning strategies, a single-factor model was specified for each question. For Learning strategies a 3-factor model was specified.

## 4.1 Enjoyment of Reading

The assessment of Enjoyment of reading included 11 items with responses recorded on a 4-point Likert scale (*strongly disagree*, *disagree*, *agree*, *strongly agree*). According to the PISA 2009 Technical Report (OECD 2012), all 11 items are conceptualized as measuring one factor. The model fit chi-square model fit statistic was $\chi^2 (88) = 6992.589, p = .00$ and indicates that the configural invariance model does not fit the data. The goodness of fit indices were *RMSEA = .124, TLI = .939,* and *CFI = .951.* Values at or below .06 for RMSEA, and at or above .95 for CFI and TLI were considered clear evidence of good fit. Values at or below .09 for RMSEA, and at or above .90 for CFI and TLI were considered suggestive evidence of good fit. The RMSEA is larger than both the .09 and .06 criteria. The TLI is somewhat smaller than .95, but larger than .90. The CFI is larger than .95. Overall the evidence is not strongly in support of the fit of the configural invariance model and may indicate that despite the

OECD recommendation of a single factor, more than one factor is required for adequate fit.

The factor loadings from the configural invariance model are shown in Table 4-1. The reference variable was “I read only if I have to”. Because its loading is specified to be 1.0, the polarity of the factor is such that higher scores on the factor indicate negative attitudes toward reading. Thus the items for which higher scores suggest more enjoyment of reading have negative loadings on the factor. Factor loadings that are pointed out for each question exhibit particularly large differences between groups and, while these loadings may or may not be the source of DIF (Millsap, 2011), the differences suggest that not all loadings are invariant across the two countries. For the USA, the loading for the second item (-1.164) has the highest absolute value, the seventh item (-.957) has higher loading for the USA as well whereas the fourth (1.835), sixth (1.502) and eighth (2.031) items have higher loadings for Turkey. The items with higher loadings for the Turkish sample indicate negative attitudes toward reading.

Estimated thresholds and their standard errors are shown in Table 4-2 and Table 4-3 for the USA and Turkey, respectively. The estimates of threshold 3 are the same for both groups as are the estimates of threshold 2 for the first items because these equalities were selected to identify the model. The thresholds that differ substantially between the two groups are pointed out for each question. Although these thresholds may or may not be the source of DIF (Millsap, 2011), the differences suggest that not all thresholds are invariant across the two countries. For question Q24, except first and fifth items, thresholds differ substantially between the two groups.

In the configural invariance model the residual variances for one of the groups (the USA in this study) are specified to be equal to one for all variables and the residual variances are estimated for the other group. The residual variances for question Q24 under the configural invariance model are shown in Table 4-4. The residual variances for first, second, fourth, sixth, eighth, and ninth items exhibit particularly large differences between groups. The results suggest that not all residual variances are invariant across the two countries.

In the metric invariance model, in which factor loadings are constrained to be equal for the USA and Turkey, the model fit chi-square model fit statistic was $\chi^2 (98) = 7803.010, p = .00.$ The goodness of fit indices were *RMSEA = .124, TLI = .939,* and *CFI = .946.* The RMSEA is larger than both the .09 and .06 criteria. The TLI and CFI are somewhat smaller than .95, but larger than .90. Overall the evidence is not strongly in support of the fit of the metric invariance model.

The model comparison chi-square for comparing the configural and metric invariance models was $\chi^2 (10) = 975.467, p = .00$ indicating significant differences in the factor loadings for the students from USA and Turkey. However, the change in the goodness of fit indices was small: *ΔRMSEA = .00, ΔCFI = -.005,* and *ΔTLI = .00.* The small change in the fit indices suggests the metric invariance model fits as well as the configural invariance model. Thus the model comparison test and the comparisons of goodness of fit indices provide conflicting evidence about whether the metric invariance model fits worse than the configural invariance model.

In the strong factorial invariance model, in which factor loadings and thresholds are constrained to be equal for the USA and Turkey, the model fit chi-square model fit

statistic was $\chi^2$ *(119) = 10615.432, p = .00*. The goodness of fit indices were *RMSEA = .132, TLI = .931,* and *CFI = .926*. The RMSEA is larger than both the .09 and .06 criteria. The TLI and CFI are somewhat smaller than .95, but larger than .90. Overall the evidence is not strongly in support of the fit of the strong factorial invariance model.

The model comparison chi-square for comparing the metric and strong factorial invariance models was $\chi^2$ *(21) = 2849.593, p= .00* indicating significant differences in the thresholds for the students from USA and Turkey. The change in the goodness of fit indices was small: *ΔRMSEA = -.008, ΔCFI = -.020,* and *ΔTLI = -.008* suggesting the strong factorial invariance model fits as well as the metric invariance model. The model comparison test and the comparisons of goodness of fit indices provide conflicting evidence about whether the strong factorial invariance model fits worse than the metric invariance model.

In the strict factorial invariance model, in which factor loadings, thresholds, and residual variances are constrained to be equal for the USA and Turkey, the model fit chi-square model fit statistic was $\chi^2$ *(130) = 10113.955, p = .00.* The goodness of fit indices were *RMSEA = .123, TLI = .940,* and *CFI = .929.* The RMSEA is larger than both the .09 and .06 criteria. The TLI and CFI are somewhat smaller than .95, but larger than .90. Overall the evidence is not strongly in support of the fit of the strict factorial invariance model.

The model comparison chi-square for comparing the strong and strict factorial invariance models was $\chi^2$ *(11) = 874.470, p = .00* indicating significant differences in the residual variances for the students from USA and Turkey. The change in the goodness of fit indices was small and positive, *ΔRMSEA = .009, ΔCFI = .003,* and *ΔTLI*

= *.009* indicating that the strict factorial invariance model fits slightly better than the strong factorial invariance model. Moreover the change in goodness of fit indices from the configural invariance model to the strict factorial invariance model was also small: *ΔRMSEA = .001, ΔCFI = -.022,* and *ΔTLI = -.001* and is evidence in support of the strict factorial invariance model.

Because TLI and CFI exceed .90 and the change in goodness of fit indices from the configural invariance model to the strict factorial invariance model was small for all three indices, factor loadings for the strict factorial invariance model are reported in Table 4-5. Comparing these results to those in Table 4-1, the loadings in the two tables would result in the same interpretation of the factor, but tend to be larger in absolute value in Table 4.5. Factor means and variances for Enjoyment of reading are shown in Table 4-6. Assuming the strict factorial invariance model is valid, these results suggest that compared to students in the United States, Turkish students report less enjoyment of reading *($z = -33.15, p < .05$).* Using the square root of the average variance to standardize the mean difference results in *-.78,* suggesting a large difference between the two groups. In addition Turkish students are more homogenous in their enjoyment of reading *($z = -13.70, p < .05$).*

## 4.2 Learning Strategies

The assessment of Learning strategies included 13 items with responses recorded on a 4-point scale (*almost never, sometimes, often, almost always*). According to the PISA 2009 Technical Report (OECD 2012) the 13 items are conceptualized to measure three factors. The first factor is Memorization strategies and is measured by four items, the second factor, also measured by four items, is Elaboration, and the third factor is Control strategies, which is measured by five items. The model fit chi-square model fit

statistic was $\chi^2$ *(124) = 5890.967, p = .00* and indicates that configural invariance model does not fit the data. The goodness of fit indices were *RMSEA = .096, TLI = .903,* and *CFI = .923.* The RMSEA is larger than both the .09 and .06 criteria. Both the CFI and TLI are somewhat smaller than .95, but larger than .90. Overall the evidence is not strongly in support of the fit of the configural invariance model and may indicate that the assignment of items to factors or the number of factors recommended by the OECD is not correct.

The factor loadings from the configural invariance model are shown in Table 4-7. Factor loadings in fifth and seventh items exhibit particularly large differences between groups and the differences suggest that all loadings are not invariant across the two countries. The loadings of the fifth (7.995) and seventh (10.110) items under the Memorization factor, also the sixth (1.675) item under the Control factor have higher loadings for Turkey whereas eleventh item has higher loading for USA (1.210). The items with extreme higher loadings for the Turkish sample are under the Memorization factor of Learning strategies and the higher loading for the USA sample is under the Control factor of the Learning strategies.

Estimated thresholds and their standard errors are shown in Table 4-8 and Table 4-9 for the USA and Turkey, respectively. In fifth, seventh and eleventh items, thresholds differ substantially between the two groups.

The residual variances for question Q27 under the configural invariance model are shown in Table 4-10. The residual variances in fifth and seventh items exhibit particularly large differences between groups. The results suggest that not all residual variances are invariant across the two countries.

The model fit chi-square for the metric invariance model was $\chi^2$ *(134) = 7519.90, p = .00.* The goodness of fit indices were *RMSEA = .104, TLI = .885* and *CFI = .901.* The RMSEA is larger than both the .09 and .06 criteria. The CFI is larger than .90 and TLI is somewhat smaller than .90. Overall the evidence is not in support of the fit of the metric invariance model.

The model comparison chi-square for comparing the configural and metric invariance models was $\chi^2$ *(10) = 1385.877, p = .00* indicating significant differences in the factor loadings for the students from USA and Turkey. The change in the goodness of fit indices was small: *ΔRMSEA = -.008, ΔCFI = -.022,* and *ΔTLI = -.018* suggesting the metric invariance model fits as well as the configural invariance model. Again, the model comparison test and the change in the goodness of fit indices provide conflicting evidence about whether the metric invariance model fits as well as the configural invariance model does.

The model fit chi-square statistic for the strong factorial invariance model was $\chi^2$ *(157) = 13507.629, p = .00.* The goodness of fit indices were *RMSEA = .129, TLI = .823,* and *CFI = .822.* The RMSEA is larger than both the .09 and .06 criteria. The TLI and CFI are smaller than .90. Overall the evidence is not in support of the fit of the strong factorial invariance model.

The model comparison chi-square for comparing the metric and strong factorial invariance models was $\chi^2$ *(23) = 5508.068, p = .00* indicating significant differences in the factor intercepts, thresholds, for the students from USA and Turkey. The change in the goodness of fit indices was moderate for RMSEA *(ΔRMSEA = -.025)* but large for

CFI and TLI *(ΔCFI = -.079),* and *(ΔTLI = -.062).* Overall, the results suggest the strong factorial invariance model fits worse than the metric invariance model.

For the strict factorial invariance model, the model fit chi-square model fit statistic was $\chi^2$ *(170) = 13090.017, p = .00* and the goodness of fit indices were *RMSEA = .122, TLI = .842,* and *CFI = .827.* The RMSEA is larger than .06 and the .09 criteria. The TLI and CFI are somewhat smaller than the both .90 and the .95 criterion. Overall there is little support for adequate fit of the strict factorial invariance model.

The model comparison chi-square for comparing the strong and strict factorial invariance models was $\chi^2$ *(13) = 437.241, p = .00* indicating significant differences in the residual variances for the students from USA and Turkey. The change in the goodness of fit indices was small and positive: *ΔRMSEA = .007, ΔCFI = .005,* and *ΔTLI = .019* indicating that the strict factorial invariance model fits slightly better than the strong factorial invariance model. The evidence about whether the strict factorial invariance model fits as well as the strong factorial invariance model is mixed. The change in goodness of fit indices from the configural invariance model to the strict factorial invariance model was moderate for RMSEA *(ΔRMSEA = -.026)* and large for CFI and TLI *(ΔCFI = -.096),* and *(ΔTLI = -.061).* The preponderance of evidence is against the fit of the strict factorial invariance model and therefore parameter estimates are not reported for the Learning strategies items.

### 4.3 Teacher Student Relations

The assessment of the Teacher student relations included five items with responses recorded on a 4-point Likert scale (*strongly disagree, disagree, agree, strongly disagree*). According to the PISA 2009 Technical Report (OECD 2012), all five

items are conceptualized as measuring one factor. The model fit chi-square model fit statistic was $\chi^2$ *(10) = 592.437, p = .00* and indicates that configural invariance model does not fit the data. The goodness of fit indices were *RMSEA = .107, TLI = .986,* and *CFI = .993.* The RMSEA is larger than both the .09 and .06 criteria. The TLI and CFI are both larger than .95. Considering the large values for TLI and CFI the evidence is somewhat in support of the fit of the configural invariance model.

The factor loadings from the configural invariance model are shown in Table 4-11. All of the items have similar factor loadings for Turkey and USA except the last item, which indicates higher loading (1.207) for the students in USA.

Estimated thresholds and their standard errors are shown in Table 4-12 and Table 4-13 for the USA and Turkey, respectively. In second and fifth items thresholds differ substantially between the two groups.

The residual variances for question Q34 under the configural invariance model are shown in Table 4-14. The residual variances in second item exhibit particularly large differences between groups. The results suggest that not all residual variances are invariant across the two countries.

The model fit chi-square for the metric invariance model was $\chi^2$ *(14) = 1602.303, p= .00.* The goodness of fit indices were *RMSEA = .150, TLI = .972,* and *CFI = .981.* The RMSEA is larger than both the .09 and .06 criteria. The TLI and CFI are larger than .95. Overall the evidence is somewhat in support of the fit of the metric invariance model.

The model comparison chi-square for comparing the configural and metric invariance models was $\chi^2$ *(4) = 875.213, p = .00* indicating significant differences in the

factor loadings for the students from USA and Turkey. The change in the goodness of fit indices was large for RMSEA *(ΔRMSEA = -.043)* and small for CFI and TLI *(ΔCFI = -.012),* and *(ΔTLI = -.014).* Thus the model comparison test and the comparisons of goodness of fit indices provide conflicting evidence about whether the metric invariance model fits worse than the configural invariance model.

The model fit chi-square statistic for the strong invariance model was $\chi^2$ *(23) = 2102.491, p = .00*. The goodness of fit indices were *RMSEA = .134, TLI = .978,* and *CFI = .975.* The RMSEA is larger than both .06 and .09 criteria. The TLI and CFI are larger than .95. Overall the evidence is somewhat in support of the fit of the strong factorial invariance model.

The model comparison chi-square for comparing the metric and strong factorial invariance models was $\chi^2$ *(9) = 737.708, p = .00* indicating significant differences in the factor intercepts, thresholds, for the students from USA and Turkey. The change in the goodness of fit indices was moderate for RMSEA *(ΔRMSEA = 0.16)* and small for CFI *(ΔCFI = -.006),* and TLI and *(ΔTLI = .006).* Again, the model comparison test and the change in the goodness of fit indices provide conflicting evidence about whether the strong factorial invariance model fits as well as the metric invariance model does.

For the strict factorial invariance, the model fit chi-square statistic was $\chi^2$ *(28) = 2920.356, p = .00* and the goodness of fit indices were *RMSEA = .143, TLI = .975,* and *CFI = .965.* The RMSEA is larger than both the .09 and .06 criteria. The TLI and CFI are larger than .95. Overall the evidence is somewhat in support of the fit of the strict factorial invariance model.

The model comparison chi-square for comparing the strong and strict factorial invariance models was $\chi^2$ *(5) = 809.185, p = .00* indicating significant differences in the residual variances for the students from USA and Turkey. The change in the goodness of fit indices was small: *ΔRMSEA = -.009, ΔCFI = -.010,* and *ΔTLI = -.003* suggesting the strict factorial invariance model fits as well as the strong factorial invariance model. The evidence about whether the strict factorial invariance model fits as well as the strong factorial invariance model is mixed. The change in goodness of fit indices from the configural invariance model to the strict factorial invariance model was large for RMSEA *(ΔRMSEA = -.036)*, moderate for CFI *(ΔCFI = -.028),* and small for TLI *(ΔTLI = -.011)*. Thus, the model comparison test and the change in the goodness of fit indices provide conflicting evidence about whether the evidence is in support of the strict factorial invariance model fit.

Because TLI and CFI exceed .95 and the change in goodness of fit indices from the configural invariance model to the strict factorial invariance model was small to moderate for TLI and CFI, factor loadings and means and variances are reported in Table 4-15. Comparing these results to those in Table 4-11, the loadings in the two tables would result in the same interpretation of the factor, but tend to be larger in Table 4.15. Factor means and variances for Teacher student relations shown in Table 4-16. Assuming the strict factorial invariance model is valid, these results suggest that compared to students in the United States, Turkish students report better teacher student relations *(z = 8.30, p < .05).* The standardized the mean difference is *.18* suggesting a small difference between the two groups. In addition United States students are more homogenous in their opinions about teacher relations *(z = 6.05,*

*p < .05)*

### 4.4 Disciplinary Climate

The assessment of Disciplinary climate included five items with responses recorded on a 4-point scale (*never or hardly ever, in some lessons, in most lessons, in all lessons*). According to the PISA 2009 Technical Report (OECD 2012), all five items are conceptualized as measuring one factor. The model fit chi-square model fit statistic was $\chi^2$ *(10) = 729.238, p = .00* and indicates that configural invariance model does not fit the data. The goodness of fit indices were *RMSEA = .119, TLI = .978,* and *CFI = .989.* The RMSEA is larger than both the .09 and .06 criteria. The TLI and the CFI are larger than .95. Considering the large values for TLI and CFI the evidence is somewhat in support of the fit of the configural invariance model.

The factor loadings from the configural invariance model are shown in Table 4-17. All of the items have higher factor loadings for Turkey compared to the USA. The largest difference is on third item with the loading (2.386) for Turkish students.

Estimated thresholds and their standard errors are shown in Table 4-18 and Table 4-19 for the USA and Turkey, respectively. All of thresholds differ substantially between the two groups.

The residual variances for question Q36 under the configural invariance model are shown in Table 4-20. All of the residual variances exhibit particularly large differences between groups. The results suggest that the residual variances are not invariant across the two countries.

The model fit chi-square for the metric invariance model was $\chi^2$ *(14) = 872.924, p = .00.* The goodness of fit indices were *RMSEA = .110, TLI = .981,* and *CFI = .987.* The

RMSEA is larger than both the .09 and .06 criteria. The TLI and CFI are larger than .95. Considering the large values for TLI and CFI the evidence is somewhat in support of the fit of the metric invariance model.

The model comparison chi-square for comparing the configural and metric invariance models was $\chi^2$ *(4) = 222.216, p = .00* and indicating significant differences in the factor loadings for the students from USA and Turkey. The change in the goodness of fit indices was small: *ΔRMSEA = .009, ΔCFI = -.002,* and *ΔTLI = .003* suggesting the metric invariance model fits as well as the configural invariance model. The model comparison test and the comparisons of goodness of fit indices provide conflicting evidence about whether the metric invariance model fits worse than the configural invariance model.

The model fit chi-square statistic for the strong invariance model was $\chi^2$ *(23) = 1602.929, p = .00.* The goodness of fit indices were *RMSEA = .117, TLI= .979,* and *CFI = .975.* The RMSEA is larger than both the .06 and .09 criteria. The TLI and CFI are both larger than .95. Considering the large values for TLI and CFI the evidence is somewhat in support of the fit of the strong factorial invariance model.

The model comparison chi-square for comparing the metric and strong factorial invariance models was $\chi^2$ *(9) = 732.900, p = .00* indicating significant differences in the factor intercepts, thresholds, for the students from USA and Turkey. The change in the goodness of fit indices was small: *ΔRMSEA = -.007, ΔCFI = -.012,* and *ΔTLI = -.007* suggesting the strong factorial invariance model fits as well as the metric invariance model. Again the model comparison test and the comparisons of goodness of fit indices

provide conflicting evidence about whether the strong factorial invariance model fits worse than the metric invariance model.

For the strict factorial invariance model, the model fit chi-square statistic was $\chi^2$ *(28) = 2164.150, p = .00* and the goodness of fit indices were *RMSEA = .123, TLI = .976,* and *CFI = .967.* The RMSEA is larger than both the .09 and .06 criteria. The TLI and CFI are both larger than .95. Considering the large values for TLI and CFI the evidence is somewhat in support of the fit of the strict factorial invariance model.

The model comparison chi-square for comparing the strong and strict factorial invariance models was $\chi^2$ *(5) = 577.263, p = .00* indicating significant differences in the residual variances for the students from USA and Turkey. The change in the goodness of fit indices was small: *ΔRMSEA = -.006, ΔCFI = -.008,* and *ΔTLI = -.003* suggesting the strict factorial invariance model fits as well as the strong factorial invariance model. The evidence about whether the strict factorial invariance model fits as well as the strong factorial invariance model is mixed. The change in goodness of fit indices from the configural invariance model to the strict factorial invariance model was small: *ΔRMSEA = -.004, ΔTLI = -.002,* and *ΔCFI = -.022,* which means evidence in support of the strict factorial invariance model.

Because TLI and CFI exceed .95 and change in goodness of fit indices from the configural invariance model to the strict factorial invariance model was small, factor loadings and means and variances are reported in Table 4-17. Comparing results to those in Table 4-17, the loadings in the two tables would result in the same interpretation of the factor, but tend to be smaller in Table 4.17. Note that because higher scores on the items indicate that the students perceive the disciplinary climate to

be bad, higher scores on the factor also indicate a bad disciplinary climate. Factor means and variances for Disciplinary climate are shown in Table 4-22. Assuming the strict factorial invariance model is valid, these results suggest that compared to students in the United States, Turkish students report a worse disciplinary climate in classrooms *(z = 3.30, p < .05).* However, the standardized mean difference is *.07* suggesting a small difference between the two groups. The variances for the factor are not significantly different *(z = -0.17, p > .05).*

### 4.5 Teachers' Stimulation of Reading Engagement

The assessment of Teachers' stimulation of reading engagement included seven items with responses recorded on a 4-point scale (*never or hardly ever, in some lessons, in most lessons, in all lessons*). According to the PISA 2009 Technical Report (OECD 2012), all seven items are conceptualized to measure one factor. The model fit chi-square model fit statistic was $\chi^2 (28) = 4817.595, p = .00$ and indicates that the configural invariance model does not fit the data. The goodness of fit indices was *RMSEA =. 184, TLI = .902,* and *CFI = .934.* The RMSEA is larger than both the .09 and .06 criteria. The CFI and TLI are somewhat smaller than .95, but larger than .90. Overall the evidence is not strongly in support of the fit of the configural invariance model and may indicate that despite the OECD recommendation of a single factor, more than one factor is required.

The factor loadings from the configural invariance model are shown in Table 4-23. The largest difference is on fourth item with the loading (.976) for Turkish students. All the other items have similar factor loadings for Turkey and USA. There are not any other large differences in the factor loadings for the groups.

Estimated thresholds and their standard errors are shown in Table 4-24 and Table 4-25 for the USA and Turkey, respectively. In fourth and seventh items, thresholds differ substantially between the two groups.

The residual variances for question Q37 under the configural invariance model are shown in Table 4-26 for the Turkish students. The residual variances in first and fourth items exhibit particularly large differences between groups. The results suggest that not all residual variances are invariant across the two countries.

The model fit chi-square for the metric invariance model was $\chi^2 (34) = 4843.351, p = .00.$ The goodness of fit indices were *RMSEA = .167*, *TLI = .919*, and *CFI = .934*. The RMSEA is larger than both the .09 and .06 criteria. The TLI and CFI are somewhat smaller than .95, but larger than .90. Overall the evidence is not strongly in support of the fit of the metric invariance model.

The model comparison chi-square for comparing the configural and metric invariance models was $\chi^2 (6) = 249.970, p= .00$ indicating significant differences in the factor loadings for the students from USA and Turkey. The change in the goodness of fit indices was moderate for RMSEA *(ΔRMSEA = .017)* and small for TLI *(ΔTLI = .017),* and CFI *(ΔCFI = .00).* The change in the goodness of fit indices was positive, indicating that the metric invariance model fits slightly better than the configural invariance model.

The model fit chi-square statistic for the strong invariance model was $\chi^2 (47) = 4822.649, p = .00$. The goodness of fit indices were *RMSEA = .142, TLI = .942,* and *CFI = .935.* The RMSEA is larger than both the .09 and .06 criteria. The TLI and CFI are larger than .90. Overall the evidence is not strongly in support of the fit of the strong factorial invariance model.

The model comparison chi-square for comparing the metric and strong factorial invariance models was $\chi^2$ *(13) = 612.230, p = .00* indicating significant differences in the thresholds for the students from USA and Turkey. The change in the goodness of fit indices was moderate for RMSEA *(ΔRMSEA = .025)* and small for CFI *(ΔCFI = .001),* and TLI *(ΔTLI = .023).* Again, the change in the goodness of fit indices was positive, indicating that the strong factorial invariance model fits slightly better than the metric invariance model.

For the strict factorial invariance model, the model fit chi-square statistic was $\chi^2$ *(54) = 6348.671, p = .00* and the goodness of fit indices were *RMSEA = .152, TLI = .933*, and *CFI = .914*. The RMSEA is larger than both the .09 and .06 criteria. The TLI and CFI are somewhat larger than .90. Overall the evidence is not strongly in support of the fit of the strict factorial invariance model.

The model comparison chi-square for comparing the strong and strict factorial invariance models was $\chi^2$ *(7) = 1631.779, p = .00* indicating significant differences in the residual variances for the students from USA and Turkey. The change in the goodness of fit indices was small: *ΔRMSEA = -.010, ΔCFI = -.021,* and *ΔTLI = -.009* suggesting the strict factorial invariance model fits as well as the strong factorial invariance model. The evidence about whether the strict factorial invariance model fits as well as the strong factorial invariance model is mixed. The change in goodness of fit indices from the configural invariance model to the strict factorial invariance model was large for RMSEA *(ΔRMSEA = .032),* small for CFI *(ΔCFI = -.020),* and moderate for TLI *(ΔTLI = .031).* Thus, the model comparison test and the change in the goodness of fit

indices provide conflicting evidence about whether the evidence is in support of the strict factorial invariance model fit.

Because TLI and CFI exceed .95 and the change in goodness of fit indices was small for CFI and moderate for TLI, factor loadings and means and variances are reported in Table 4-27. Comparing these results to those in Table 4-23, the loadings in the two tables would result in the same interpretation of the factor, but tend to be larger in Table 4.27. Factor means and variances for Teachers' stimulation of reading engagement are shown in Table 4-28. Assuming the strict factorial invariance model is valid, these results suggest that compared to students in the United States, Turkish students report more teachers' stimulation of reading engagement *($z = 6.55, p < .05$)*. The standardized mean difference is *.14* suggesting a small difference between the two groups. In addition United States students are more homogenous in their teachers' stimulation of reading engagement *($z = 3.43, p < .05$).*

### 4.6 Teachers' Use of Strategies

The assessment of Teachers' use of structuring and scaffolding strategies and included nine items with responses recorded on a 4-point scale (*never or hardly ever, in some lessons, in most lessons, in all lessons*). According to the PISA 2009 Technical Report (OECD 2012), all nine items are conceptualized to measure one factor. Estimation of the configural invariance would not converge. Consequently, the factor loadings, thresholds and residual variances are not reported for the configural variance model.

The model fit chi-square for the metric invariance model was $\chi^2 (62) = 3274.944$, $p = .00$ and indicates that the metric invariance model does not fit the data. The

goodness of fit indices were *RMSEA = .102*, *TLI = .958*, and *CFI = .964*. The RMSEA is larger than both .06 and .09 criteria. The TLI is equal to and CFI is larger than .95. Considering the large values for TLI and CFI the evidence is somewhat in support of the fit of the metric invariance model.

Estimated thresholds and their standard errors are shown in Table 4-29 and Table 4-30 for the USA and Turkey, respectively. In second, third, fourth, sixth, seventh, eighth and ninth items thresholds differ substantially between the two groups.

The residual variances for question Q38 under the metric invariance model are shown in Table 4-31. All of the residual variances except ninth item exhibit particularly large differences between groups. The results suggest that not all residual variances are invariant across the two countries.

The model fit chi-square statistic for the strong invariance model was $\chi^2 (79) = 4527.973, p = .00.$ The goodness of fit indices were *RMSEA = .106*, *TLI = .954*, and *CFI = .950*. The RMSEA is larger than.06 and .09 criteria. The CFI is equal to and the TLI is larger than .95. Overall the evidence is not strongly in support of the fit of the strong factorial invariance model.

The model comparison chi-square for comparing the metric and strong factorial invariance models was $\chi^2 (17) = 1254.915, p = .00$ indicating significant differences in the thresholds for the students from USA and Turkey. The change in the goodness of fit indices was small: $\Delta RMSEA = -.004$, $\Delta CFI = -.014$, and $\Delta TLI = -.004$ suggesting the strong factorial invariance model fits as well as the metric invariance model. The model comparison test and the comparisons of goodness of fit indices provide conflicting

evidence about whether the strong factorial invariance model fits worse than the metric invariance model.

For the strict factorial invariance model, the model fit chi-square statistic was $\chi^2$ *(88) = 6729.035, p= .00* and the goodness of fit indices were *RMSEA = .123*, *TLI = .938*, and *CFI = .925.* The RMSEA is larger than both the .09 and .06 criteria. Both the TLI and CFI are somewhat smaller than .95, but larger than .90. Overall the evidence is not strongly in support of the fit of the strict factorial invariance model.

The model comparison chi-square for comparing the strong and strict factorial invariance models was $\chi^2$ *(9) = 1884.642, p = .00* indicating significant differences in the residual variances for the students from USA and Turkey. The change in the goodness of fit indices was moderate for RMSEA *(ΔRMSEA = -.017)*, and small for CFI *(ΔCFI = -.025)*, and TLI *(ΔTLI = -.016)*. The evidence about whether the strict factorial invariance model fits as well as the strong factorial invariance model is mixed. The change in goodness of fit indices from the metric invariance model to the strict factorial invariance model was moderate for RMSEA *(ΔRMSEA = -.021),* and CFI *(ΔCFI = -.039),* and small for TLI *(ΔTLI = -.020).* Thus, the model comparison test and the change in the goodness of fit indices provide conflicting evidence about whether the evidence is in support of the strict factorial invariance model fit.

Because TLI and CFI exceed .90 and the change in goodness of is no larger than moderate for any of the three indices, factor loadings are reported in Table 4-32. The seventh item (1.571) has the highest, and the sixth item (.969) has the lowest factor loading. Note that for this question, factor loadings from configural invariance model would not calculated, and cannot be compared with the loadings from strict factorial

invariance model. Factor means and variances for Teachers' use of structuring and scaffolding strategies are shown in Table 4-33. Assuming the strict factorial invariance model is valid, these results suggest that compared to students in the United States, Turkish students report less teachers' use of strategies *(z = 14.10, p < .05)*. Using the square root of the average variance to standardize the mean difference results in *.30*, suggesting a moderate difference between the two groups. In addition United States students are more homogenous in their teachers' use of strategies *(z = 10.87, p < .05).*

Table 4-1. Factor loadings and standard errors for enjoyment of reading

| Item | USA | | Turkey | |
|---|---|---|---|---|
| | Factor loading | Standard error | Factor loading | Standard error |
| 1. I read only if I have to | 1.000* | | 1.000* | |
| 2. Reading is one of my favourite hobbies | -1.164 | 0.030 | -0.911 | 0.048 |
| 3. I like talking about books with other people | -0.945 | 0.026 | -0.893 | 0.047 |
| 4. I find it hard to finish books | 0.478 | 0.014 | 1.835 | 0.169 |
| 5. I feel happy if I receive a book as a present | -0.904 | 0.025 | -0.959 | 0.051 |
| 6. For me, reading is a waste of time | 0.994 | 0.027 | 1.502 | 0.115 |
| 7. I enjoy going to a bookstore or a library | -0.957 | 0.024 | -0.646 | 0.034 |
| 8. I read only to get information that I need | 0.730 | 0.019 | 2.031 | 0.208 |
| 9. I cannot sit still and red for more than a few minutes | 0.630 | 0.018 | 0.876 | 0.068 |
| 10. I like to express my opinions about books I have read | -0.711 | 0.019 | -0.745 | 0.040 |
| 11. I like to exchange books with my friends | -0.859 | 0.023 | -0.932 | 0.049 |

* Fixed parameter, no standard error is reported.

Table 4-2. Thresholds and standard errors for USA for enjoyment of reading

| Item | Threshold 1 | Standard error | Threshold 2 | Standard error | Threshold 3 | Standard error |
|---|---|---|---|---|---|---|
| 1 | -1.707 | 0.039 | 0.008 | 0.031 | 1.802 | 0.037 |
| 2 | -1.108 | 0.035 | 1.032 | 0.038 | 2.651 | 0.052 |
| 3 | -1.207 | 0.032 | 0.415 | 0.031 | 2.464 | 0.047 |
| 4 | -0.817 | 0.023 | 0.608 | 0.022 | 1.722 | 0.031 |
| 5 | -0.949 | 0.030 | 0.516 | 0.030 | 2.454 | 0.046 |
| 6 | -0.780 | 0.032 | 1.154 | 0.035 | 2.283 | 0.042 |
| 7 | -1.559 | 0.033 | -0.151 | 0.031 | 1.699 | 0.036 |
| 8 | -1.496 | 0.031 | 0.102 | 0.026 | 1.721 | 0.033 |
| 9 | -0.696 | 0.025 | 0.767 | 0.026 | 1.724 | 0.032 |
| 10 | -1.389 | 0.030 | -0.033 | 0.026 | 1.676 | 0.033 |
| 11 | -0.972 | 0.029 | 0.652 | 0.030 | 2.127 | 0.040 |

Table 4-3. Thresholds and standard errors for Turkey for enjoyment of reading

| Item | Threshold 1 | Standard error | Threshold 2 | Standard error | Threshold 3 | Standard error |
|---|---|---|---|---|---|---|
| 1 | -1.729 | 0.091 | 0.008 | 0.031 | 1.802 | 0.037 |
| 2 | -0.548 | 0.071 | 0.793 | 0.047 | 2.651 | 0.052 |
| 3 | -0.361 | 0.067 | 0.737 | 0.045 | 2.464 | 0.047 |
| 4 | -4.766 | 0.446 | -1.352 | 0.202 | 1.722 | 0.031 |
| 5 | -0.956 | 0.091 | 0.246 | 0.059 | 2.454 | 0.046 |
| 6 | -1.689 | 0.171 | 1.098 | 0.072 | 2.283 | 0.042 |
| 7 | -0.574 | 0.058 | 0.299 | 0.037 | 1.699 | 0.036 |
| 8 | -5.278 | 0.542 | -1.350 | 0.227 | 1.721 | 0.033 |
| 9 | -2.383 | 0.167 | -0.225 | 0.077 | 1.724 | 0.032 |
| 10 | -1.201 | 0.095 | -0.284 | 0.062 | 1.676 | 0.033 |
| 11 | -0.796 | 0.088 | 0.277 | 0.055 | 2.127 | 0.040 |

Table 4-4. Residual variances for Turkey on enjoyment of reading

| Item | Residual variance | Standard error |
|---|---|---|
| 1 | 2.253 | 0.170 |
| 2 | 0.989 | 0.061 |
| 3 | 0.749 | 0.047 |
| 4 | 5.344 | 0.766 |
| 5 | 1.383 | 0.091 |
| 6 | 2.839 | 0.293 |
| 7 | 0.487 | 0.032 |
| 8 | 6.461 | 1.033 |
| 9 | 3.284 | 0.305 |
| 10 | 1.113 | 0.084 |
| 11 | 1.235 | 0.087 |

Table 4-5. Factor loadings and standard errors for enjoyment of reading in the strict factorial invariance model

| Item | Factor loading | Standard error |
|---|---|---|
| 1. I read only if I have to | 1.000* | |
| 2. Reading is one of my favourite hobbies | -1.210 | 0.024 |
| 3. I like talking about books with other people | -1.092 | 0.022 |
| 4. I find it hard to finish books | 0.613 | 0.013 |
| 5. I feel happy if I receive a book as a present | -1.017 | 0.021 |
| 6. For me, reading is a waste of time | 1.075 | 0.022 |
| 7. I enjoy going to a bookstore or a library | -1.051 | 0.021 |
| 8. I read only to get information that I need | 0.828 | 0.017 |
| 9. I cannot sit still and red for more than a few minutes | 0.621 | 0.014 |
| 10. I like to express my opinions about books I have read | -0.821 | 0.017 |
| 11. I like to exchange books with my friends | -0.992 | 0.021 |

*Fixed parameter, no standard error is reported.

Table 4-6. Factor means and variances for enjoyment of reading in the strict factorial invariance model

| USA | | | | TURKEY | | | |
|---|---|---|---|---|---|---|---|
| Factor mean | Standard error | Factor variance | Standard error | Factor mean | Standard error | Factor variance | Standard error |
| 0.000* | | 1.791 | 0.061 | -0.895 | 0.027 | 0.847 | 0.032 |

*Fixed parameter, no standard error is reported.

Table 4-7. Factor loadings and standard errors for learning strategies

| Item | USA Factor loading | Standard error | TURKEY Factor loadings | Standard error |
|---|---|---|---|---|
| | Memorization | | | |
| 1. When I study, I try to memorize everything that is covered in the text | 1.000* | | 1.000* | |
| 3. When I study, I try to memorize as many details as possible | 1.314 | 0.048 | 1.068 | 0.087 |
| 5. When I study, I read the text so many times that I can recite it | 0.929 | 0.033 | 7.995 | 1.631 |
| 7. When I study, I read the text over and over again | 1.067 | 0.037 | 10.110 | 3.803 |
| | Elaboration | | | |
| 4. When I study, I try to relate new information to prior knowledge acquired on other subjects | 1.000* | | 1.000* | |
| 8. When I study, I figure out how the information might be useful outside school | 0.878 | 0.030 | 1.018 | 0.043 |
| 10. When I study, I try to understand the material better by relating it to my own experiences | 1.014 | 0.034 | 0.884 | 0.035 |
| 12. When I study, I figure out how the text information fits in with what happens in real life | 1.048 | 0.035 | 1.074 | 0.044 |
| | Control | | | |
| 2. When I study, I start by figuring out what exactly I need to learn | 1.000* | | 1.000* | |
| 6. When I study, I check if I understand what I have read | 1.174 | 0.036 | 1.675 | 0.089 |
| 9. When I study, I try to figure out which concepts I still haven't really understand | 1.184 | 0.036 | 1.317 | 0.059 |
| 11. When I study, I make sure that I remember the most important points in the text | 1.210 | 0.039 | 0.576 | 0.035 |
| 13. When I study and I don't understand something, I look for additional information to clarify this | 0.965 | 0.030 | 1.325 | 0.066 |

* Fixed parameter, no standard error is reported.

Table 4-8. Thresholds and standard errors for USA for learning strategies

| Item | Threshold 1 | Standard error | Threshold 2 | Standard error | Threshold 3 | Standard error |
|---|---|---|---|---|---|---|
| 1 | -1.677 | 0.035 | 0.197 | 0.025 | 1.418 | 0.032 |
| 2 | -2.205 | 0.043 | -0.727 | 0.027 | 0.823 | 0.026 |
| 3 | -2.403 | 0.056 | -0.680 | 0.032 | 0.996 | 0.033 |
| 4 | -1.687 | 0.039 | -0.017 | 0.028 | 1.534 | 0.037 |
| 5 | -0.431 | 0.024 | 0.960 | 0.028 | 1.994 | 0.039 |
| 6 | -2.068 | 0.043 | -0.332 | 0.028 | 1.204 | 0.031 |
| 7 | -1.377 | 0.033 | 0.160 | 0.026 | 1.363 | 0.032 |
| 8 | -0.525 | 0.027 | 1.003 | 0.030 | 2.161 | 0.043 |
| 9 | -1.915 | 0.040 | -0.134 | 0.028 | 1.480 | 0.032 |
| 10 | -1.125 | 0.032 | 0.511 | 0.030 | 1.913 | 0.040 |
| 11 | -2.467 | 0.052 | -0.994 | 0.033 | 0.538 | 0.028 |
| 12 | -1.131 | 0.033 | 0.658 | 0.031 | 2.125 | 0.043 |
| 13 | -1.549 | 0.032 | -0.046 | 0.025 | 1.226 | 0.028 |

Table 4-9. Thresholds and standard errors for Turkey for learning strategies

| Item | Threshold 1 | Standard error | Threshold 2 | Standard error | Threshold 3 | Standard error |
|---|---|---|---|---|---|---|
| 1 | -1.954 | 0.100 | 0.197 | 0.025 | 1.418 | 0.032 |
| 2 | -2.057 | 0.070 | -0.727 | 0.027 | 0.823 | 0.026 |
| 3 | -1.976 | 0.162 | -0.309 | 0.068 | 0.996 | 0.033 |
| 4 | -1.726 | 0.080 | -0.017 | 0.028 | 1.534 | 0.037 |
| 5 | -15.827 | 3.441 | -5.584 | 1.439 | 1.994 | 0.039 |
| 6 | -2.735 | 0.171 | -0.866 | 0.085 | 1.204 | 0.031 |
| 7 | -26.279 | 10.136 | -11.216 | 4.588 | 1.363 | 0.032 |
| 8 | -1.483 | 0.086 | 0.615 | 0.042 | 2.161 | 0.043 |
| 9 | -2.017 | 0.103 | -0.144 | 0.046 | 1.480 | 0.032 |
| 10 | -0.928 | 0.060 | 0.669 | 0.034 | 1.913 | 0.040 |
| 11 | -0.967 | 0.062 | -0.194 | 0.024 | 0.538 | 0.028 |
| 12 | -1.741 | 0.093 | 0.397 | 0.044 | 2.125 | 0.043 |
| 13 | -2.078 | 0.117 | -0.242 | 0.051 | 1.226 | 0.028 |

Table 4-10. Residual variances for Turkey on learning strategies

| Item | Residual variance | Standard error |
|---|---|---|
| 1 | 1.265 | 0.099 |
| 2 | 0.945 | 0.051 |
| 3 | 0.631 | 0.195 |
| 4 | 1.142 | 0.073 |
| 5 | 38.843 | 14.902 |
| 6 | 1.289 | 0.122 |
| 7 | 57.113 | 41.473 |
| 8 | 2.011 | 0.130 |
| 9 | 1.051 | 0.074 |
| 10 | 0.993 | 0.063 |
| 11 | 0.201 | 0.022 |
| 12 | 1.633 | 0.106 |
| 13 | 1.380 | 0.111 |

Table 4-11. Factor loadings and standard errors for question for teacher student relations

| Item | USA Factor loadings | Standard errors | TURKEY Factor loadings | Standard errors |
|---|---|---|---|---|
| 1. I get along well with most of my teachers | 1.000* | | 1.000* | |
| 2. Most of my teachers are interested in my well-being | 1.133 | 0.031 | 0.951 | 0.101 |
| 3. Most of my teachers really listen to what I have to say | 1.280 | 0.036 | 1.266 | 0.049 |
| 4. If I need extra help, I will receive it from my teachers | 0.974 | 0.025 | 0.877 | 0.041 |
| 5. Most of my teachers treat me fairly | 1.207 | 0.036 | 0.651 | 0.027 |

* Fixed parameter, no standard error is reported.

Table 4-12. Thresholds and standard errors for USA for teacher student relations

| Item | Threshold 1 | Standard error | Threshold 2 | Standard error | Threshold 3 | Standard error |
|---|---|---|---|---|---|---|
| 1 | -3.190 | 0.058 | -2.119 | 0.043 | 1.023 | 0.032 |
| 2 | -3.345 | 0.064 | -1.582 | 0.040 | 1.729 | 0.041 |
| 3 | -3.290 | 0.066 | -1.227 | 0.041 | 2.062 | 0.050 |
| 4 | -3.148 | 0.056 | -1.957 | 0.040 | 1.072 | 0.031 |
| 5 | -3.604 | 0.072 | -2.291 | 0.052 | 1.388 | 0.040 |

Table 4-13. Thresholds and standard errors for Turkey for teacher student relations

| Item | Threshold 1 | Standard error | Threshold 2 | Standard error | Threshold 3 | Standard error |
|---|---|---|---|---|---|---|
| 1 | -3.806 | 0.103 | -2.119 | 0.043 | 1.023 | 0.032 |
| 2 | -7.475 | 0.425 | -4.114 | 0.253 | 1.729 | 0.041 |
| 3 | -4.142 | 0.175 | -1.614 | 0.092 | 2.062 | 0.050 |
| 4 | -3.235 | 0.163 | -1.877 | 0.103 | 1.072 | 0.031 |
| 5 | -1.988 | 0.085 | -0.621 | 0.044 | 1.388 | 0.040 |

Table 4-14. Residual variances for Turkey on teacher student relations

| Item | Residual variance | Standard error |
|---|---|---|
| 1 | 1.652 | 0.078 |
| 2 | 4.217 | 0.408 |
| 3 | 1.540 | 0.107 |
| 4 | 1.077 | 0.090 |
| 5 | 1.224 | 0.082 |

Table 4-15. Factor loadings and standard errors for teacher student relations in the strict factorial invariance model for teacher student relations

| Item | Factor loading | Standard error |
|---|---|---|
| 1. I get along well with most of my teachers | 1.000* | |
| 2. Most of my teachers are interested in my well-being | 1.386 | 0.028 |
| 3. Most of my teachers really listen to what I have to say | 1.523 | 0.031 |
| 4. If I need extra help, I will receive it from my teachers | 1.228 | 0.023 |
| 5. Most of my teachers treat me fairly | 1.066 | 0.021 |

* Fixed parameter, no standard error is reported.

Table 4-16. Factor means and variances for teacher student relations in the strict factorial invariance model

| USA | | | | TURKEY | | | |
|---|---|---|---|---|---|---|---|
| Factor mean | Standard error | Factor variance | Standard error | Factor mean | Standard error | Factor variance | Standard error |
| 0.000* | | 1.618 | 0.058 | 0.257 | 0.031 | 2.235 | 0.085 |

*Fixed parameter, no standard error is reported.

Table 4-17. Factor loadings and standard errors for disciplinary climate

| Item | USA Factor loading | Standard error | TURKEY Factor loading | Standard error |
|---|---|---|---|---|
| 1. Students don't listen to what teachers says | 1.000* | | 1.000* | |
| 2. There is noise and disorder | 1.176 | 0.035 | 2.098 | 0.122 |
| 3. The teachers has to wait a longtime for students to quiet down | 1.391 | 0.041 | 2.386 | 0.140 |
| 4. Students cannot work well | 1.161 | 0.035 | 1.754 | 0.093 |
| 5. Students don't start working for a long time after the lesson begins | 1.198 | 0.034 | 2.085 | 0.111 |

*Fixed parameter, no standard error is reported.

Table 4-18. Thresholds and standard errors for USA on disciplinary climate

| Item | Threshold 1 | Standard error | Threshold 2 | Standard error | Threshold 3 | Standard error |
|---|---|---|---|---|---|---|
| 1 | -1.438 | 0.035 | 1.095 | 0.033 | 2.490 | 0.048 |
| 2 | -1.522 | 0.038 | 1.025 | 0.035 | 2.629 | 0.052 |
| 3 | -0.950 | 0.037 | 1.599 | 0.044 | 3.221 | 0.068 |
| 4 | -0.392 | 0.031 | 1.986 | 0.046 | 3.288 | 0.067 |
| 5 | -0.565 | 0.032 | 1.615 | 0.040 | 3.005 | 0.058 |

Table 4-19. Thresholds and standard errors for Turkey on disciplinary climate

| Item | Threshold 1 | Standard error | Threshold 2 | Standard error | Threshold 3 | Standard error |
|---|---|---|---|---|---|---|
| 1 | -3.133 | 0.237 | 1.095 | 0.033 | 2.490 | 0.048 |
| 2 | -5.158 | 0.576 | 0.327 | 0.159 | 2.629 | 0.052 |
| 3 | -5.207 | 0.580 | 0.193 | 0.199 | 3.221 | 0.068 |
| 4 | -3.873 | 0.368 | 0.675 | 0.127 | 3.288 | 0.067 |
| 5 | -4.152 | 0.459 | 0.569 | 0.147 | 3.005 | 0.058 |

Table 4-20. Residual variances for Turkey on disciplinary climate

| Item | Residual variance | Standard error |
|---|---|---|
| 1 | 2.140 | 0.219 |
| 2 | 3.344 | 0.518 |
| 3 | 6.601 | 0.966 |
| 4 | 5.661 | 0.651 |
| 5 | 4.383 | 0.601 |

Table 4-21. Factor loadings and standard errors for disciplinary climate in the strict factorial invariance model

| Item | Factor loading | Standard error |
|---|---|---|
| 1. Students don't listen to what teachers says | 1.000* | |
| 2. There is noise and disorder | 1.736 | 0.039 |
| 3. The teachers has to wait a longtime for students to quiet down | 1.606 | 0.034 |
| 4. Students cannot work well | 1.302 | 0.028 |
| 5. Students don't start working for a long time after the lesson begins | 1.595 | 0.034 |

*Fixed parameter, no standard error is reported.

Table 4-22. Factor means and variances for disciplinary climate in the strict factorial invariance model

| USA | | | | TURKEY | | | |
|---|---|---|---|---|---|---|---|
| Factor mean | Standard error | Factor variance | Standard error | Factor mean | Standard error | Factor variance | Standard error |
| 0.000* | | 1.244 | 0.047 | 0.079 | 0.024 | 1.232 | 0.050 |

*Fixed parameter, no standard error is reported.

Table 4-23. Factor loadings and standard errors for teachers' stimulation of reading engagement

| Item | USA Factor loading | Standard error | TURKEY Factor loading | Standard error |
|---|---|---|---|---|
| 1. The teacher asks students to explain the meaning of the text | 1.000* | | 1.000* | |
| 2. The teacher asks questions that challenge students to get a better understanding of a text | 1.218 | 0.033 | 1.253 | 0.058 |
| 3. The teacher gives students enough time to think about their answers | 0.761 | 0.020 | 0.866 | 0.045 |
| 4. The teacher recommends a book or author to read | 0.677 | 0.020 | 0.976 | 0.044 |
| 5. The teacher encourages students to express their opinion about a text | 1.079 | 0.027 | 0.980 | 0.047 |
| 6. The teacher helps students relate the stories they read to their lives | 0.934 | 0.025 | 0.915 | 0.038 |
| 7. The teacher shows students how the information in texts builds on what they already know | 1.033 | 0.026 | 0.820 | 0.034 |

*Fixed parameter, no standard error is reported.

Table 4-24. Thresholds and standard errors for USA on teachers' stimulation of reading engagement

| Item | Threshold 1 | Standard error | Threshold 2 | Standard error | Threshold 3 | Standard error |
|---|---|---|---|---|---|---|
| 1 | -2.667 | 0.048 | -0.782 | 0.031 | 1.071 | 0.031 |
| 2 | -3.323 | 0.063 | -1.136 | 0.037 | 1.073 | 0.034 |
| 3 | -2.337 | 0.042 | -0.714 | 0.026 | 0.872 | 0.026 |
| 4 | -1.151 | 0.026 | 0.226 | 0.023 | 1.352 | 0.028 |
| 5 | -2.534 | 0.045 | -0.679 | 0.031 | 1.035 | 0.031 |
| 6 | -1.737 | 0.034 | -0.024 | 0.027 | 1.463 | 0.032 |
| 7 | -2.295 | 0.040 | -0.371 | 0.030 | 1.287 | 0.032 |

Table 4-25. Thresholds and standard errors for Turkey on teachers' stimulation of reading engagement

| Item | Threshold 1 | Standard error | Threshold 2 | Standard error | Threshold 3 | Standard error |
|---|---|---|---|---|---|---|
| 1 | -2.869 | 0.095 | -0.782 | 0.031 | 1.071 | 0.031 |
| 2 | -3.212 | 0.171 | -1.031 | 0.076 | 1.073 | 0.034 |
| 3 | -2.228 | 0.129 | -0.609 | 0.057 | 0.872 | 0.026 |
| 4 | -2.330 | 0.114 | -0.238 | 0.048 | 1.352 | 0.028 |
| 5 | -2.367 | 0.126 | -0.513 | 0.053 | 1.035 | 0.031 |
| 6 | -1.543 | 0.078 | 0.094 | 0.037 | 1.463 | 0.032 |
| 7 | -1.476 | 0.073 | 0.037 | 0.033 | 1.287 | 0.032 |

Table 4-26. Residual variances for Turkey on teachers' stimulation of reading engagement

| Item | Residual variance | Standard error |
|---|---|---|
| 1 | 1.551 | 0.083 |
| 2 | 1.161 | 0.100 |
| 3 | 1.041 | 0.097 |
| 4 | 1.563 | 0.111 |
| 5 | 0.848 | 0.071 |
| 6 | 0.857 | 0.057 |
| 7 | 0.696 | 0.047 |

Table 4-27. Factor loadings and standard errors for teachers' stimulation of reading engagement in the strict factorial invariance model

| Item | Factor loading | Standard error |
|---|---|---|
| 1. The teacher asks students to explain the meaning of the text | 1.000* | |
| 2. The teacher asks questions that challenge students to get a better understanding of a text | 1.524 | 0.031 |
| 3. The teacher gives students enough time to think about their answers | 1.008 | 0.020 |
| 4. The teacher recommends a book or author to read | 0.924 | 0.019 |
| 5. The teacher encourages students to express their opinion about a text | 1.357 | 0.025 |
| 6. The teacher helps students relate the stories they read to their lives | 1.212 | 0.024 |
| 7. The teacher shows students how the information in texts builds on what they already know | 1.260 | 0.024 |

*Fixed parameter, no standard error is reported.

Table 4-28. Factor means and variances for teachers' stimulation of reading engagement in the strict factorial invariance model

| USA | | | | TURKEY | | | |
|---|---|---|---|---|---|---|---|
| Factor mean | Standard error | Factor variance | Standard error | Factor mean | Standard error | Factor variance | Standard error |
| 0.000* | | 1.160 | 0.039 | 0.157 | 0.024 | 1.383 | 0.052 |

*Fixed parameter, no standard error is reported.

Table 4-29. Thresholds and standard errors for USA on teachers' use of strategies

| Item | Threshold 1 | Standard error | Threshold 2 | Standard error | Threshold 3 | Standard error |
|---|---|---|---|---|---|---|
| 1 | -2.551 | 0.045 | -1.012 | 0.028 | 0.692 | 0.026 |
| 2 | -2.708 | 0.047 | -0.866 | 0.031 | 0.986 | 0.030 |
| 3 | -2.801 | 0.048 | -0.864 | 0.031 | 1.029 | 0.030 |
| 4 | -2.491 | 0.042 | -0.802 | 0.030 | 0.791 | 0.028 |
| 5 | -2.858 | 0.049 | -0.950 | 0.033 | 0.802 | 0.030 |
| 6 | -2.678 | 0.052 | -1.446 | 0.031 | -0.208 | 0.023 |
| 7 | -3.452 | 0.066 | -1.496 | 0.039 | 0.342 | 0.030 |
| 8 | -2.624 | 0.045 | -0.825 | 0.031 | 0.861 | 0.029 |
| 9 | -1.366 | 0.027 | 0.044 | 0.023 | 1.112 | 0.026 |

Table 4-30. Thresholds and standard errors for Turkey on teachers' use of strategies

| Item | Threshold 1 | Standard error | Threshold 2 | Standard error | Threshold 3 | Standard error |
|---|---|---|---|---|---|---|
| 1 | -6.149 | 0.227 | -1.012 | 0.028 | 0.692 | 0.026 |
| 2 | -6.833 | 0.255 | -2.842 | 0.120 | 0.986 | 0.030 |
| 3 | -6.435 | 0.243 | -2.309 | 0.106 | 1.029 | 0.030 |
| 4 | -6.399 | 0.245 | -2.480 | 0.108 | 0.791 | 0.028 |
| 5 | -6.539 | 0.257 | -2.624 | 0.121 | 0.802 | 0.030 |
| 6 | -5.455 | 0.252 | -2.175 | 0.102 | -0.208 | 0.023 |
| 7 | -7.505 | 0.306 | -3.331 | 0.144 | 0.342 | 0.030 |
| 8 | -6.432 | 0.247 | -2.787 | 0.121 | 0.861 | 0.029 |
| 9 | -3.573 | 0.134 | -1.174 | 0.059 | 1.112 | 0.026 |

Table 4-31. Residual variances for Turkey on teachers' use of strategies

| Item | Residual variance | Standard error |
|---|---|---|
| 1 | 7.007 | 0.535 |
| 2 | 4.487 | 0.323 |
| 3 | 4.551 | 0.325 |
| 4 | 4.488 | 0.330 |
| 5 | 7.252 | 0.546 |
| 6 | 4.458 | 0.428 |
| 7 | 5.812 | 0.473 |
| 8 | 5.038 | 0.368 |
| 9 | 1.993 | 0.325 |

Table 4-32. Factor loadings and standard errors for teachers' use of strategies in the strict factorial invariance model

| Item | Factor loading | Standard error |
|---|---|---|
| 1. The teacher explains beforehand what is expected of the students | 1.000* | |
| 2. The teacher checks that students are concentrating while working on the reading assignment | 1.467 | 0.028 |
| 3. The teacher discusses students' work, after they have finished the reading assignment | 1.480 | 0.028 |
| 4. The teacher tells students in advance how their work is going to be judged | 1.386 | 0.027 |
| 5. The teacher asks whether every student has understood how to complete the reading assignment | 1.410 | 0.028 |
| 6. The teacher marks students' work | 0.969 | 0.022 |
| 7. The teacher gives students the chance to ask questions about the reading assignment | 1.571 | 0.032 |
| 8. The teacher poses questions that motivate students to participate actively | 1.442 | 0.028 |
| 9. The teacher tells students how well they did on the reading assignment immediately after | 1.111 | 0.023 |

*Fixed parameter, no standard error is reported.

Table 4-33. Factor means and variances for teachers' use of strategies in the strict factorial invariance model

| USA | | | | TURKEY | | | |
|---|---|---|---|---|---|---|---|
| Factor mean | Standard error | Factor variance | Standard error | Factor mean | Standard error | Factor variance | Standard error |
| 0.000* | | 0.856 | 0.031 | -0.296 | 0.021 | 1.030 | 0.039 |

*Fixed parameter, no standard error is reported.

# CHAPTER 5
# DISCUSSION

As noted by OECD (2012), it is important to have evidence of measurement equivalence in cross-national research (see also, Hui & Triandis, 1985; Hambleton & Kanjee, 1995; Steenkamp & Baumgartner, 1998; Wolf, 1998; Hambleton, 2005). The purpose of this study was to investigate measurement invariance for several questions on the PISA 2009 student questionnaire. Questions were selected for investigation only if OECD (2012) reported the number of latent variables measured by the items that comprise the question. Multiple-group confirmatory factor analysis was used to investigate measurement invariance for the USA and Turkey. A sequential procedure presented by Millsap (2011) was used for invariance testing. Four models were fit to the data for each question: the configural, metric, strong factorial, and strict factorial invariance models. The fit of the measurement models were assessed in two ways:

1. By using chi-square goodness of fit statistics and chi-square model comparison tests.
2. By using goodness of fit indices.

This chapter includes a summary of the results, and the interpretation of the findings, and discussion.

## 5.1 Summary

Evaluation of the models by the chi-square goodness of fit test indicated that all models should be rejected. In addition comparison of models by the chi-square model comparison test indicated that for each model pair (configural and metric invariance models, metric and strong factorial invariance models, strong factorial and strict factorial invariance models) the more restricted model (listed second in each pair) should be rejected in favor of the less restricted model. This indicates that adequately fitting

models must be established and DIF studies on the questions in the 2009 PISA questionnaire should be undertaken once these models are established.

Table 5-1 contains the goodness of fit indices for all questions and models. The criteria for adequate fit using RMSEA were a stringent criterion, RMSEA less than .06 and a relaxed criterion, RMSEA less than .09. The results indicated that for all questions and all models RMSEA was larger than .09 and indicates the all models should be rejected. Thus, based on RMSEA, none of the models proposed in OECD (2012) adequately fit the data. Again, adequately fitting models must be established and DIF studies should be undertaken once these models are established.

Using CFI and TLI, the criteria for adequate fit were a stringent criterion, CFI and TLI greater than .95 and a relaxed criterion, CFI and TLI greater than .90. For all questions, except Learning strategies, CFI and TLI indicated that by either the strict or relaxed criterion all models fit the data. Further, for all questions except Learning strategies, the change in CFI and TLI was fairly small. The largest decline in fit was .025 for CFI and .014 for TLI. These results supported the fit of the strict factorial invariance model and thus that there is no DIF for questions other than Learning strategies.

The CFI and TLI results for Learning strategies indicated that the configural and metric invariance models adequately fit the data, but the strong factorial and strict factorial invariance models do not. Thus according to the CFI and TLI results, the items on Learning strategies exhibit DIF and the DIF is due either to the thresholds or to the residual variances or to both.

For all questions, except Learning strategies, factor loadings, thresholds, and residual variances from the configural invariance model were reported and compared across groups. For every question, substantial differences were reported for some factor loadings, thresholds, and residual variances. As noted by Millsap (2011) these differences cannot pinpoint where DIF occur, but they certainly suggest the presence of DIF for all questions.

As noted previously, CFI and TLI supported the claim that the items on all questions except Learning strategies fit the strict factorial invariance model. Therefore means and variances for Turkey and the USA were compared on all questions except Question Learning Strategies. These differences in means and variances are summarized in the following paragraphs.

**Enjoyment of reading.** Factor means differed significantly and suggested a large difference in enjoyment of reading between students in Turkey and the United States. The results indicated that the mean score for Turkish students on the enjoyment of reading factor is lower than that for American students. That is, students in Turkey generally reported less enjoyment experiences than students in United States.

**Teacher student relations.** There were significant differences in factor means for Turkey and the USA, but the results suggested a small difference in teacher student relations between Turkish students and students the United States. The results indicated that the mean score for Turkish students on the teacher student relations factor is higher than that for American students and means that students in Turkey generally reported better teacher relations than students in United States.

**Disciplinary climate.** Factor means differed significantly and suggested a small difference in the perception of the disciplinary climate in classrooms for Turkish and American students. The results indicated that the mean score for Turkish students on the disciplinary climate factor is higher than that for American students. That is, students in Turkey generally reported worse disciplinary climate than students in United States.

**Teachers' stimulation of reading engagement.** Significant differences were found for the factor means and suggest a small difference in perception of teachers' simulation of reading engagement for students in Turkey and the United States. The results indicated that the mean score for Turkish students on the teachers' simulation of reading engagement factor is higher than that for American students. That is, students in Turkey generally reported more stimulation of engagement than students in United States.

**Teachers' use of strategies.** Factor means differed significantly and suggested a moderate difference in Turkish and American students' perception of teachers' use of strategies. The results indicate that the mean score for Turkish students on the teachers' use of strategies factor is lower than that for American students. That is, students in Turkey generally reported less use of strategies than students in United States.

## 5.2 Conclusions

The detection of DIF between Turkey and the United States was examined by using multiple group confirmatory factor analysis. By investigating the fit of the measurement invariance models across the two countries, the results showed suggestive evidence of DIF between the countries; however the evidence about DIF is ambiguous due to some methodological problems.

First, there is a lack of clear criteria for assessing model fit in confirmatory factor analysis (Browne & Cudeck, 1993; Hu & Bentler, 1999; Kline, 2005). According to Kline (2005), various model fit indices are reported in the structural equation modeling literature, and the minimal sets of fit indices that should be reported are the model chi-square, RMSEA, CFI, and the standardized root mean squared residual (SRMR). Unfortunately SRMR is not available for the approach used in M*plus* when data on polytomous items are analyzed. In addition, there are problems with relying solely on goodness of fit indices. Furthermore, criteria for judging adequacy of fit by using RMSEA, CFI and TLI have not been not firmly established (Hu & Bentler, 1999; Browne & Cudek, 1993). For this study, .06 and .09 were used as criteria for RMSEA, and .95 and .90 criteria were used for CFI and TLI. While RMSEA indicated inadequate fit, CFA and TLI suggested adequate fit in many questions. Also criteria for small, moderate and large changes in goodness of fit indices across invariance models were set forth for use in this study. But these criteria have not been validated. Furthermore, chi-square goodness of fit statistics and chi-square model comparison tests results are dependent on the validity of the graded response model and, in particular, the used of the normal ogive approach to the graded response model. The evaluation of measurement invariance with CFA involves the comparison of relative fits with the chi-square values, but the chi-square value is affected by the validity of the assumption that the graded response model fits the data well. Also evaluation of measurement invariance is affected by sample size. Specifically large sample sizes may lead to rejection of a model even though lack of fit of the model is small and may favor a complex model over

a simpler model even when the complex model fits only marginally better than the simpler model (Kline, 2005).

### 5.3 Limitations and the Need for Additional Research

The purpose of this study was to investigate measurement invariance by using factor analysis to determine if there was evidence of DIF between the United States and Turkey in six specific questions from the PISA 2009 student questionnaire. Results for two approaches to selecting well-fitting models were presented: use of (a) model fit and model comparison tests and (b) model fit indices and comparison of model fit indices. Using both approaches clear evidence of non-invariant measurement models and, therefore, of DIF emerged for the learning strategies question. For the other questions, conclusions about DIF depended on which approach was used. Additional research using other approaches to studying DIF may help clarify whether items on the six questions function similarly for students in the United States and Turkey. As reported in the literature review, the following approaches are alternatives to factor analysis for investigating DIF on polytomous items: item response theory using the graded response model, the generalized partial credit model, or the partial credit model; ordinal logistic regression models; and generalized Mantel-Haenszel methods. According to the PISA 2009 technical report (OECD 2012), the partial credit model was used to select the models for the six questions that were investigated in this thesis. Therefore, investigation using the partial credit model may be particularly attractive.

Students who participated in PISA were selected by using a two-stage stratified sample design. The analyses conducted in this study did not take into account the stratified cluster sample. According to the Stapleton (2006), in structural equation modeling, it is usually assumed that simple random sampling is used and the data are

distributed independently, even though most large-scale datasets are collected using sample designs other than simple random sampling. Therefore, in order to analyze, multilevel regression methods and multilevel structural equation modeling approaches were proposed (Stapleton, 2006; Stapleton, 2008; Wu & Kwok, 2012). du Toit and du Toit (2008) pointed out that by ignoring the cluster sampling, incorrect parameter estimates, standard errors, and inappropriate fit statistics may be obtained (see also, Stapleton, 2006; Stapleton, 2008; Wu & Kwok, 2012). Stapleton (2008) discussed the problems associated with the analysis of data that come from complex sampling designs, and concluded that the model fit information with the chi-square statistic rejection rates could show differences up to 70% which leads to acceptance of the null hypothesis more often. Moreover, the estimates of residual and disturbance variances were negatively biased and the standard errors of all estimates were positively biased. Similarly, Wu and Kwok (2012) reported that ignoring higher-level data structures violated the assumption of independency under simple random sampling and concludes that the model chi-square test statistic and CFI can only offer partial information if there is a model misspecification in analyzing the clustered data (Wu & Kwok, 2012). Therefore analysis of the data using a method appropriate for the cluster sample should be conducted to determine if the conclusions about DIF are affected.

Studies emphasizing the relationship between attitudes and success of students have been conducted in previous PISA cycles (Anil, 2011; Bybee & McCrae, 2011; Drechsel, Carstensen & Prenzel, 2011; Hopfenbeck & Maul, 2011). Moreover, according to OECD, policy adjustments have been planned both in the USA and Turkey (National Commission on Teaching and America's Future, 1996; National Science

Board, 1999; National Research Council, 2001; Meb, 2010). It is likely that studies relating student achievement to questionnaire constructs will be conducted using PISA 2009 data and policy changes will be based on these analyses will be proposed. Given this likelihood and noting concern expressed in OECD (2012) about cross-country validity of measures derived from questionnaires, additional DIF studies are important.

Table 5-1. Goodness of fit indices for all questions and models

| Question | Model | RMSEA | CFI | TLI |
|---|---|---|---|---|
| Enjoyment of reading | Configural | 0.124 | 0.951 | 0.939 |
| | Metric | 0.124 | 0.946 | 0.939 |
| | Strong | 0.132 | 0.926 | 0.931 |
| | Strict | 0.123 | 0.929 | 0.940 |
| Learning strategies | Configural | 0.096 | 0.923 | 0.903 |
| | Metric | 0.104 | 0.901 | 0.885 |
| | Strong | 0.129 | 0.822 | 0.823 |
| | Strict | 0.122 | 0.827 | 0.842 |
| Teacher student relations | Configural | 0.107 | 0.993 | 0.986 |
| | Metric | 0.150 | 0.981 | 0.972 |
| | Strong | 0.134 | 0.975 | 0.978 |
| | Strict | 0.143 | 0.965 | 0.975 |
| Disciplinary climate | Configural | 0.119 | 0.989 | 0.978 |
| | Metric | 0.110 | 0.987 | 0.981 |
| | Strong | 0.117 | 0.975 | 0.979 |
| | Strict | 0.123 | 0.967 | 0.976 |
| Teachers' stimulation of reading engagement | Configural | 0.184 | 0.934 | 0.902 |
| | Metric | 0.167 | 0.934 | 0.919 |
| | Strong | 0.142 | 0.935 | 0.942 |
| | Strict | 0.152 | 0.914 | 0.933 |
| Teachings' use of strategies* | Metric | 0.102 | 0.964 | 0.958 |
| | Strong | 0.166 | 0.950 | 0.954 |
| | Strict | 0.123 | 0.925 | 0.938 |

* Estimation procedure did not converge for the configural invariance models.

LIST OF REFERENCES


Agresti, A. (2002). *Categorical data analysis*. Hoboken, NJ: John Wiley.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Angoff, W. H. (1993). *Perspectives on differential item functioning methodology.* In P. W. Holland & H. Wainer (Eds.). Hillsdale, NJ: Erlbaum.

Anil, D. (2011). Investigation of factors influencing Turkey's PISA 2006 science achievement with structural equation modeling. *Educational Sciences: Theory & Practice, 11*(3), 1261-1266.

Browne, M. W. & Cudeck, R. (1993). *Alternative ways of assessing model fit*. In: Bollen, K. A. & Long, J. S. (Eds.). Beverly Hills, CA: Sage.

Bybee, R., & McCrae, B. (2011). Scientific literacy and student attitudes: Perspectives from PISA 2006 science. *International Journal of Science Education, 33*(1), 7-26.

Camilli, G., & Congdon, P. (1999). Application of a method of estimating DIF for polytomous test items. *Journal of Educational and Behavioral Statistics, 4,* 323-341.

Chang, H. H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33,* 333-353.

Drechsela, B., Carstensen, C., & Prenzel, M. (2011). The role of content and context in PISA interest scales: A study of the embedded interest items in the PISA 2006 science assessment. *International Journal of Science Education, 33*(1), 73-95.

French, A.W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement, 33*, 315-332.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10,* 229-244.

Hambleton, R. K., & Kanjee, A. (1995). Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptations. *European Journal of Psychological Assessment, 11,* 147-157.

Hambleton, R. K. (2005). *Issues, designs, and technical guidelines for adapting tests in multiple languages*. In R. K. Hambleton, P. Merenda, & C. D. Spielberger (Eds.). Hillsdale, NJ: Lawrence Erlbaum.

Hambleton, R. K., Sireci, S. G. Patsula, L. (2005). *Statistical methods for identifying flaws in the test adaptation process.* In R. K. Hambleton, P. Merenda, & C. D. Spielberger (Eds.). Hillsdale, NJ: Lawrence Erlbaum.

Holland, P. W., & Thayer, D. T. (1988). *Differential item performance and the Mantel-Haenszel procedure.* In H. Holland & H. I. Braun (Eds.). Hillsdale, NJ: Lawrence Erlbaum.

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

Hopfenbeck, T., & Maul, A. (2011). Examining evidence for the validity of PISA learning strategy scales based on student response processes. *International Journal of Testing, 11*, 95-121.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1- 55.

Hui, H., & Triandis, H. (1985). Measurement in cross-cultural Psychology: A review and comparison of strategies. *Journal of Cross-Cultural Psychology, 16,* 131-152.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36,* 409-426.

Kim, S.H., & Cohen, A. S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement, 22,* 345-355.

Kline, R.B. (2005). *Principles and practice of Structural Equation Modeling.* NY: Guilford.

Kristjansson, E., Aylesworth, R., McDowell, I., & Zumbo, B. D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement, 65,* 933-953.

Lord, F. (1980). *Applications of Item Response Theory to practical testing problems.* Hillsdale: Lawrence Erlbaum Associates.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. Journal of the American Statistical Association, 58, 690-700.

Mantel, N., & Haenszel, W. M. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

Mellenbergh, G. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics 7*, 105-118.

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

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement, 30*(2), 107-122.

Milli Egitim Bakanligi. (2010). *PISA 2009 projesi ulusal on raporu*. Ankara, Turkey: Egitimi Arastirma ve Gelistirme Baskanligi.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York: Routledge.

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

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Muthén, B. O., & Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus* (M*plus* Web Note No. 4). Retrieved April 28, 2005, from http://www.statmodel.com/mplus/examples/webnote.html.

Muthén, B., & Christofferson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika, 46*, 407-419.

National Science Board. (1999). *Preparing our children: mathematics and science education in the national interest.* (NSB 99-31). Washington, D.C.: National Science Board.

National Commission on Teaching and America's Future. (1996). *What matters most: teaching for America's future.* Washington, D.C.: U.S. Department of Education.

National Research Council. (2001). *Educating teachers of science, mathematics, and technology: new practices for the new millennium.* Washington, D.C.: Committee on Science and Mathematics Teacher Preparation, National Academy Press.

OECD (2002). *PISA 2000 technical report.* PISA: OECD Publishing.

OECD (2005). *PISA 2003 technical report*. PISA: OECD Publishing.

OECD (2009). *PISA 2009 Assessment framework key competencies in reading, mathematics and science.* PISA: OECD Publishing.

OECD (2012). *PISA 2009 technical report.* PISA: OECD Publishing.

Oshima, T. C., & Morris, S. B. (2008). An NCME instructional module on Raju's differential functioning of items and tests (DFIT). *Educational Measurement: Issues and Practice, 27*, 43-50.

Penfield, R. D., & Algina, J. (2003). Applying the Liu-Agresti estimator of the cumulative common odds ratio to DIF detection in polytomous items. *Journal of Educational Measurement, 40*(4), 353-370.

Penfield, R. D., Alvarez, K., & Lee, O. (2009). Using a taxonomy of differential step functioning to improve the interpretation of DIF in polytomous items: An illustration. *Applied Measurement in Education, 22,* 61-78.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph No. 17.

Samejima, F. (1999). *General graded response model*. Annual meeting of National Council on Measurement in Education. Montreal, Quebec, Canada.

Spray, J., & Miller, T. (1994). *Identifying nonuniform DIF in polytomously scored test items. ACT* Research Report Series 94-1.

Stapleton, M.L. (2006). An assessment of practical solutions for structural equation modeling with complex sample data. *Structural Equation Modeling, 13(1),* 12-58.

Stapleton, M.L. (2008). Variance estimation using replication methods in structural equation modeling with complex sample data. *Structural Equation Modeling, 15,* 183-210.

Steenkamp, J. B., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25,* 78-90.

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

Thissen, D., Steinberg, L. and Wainer, H. (1988). *Use of item response theory in the study of group differences in trace lines.* In: H. Wainer and H. Braun (Eds). Hillsdale: Lawrence Erlbaum Associates.

Tian, F. (1999). *Detecting DIF in polytomous item responses (*Unpublished doctoral dissertation). University of Ottawa, Ottawa, Canada.

van de Vijver, F. J. R., & Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 47,* 263-279.

Vaughn, B. K. (2006). *A hierarchical generalized linear model of random differential item functioning for polytomous items: A bayesian multilevel approach* (Unpublished Doctoral dissertation). Florida State University, Tallahassee, FL.

Wolf, R. M. (1998). Validity issues in international assessments. International *Journal of Educational Research, 29*, 491-501.

Wu, J., & Kwok, O. (2012). Using SEM to analyze complex survey data: A comparison between design-based single-level and model-based multilevel approaches. *Structural Equation Modeling, 19*, 16-35.

Zumbo, B. D. (1999). *A handbook on the theory and methods for differential item functioning: Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*(3), 233-251.

# BIOGRAPHICAL SKETCH

Derya Evran was born in Adana, Turkey. She received her Bachelors of Arts degree in Faculty of Education, Science Teaching from Hacettepe University, Turkey. She received her Master of Arts in education degree from the program of Research and Evaluation Methodology at University of Florida in 2012.