

THE DIFFERENTIAL ITEM FUNCTIONING (DIF) ANALYSIS OF  
MATHEMATICS ITEMS IN THE INTERNATIONAL ASSESSMENT  
PROGRAMS

A THESIS SUBMITTED TO  
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES  
OF  
MIDDLE EAST TECHNICAL UNIVERSITY

BY

HÜSEYİN HÜSNÜ YILDIRIM

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR  
THE DEGREE OF DOCTOR OF PHILOSOPHY  
IN  
SECONDARY SCIENCE AND MATHEMATICS EDUCATION

APRIL 2006

Approval of the Graduate School of Natural and Applied Sciences

---

Prof. Dr. Canan ÖZGEN  
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of  
Doctor of Philosophy

---

Prof. Dr. Ömer GEBAN  
Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully  
adequate, in scope and quality, as a thesis for the degree of Doctor of Philosophy

---

Prof. Dr. Giray BERBEROĞLU  
Supervisor

Examining Committee Members

Prof. Dr. Petek AŞKAR (HU, CEIT)

---

Prof. Dr. Giray BERBEROĞLU (METU, SSME)

---

Prof. Dr. Doğan ALPSAN (METU, SSME)

---

Prof. Dr. Ömer GEBAN (METU, SSME)

---

Prof. Dr. Nizamettin KOÇ (AU, EDS)

---

**I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.**

Name, Last name: Hüseyin Hüsnü YILDIRIM

Signature :

## ABSTRACT

### THE DIFFERENTIAL ITEM FUNCTIONING (DIF) ANALYSIS OF MATHEMATICS ITEMS IN THE INTERNATIONAL ASSESSMENT PROGRAMS

Yıldırım, Hüseyin Hüsnü

Ph.D., Department of Secondary Science and Mathematics Education

Supervisor: Prof. Dr. Giray BERBEROĞLU

April 2006, 154 pages

Cross-cultural studies, like TIMSS and PISA 2003, are being conducted since 1960s with an idea that these assessments can provide a broad perspective for evaluating and improving education. In addition countries can assess their relative positions in mathematics achievement among their competitors in the global world. However, because of the different cultural and language settings of different countries, these international tests may not be functioning as expected across all the countries. Thus, tests may not be equivalent, or fair, linguistically and culturally across the participating countries. In this context, the present study aimed at assessing the equivalence of mathematics items of TIMSS 1999 and PISA 2003 across cultures and languages, to find out if mathematics achievement possesses any culture specific aspects.

For this purpose, the present study assessed Turkish and English versions of TIMSS 1999 and PISA 2003 mathematics items with respect to, (a) psychometric characteristics of items, and (b) possible sources of Differential Item Functioning (DIF) between these two versions. The study used Restricted Factor Analysis, Mantel-Haenszel Statistics and Item Response Theory Likelihood Ratio methodologies to determine DIF items.

The results revealed that there were adaptation problems in both TIMSS and PISA studies. However it was still possible to determine a subtest of items functioning fairly between cultures, to form a basis for a cross-cultural comparison.

In PISA, there was a high rate of agreement among the DIF methodologies used. However, in TIMSS, the agreement rate decreased considerably possibly because the rate of differentially functioning items within TIMSS was higher, and differential guessing and differential discriminating were also issues in the test.

The study also revealed that items requiring competencies of reproduction of practiced knowledge, knowledge of facts, performance of routine procedures, application of technical skills were less likely to be biased against Turkish students with respect to American students at the same ability level. On the other hand, items requiring students to communicate mathematically, items where various results must be compared, and items that had real-world context were less likely to be in favor of Turkish students.

Keywords: Differential Item Functioning, Restricted Factor Analysis, Mantel-Haenszel Method, Item Response Theory Likelihood Ratio Analysis, Third International Mathematics and Science Study (TIMSS), Programme for International Student Assessment (PISA).

## ÖZ

### ULUSLARARASI DEĞERLENDİRME ÇALIŞMALARINDA KULLANILAN MATEMATİK SORULARININ MADDE YANLILIĞI YÖNTEMLERİ KULLANILARAK İNCELENMESİ

Yıldırım, Hüseyin Hüsnü

Doktora., Ortaöğretim Fen ve Matematik Alanları Eğitimi Bölümü

Tez Yöneticisi: Prof. Dr. Giray BERBEROĞLU

Nisan 2006, 154 sayfa

Ülkeler kendi matematik başarılarını, küreselleşen dünyadaki rakipleriyle karşılaştırma ihtiyacı duymaktadırlar. Bu bağlamda, TIMSS ve PISA 2003 gibi, 1960lı yıllardan beri yapılan kültürler-arası çalışmalar, ülkelerdeki matematik eğitimini değerlendirmek ve geliştirmek için geniş bakış açıları sağlamaktadır. Ancak farklı ülkelerdeki farklı kültürel yapılar ve bu ülkelerin çoğunlukla farklı dilleri konuşuyor olmaları, uluslararası sınavlarda kullanılan testlerle, bunların ilgili ülke dillerine uyarlanmış hallerinin farklı kültürler arasında aynı şekilde çalışıp çalışmadığının, veya başka bir deyişle denk olup olmadığının incelenmesini gerekli kılmıştır.

Bu bağlamda, söz konusu çalışma TIMSS-1999 ve PISA 2003 uluslararası sınavlarının matematik başarı testlerinin kültürler arası denkliliğini, ve matematik başarısı kavramının altında kültürler has özgül yapılar olup olmadığını araştırmıştır. Bu amaçla, TIMSS-1999 ve PISA 2003 Türkçe ve İngilizce versiyonlarındaki matematik başarı testi maddeleri, (a) farklı dildeki test maddelerinin psikometrik özellikleri, ve (b) maddelerin farklı dil testler arasında yanlı çalışmasının muhtemel sebepleri açılarından değerlendirilmiştir. Yanlı çalışan maddelerin tespitinde, Sınırlandırılmış Faktör Çözümlemeleri, Mantel-Haenszel Yöntemi, ve Madde Tepki Kuramı En Çok Olabilirlik Oran Analizi yöntemleri kullanılmıştır.

Sonuçlar, hem TIMSS hem de PISA çalışmalarında kullanılan bazı maddelerde kültürel denklik açısından problemler olduğunu ortaya koymuştur. Ancak her iki çalışmada kültürler arası karşılaştırmayı mümkün kılacak uygun madde alt grupları vardır.

PISA çalışmasında, madde yanlılığı tespitinde kullanılan farklı metotların sonuçları arasında yüksek derecede uyum tespit edilmiştir. Ancak, TIMSS çalışmasında, yanlı çalışan madde sayısının fazla olması, maddelerin tahmin ve ayırt edicilik indislerindeki farklılıklar bu uyumu kayda değer ölçüde bozmuştur.

Çalışma ayrıca, bilgi düzeyinde ve rutin işlem becerisi gerektiren soruların, aynı yetenekteki Türk öğrencilere kıyasla Amerikalı öğrencileri kayırma ihtimalinin daha az olduğunu, buna karşılık birden çok durumun karşılaştırılıp bir karara varılmasını ve bu kararın ifade edilmesini gerektiren gerçek yaşamla ilgili soruların ise Türk öğrencileri kayırma ihtimalinin daha az olduğunu ortaya koymuştur.

Anahtar Kelimeler: Madde Yanlılığı, Sınırlandırılmış Faktör Çözümlemeleri, Mantel-Haenszel Yöntemi, Madde Tepki Kuramı En Çok Olabilirlik Oran Analizi, Üçüncü Uluslararası Matematik ve Fen Bilgisi Çalışması (TIMSS), Uluslararası Öğrenci Başarısını Belirleme Programı (PISA).

To Handan, Mesim and Fazıl YILDIRIM



## ACKNOWLEDGEMENTS

My grateful thanks are due to Prof. Dr. Giray Berberoğlu from whom my inspiration came.

Additionally, many thanks to everyone who may have contributed to this dissertation.

## TABLE OF CONTENTS

PLAGIARISM .....	iii
ABSTRACT .....	iv
ÖZ .....	v
ACKNOWLEDGEMENT .....	ix
TABLE OF CONTENTS .....	x
CHAPTER	
1. INTRODUCTION .....	1
1.1 PURPOSE OF THE STUDY .....	7
1.2 DEFINITION OF TERMS .....	9
1.3 SIGNIFICANCE OF THE STUDY .....	10
2. REVIEW OF RELATED LITERATURE .....	11
2.1 CROSS-CULTURAL COMPARISONS .....	11
2.1.1 A CLASSIFICATION OF INFERENCES .....	15
2.2 PSYCHOMETRIC ANALYSIS OF EQUIVALENCE .....	17
2.2.1 IRT AND DIF .....	21
2.3 DIFFERENTIAL PERFORMANCE IN MATHEMATICS .....	23
2.3.1 CAUSES OF DIF IN MULTILANGUAGE ASSESSMENTS .....	24
2.4 HOW TO DEAL WITH NONEQUIVALENT DATA .....	25
3. METHODS .....	27
3.1 POPULATION AND SAMPLE .....	27
3.1.1 TIMSS 1999 .....	27
3.1.2 PISA 2003 .....	29
3.2 INSTRUMENTS .....	29
3.2.1 TIMSS 1999 .....	29
3.2.2 PISA 2003 .....	33
3.2.3 TRANSLATION PROCESS .....	34
3.3 ANALYSIS OF DATA .....	35

3.3.1 DESCRIPTIVE SUMMARY .....	35
3.3.2 CONSTRUCT EQUIVALENCE .....	36
3.3.3 DIFFERENTIAL ITEM FUNCTIONING .....	43
3.3.4 USING ANCHOR ITEMS IN IRT-LR .....	53
3.3.5 AGREEMENT OF THE DIF PROCEDURES .....	55
3.3.6 DISENTANGLING SOURCES OF DIF .....	56
4. RESULTS .....	61
4.1 DESCRIPTIVE SUMMARY .....	61
4.2 CONSTRUCT EQUIVALENCE .....	62
4.2.1 EXPLORATORY FACTOR ANALYSIS .....	62
4.2.2 CONFIRMATORY FACTOR ANALYSIS .....	64
4.3 DIFFERENTIAL ITEM FUNCTIONING .....	70
4.3.1 RESTRICTED FACTOR ANALYSIS .....	71
4.3.2 IRT LIKELIHOOD RATIO ANALYSIS .....	73
4.3.3 MANTEL-HAENSZEL AND ANCHOR ITEMS .....	76
4.3.4 COMPARISON OF THE RESULTS OF DIF ANALYSES ...	83
4.4 SOURCES OF DIF .....	88
5. CONCLUSION .....	90
5.1 CONSTRUCT EQUIVALENCE .....	90
5.2 ITEM LEVEL ANALYSES .....	93
5.2.1 RFA VERSUS M-H .....	93
5.2.2 MH VERSUS IRT-LR .....	93
5.2.3 RFA VERSUS IRT-LR .....	95
5.2.4 SCALE LEVEL ANALYSIS VERSUS ITEM LEVEL ANALYSIS .....	96
5.2.5 THE EFFECTS OF PURIFYING MATCHING CRITERION ON M-H RESULTS .....	97
5.2.6 THE EFFECTS OF USING ANCHOR ITEMS ON IRT-LR RESULTS .....	98
5.3 POSSIBLE SOURCES OF DIF .....	99
5.4 LIMITATIONS OF THE STUDY .....	101

5.5 FUTURE DIRECTIONS .....	102
REFERENCES.....	104
APPENDICES	
A. PERCENTAGE OF RECODED ITEMS IN PISA AND TIMSS BOOKLETS .....	115
B. PROPORTION CORRECTS OF THE PISA AND TIMSS .....	117
C. ROTATED FACTOR LOADINGS FOR TURKISH AND AMERICAN PISA AND TIMSS DATA .....	118
D. ROTATED COMPONENT MATRIX OF PISA AND TIMSS FROM THE TOTAL SAMPLE OF USA AND TURKEY.....	120
E. FACTOR LOADINGS AND ERROR VARIANCES OF SELECTED ITEMS OF PISA AND TIMSS DATA .....	122
F. THE SIMPLIS SYNTAX USED TO TEST THE STRICT INVARIANCE MODEL.....	126
G. ESTIMATIONS OF THE INTERCEPTS, FACTOR LOADINGS AND ERROR VARIANCES IN THE FINAL MODELS OF PISA AND TIMSS .....	127
H. PRELIS AND LISREL SYNTAXES USED IN RFA .....	129
I. EZDIF AND IRTLRDIF PROGRAM OUTPUTS.....	131
J. RELEASED TURKISH DIF ITEMS IN PISA .....	135
K. RELEASED TURKISH DIF ITEMS IN TIMSS.....	142
L. RELEASED ENGLISH DIF ITEMS IN PISA.....	145
M. RELEASED ENGLISH DIF ITEMS IN TIMSS .....	151
CURRICULUM VITAE .....	154

# CHAPTER I

## INTRODUCTION

Beyond being a fascinating game for pure mathematicians, mathematics has always been considered as an important subject in society as well (King, 1998). It had roots in ancient Egypt and Babylonia, and then grew rapidly in ancient Greece. Mathematics written in ancient Greek was translated into Arabic, and later some of this mathematics was translated into Latin and became the mathematics of Western Europe. Over a period of several hundred years, it became the mathematics of the world (Ülger, 2003). This mathematics of the world has also served nearly all other branches of the science, like physical and life sciences, social sciences etc, and has been one of the most important factors which affected all industrialized countries to experience a shift from an industrial to an information society. And what is more is that; globalization is forcing all countries to experience the same shift.

Use of calculators, computers, (for example, even this current study would not have been completed without the use of computers) and other technology has changed not only the nature of sciences but also nature of business, industry and government as well (National Council of Teachers of Mathematics (NCTM), 1996). Information is the new capital, and today it is the use of communication and computer technology determining the pace of economic change. Mathematics is the language of all these technologies and the new information society needs mathematically literate workers to deal with all these technological processes. That accounts a great amount of the concern on “success in mathematics education”, which is considered to be one of the most important issues a society should satisfy.

All these economical issues also forced mathematics education to experience a shift from “problems for mathematics” perspective to that of “mathematics for problems” perspective (OECD, 2003a).

Most of the national or international educational organizations revised the definitions of the concepts within their frameworks to account for this shift as well. For example, NCTM (1996) specified ‘mathematical literacy’ as a new social goal for education and stated that students should learn to value mathematics, they should become mathematical problem solvers, they should learn to communicate mathematically, and they should learn to reason mathematically. This definition is in line with the basic theme of the PISA 2003 as well (OECD, 2005).

As globalization forces all the countries to experience the same processes stated above, the subject of mathematics education transcends cultural boundaries and its importance is universally recognized. That is one of the points to explain why NCTM’s specifications are also valued by other countries’ national education policies as well. Now it seems to be an international agreement that mathematical literacy equips pupils with a uniquely powerful set of tools to understand and change the world. These tools include logical reasoning, problem-solving skills, and the ability to think in abstract ways (NCTM, 1996). All these are in fact due to the importance of mathematics in everyday life, in many forms of employment, in science and technology, in medicine, in the economy, in the environment and development, and in public decision-making.

Therefore, mathematics has been a critical filter for most of career choices at the university level because not only the hard sciences, but also rapidly expanding areas such as health care, commerce, and computing sciences also require mathematics. Many parents, students, and teachers understand that mastering mathematics is a gateway to university. From this perspective, it seems nothing has changed since the time of Plato’s Academia with words over the doors: “Let no one destitute of geometry enter my doors.”

However, it is very interesting that beyond this international importance of mathematics education, students’ difficulties in understanding mathematical concepts and principles also seem to be universal (McKnight & Valverde, 1999).

Mathematics is a subject that most of the students identify as their least favorite. It is a barrier to some students' success in the school, as well as to admission into universities. The discrepancy between students' disdain for mathematics, and society's growing demand for mathematical competency presents a challenge both for students, in seeking employment, and for society as a whole, in meeting its needs for a mathematically-literate workforce.

Failure on education, specifically on mathematics education, is considered to be a fatal gap in a nation's future. For instance, the report "A Nation at Risk" (NCEE, 1983) stated that educational foundations in the USA were not effective in reflecting the needs of society to the young, and this was regarded to be such an important issue to threaten the future of America as a Nation.

Achievement level of students in mathematics was also reported in this study. For example, it was reported that The College Board's Scholastic Aptitude Tests (SAT) demonstrated a virtually unbroken decline from 1963 to 1980. Average mathematics scores dropped nearly 40 points. The study also reported that many of the students did not possess the higher order intellectual skills that were expected from them. Nearly 40 percent of 17-years-olds students could not draw inferences from written material; and only one-third could solve a mathematics problem requiring several steps.

Also, in Turkey there are national studies investigating the achievement level of Turkish students in mathematics (Ersoy & Erbaş, 2000; Dede & Argün, 2003). Unfortunately these studies also demonstrate that students have difficulties in learning mathematics. In addition a very recent study of Ministry of National Education of Turkey on 4<sup>th</sup> to 8<sup>th</sup> grade levels indicates that, in mathematics, students can achieve no more than 50% of the curricular objectives. In particular, this percent drops to 25% in some subjects such as ratio and proportion (EARGED, 2003).

Also, beyond these national studies, some international associations, such as International Association for the Evaluation of Educational Achievement (IEA) or Organization for Economic Co-operation and Development (OECD) conduct assessment programs since 1960s to monitor the educational processes for an in-depth understanding of how various factors affect these processes, and to provide a common basis for cross-national achievement studies in different subject areas. Again the results from these programs revealed the low achievement level of most of the students in mathematics achievement test (McKnight & Valverde, 1999).

Among those programs, Third International Mathematics and Science Study Repeat (TIMSS-R or TIMSS 1999), conducted by IEA in 1999 and Programme for International Student Assessment (PISA) conducted by OECD in 2003 are the ones Turkey also participated. Unfortunately, in these studies Turkish students also showed so low performance that this might be considered as an alert calling a reform movement in mathematics education.

An idea that these international assessment programs can provide a broad perspective for evaluating and improving education is increasing the number of countries participating in these large-scale programs. Analyzing the data collected in these large-scale comparative studies of educational achievement may enable to understand the educational processes, and in addition the comparative aspect of these studies may provide a priceless advantage of identifying new issues relevant to reform movements in educational system. Also, analysis within and across countries may determine the links among students' achievements, teachers' instructional practices, and content of the curriculum (Robitaille & Beaton, 2002).

However, to serve these international studies to the stated purposes above, it must first be assured that the tests used in the studies are fair among different countries or cultures (Poortinga, 1989; Kleime & Baumert, 2001). In other words it must be assured that the tests measure some common construct in different cultures to form a basis for comparison.



In these cross-cultural studies, since the tests are administered in different countries it is not possible to use a single common form but translations. To this reason, tests are usually being constructed in one language, which is called the source language, and then translated to languages of other participating countries, which are called target languages.

However, because of the different cultural and language settings of different countries, these translated tests may not be functioning in the same way in all cultures, which is also called that tests may not be equivalent or tests may not be fair among different cultures (Allalouf, Hambleton & Sireci, 1999; Ercikan, 1998; Ercikan, 2002; Robin, Sireci & Hambleton, 2003; Hui & Triandis, 1989; Bontempo, 1993; Hulin & Mayer, 1986). For example, the reasons like producing different connotations due to the translation, or affecting the degree of difficulty of key vocabulary may imply that the items are interpreted differently in different countries (Poortinga & van de Vijver, 1987). On the other hand, beyond translation effects, such as producing words having different connotations in different languages, other cultural and curricular differences between countries may lead to different response styles or response patterns in different countries. Presence of these factors with the potential of affecting item equivalence can cause problems in comparability of items in different languages (Sireci & Berberoğlu, 2000; Arim & Ercikan, 2005).

This issue can also be defined from a multidimensional perspective as well. That is, in the case of international assessments, different groups of examinees may have different multidimensional ability distributions due to language, cultural and curriculum differences (Ercikan, 1998). In addition, if test items are capable of measuring these multiple dimensions, then using any unidimensional scaling procedure may produce item bias (Ackerman, 1992). From this perspective, a test item functioning differentially between two groups is an item measuring a secondary dimension that favors one of the groups after controlling the main dimension that the test is intended to measure. (Camilli & Shepard, 1994).

In large-scale assessments it is very difficult to get a perfect unidimensional test. Even the tests fitting a unidimensional model also possess a misfit variance that is usually regarded as a negligible specification error. In fact this negligible error component may contain a systematic component reflecting multidimensionality that is not considered in the scaling process (Klieme & Baumert, 2001). Differential item functioning (DIF) analyses aim to reveal these possible hidden dimensions. In other words, DIF methods measure violations from unidimensionality (Dorans & Holland, 1993). There are many DIF methods that look for evidences of differential performance of subgroups to detect biases stated above (Sireci, 1997; Allalouf, Hambleton & Sireci, 1999). DIF statistics are indices providing such evidences to be interpreted (van de Vijver & Poortinga, 1982; Osterlind, 1983; Ellis, 1989; Zumbo 2003; Sireci & Berberoglu, 2000).

Traditionally, DIF analysis is a three-step procedure. Differential item functioning analyses starts with checking the conceptual equivalence of the instrument in two groups. That is checking whether the items measure the same latent variables in all groups. Conceptual equivalence is a prerequisite for considering item equivalence (Hui & Triandis, 1985). Then come the statistical analyses at the item level to determine whether the item under investigation measures another dimension than that of the intended to be measured dimension in the test. However, these statistical analyses are capable of determining this unintentionally measured dimension only if that dimension favors one of the groups after controlling the performance of the students on a matching variable, which is usually the test score (Camilli & Shepard, 1994). Finally through qualitative reviews it is to be decided whether the multidimensionality signaled by DIF (if the second step signals DIF) is a legitimate part of the test content. This final process may reveal “emic” (culture or group specific) and “etic” (common to both cultures) aspects of the construct intended to be measured by the test (Hui & Triandis, 1985). The item is “biased” only when the judgmental reviews expose that DIF signals an illegitimate part of the test content (Thissen, Steinberg & Gerrard, 1986).

Unfortunately the steps of DIF analyses described above are not free of error. Namely, beyond the error associated with the random sampling, the reliability and validity of DIF statistics are affected by many other factors. Test characteristics like the range of item difficulties, population characteristics like the distribution of abilities, choice of computer algorithm, or the extent of DIF in a test are some of the factors that may affect reliability of the DIF indices (Camilli & Shepard, 1994). Choosing more than one DIF statistics and dealing with items only when they show DIF in both analyses would reduce the error rate to a certain extent.

In this context, there were two major purposes in this study. In the first phase the equivalence of the Turkish and English versions of the TIMSS 1999 and PISA 2003 were investigated through the use various DIF techniques. In the second phase possible sources of DIF were investigated. In addition, it was also aimed to determine whether mathematics achievement and mathematics literacy possessed any culture specific aspect in the item content.

## 1.1 Purpose of the Study

The main purpose of this current study was, in addition to investigating the cross-cultural equivalence of TIMSS and PISA, providing an overview of some statistical methods in empirically assessing flawed items due to test translation in the context of mathematics achievement testing. The current dissertation described a comprehensive approach for investigating the cross-cultural equivalence of mathematics tests of TIMSS 1999 and PISA 2003.

To this purpose, the present study assessed Turkish and English versions of TIMSS 1999 and PISA 2003 mathematics items with respect to, (a) comparability of the constructs measured by the tests (that is scale-level analysis) (b) equivalence of the scaling metrics that relate the items and the constructs that they intend to measure (that is item-level analysis), and (c) possible sources of DIF between these language versions.

In determining the performance of students at the item level, item means were calculated through the use of polychoric correlation matrices. To measure the construct equivalence, both multi-group confirmatory factor analysis (MGFA) and principal component analysis (PCA) methods were used. These are confirmatory and exploratory techniques, respectively. For the DIF analyses to assess whether the measurement equivalence existed between groups at the item level, restricted factor analysis (RFA), item response theory likelihood ratio analysis (IRT-LR), and Mantel-Haenszel (M-H) techniques were used. There were also three additional questions for DIF analyses. First the present study compared the results from three different DIF methods. Second it was investigated whether results from scale-level analyses manifest themselves in item-level analyses. Finally the effects of using anchor items in IRT-LR method, and purification in M-H were examined. Possible sources of DIF in items flagged by statistical analyses were also examined in the study.

Three research questions in the study are:

- 1- Do the mathematics tests of TIMSS 1999 and PISA 2003 have the same factor structure across U.S. and Turkish groups?
- 2- Are the original and adapted test items equivalent?
  - a. How consistently do RFA, M-H, and IRT-LR agree?
  - b. Do the results from scale-level analyses manifest themselves in item-level analyses?
  - c. Does the use of anchor items affect the IRT-LR results?
  - d. Does the purification of the matching criterion affect M-H results?
- 3- What are the possible sources making items function differentially across U.S and Turkish groups?

## 1.2 Definition of Terms

Although defined in detail in the related chapters, some specific terms are also briefly defined in this section.

DIF - “An item shows DIF if individuals having the same ability, but from different groups, do not have the same probability of getting an item right” (pp.110) (Hambelton, Swaminathan, & Rogers, 1991).

Mathematics Achievement - is, in general can be defined as the measurement of succeeding in reaching an aim related to mathematics learning, within the context of this study, mathematics achievement is what the TIMSS 1999 mathematics achievement test measures.

Mathematics Literacy - is an individual’s capacity to identify and understand the role that mathematics play in the world, to make well-founded judgments to use and engage with mathematics in ways that meet the needs of that individual’s life as a constructive, concerned and reflective citizen (OECD, 2003a). Within the context of this study, mathematics literacy is what the PISA 2003 mathematics test measures.

Scale-Level Analyses - are exploratory and confirmatory factor analyses to determine construct equivalence across groups.

Item-Level Analyses - are differential item functioning analyses to examine whether individuals having the same ability, but from different groups, also have the same probability of getting an item right.

Anchor Items - are the items that are used to form a baseline model within the context of IRT-LR. The parameters of these items are fixed to be equal in reference and focal groups.

Reference and Focal Groups are also named as majority and minority groups, respectively. Reference group provides a standard of comparison; it is the focal group, which is of primary interest. In the context of cross-cultural studies, the test is originally developed in Reference group’s language and then translated to the focal group’s language. USA is the reference group, and Turkey is the focal group in this current study.

### 1.3 Significance of the Study

The purpose of this study was twofold: First, this study is a comprehensive DIF analysis for large-scale mathematics tests across different countries. To this reason, this study produced hypotheses to identify the sources of translation DIF. These hypotheses may be used in further studies to be confirmed, to create a body of tested hypothesis that may be used to develop guidelines for reducing DIF in translated tests. Confirmed hypotheses over studies may lead to a better understanding of causes of DIF.

Second, this study provided an overview of some statistical methods for empirically assessing DIF items due to test translation in the context of mathematics testing. In this context this study provided empirical evidences comparing the potential of some DIF methods mentioned previously.

It is worth adding that DIF represents the differential functioning of a test item in the context of other test items. In other words, DIF is not an intrinsic property of a test item. This means that each different context deserves a distinct DIF analysis. In addition, identifying DIF test items in a specific context is of significant importance to test developers and test policy-makers for validity purposes.

Finally, having a claim of being a member of the European Association, it seems that it is inevitable for Turkey to refrain from such international assessments in the future, so the framework provided by this present study in assessing mathematics items can be used in the future cross-cultural studies as well.

## CHAPTER II

### REVIEW OF RELATED LITERATURE

This chapter starts by defining ‘measurement equivalence’ in cross-cultural studies. Studies explaining theory of measurement equivalence, including cause of inequivalence, and ways of checking equivalence are reviewed. In addition, review of DIF methods used, results from DIF studies investigating mathematics items are given at the end of the chapter.

#### 2.1 Cross-Cultural Comparisons

In cross-cultural studies, scales not only revealing the different cultures’ relative positions on a common construct but also reflecting cultural uniqueness are of special importance (Hui & Triandis, 1985; Hulin, 1987). It is not possible comparing different countries until assuring that the instrument and the scaling process provide such an equivalent scale. This section defines the theoretical framework of this equivalence issue.

Three hierarchical equivalence levels are defined in the literature (van de Vijver & Tanzer, 1997). At the first level is the structural equivalence, which assures that the same construct is measured in each group. Given that the instrument measures the same construct in both groups, scales on which the scores are reported must have the same unit across populations to have measurement unit equivalence. Finally, for a full scale equivalence scores must have the same origin in all populations, in addition to the same measurement unit.

The inferences to be made from cross-cultural studies are also limited by the level of the equivalence the scale provides. Following paragraphs give a detailed explanation of equivalence in these perspectives.

Equivalence of scale unit is one of the requirements for comparing different cultures. For a comparison between two or more groups, attributes being compared and scale units indicating these common attributes must be the same (Poortinga, 1989; van de Vijver & Tanzer, 1997). For example, comparing length of group A with weight of group B is not possible, because length and weight are not same attributes. In addition, comparing length of group A, measured in inches with the length of group B measured in centimeters is also not possible, because scale units are not the same.

In psychological measures, because the attribute being measured is not directly observable it is given a special name, “construct”. Then, the equivalence of the attribute measured can be re-specified as; the construct to be compared must possess the same properties and meaning in both cultures, which is called construct equivalence (Hui & Triandis, 1983). In addition, the scale representing this construct must also be the same. Such an identical scale is called comparison scale. In other words, comparison scale can be regarded as a measurement scale on which equivalence is assured.

The problem due to the lack of equivalence in the construct being measured is called construct bias (van de Vijver & Tanzer, 1997). Construct bias may be due to various reasons, such as poor sampling of all relevant behaviors, or partial correspondence of the construct over cultures. If the construct being measured does not possess the same set of behaviors in both cultures, such as ‘respect to parents’ in USA and in Turkey, this will be a threat to equivalence of measurement across cultures. That is why Allouf, Hambleton and Sireci (1999) argue that equivalence of test structure in each language version should be assessed before conducting further statistical analysis to check item equivalence. Sireci (1997) also states that before linking tests it must be demonstrated that the constructs measured by different language tests are comparable.

The next step deals with information collection process about the construct of interest. This process deals with collecting a set of behaviors through the use of tests, interviews etc. and then making generalizations from this set of behaviors to the set of all behaviors of the construct under investigation.



Measurement scale is the scale where this collected information is expressed (Poortinga, 1989). Lack of equivalence at various steps of this process leads to method bias. For example, if samples are not comparable because of the difference of educational background of individuals in different groups (which is called sample bias, a special type of method bias), this can have a ruining effect on validity of cross-cultural comparisons. In addition, if, for example, there is a communication problem between the interviewers and interviewees in one of the cultures (administration bias), or if in a Likert-scale one of the cultures tends to avoid the extreme cases (instrument bias), there will be a threat on the validity of inferences, or generalizations to the domain intended to be measured by the test. Further discussion about this generalization process is given in the section of 'A Classification of Inferences' in this chapter.

Given that there are no construct and method biases, it must also be assured that there are no distortions at the item level to satisfy cross-cultural equivalence (Borsboom, Mellenbergh & van Heerden, 2002). If, for example, there is a poor item translation, or an item invokes additional abilities in one of the groups, then the item may not be equally difficult for equal ability individuals from different groups. In other words, item may be biased.

Different type of biases defined above may distort the relation between measurement and comparisons scales. Poortinga (1989) defines equivalence of cross-cultural data as the situation where an observed cross-cultural difference on a measurement scale is matched by a corresponding difference on the comparison scale. This relation is specified due to three levels of measurement scale identity. Namely, 1) Same scale origin and same metric, 2) Same metric, and 3) Same metric after linear transformation (Van de vijver & Tanzer, 1997; Poortinga, 1989).

If a cross-cultural study aims at comparing the score levels obtained in different cultures, it has to assure that the scale on which the scores are expressed has the same zero-point (origin) and the same scale units (same metric). This is also called as scalar equivalence, which assumes completely bias-free measurement. Construct, method and item bias can seriously threat the scalar equivalence.

At this point, it may be useful to state that, invariance of scale means that, parameters of a mathematical function describing the scale have the same value across cultures. It does not imply that the score distributions also have to be the same (Hulin, Drasgow & Komocar, 1982). That is why IRT provides a very suitable framework for cross-cultural studies to define an invariant scale, because the parameters defined with respect to IRT are independent of the score distribution of the cultures (Hulin, 1987; Ellis, Becker & Kimmel, 1993). This is further discussed in the related section in this chapter.

If the scale have the same metric across cultures but fail to have the same origin, then this scale identity can assure a valid cross-cultural comparison of only the relative differences between pairs of mean scores obtained in different cultures.

If no direct comparisons are intended between groups but only structural relationships between variables are of interest, then it is not necessary that a measurement scale should have the same metric across cultures (Poortinga, 1989). That is, neither method nor item bias will be a threat to cross-cultural equivalence at this level of comparison. It is enough to assure that there is no construct bias.

The three levels of invariance stated above are hierarchically ordered. That is, invariance as defined at the first level presumes the second and third level as well. Poortinga (1989) summarizes these issues as, “Which psychometric properties of data can be validly compared depends on which parameters of measurement scales can be taken as invariant across cultures.” (pp. 740).

The following case can exemplify the theoretical issues stated up to this section.

Comparing American and Turkish students with respect to mathematics achievement in TIMSS 1999 mathematics test requires considering all the bias types to determine if this measurement provides a scalar equivalence. Assuming that possible sources of method bias were controlled through extensive training of indicators, detailed manual for administration etc. there remain two additional issues to be investigated; First issue deals with whether mathematics achievement means the same thing for both cultures. That is, whether the notion of mathematics achievement can serve as a comparison scale.

Second issue concerns whether score differences on the mathematics test reflect corresponding differences in the achievement level. That is whether the relationship between the mathematics achievement (ie. comparison scale) and the scores in mathematics achievement test of TIMSS 1999 (ie. scale of measurement) is the same in the two groups. Methods in investigating these issues are also given in this chapter.

### 2.1.1 A Classification of Inferences

Beyond this stated logic of cross-cultural comparison comes another important issue: The validity of the possible inferences to be made from the data. What is done in measuring is very roughly, making inferences based on a domain of behavior provided by the measurement instrument to the psychological domain of interest, which includes all possible behaviors determining the domain (Gulliksen, 1950). These inferences can be seen as generalizations to the domain of interest. Although Poortinga (1989) defines three levels of generalizations, only the first two that is related with the scope of this study will be given.

These levels of generalizations are determined with respect to the attributes to be included in a psychological domain. If generalization is to be made to a domain that is defined in terms of observable psychological attributes of individuals, this is called measurement as sample or measurement at the first level. In this case the domain is well defined, i.e. an instrument can form a representative sample of the domain. For example arithmetic skills is a domain where its behaviors can clearly be defined, which makes it is easy to construct an instrument including representative sample of the behaviors of the concept of mathematical skills.

However, translated versions of an instrument measuring arithmetic skills may still be inequivalent at this first level of measurement. There are two possible reasons causing this inequivalence. First one is the difficulty in meeting the requirements of representativeness. For example although a test is representative of the domain in one culture, its translated version may contain some unexpected difficulties because of some wording problems (Ercikan, 1998; Ercikan, 2002).

Bias in administration procedure is the second possible threat. For example lack of familiarity of individuals of one of the cultures with a specific item format may produce some inequivalence. Both of these reasons are called as measurement artifacts (Allalouf, Hambleton & Sireci, 1999; Wolf, 1998).

At the second level are the generalizations to domains that are defined in terms of unobservable psychological attributes of individuals. Mathematics literacy can be an example for this domain. It is difficult to distinguish between the specific behaviors that do and do not belong to the domain of ability. A generalization at this second level is called measurement as index (Poortinga, 1989). Although the attributes of the construct are unobservable at this second level, to make a cross-cultural comparison we have to limit the range of possible responses to an item. Otherwise no rationale exists for a valid comparison. Hulin (1987) assures this limitation within his definition of linguistic equivalence, which he states as the goal of translation of psychological scales.

Hulin (1987) argues that the goal of linguistic equivalence is to provide an equivalent structure to the material to achieve equivalent stimuli. In this context he states that, "...psychometrically equivalent items (stimuli) evoke a specified response, from the set of permissible responses, with the same probability among individuals with equivalent amounts of the characteristic assessed by the item or scale comprising the items."(pp. 123). The importance of this definition is that, it limits the range of possible responses. The definition of equivalence is relative to a restricted set of responses and excludes unobservable behaviors.

In this perspective, methods analyzing differential item functioning (DIF) provide a basis to check whether the items evoke equivalent (not equal) stimuli across groups. If methods point out inequivalence of some items at this second level of measurement, in addition to measurement artifacts, inappropriateness of a measurement as an index of a certain domain in one of the cultures may be a source as well. However, this inequivalence may also be pointing a cultural uniqueness or an emic concept in one of the cultures (Hulin & Mayer, 1986). That is, these items should not be regarded directly as threats to measurement equivalence.

Hambleton and Kanjee (1995) have stated that, when there is substantial overlap between items administered in different groups then culture specific items (emic items) may well enhance the validity of the instrument in that culture.

## 2.2 Psychometric Analysis of Equivalence

This section gives a brief summary of strategies for identifying and dealing with bias in cross-cultural assessment. Giving the framework of equivalence of Hulin's (1987) translation equivalence or Poortinga's levels of generalization, the second dimension to be discussed is the statistics that provide information for empirical analyses for assessing equivalence (Poortinga, 1989). Statistical analyses to be discussed are factor analyses, item bias analyses, and regression analyses.

To investigate whether the test measures the same psychological construct across all studied groups, exploratory factor analysis, confirmatory factor analysis, and multidimensional scaling are the methods used in the studies (Sireci, Bastari & Allalouf, 1998).

The most widely used exploratory factor analysis (EFA) method is principal component analysis (van de Vijver & Poortinga, 1997; Arim & Ercikan, 2005). The studies using EFA perform separate factor analyses for each group and then compare the results. However, there are no statistical tests determining the degree of resemblance of factor structures among groups.

On the other hand, confirmatory factor analysis (CFA) provides statistical tests of model fit for the studies handling multiple groups simultaneously ( Reise, Wdaman & Pugh, 1993; Gierl, 2004; Zumbo, 2003, Jöreskog, 1971). In addition availability of statistical packages like LISREL (Jöreskog & Sörbom, 1993; 2001; 2002) increased the popularity of this technique as well. This methodology also enables checking different forms of invariance, from weak to strict invariance (Zumbo, 2003). Strict invariance is assured when a measurement model is reproduced in both groups including magnitude of factor loadings and error variances. Weak invariance only reproduces the same dimensionality in both groups but not the same magnitudes for the parameters.

Multidimensional scaling (MDS) is an alternative for EFA, providing subject weights reporting differences among the groups with respect to dimensional structure (Sireci & Geisinger, 1995; Meara, Robin & Sireci, 2000).

It is also important to note that all these analyses deal with structural or construct level equivalence, and they cannot provide enough evidence assuring the equivalence of scale metric or origin. For example item level bias can still be present even when an equivalent construct is present (Zumbo, 2003). So any study comparing the scores across cultures needs more evidence than factor analyses provide.

Item level analyses have the power of providing such additional evidences related with the equivalence of the measurement scale. These analyses investigate whether cross-cultural differences in scores on an item are in line with the expectations based on the other items of the test. If not, that item may be biased.

Three important terms to be discussed before giving the methodologies suitable at this level are: item impact, DIF, and item bias. Item impact is a significant group difference on an item, which may be due to true group differences or item bias. DIF analyses are conducted to reveal this phenomenon by matching the examinees on the ability being measured. If the examinees of equal ability from different groups do not respond similarly to an item, that item is said to be functioning differentially between groups. However a further qualitative judgment is still required to label an item as biased against a certain group (Camilli & Shepard, 1994).

Statistical techniques within the framework of item bias analyses are DIF techniques, which may be classified as Classical Test Theory CTT-based methods, Factor Analysis FA-based methods,  $\chi^2$ -based methods, and Item Response Theory IRT-based methods (Benito & Ara, 2000; Sireci & Allalouf, 2003). What differentiates these techniques is mainly the way they match individuals of equal ability.

Delta Plot is an example for CTT-based methods (Angoff & Ford, 1973). This method deals with difference between the difficulty parameter estimates obtained from different groups.

It must also be added that Camilli and Shepard (1994) do not offer this methods to be used.

Restricted Factor Analysis (RFA) is a FA-based method. The most promising aspect of RFA is its flexibility in considering several potential violators simultaneously (Oort, 1992). A detailed explanation of RFA is given in the next chapter.

Mantel-Haenszel (MH) (Holland & Thayer, 1988), and Logistic Regression (LR) (Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990) procedures are  $\chi^2$  Methods. They both use total test scores in matching test takers from different groups.

Finally IRT likelihood ratio test is an IRT-based method details of which are given in the next chapter (Thissen, Steinberg & Wainer, 1988; 1993). This method can deal with both dichotomous and polytomous data according to the IRT model chosen. In this method the underlying ability scale forms the basis for matching.

As a final note about DIF methodologies, it must be added that these techniques assume random distribution of bias effect in groups. But it must also be taken into consideration that, especially in cross-cultural studies an effect like for example social desirability, or an improper translation may cause a systematic bias affecting all the scale.

It is clear from the definition that bias analyses cannot detect this type of overall bias, because these techniques compare item performance to that of all the scale. As a systematic bias affects both the item and the scale there may not be relative difference between these two then. In fact this is a very extreme case, but this probability of failing to detect an overall bias should be taken into consideration in cross-cultural studies especially for those studies intending to compare the scores across groups (Hulin, 1987; Poortinga, 1989). This issue is of special importance for IRT-based studies (Sireci & Berberoğlu, 2000).

Finally, regression analysis is the most powerful statistic that can provide evidence indicating an equivalent scale with same origin and metric. Unfortunately it has got a prerequisite that is difficult to satisfy, namely finding a criterion measurement that is free of bias.

Assuming such a criterion measurement exists, regression function relating test scores to criterion scores has to be same for the groups to be compared. Such evidence of full-invariance is sufficient for comparing scores across cultures (Poortinga, 1989).

These different DIF methodologies also have different methods in matching the individuals of the same ability, different computer algorithms, different estimation methodologies etc. This makes comparison of these methodologies a special concern in the field of DIF studies. There is a considerable amount of literature in this area (Gao & Wang, 2005; Rogers & Swaminathan, 1993; Gierl, Jodoin & Ackerman, 2000; Benito & Ara, 2000).

Benito and Ara (2000) have reported that MH is the best, in the sense that it has smaller error rates and high power, with respect to LR, RFA and IRT-based procedures. In addition they have reported that in general non-IRT methods performed better than IRT methods. However, they also reported a disadvantage for MH as well, namely it cannot detect non-uniform DIF. They have also suggested the use of RFA.

Gierl et al. (2000) have compared LR and MH with respect to Type I error rate and power when there were large numbers of DIF items in the test. They reported using purification, that is conducting the analysis for the second time by excluding the DIF items determined in the first run, in MH provides an additional advantage in controlling Type I errors.

On the other hand Rogers and Swaminathan (1993) have reported that using purification did not change the MH results. However, they in addition reported that percent of DIF items did not affect the MH results because of the use of purification method. Another interesting result they declared was that, MH may fail to detect DIF in moderately difficult items.

These studies also have reported that different DIF methodologies usually produce divergent results (Gao & Wang, 2005). For this reason the studies suggest using more than one DIF methodology in the analyses.



Within these DIF methodologies IRT-based methods deserve an additional concern, as they have a relatively complicated mathematical framework. In addition, the development of the definition of DIF, and its correspondence to that of ICC can easily be seen within IRT framework.

### 2.2.1 IRT and DIF

Item Response Theory is a theory mathematically linking item responses to underlying latent traits (Lord & Novick, 1968; Lord, 1980; Hambleton, Swaminathan & Rogers, 1991). For each item, this theory specifies three parameters that define a S-shaped logistic curve, which is called item characteristic curve (ICC), linking probabilities of specified responses to position of individuals in the latent trait (or ability)  $\theta$ .

Item Characteristic Curves, such as the one in Figure 3.2 of the next chapter, can be interpreted as nonlinear regression lines describing the relationship between examinees' item performance and the set of traits underlying item performance. What is important is that this relationship does not depend on the distributions of  $\theta$  in the sample (Hambleton, Rogers & Swaminathan, 1991).

On the other hand, DIF analysis searches whether there exists multidimensionality in an item, which is unique to one of the groups (Camilli & Shepard, 1994). This means that, DIF analysis checks whether there exists a special dimension specified by an item in only one of the groups. If this is the case it is called that the item functions differentially between two groups.

Then further investigation is required to search whether this dimension is relevant or irrelevant to the construct being measured. Only if irrelevant, then the item can be called as biased (Hambleton et al., 1991). However, especially at the beginning of the DIF studies, differentiation between the meanings of item bias and DIF has not been made in most of the studies (Camilli & Shepard, 1994).

In recent studies, DIF has been defined as the situation for an item where two different groups have different mean performances on the item.

However this difference in the performances may have been a result of real between-group difference in ability. Then came a stronger definition. ‘An item shows DIF if individuals having the same ability, but from different groups, do not have the same probability of getting the item right’ (Hambleton et al., 1991). Unfortunately this definition neither was free of problem. Now the problem of how to determine individuals having the same ability was the new concern.

In classical test theory, ability is expressed as the expected value of observed performance on the test (Crocker & Algina, 1986). When the test is hard, the examinee will appear to have low ability. Also if the probability of getting the item right is defined as the proportion correct scores, then this probability depends on the ability of examinees taking the test. In this context, in the case of translated tests, if the translation process produces a harder version of the test because of some wording problems or cultural differences, it will itself be a problem to determine the individuals having the same ability. Fortunately, IRT seems to provide a solution to this problem by two of its desirable features; namely describing item characteristics that are not group dependent and describing test-independent examinee proficiency. These are called as the invariance properties of IRT (Lord, 1980).

As stated above, ICCs do not depend on the distributions of  $\theta$ , or in other words; ICCs for an item based on responses from two different groups are invariant up to a linear transformation (Poortinga, 1989). It is this invariance property that enables interpretation of the measurement properties of the items among cultures.

In this ICC context, DIF may be restated as the situation where an item response function across different subgroups is not identical (Hambleton et al., 1991). From the same perspective, Drasgow (1984) also defines measurement equivalence as “Equivalent measurement is obtained when the relations between observed test scores and the latent attribute measured by the test are identical across subpopulations.” (pp.164).

From these definitions, invariant item parameters across languages and subcultures would be interpreted as evidence of the equivalence of meanings of items and their psychometric characteristics. However, what is worth to state is that, these properties are related with the populations.

This means that, in order to use IRT in cross-cultural studies, a group of individuals from a different culture who speak the target language should be regarded as a different subpopulation of the population who responded to the items in their original source language (Hulin 1987).

However, it should also be considered that Sireci (1997) and Sireci and Berberoğlu (2000) questions whether item parameter invariance property of IRT holds over samples derived from different language groups, in which a systematic bias would affect the scales.

### 2.3 Differential Performance in Mathematics

A considerable amount of DIF studies in mathematics usually deal with gender or SES differences. Item content, item type and cognitive complexity are the most common factors on which differential performances are investigated (Scheuneman & Grima, 1997; Harris & Carlton, 1993; Berberoglu, 1995; Yurdugül & Aşkar, 2004a; Yurdugül & Aşkar, 2004b).

Most of the studies have reported overall differences between males and females, usually in favor of males especially when the cognitive complexity increases and the content changes from arithmetic to algebra and geometry (Engelhard, 1990). In addition these differences seem to be replicable over different cultures.

However, Berberoğlu (1995) have reported contradictory results. He examined the Turkish students with IRT-based DIF methods and have concluded that verbal items and items requiring spatial ability favored females, whereas items requiring computational skills favored males. In addition he also have reported the advantage of high SES groups. Another interesting result of this study was reporting gender DIF as a function of SES.

On the other hand Doolittle and Cleary (1987) have reported that males were better on word problems, and algebra items favored females.

Beller and Gafni (1996) have reported that measurement items and items involving problem solving favored males. They have also considered the age levels and concluded that gender related DIF becomes larger as age increases.

In addition to the gender related studies, there are also studies in the literature concerning race and ethnicity related DIF as well. For example, Harris and Carlton (1993) have reported that Whites had an advantage over African Americans in items with figures, applied and realistic problems, and word problems. On the other hand Whites were disadvantaged in Algebra, items with a variable, curriculum-like items, and abstract items.

Scheuneman and Grima (1997) in addition have reported that Whites were advantaged over Hispanics in word problems as in the previous case. Data-interpretation and realistic problems were the other items favoring Whites. In the same study Geometry, Algebra, Spatial items, items with figures were favoring the Hispanics. In addition they have reported that female and black examinees appear to find mathematics word problems relatively difficult.

It should also be mentioned that some of the DIF results specified in the literature are not stable. This may be due to the relation of DIF results with the content being investigated, test item format, or the DIF detection methods used. Because all these have an affect on DIF results, they also decrease the stability of results obtained in different settings.

### 2.3.1 Sources of DIF in Multilanguage Assessments

Multilanguage assessments are administered in more than one language, after adaptation of the test to the related group. However, inappropriate translations, cultural or curricular differences may affect the equivalence of items between groups (Hulin, 1987; Botempo, 1993). For example Ercikan (2002) have reported that adaptation related differences affected 27% of mathematics items of TIMSS and curricular differences affected 23% of mathematics items of TIMSS.

Allalouf, Hambleton and Sireci (1999) provided one of the most comprehensive classifications of causes of DIF in translated verbal items.

They report four main causes for DIF in translated instruments: (1) Changes in difficulty of Words or Sentences specifies the situation in which some words became easier or more difficult after the translation. (2) Changes in content may be due to an incorrect translation changing the meaning of an item. Gierl and Khalig's (2000) category of "Omissions or Additions that Affect Meaning" also deals with same issue. (3) Changes in format are the cases, for example when a sentence become much longer after the translation. Gierl and Khalig (2000) also include the changes in punctuation, capitalization etc. in this category. (4) Differences in cultural relevance is the last category. In this case the items remain same, however it's the cultural content of the item that causes DIF. For example content of a sentence completion item may be more familiar for one of the groups. Gierl and Khalig define this category as, "Differences in words or expressions inherent to a language".

Scheuneman and Grima (1997), on the other hand provides an additional cognitive perspective to the classification of possible sources of group performance differences in mathematics items. They specify three categories. Namely, (1) the cognitive nature of the task presented to the examinee, (2) mathematical content of the item, and (3) the surface properties of the item such as item format etc.

## 2.4 How to Deal With Inequivalent Data

Finally to conclude this chapter it should also be mentioned what to do if the DIF analyses specify inequivalent items between groups. In general, in dealing with the inequivalent data, four different methods are possible, two of which may be considered as the extreme cases. One of the extreme cases is precluding the comparison when an evidence of inequivalence is faced. On the other side lies the second extreme case, which is completely ignoring the inequivalence. Poortinga (1989) states that this is usually a result of the wrong assumption that face validity of an instrument assures equivalence.

Reduction of equivalence is the third option in which biased items are eliminated. For example Hulin (1987) explains a method of eliminating the nonequivalent items and then reestimating the abilities with the rest of the items and again testing all items for equivalence and eliminating nonequivalent items and so on until no nonequivalent items are found. On the other hand, Roznowski and Reith (1999) provide an interesting perspective with result of their study that is not supporting the assumption of, 'differentially functioning items should be deleted in order to get a fair measurement'. They suggest that after determining the differentially functioning items it must be the following concern to test whether elimination of these items contributed the quality of the test, because differential item functioning is an item level analysis whereas individuals are usually compared at the test level. Also, biased items may be pointing out some cross-cultural differences that may require further investigations. In the case of these items, this potential source of information will be lost. In addition, this elimination method may deform the content validity of the instruments.

In this context, fourth method, interpreting the inequivalence, seems to be the most effective one, because no information is lost in this method. In this method inequivalence itself is considered as potential informative about the nature of cross-cultural differences. Purpose of these studies are not comparing the scores any more but treating biased versus unbiased items as a dichotomous variable and explaining the cross-cultural differences with this variable (Hulin, 1987).

This section is concluded with the discussion of etic and emic concepts. If characteristics that an item measures are relevant to the trait in the source but not in the target culture, such culturally specific characteristics or concepts are referred to as emic, and in contrast culturally general concepts are called as etic concepts (Hui & Triandis, 1985; Hulin & Mayer, 1986; Hulin, 1987). The general aim in cross-cultural studies is increasing the sensitivity and cultural relevance of the instrument for both cultures, but at the same time retaining the psychometric equivalence. For this aim, etic items can be used as anchor items in linking the two language forms of a test to generate scales that not only reflect cultural uniqueness but at the same time satisfying the psychometric equivalence (Sireci, 1997)

## CHAPTER III

### METHODS

In this chapter, three major aspects of the study; (1) samples, (2) instruments and (3) procedures, are described.

#### 3.1 Population and Sample

This study examined American and Turkish data from two international studies: (1) TIMSS 1999 and (2) PISA 2003.

##### 3.1.1 TIMSS 1999

In IEA studies, international desired population and national desired population are defined. International desired population is the target population for all countries. Particularly for TIMSS 1999 international desired population is defined as, “All students enrolled in the upper of the two adjacent grades that contain the largest proportion of 13-year-olds at the time of testing” (Foy & Joncas, 2000). In addition, all participating countries are expected to define their national desired population to correspond as closely as possible to the definition of TIMSS 1999 international desired population. There may be differences between these two definitions because sometimes National Research Coordinators (NRCs) had to make changes, for example some countries had to restrict non-native language speakers.

The basic sample design for TIMSS 1999 is generally referred to as a two-stage stratified cluster sample design. The first stage consisted of a sample of schools, which may be stratified; the second stage consisted of a single classroom selected at random from the target grade in sampled schools.

More technically, the sample-selection method used for first-stage sampling was based on a systematic probability-proportional-to-size (PPS) technique. In this technique the probability of selection for a school was proportional to the number of eighth-grade students in the school (Foy & Joncas, 2000). In the second sampling stage, classrooms of students were sampled. Generally, in each school, one classroom was sampled from the target grade. It is worth adding that in the second sampling stage the sampling units were classrooms, whereas the fundamental sampling units were students. However if each student is a member of one and only one of the classes in a school from which the sampled classes were to be selected then this assures taking students as sampling units in the study.

TIMSS 1999 assessed eighth grade students' mathematics and science achievement in 38 countries. In each participating country, approximately 150 government or state schools were randomly selected for the assessment. In each school, one or two mathematics classrooms of eighth-grade students were randomly selected for a total of about 3500 eighth-grade students in each country.

7841 Turkish, 42.1% females, and 9072 American, 50.9% females, students answered one of the eight booklets of the study. In particular, this study selected subgroup of students who answered the 7<sup>th</sup> booklet of the study, which was the booklet including maximum number of released items.

### 3.1.2 PISA 2003

Very much like the same sampling process was conducted in OECD's PISA study as well. PISA 2003 target population was 15-year-old students in grades 7 and higher from 41 countries. A two-stage stratified sampling design was used in selecting a minimum of 150 schools and 4500 students in each participating country (OECD, 2005).

4855 Turkish, 45% females, and 5456 American, 55% females, students answered one of the thirteen booklets of the study. However, the coverage of 15-years-old was in Turkey.



In particular, this study selected subgroup of students who answered the 2<sup>nd</sup> booklet of the study, which was the booklet including maximum number of released items. The demographic information for both subgroups is given in Table 3.1

Table 3.1 Demographics of the subgroups

	TIMSS 1999 (7 <sup>th</sup> booklet)		PISA 2003 (2 <sup>nd</sup> booklet)	
	Female	Male	Female	Male
American	562 (51%)	548 (49%)	202 (48%)	223 (52%)
Turkish	411 (42%)	569 (58%)	165 (42%)	226 (58%)

## 3.2 Instruments

This study examined the mathematics items in the 7<sup>th</sup> booklet of TIMSS 1999 and mathematics items in the 2<sup>nd</sup> booklet of PISA 2003 achievement tests. Detailed explanation about test designs is given below.

### 3.2.1 TIMSS 1999

IEA conducted its third international study in three steps. First step was conducting the Third International Mathematics and Science Study (TIMSS 1995) in 1991–1998. Third International Mathematics and Science Study-Repeat (TIMSS 1999) in 1997-2001 followed the TIMSS 1995 study. This study is also known as TIMSS-R. Trends in Mathematics and Science Study (TIMSS 2003), was the final step of the TIMSS series that was completed in 2000—2004. Turkey participated in the TIMSS 1999 study, and the study was conducted in May 1999 in Turkey. This current study used the data of TIMSS 1999 mathematics achievement test.

TIMSS focused on curriculum as an important factor in explaining student achievement (Gonzalez & Miles, 2001).

From this perspective curriculum has been considered in three dimensions. The first dimension is the intended curriculum, which is what society would like to see taught. This dimension is usually specified in national educational policies of the countries. What is actually taught is the implemented curriculum, which is investigated through teacher questionnaires, and finally the last dimension is the attained curriculum, which is determined by what the students learn.

The organization and coverage of the intended curriculum were investigated through curriculum questionnaires that were completed by National Research Coordinators. Then from this information, three dimensions were determined to be contained in the achievement test, namely; Content, Performance Expectation, and Perspectives. These three aspects, and categories under these aspects for the mathematics achievement test are given in the Table 3.2, which was adapted from Gonzalez and Miles (2001; p1-8).

The content aspect represents the subject matter content of school mathematics. The performance expectations aspect describes, in a non-hierarchical way, the many kinds of performance or behavior that might be expected of students in school mathematics. The perspectives aspect focuses on the development of students' attitudes, interest, and motivation in the subjects.

This framework was developed for the entire span of curricula from the beginning of schooling through the completion of secondary school. Whereas only the aspects related with the eighth-grade curriculum in all the participating countries are reflected in the eighth-grade TIMSS assessment.

Table 3.2 Three aspects of TIMSS 1999 Mathematics Achievement Test

Content	Performance Expectations	Perspectives
Numbers	Knowing	Attitudes
Measurement	Using Routine Procedures	Careers
Geometry	Investigating and Problem Solving	Participation
Proportionality	Mathematical Reasoning	Increasing Interest
Functions, Relations, and Equations	Communicating	Habits of Mind
Data Representation		
Probability and Statistics		
Elementary Analysis, Validation and Structure		

In this context 162 mathematics items – 61 of which is from Fractions and Number Sense, 24 from Measurement, 21 from Data Representation, Analysis and Probability, 21 from Geometry, and 35 from Algebra – in five content areas were specified in the TIMSS 1999 mathematics achievement test. The subjects under these content areas were as follows: Fractions and number sense included whole numbers, fractions and decimals, integers, exponents, estimation and approximation, proportionality. The measurement area included standard and non-standard units, common measures, perimeter, area, volume, and estimation of measures. Data representation, analysis, and probability included representing and interpreting tables, charts, and graphs; range, mean; informal likelihood, simple numerical probability. Geometry included points, lines, planes, angles, visualization, triangles, polygons, circles, transformations, symmetry, congruence, similarity, and constructions. And finally algebra section included number patterns, representation of numerical situations, solving simple linear equations, operations with expressions, representations of relations and functions.

About one-third of the test time of TIMSS was devoted to free-response items. To ensure reliable scoring procedures detailed guidelines were prepared. In general, free-response items were evaluated in one of completely incorrect, partially correct or complete correct statuses.

To ensure broad subject matter coverage a rotated design was used (Adams & Gonzalez, 1996). In this method, items in the item pool were first assigned to one of 26 mutually exclusive groups, or ‘clusters’. The clusters of items were then systematically assigned to eight test booklets, and this eight student booklets were distributed systematically in each classroom, one per student (Gonzalez & Miles, 2001). Even though no student has responded to the entire item pool this method produces a reliable estimates of the performance of the population on all the items.

This present study investigated one of these eight booklets, the 7<sup>th</sup> booklet which includes more number of released items than that of others. Number of TIMSS 1999 mathematics test items by type and reporting category in 7<sup>th</sup> booklet is given in Table 3.3.

Table 3.3: TIMSS 1999 Mathematics Test Items of 7<sup>th</sup> Booklet by Type and Reporting Category

Reporting Category	Item Type			Number of Items
	Multiple-Choice	Short-Answer	Extended-Response	
Fractions and Number Sense	12	2	-	14
Measurement	2	1	-	3
Data Representation, Analysis and Probability	6	-	1	7
Geometry	5	-	-	5
Algebra	6	3	1	10
Total	31	6	2	39

### 3.2.2 PISA 2003

PISA 2003 is an OECD project. The first step of the project, which surveyed reading, mathematical and scientific literacy, with a primary focus on reading, was conducted in 2000. Reading literacy, mathematical literacy, scientific literacy and problem solving were the four domains covered in PISA 2003, in which also Turkey participated, with a primary focus on mathematical literacy.

Unlike TIMSS, PISA study assessed how well the students can use what they have learned, in real-life situations. This performance, named mathematical literacy, was defined as individual's capacity of not only identifying and understanding the role that mathematics plays in the world but also using mathematics (OECD 2003a). This is a "mathematics for problems" perspective, which can be regarded as the contradiction of traditional "problems for mathematics" perspective.

PISA study specified this framework with three components; (1) context, (2) content, and (3) competencies. Real life situations determined the context of PISA's framework.

Somewhat different from the curricular approach, content of the study was determined by generic terms; quantity, space and shape, change and relationships, and uncertainty. Finally, in the process of solving real-life problems, referred as mathematisation, mathematical competencies that students should possess were grouped under reproduction, connection and reflection clusters (OECD, 2003a).

Within each domain, students' performance in using their knowledge and skills in order to meet the real-life challenges was assessed. 85 mathematics items, in multiple choice, short answer and extended response types, were used in the study (OECD 2005).

PISA study also used rotated design to produce 13 booklets, one of which is randomly assigned to each of sampled students. This present study investigated one of these thirteen booklets, the 2<sup>nd</sup> booklet which includes more number of released items than that of others. Number of PISA2003 mathematics test items by content category and item format in 2<sup>nd</sup> booklet is given in Table 3.4.

Table 3.4 PISA 2003 Mathematics Test Items of 2<sup>nd</sup> Booklet by Format and Content Category

Content Category	Item Format			Number of Items
	Multiple-Choice	Closed-Constructed	Open-Constructed	
Space and Shape	3	6	1	10
Quantity	1	8	1	10
Change and Relationships	1	5	2	8
Uncertainty	5	3	-	8
Total	10	22	4	36

### 3.2.3 Translation Process

To minimize the semantic, psychometric, and linguistic differences between the source and translated language versions of the tests, strict verification procedures to assure translation equivalence were followed.

TIMSS 1999 instruments were developed in English, translated into 33 other languages by following explicit guidelines for translation and adaptation. Professional translators, in consultation with subject matter experts, in National Centers tried to assure that meaning and difficulty of items did not change between source and target versions. In addition a series of statistical checks to detect items performing differently were carried on (Gonzalez & Miles, 2001).

PISA 2003 study implemented stricter verification procedures. Two parallel source versions, English and French, were developed to provide a chance for the countries in translating each of the two versions in their language and then reconciling them into one national version. National translators were recommended guidelines and trained on translation procedures. In addition, international professional translators verified the national versions against source versions (OECD 2005).

A double translation procedure was used in both studies. This procedure requires two independent translations from the source language and followed by a reconciliation of a third translator. PISA has an additional advantage of double translation from two different languages.

American version of PISA instrument was adapted from the English source, and the Turkish version was double translated from the English source.

### 3.3 Analysis of Data

In this section, a description of data selection procedures, information about recoding, and rationale of selecting subgroup of items are given. In addition, statistical and judgmental procedures used in the analyses are defined.

Investigating cross-cultural equivalence requires hierarchical analyses to determine whether the tests forms are free of construct, method, and item biases (Hui & Triandis, 1983; 1985).

Exploratory and Confirmatory Factor Analyses were used in complimentary fashion to check whether there was a substantial overlap of the construct measured by the tests across cultures, or in other words whether the same psychological construct was measured across groups (van de Vijver & Tanzer, 1997; Sireci, Bastari & Allalouf, 1998).

Restricted Factor Analysis, Mantel-Haenzsel, Item Response Theory Likelihood Ratio tests are used to detect item level differentiations. In addition the effect of using anchor items, and the effect of purification were investigated. The following pages of this section presents details of how these analyses were conducted, including judgmental reviews of the items which were detected as functioning differentially across groups.

#### 3.3.1 Descriptive Summary

A detailed coding process was conducted in TIMSS and PISA projects by the related stakeholders. Prior to the release, the data was cleaned as a component of quality control and assurance program works (OECD, 2003a; Gonzalez & Miles,

2001). Well-defined procedures were developed for reliably evaluating student responses. In addition, this current study conducted the following steps.

In PISA 2003 and TIMSS 1999, items not responded although it was expected to be, non-reached items and items in which more than one alternative selected were coded as missing values. This differentiation was optimal for parameter estimation within item response theory. However within the context of differential item functioning analyses, these items were recoded as incorrect answers.

It should also be mentioned that for free response items scored on 0 to 2 scale, these are the items mentioned as PCR on Appendix A, the scores were rescaled to a scale of 0 to 1 prior to analyses so that the interpretation of the statistics would be the same for all item types. Partially correct answers were treated as correct answers. The number of partially correct cases recoded as correct is given within parentheses by the names of items for Turkey and USA respectively: For TIMSS; m022262c (38,52) and m022256 (353,253) and for PISA; m124q03t (114,185), m150q02t (111,180), m462q01t (17,45) and m520q01t (40,38).

### 3.3.2 Construct Equivalence

Exploratory and confirmatory factor analyses were used to evaluate construct equivalence of assessments across Turkish and American groups. In addition the results from this step were used to form a basis on which DIF hypotheses based (Gierl, 2005; Williams, 1997). Analyses conducted in this section, details of which are given in the following pages, can be summarized as; a) checking whether individual EFAs in both groups produced equal factor structures, and produced any evidences indicating the unidimensionality of the data, b) providing a statistical test of unidimensionality through CFA, and when the data were shown to be not unidimensional, selecting a subset of items through EFA in pooled data of both groups into one data file, c) conducting multi-group CFA.



## Exploratory Factor Analysis

Separate principal component analysis (PCA) was performed for Turkish and American groups. Searching evidence of construct equivalence, rotated components of the PCA were compared to check whether factor loadings were similar (Sireci, Bastari & Allaouf, 1998). The results were analyzed qualitatively to determine a basis for confirmatory factor analyses as well.

PCA was conducted using SPSS (version 10.0). Factors of eigenvalues greater than 1 were extracted, and components rotated with varimax rotation were compared (George & Mallery, 2003; Zwick & Velicer, 1986).

## Confirmatory Factor Analysis

CFA was used for two purposes. First, the models inferred through the exploratory analyses were tested individually for each group. At the second step, multiple group analyses were carried on to check whether the groups' data possessed a common structure. A framework of the details is given below.

PRELIS 2.72 and LISREL 8.72 programs and SIMPLIS command language were used in conducting confirmatory factor analysis (CFA). LISREL is a computer program (Jöreskog & Sörbom, 2001; 2002) performing structural equation modeling. On the other hand the SIMPLIS command language has the advantage of moving away from the matrix formulation of the LISREL model to a more natural language to define LISREL models (Kelloway, 1998).

For LISREL analyses, factor structures were specified through measurement models to define how the latent variables or hypothetical constructs were measured in terms of the observed variables. Latent variables are indirectly observable or measured variables. In other words, they are the variables that can be indirectly measured through observable variables such as items in a test (Schumacker & Lomax, 1996). Mathematical Literacy and Mathematics Achievement were two latent variables defined within the context of PISA and TIMSS, respectively.

The relationships between the observed variables and the latent variables are described on the basis of the factor loadings, which are in fact regression coefficients. By the factor loadings, the information about the extent to which a given observed variable is able to measure the latent variable is provided. These coefficients serve as validity coefficients. In addition; the measurement errors for the observed variables is a basis for reliability coefficients (Schumacker & Lomax, 1996).

In TIMSS and PISA the observed variables, i.e. the items of the tests, are ordinal, which do not have origins or units of measurements (metric). To account for this, PRELIS assigns a metric to these ordinal variables assuming that there is an underlying continuous variable for each ordinal variable having parameters corresponding to the categories of the ordinal variable, which are called thresholds. PRELIS then estimates the polychoric correlations among these underlying variables, in other words tetrachoric correlations as the variables are dichotomous, and their asymptotic covariance matrix. Then, using this estimated matrix of polychoric correlations and corresponding asymptotic covariance matrix, LISREL program calculates not only the value of factor loadings and measurement errors but also goodness of fit indices (Jöreskog, 2005).

These fit indices determine the degree to which the specified structural equation model fits the sample data. The differences between the observed and model-implied correlation (or covariance) matrix are considered by the program in calculating the indices. Indices used to investigate model-data-fit in this current study is given below.

A non-significant Chi-Square ( $\chi^2$ ), which specifies statistical fit, implies non-significant difference between the covariance matrix implied by the model and the population covariance matrix, which means that the population covariance matrix can be reproduced by the model (Kelloway, 1998).

A point worth specifying is that the  $\chi^2$  criterion is very sensitive to sample size. When the sample size increases, generally above 200, the  $\chi^2$  criterion tends to indicate a significant probability level (Schumacker & Lomax, 1996).

For this reason, an adjustment can be used. Normed Chi-Square (NC) is the adjusted Chi-Square on ratio of the  $\chi^2$  and its degrees of freedom.  $\chi^2 / df$  ratios of less than 5 indicate a good fit to the data, like ratios between 2 and 5. Moreover,  $\chi^2 / df$  ratios of less than 2 indicate over fitting (Kelloway, 1998).

On the other hand, there are indices investigating model-data-fit beyond statistical fit. Goodness-of-Fit Index (GFI) is based on the ratio of the sum of the squared differences between the observed and reproduced matrices to the observed variances. (Schumacker & Lomax, 1996). The range of the GFI is from 0 to 1. The values exceeding 0.9 indicates a good fit to the data (Kelloway, 1998).

Adjusted Goodness-of-Fit Index (AGFI) is the adjusted GFI for the degrees of freedom of a model relative to the number of variables (Schumacker & Lomax, 1996). The AGFI also has a range from 0 to 1, with values 0.9 indicating a good fit to the data (Kelloway, 1998). The fit of two different models with the same data or the fit of models with different data can be compared by using the GFI and AGFI indices (Schumacker & Lomax, 1996).

Root-Mean-Square Residual (RMR) is another fit indices. The RMR is the square root of the mean of the squared differences between the implied and observed covariance matrices. A good fit is indicated by the low values of RMR whose lower bound is 0. Because of the difficulty of determining what a low value is, the standardized RMR is provided by LISREL. The standardized RMR (SRMR) has a lower bound of 0 and an upper bound of 1. For the interpretation of indicating a good fit to the data, values less than 0.05 are generally accepted (Kelloway, 1998).

Root-Mean-Squared Error of Approximation (RMSEA) is computed on the basis of the analysis of residuals. Smaller values of RMSEA indicate a better fit to the data. According to Steiger (1990), values below 0.10 indicate a good fit, values below 0.05 indicate a very good fit and the rarely obtained values below 0.01 indicate an outstanding fit to the data.

RMSEA also provides 90% confidence intervals for the point estimate. In addition, a test of the significance of the RMSEA is provided by the LISREL (Kelloway, 1998).

Normed Fit Index (NFI) is based on the percentage improvement in fit over the baseline independence model (Bentler & Bonett. 1980). The NFI has a lower bound 0 and an upper bound of 1. A NFI of 0.90 means that the model is 90% better fitting than the null model. In spite of the widely usage of the NFI, it has a disadvantage of underestimating the fit of the model with small samples (Kelloway, 1998). Non-Normed Fit Index (NNFI) is also calculated to come over this disadvantage. The NNFI is the adjusted NFI for the number of degrees of freedom in the model. For a better fitting model, higher values of NNFI of 0.90 indicate a good fit of the model to the data (Kelloway, 1998).

Comparative Fit Index (CFI) is the last indices used in the study. The CFI is proposed by Bentler (1990) on the basis of non-central  $\chi^2$  distribution. The range of CFI is from 0 to 1, with the values exceeding 0.90 indicating a good fit to the data.

RMSEA, RMR, CFI and NNFI were also used in the original model data fit analyses conducted by the contributors of PISA study (OECD, 2005).

At the second step of the CFA, items selected through the studies at the first step were further analyzed to check whether they measured the same latent constructs in all countries.

Multi-group confirmatory factor analyses investigate existence of an equivalent construct in the original and translated tests. In addition it also investigates to what degree that construct is being measured equivalently among groups (Zumbo, 2003).

The model tested in multigroup analyses between American and Turkish groups is given below in equation (1) as in standard LISREL notation.

$$\vec{\mathbf{x}} = \vec{\boldsymbol{\tau}}_{\mathbf{x}} + \vec{\boldsymbol{\Lambda}}_{\mathbf{x}} \vec{\boldsymbol{\xi}} + \vec{\boldsymbol{\delta}} \quad (1)$$

The superscripts indicate that variables were in matrix or vector styles.  $\vec{\mathbf{x}}$  is a vector of underlying variables of the items. Because the items in the test are in ordinal scale, underlying variables specified by threshold values are estimated to get a scale having origin and unit of measurement.

$\vec{\tau}$  is a vector of intercepts, which includes magnitudes to determine individuals' performance on an item when the effect of their position with respect to the construct measured by the test is neglected.  $\vec{\Lambda}$  is a vector of factor loadings indicating the effect of the construct on item performances,  $\vec{\xi}$  is a vector of latent variables or in other words constructs measured in the test and finally  $\vec{\delta}$  is a vector of measurement errors.

If the measurement model as specified in one group is completely reproduced in the other including the magnitude of the factor loadings, intercepts and error variances, there exists a strict structural invariance between groups. Strict invariance points out not only the existence of an equivalent construct in the original and translated tests but also assures that the construct is being measured equivalently among groups.

On the other hand, if only the overall pattern of the model exists in both groups but neither the magnitudes of the factor loadings nor error variances are equivalent then there is a weak invariance between groups, which means that although there exists an equivalent structure between groups, it is not measured equivalently between groups. Hence, weak invariance still provides a basis to carry on item-level analyses (Zumbo, 2003; Reise, Widaman & Pugh, 1993). In empirical studies, construct equivalence met is mostly between these two extreme, namely strict and weak invariance ends.

It is regarded that error variances are sample specific whereas intercepts and factor loadings are attributes of the variables (Reise, Widaman & Pugh, 1993). Therefore, in this current study the strict model was specified to estimate equal intercepts and factor loadings but error variances were allowed to differ between groups.

In this context, structural (also called factorial) invariance of TIMSS and PISA items were tested via multi-group confirmatory factor analyses using PRELIS and LISREL programs (Jöreskog & Sörbom, 2001; 2002).

At the PRELIS step, set of thresholds from the combined sample of American and Turkish data was estimated individually for TIMSS and PISA study.

Then, to define a common scale for the underlying variables in both countries, thresholds to be calculated for each individual country were fixed at these estimated thresholds in running the PRELIS program to get the mean vector, the covariance and the asymptotic covariance matrices of the underlying variables for American and Turkish groups.

At the LISREL step, these mean vectors, the covariance matrices, and the asymptotic covariance matrices from the PRELIS program were used in multi-group analyses.

This specified method has an advantage of dealing with ordinal data. As the ordinal data possess neither a metric nor an origin, calculating means of these variables or correlations among these variables is meaningless. On the other hand, estimation of thresholds for each underlying continuous variable corresponding to an observed ordinal variable, and using these values as if the latent variables had been observed solves the metric problem. In addition, fixing these thresholds in both groups provides a common metric for group comparisons (Jöreskog, 2005).

In TIMSS and PISA studies individually, a model was forced to be reproduced in both countries including all its parameters but the error variances. Then, in addition to the goodness-of-fit indices, modification indices (MI's) produced by LISREL were investigated in deciding the parameters to be allowed to differ between countries.

LISREL program specifies in its output additional parameters to be estimated in order to get a better model-data fit. MI's are quantities showing this improvement (Jöreskog, 2005). In other words, MI for each parameter is an estimate of the decrease in chi-square that would occur if the parameter was set free to be estimated.

The distribution of MI is approximately chi-square with one degree of freedom. In this current study, the largest of all MI's indicating to set an intercept or a factor loading free to be estimated in Turkish and American groups was considered. If it was significant, the associated parameter was set to be free, and the new model was fitted once more. Then subsequently the largest MI was considered again in the same manner as before. This process was repeated until no significant MI's were produced.

In determining the significance of MI's an adjustment procedure was used. Because the program calculates  $2 \times n$  MI's for an  $n$  item test (one for intercept and one for factor loading of each item), there is a problem of chance capitalization. In addition, MI is influenced by sample size. Given a sufficiently large sample size, some of the MI's will be statistically significant. To account for these Type 1 error promotions the following adjustment procedure offered by Oort (1992) was used.

Let  $\text{Chi}^2$ ,  $\text{df}$ , and  $\text{MI}$  also represent the estimated magnitudes of the corresponding Chi-square value for overall fit, degree of freedom and the largest modification index for an intercept or factor loading parameter, respectively. It is expected that setting free the parameter indicated by the largest MI would yield a new model with Chi-square value of  $(\text{Chi}^2 - \text{MI})$  and  $(\text{df} - 1)$  degrees of freedom. Adjusted modification index (AMI) is an estimate of MI in case the relaxed model would have resulted in perfect fit. As, with a perfect fit, Chi-square has an expected value that is equal to the degrees of freedom, the Chi-square value of the released model should have to be multiplied by  $((\text{df} - 1) / (\text{Chi}^2 - \text{MI}))$  to make it equal with the degrees of freedom. To get AMI, MI is multiplied by the same factor, as given in equation 2.

AMI were then compared to the critical  $\text{Chi}^2$  ( $\text{df}=1$ ) value for the 1% level of significance.

$$\text{AMI} = \frac{(\text{df} - 1)}{(\text{Chi}^2 - \text{MI})} \times \text{MI} \quad (2)$$

### 3.3.3 Differential Item Functioning

Restricted Factor Analysis (RFA), Mantel-Haenszel (M-H) and Item Response Theory Likelihood Ratio (IRT-LR) methods were used in the DIF analyses of the items selected through the works in the previous section. In addition, comparison of DIF results among DIF methodologies, the effect of using anchor items in IRT-LR, the effect of purification in (M-H) were investigated in the study.

## Restricted Factor Analysis

DIF is an item level analysis. A modified form of the model given by equation 1 for an individual item specifies the model used in RFA as well. For an item, measuring trait  $\xi$ , to be determined as functioning differentially with respect to a violator  $\psi$ , conditional distribution of this item scores given  $\xi$ , should be different from the conditional distribution of this item scores given  $\xi$  and  $\psi$  (Oort, 1992). From this definition the model in equation 1 is modified as in equation 3, to be used in detection of DIF in RFA.

$$X_i = \tau_i + \Lambda_i \xi + \Delta_i \psi + \delta_i \quad (3)$$

In the equation,  $X_i$ , the score  $X$  of a randomly selected subject on item  $i$ , is modeled in terms of scores on main construct  $\xi$  intended to be measured and scores on the potential violator  $\psi$ .  $\Lambda_i$  and  $\Delta_i$  are the respective population regression coefficients of item on  $\xi$  and  $\psi$ . In addition,  $\delta_i$  is the residual factor and  $\tau_i$  is the intercept.

In terms of the model given in equation 3, an item is said to be functioning differentially with respect to  $\psi$ , if there is a direct effect of violator  $\psi$  on item  $i$ , that is,  $\Delta_i \neq 0$ . It should also be noted that more than one potential violator can be included in equation 3. However, within the scope of current study only grouping variable was included as a potential violator.

The data files for RFA were prepared by pooling the Turkish and American data into a common file, individually for PISA and TIMSS. The “country” variable, coded 0 for Turkish and 1 for American groups, was also included as an additional variable in the data file.

Using these data files for PISA and TIMSS, polychoric correlations and their corresponding asymptotic covariance matrix were estimated via PRELIS program. As the variables included in original data files were ordinal, these estimated matrices provide a metric and a unit of measurement to carry out the further analyses (Jöreskog, 2005).



At the LISREL step, the model specifies in equation 3 was tested for PISA and TIMSS individually. Polychoric correlations and their corresponding asymptotic covariance matrices, estimated with PRELIS, were used in this step. Figure 3.1 gives the graphical display of the model tested in the analyses.

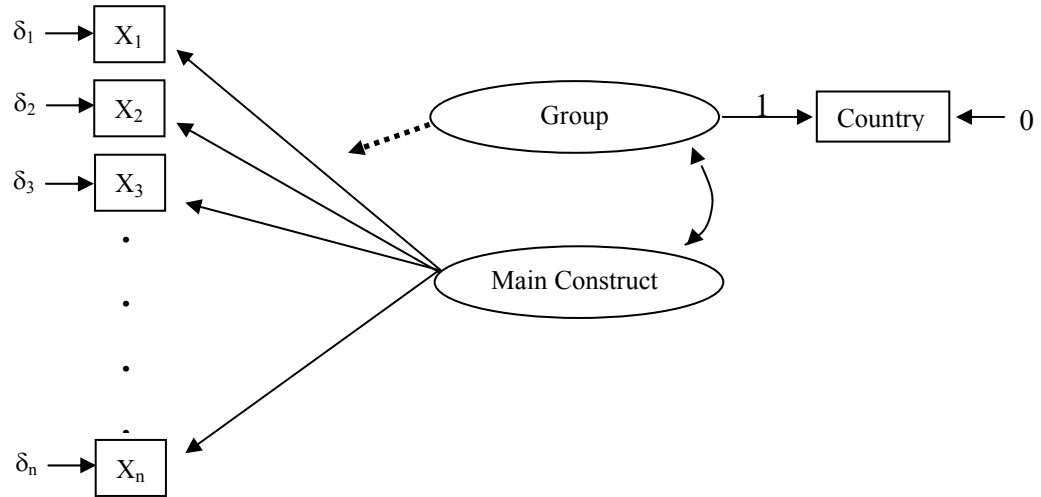


Figure 3.1 Graphical Display of RFA Model Used in Analyses.

In RFA analyses, DIF was detected by the regression coefficients of the items on the “Group” variable, the potential violator. As the “Group” variable was determined only by observed “Country” variable, measurement error was fixed at 0, and factor loading of “Country” on “Group” was fixed at 1 to replace “Group” by “Country”.

Analyses started with the Null Model, in which all of the regression coefficients of the items on the “Group” variable were fixed at zero. Starting from the item with the largest MI, it was investigated whether there would be a significant increase in the model-data fit, if the corresponding factor loading of the item would have been set to be freely estimated. Significance analyses were carried on using AMI’s given in Equation 2.

If the largest of all AMI's were significant, the corresponding item was removed from the test and the Null Model was fitted on the reduced data set for the subsequent analysis. This process was repeated until all DIF items were removed from the tests.

Expected Parameter Changes (EPC) for each item determined as showing DIF, was investigated to determine the direction of DIF. EPC is an estimate of the magnitude of the parameter fixed to be 0 in the model if it was allowed to be freely estimated in the model. Positive values of EPC indicate that item is more attractive to subjects with high scores on the potential variable, American students in our case as the highest score of 1 in "Country" variable indicates USA.

In RFA analyses as all the data from different groups are combined in a single data set, no additional concern of getting a common scale for comparisons is required.

### Item Response Theory Likelihood Ratio Analysis

Item Response Theory (IRT) provides models mathematically linking item responses to underlying latent traits measured by the test, through some specified models (Lord & Novick, 1968; Lord, 1980).

One of the most popular of these models is (one, two or three parameter) logistic model. In this model, the construct measured by the test, named as "ability" within the context of IRT and called  $\theta$ , with some additional item parameters are used to specify the performance of a randomly selected individual on the item (Hambleton, Swaminathan & Rogers, 1991).

Three-parameter logistic model is given in equation 4 defining the probability of a correct response of an individual at the  $\theta$  level to the  $i^{\text{th}}$  item.

$$P_i(\theta) = c_i + \frac{(1 - c_i)}{1 + \exp[-Da_i(\theta - b_i)]} \quad (4)$$

In the equation determining the probability of a randomly selected individual getting an item  $i$  correct,  $P_i(\theta)$ ,  $c_i$  is the lower asymptote of the item characteristic curve, also called as pseudo-chance level parameter and specifies the probability on a very low ability student getting the item correct;  $b_i$  is the item difficulty expressed in the same metric as  $\theta$  and represents the point on the  $\theta$  scale at which examinees have a 50% chance of answering the item correct; and  $a_i$  is the item discrimination parameter (or slope), which is proportional to the relationship between item response and  $\theta$ .  $D$  is a scaling constant usually set equal to 1,702 in order to take advantage of certain relations between logistic item response models and normal ogive models of item characteristic curves (Thissen, Steinberg & Wainer, 1993).

Equation 4 specifies the three-parameter IRT model. If  $c_i$  in this equation is set equal to zero we get the two-parameter model, and in addition if it is assumed that all items in the test have the same discrimination, that is fixing  $a_i$  for all items in a test, we get the one-parameter model (Lord, 1980).

Item Characteristic Curve (ICC) is a graphical representation of the relationship given in Equation 4. As an example, Figure 3.2 gives an ICC produced by BILOG-MG (Version 3.0) for an item (du Toit, 2003).

The S shaped trace line in the figure is ICC specifying the probability of an individual at various ability levels answering an item of discrimination 1.012, difficulty 0.317 and chance level 0.267 correct. Confidence intervals of performance estimations through the model and observed performances (specified by points) are also given on ICC to determine model-data fit.

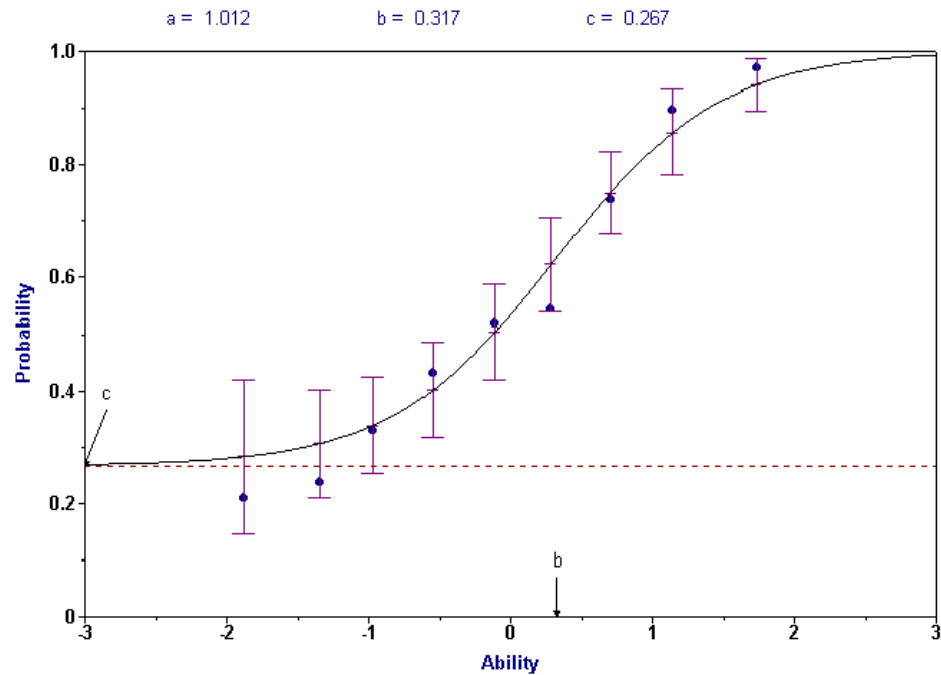


Figure 3.2 An Item Characteristic Curve

Beyond many uses, ICC curve provides a framework for DIF analyses as well. If the trace lines of an item estimated from different groups are also different, provided that the parameters are on the same scale, that item is said to be functioning differentially between groups.

Thissen, Steinberg and Wainer (1993) give a comprehensive definition of DIF in terms of ICC as, “The value of the trace line at each level of  $\theta$  is the conditional probability of a correct response given that ability or proficiency. If we are considering the possibility that an item may function differently (exhibit DIF) for some focal (F) group relative to some (other) reference (R) group, then in the context of IRT we are considering whether the trace lines differ for the two groups.” (p. 68)

On the other hand, as an ICC is determined by item parameters, such as  $a$ ,  $b$  and  $c$  in the three-parameter model, considering whether the trace lines differ for the two groups coincides with considering whether the item parameters differ for the two groups. In this context, Item Response Theory Likelihood Ratio (IRT-LR) method is a way using chi-squares to test the null hypothesis of no group differences in ICCs or equivalently in item parameters (Thissen, Steinberg & Wainer, 1988; 1993; Thissen 2001). Information about the IRT-LR process is given next.

IRT-LR method uses “Likelihood” magnitudes, which represent the likelihood of the data given the parameter estimates of a model, in comparing compact and augmented models. The compact model (Model C) restricts some parameters of items to be equal in both groups. The augmented model (Model A) includes all of the parameters of the compact model, but in addition allows at least one of the restricted item parameters of compact model to vary between groups. Then, likelihood ratio test is used to test the difference between the two models. Testing whether it is worth to estimate two different parameters for an item from two different groups, in the augmented model, is also testing whether this item functions differentially between groups.

In likelihood ratio (LR) test, the statistic to be tested is the difference between the negative twice loglikelihoods of the compact and augmented models, that is,  $-2(\text{Loglikelihood [Model C]} - \text{Loglikelihood [Model A]})$ , denoted by  $G^2$ .  $G^2$  is distributed as  $\text{Chi}^2$ . Degree of freedom of this distribution is the difference between the number of parameters in Model C and Model A (Thissen, Steinberg & Wainer, 1988; 1993).

IRT-LR test was used to be conducted through the computer application MULTILOG (du Toit, 2003). However, as this program requires multiple runs it is relatively difficult to carry on the analyses with MULTILOG. To provide a solution to this issue, Thissen (2001) has recently developed a new application IRTLRF v.2.0b for the detection of DIF using IRT-LR test. This program uses marginal maximum likelihood method in estimating the item parameters (Bock & Aitkin, 1982).

A summary of the procedures, described by Thissen (2001), used in IRTLRDIF program to detect DIF is given in Figure 3.3. Released parameters to be estimated freely between groups in each model, are also specified in the figure.

The program starts with Model-A constraining all parameters to be equal for the two groups. This very first step also produces  $\mathcal{LL}_{AllEqual}$ , the loglikelihood for all item parameters constrained equal. Then DIF analyses start for each item I in turn, if no designated anchor is specified (the anchor specified procedure is explained in the next section).

In Model-B all item parameters are constrained to be equal except those of item I. The program then calculates the  $G^2$  value, that is  $(-2(\mathcal{LL}_{AllEqual} - \mathcal{LL}_{I Not Equal}))$  between Model-A and Model-B, for an overall test of significance of DIF, considering all parameters of item I.

For the program to continue analyzing the item parameters causing DIF,  $G^2$  value should exceed 3.84, the critical value of the  $\chi^2$  distribution for one degree of freedom at 5% alpha level. Because there is no possibility that any of the single degree of freedom hypothesis tests reach significance if the item-level omnibus test does not exceed 1 degree of freedom critical value (Thissen, 2001).

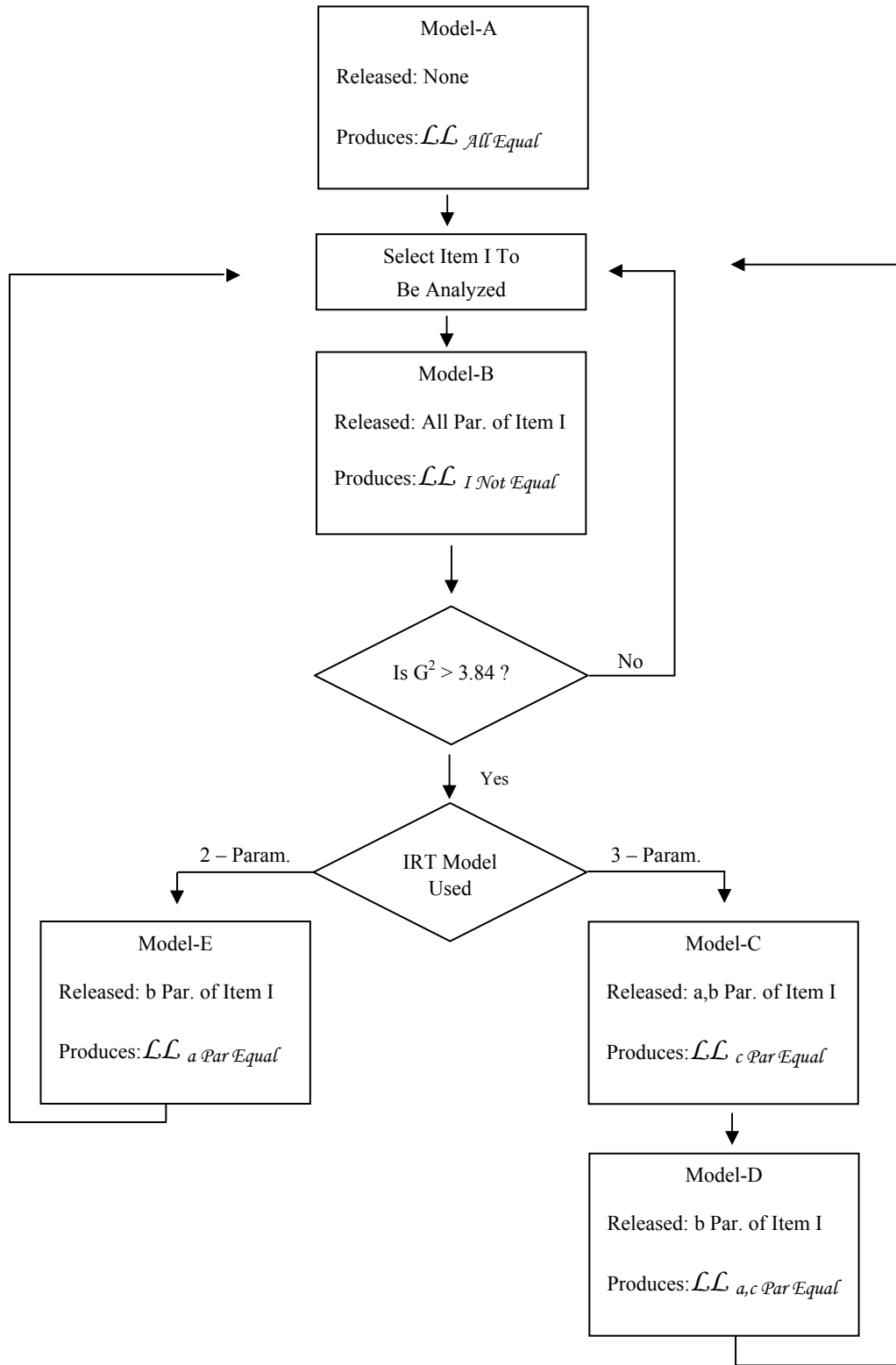


Figure 3.3. DIF-detection Algorithm in IRTL RDIF

In single degree of freedom tests analyzing equivalence of single item parameter at a time, first equivalence of the guessing parameter  $c$ , then the slope parameter  $a$ , and finally the threshold parameter  $b$  should be tested. IRTL RDIF program takes this into consideration.

In three-parameter model, for the test of  $c$ -DIF, the program compares Models C and B given in figure 3.2 through LR test. In the same manner, Models D and C for the test of  $a$ -DIF conditional on equal  $c$  parameters, and Models A and D for the test of  $b$ -DIF conditional on equal  $c$  and  $a$  parameters for the two groups are compared. On the other hand, in the two-parameter model, for the test of  $a$ -DIF, Models E and B, and for the test of  $b$ -DIF conditional on equal  $a$  parameters Models A and E are compared through LR tests.

In all these comparisons, IRTL RDIF program fixes the mean and standard deviation of the reference group, (USA within the context of the current study) at 0 and 1, respectively and estimates the mean and standard deviation of the focal group (Turkey within the context of the current study). In testing the models stated above, Benjamini-Hochberg (1995) procedure was used to reduce the false discovery rate due to multiple comparisons (Williams, Jones & Tukey, 1999). To this reason, observed  $p$  values corresponding the  $G^2$  differences between the compared groups were calculated first. Then this observed  $p$  values were ranked from largest to smallest. All possible number of comparisons, such as 60 in a 20-item test with respect to two-parameter model, was attached as the rank of the greatest observed  $p$  value, and this rank was decreased one for each subsequent  $p$  value. These ranks were used to calculate the adjusted critical  $p$  values (AC) as given in equation 5. In a model comparison, corresponding hypothesis was rejected when the observed  $p$  value was less than the AC.

$$AC = (\text{level of significance}) \times \frac{\text{rank of observed } p \text{ value}}{\text{number of comparisons}} \quad (5)$$



### 3.3.4 Using Anchor Items in IRT-LR

From a unidimensional perspective, DIF analysis measures violations from unidimensionality. To detect this violation, DIF methodologies starts with matching the students having the same position with respect to the main dimension measured by the test. As the more valid this matching process the more valid the DIF results, there is a special concern in DIF studies to enhance this matching process. Purification of the matching criterion (Allalouf, Hambleton & Sireci, 1999; Dorans & Holland, 1993; Camilli & Shepard, 1994), determining a valid sector (Ackerman, 1992), or a valid subtest (Shealy & Stout, 1993), using anchor items (Williams, 1997) all deal for a better specification of the main construct that the test is intended to measure. In this context, to increase the effectiveness of IRT-LR, this current study also detected anchor items by using a methodology outside of IRT calibration model (Sireci, 1997), and then reconducted the IRT-LR procedure as defined in the previous section, but using the anchor items specified in this section. The details of this process are given next.

IRT-LR DIF detection method in the previous section used all the items except from the investigated one as anchor items, that is, all the parameters of these items were fixed among groups. In this section using the methodology offered by Williams (1997), a group of anchor items were determined that are separate from the items under investigation and then only these items were used in matching the students in IRT-LR.

In determining the anchor items, it was required that the anchor items have a large item discrimination values and a wide range of item difficulties, have small and nonsignificant Mantel-Haenszel statistics, and have high factor loadings and small error variances in EFA. It was also assured that the mean differences between groups in anchor items were consistent with the mean differences in the entire set of items. Mantel-Haenszel procedure was conducted using the EZDIF computer program produced by N.G.Waller (2005). Thick matching was used in determining the students of the same ability (Donoghue & Allen, 1999).

## Mantel-Haenszel

Mantel-Haenszel (MH) is a nonparametric DIF method (Dorans & Holland, 1993). MH yields a  $\chi^2$  test to test the null hypothesis that there is no relation between group membership and test performance on an item for the individuals of the same ability.

For a studied item, MH creates two-by-two contingency tables for each score categories, indicating the number of individuals in reference and focal groups providing correct or wrong answers to the studied item. The method then, calculates the ratio of the odds that a reference group examinee will answer the item correct to the odds that a focal group examinee will answer the item correct. This ratio is also called as odds ratio and denoted by  $\alpha$ . Mantel-Haenszel method tests the hypothesis claiming that  $\alpha = 1$ .

In determining the score categories, percent total method, a thick matching procedure allocating similar number of examinees to each level of the matching variable was used (Donoghue & Allen, 1993). To this purpose, 20<sup>th</sup>, 40<sup>th</sup>, 60<sup>th</sup> and 80<sup>th</sup> percentiles of the total score of the combined Turkish and American data were calculated, for TIMSS and PISA individually. Then these score categories were provided to the EZDIF program to determine 5 score categories (Waller, 2005).

Mantel-Haenszel also provides an effect size measure MH D-DIF given in equation 6. “DIF free” items are those with MH D-DIF values not significantly different from zero. Items with MH D-DIF values that are significantly different from zero are denoted as “Low DIF” items, provided that the absolute value of MH D-DIF is also equal or greater than 1 but less than 1.5. For an item to be categorized as “High DIF”, MH D-DIF value should be at least 1.5 and significantly greater than 1. In the EZDIF output these classification is denoted as A-, B-, and C- DIF respectively. It should also be added that a positive value of MH D-DIF specifies items favoring the focal group (Holland & Thayer, 1988).

$$\text{MH D-DIF} = -2.35 \ln(\alpha) \quad (6)$$

EZDIF program performs single degree of freedom  $\chi^2$  test at 5% level in determining the significance of the statistics. However, to specify a common basis with the other DIF methodologies, and to control the possible inflation in the Type I error rate due to multiple hypotheses tested, Benjamini-Hochberg (1995) procedure was also used to calculate adjusted critical values according to the equation 5 at 1% level of significance.

After getting the results of MH, items with small and nonsignificant Mantel-Haenszel statistics were further investigated to check whether they also had large item discrimination values and wide range of item difficulties, and had high factor loadings and small error variances in EFA. It was also studied whether the mean differences between groups in anchor items were in consistent with the mean differences in the entire set of items. It was not until the items satisfy these conditions to determine them as anchor items.

The anchor items determined through the process stated above were then specified in the IRTLRDIF program (Thissen, 2001). This program again uses the same algorithm given in Figure 3.3 with only a difference in calculating the index of " $\mathcal{LL}_{AllEqual}$ " value. The program calculates this value using all the items when no anchor item is specified. However, when anchor items are specified, program calculates a new " $\mathcal{LL}_{AllEqual}$ " value repeatedly for each studied item from a model fixing the anchor items' and the studied item's parameters between groups.

The effect of using anchor items was investigated by comparing the results from IRT-LR analysis using anchor items and IRT-LR analysis not using anchor items.

### 3.3.5 Agreement of the DIF Procedures

There are many results that may affect reliability of DIF indices. For example, as DIF statistics are difference measures, i.e. they analyze the difference between parameters from different groups in determining whether an item functions differentially between groups, this process may add possible error components from individual estimates up to the DIF statistics. In addition, DIF indices deal with

single item behaviors, not the more stable aggregation of a set of items (Camilli & Shepard, 1994).

On the other hand, beyond these errors, test characteristics such as range of item difficulties, population characteristics such as average difference in performance, choice of statistical model in DIF analyses, choice of computer algorithm, or the extent of DIF items in a test all affect the reliability or validity of DIF statistics (Camilli & Shepard, 1994).

To this reason it is a common practice in the scope of DIF analyses to use different methodologies, and to flag an item only when it shows DIF in all the methodologies used in the analysis. This current study also compared the results from the previously mentioned methodologies for not only determining the DIF items but also specifying the agreement rate of the methodologies in the context of DIF analyses of mathematics items.

In investigating the agreement of the results, agreement rates specifying the rate of the items showing either DIF or no DIF in both analyses were calculated. In addition to compare the methodologies in terms of being liberal or conservative in determining DIF items, the percentages indicating the items determined as showing DIF and not showing DIF within each methodology were also calculated.

### 3.3.6 Disentangling Sources of DIF

Through the analysis of the studies in the literature that try to identify the sources of DIF, the following dimensions given in Table 3.5 were specified. These dimensions can be classified under three headings: 1) Curricular Differences, 2) Cultural Differences, and 3) Translation Fidelity.

Table 3.5 Possible Sources of Group Performance Differences

<u>COGNITIVE EXPECTATION</u>		<u>MATHEMATICAL CONTENT</u>		<u>VERBAL PROPERTIES OF ITEM</u>	
		<u>Curriculum Related</u>		<u>Item Contents</u>	
TIMSS	PISA	TIMSS	PISA		
Representing	Reproduction	Fractions and Number Sense	Space and Shape	Items With Figures	Difference in Cultural Relevance
Solving	Connection	Algebra	Change and Relationships	Realistic Problems	Changes in Format
Using more complex procedures	Reflection	Measurement	Uncertainty	Word Problems	Changes in Content
Recalling mathematical objects and properties		Geometry	Quantity	Data- Interpretation	Changes in Difficulty of Words
Performing routine procedures		Data Representation, Analysis, and Probability		Curriculum- Like Items	
Predicting				Easy Items	
				Difficult Items	
				Items With Variables	

### Curricular Differences

In order to investigate whether the DIF in an item was due to the cognitive expectations or curricular differences, individual cluster tables of DIF items with respect to cognitive expectations and curricular subjects were generated (Ercikan, 2002).

Clustering of DIF items in a content area, such as measurement, or in a cognitive expectation, such as reasoning, was interpreted as an evidence of association between DIF and the subject area, or cognitive expectation. It should also be added that, to provide an information rich case, in determining the cluster tables an item was flagged only when it showed DIF in at least two of the three methodologies used in the analysis.

In addition, how the items presented the tasks were also given a special importance. It was investigated whether the items favoring the same group also had a common property such as including mathematical notations rather than words to convey information, using graphs in the items as a source of information, or being curriculum-like items etc. (Harris & Carlton, 1993).

### Cultural Differences

To determine whether the verbal properties would account for the DIF in items, some subjective analyses were conducted. Two bilingual researchers, one of which was the author of this dissertation, experienced both in mathematics teaching and DIF analyses, reviewed the DIF items independently with respect to the dimensions specified in Table 3.5. Then they had a discussion and tried to come up with a consensus on the possible sources of DIF in the items. Common judges about the sources of DIF in each item were reported in the study.

Analyses to detect difference in cultural relevance investigated the cases where the item remained exactly the same in both groups but because of the differential relevance of the item content to culture, item was more familiar to one of the groups (Allalouf, Hambleton & Sireci, 1999).

Gierl and Khalig (2000), also describe this issue specifying that differences in words or structure of items that are inherent to the language are likely to cause differential performance among groups. Within this context, it was investigated whether the item contained words which have unfamiliar meanings for a group, and contain distracters which may be especially attractive to a group as well.

## Translation Fidelity

In general all adaptation problems can be investigated under this heading (Bontempo, 1993; Hulin, 1987). Specifically, Changes in format, identify the cases where the translation procedure makes the translated sentence much longer, change the punctuation, capitalization, or typeface (Allalouf, et al., 1999; Gierl & Khalig, 2000). Number of words, number of three-syllable words and number of sentences in an item was compared between source and target forms as well (Scheuneman & Grima, 1997). In the present study, a subjective decision of whether these differences provided a clue for correct answer was regarded as the source of DIF.

Change in content, on the other hand, deals with incorrect translations. Omission, or additions of words or phrases that affect the meaning, or inadequate adaptation of keywords were investigated in this category (Allalouf, et al., 1999; Gierl & Khalig, 2000; Ercikan, 2002). Use of quantitative language was also given a special importance in this category. Scheuneman and Grima (1997) define quantitative language as the words indicating the operations to be performed, such as add, average, arithmetic mean etc.

It was investigated whether the quantitative words provided similar stimuli in both groups. It was also controlled whether translation produced easier or more difficult words. For example commonness of a vocabulary in a given context may affect the difficulty of a word differentially among groups (Hambleton & Patsula, 2000).

In subjective analyses the DIF items were also reviewed from two additional dimensions not given in Table 3.5, as well. First, specifics of multicontext model were used to determine the cultural aspects of Turkish and American groups, with respect to the DIF items (Li, Cohen & Ibarra, 2004). Considering the properties of the DIF items, for example whether they were real-world problems, whether they required analytic or comprehensive thinking etc., cultural context of the countries were also tried to be determined. Individuals from low-context cultures are context independent. On the other hand, information without context is meaningless for individuals from high-context cultures.

In addition, analytical thinking, inductive reasoning, following directions, examining ideas rather than real-world applications are other characteristics of low-context cultures. On the other hand, deductive reasoning, being process oriented rather than task oriented, perceiving facts as complete units embedded in the context of situations are characteristics of high-context cultures.

Secondly, Steinberg's (2001) claim that differences in slope and threshold parameters of LRT-LR analysis indicate the meaning of the item, and the endorsement rates differ between groups, respectively, was also considered in addition to that of Jöreskog's, (2005) indicating that the difference in intercepts of MGFA means that, there is more difference in that item, between groups, than can be accounted by the difference between the latent construct.



## CHAPTER IV

### RESULTS

This chapter is divided into four sections. The first section reports characteristics of the selected mathematics items of TIMSS 1999 and PISA 2003 studies. Means, standard deviations and reliabilities are provided to present a picture of data structure. Results of construct equivalence analyses are presented in the second section. This section includes results from both exploratory and confirmatory analyses. Differential item functioning analyses' results, via Restricted Factor Analysis (RFA), Mantel Haenszel (M-H), and Item Response Theory Likelihood Ratio (IRT-LR), methods are given in the third section. Comparison of DIF results among different methodologies, the effect of using anchor items in IRT-LR, and purification of the matching criterion in M-H are also included in this section. Finally the fourth section presents the judgmental reviews of the items to determine the causes of DIF.

#### 4.1 Descriptive Summary

Items not responded although it was expected to be, non-reached items and items in which more than one alternative selected were recoded as incorrectly answered items. The percentages of the recoded items are given in Appendix A.

Table 4.1 gives the scale statistics for 7<sup>th</sup> booklet of TIMSS and 2<sup>nd</sup> booklet of PISA. The results indicate that, USA students performed better than Turkish students in both tests. In addition the difference between American and Turkish students in TIMSS is larger than that of PISA. Especially, the positively skewed distribution of TIMSS indicates Turkey has more scores than USA toward the lower end of the scale. Difficulties of each item (proportion corrects) are given in Appendix B.

It should also be mentioned that three items, m413q01, m438q01 and m505q01, coded 0 for all the cases of USA were excluded from the analyses.

Table 4.1 Scale statistics for Mathematics Tests of PISA and TIMSS

SCALE STATISTICS	TIMSS 1999 (39 Items)		PISA 2003 (33 Items)	
	TUR	USA	TUR	USA
# Examinees	980	1110	391	425
Mean	16.263	23.256	12.113	14.718
Std. Dev.	6.833	9.114	6.857	6.987
Skew	0.600	-0.153	0.546	0.122
Kurtosis	-0.305	-1.082	-0.251	-0.848
Alpha	0.842	0.920	0.891	0.885
Mean PC <sup>1</sup>	0.417	0.596	0.367	0.446
Mean Biserial <sup>2</sup>	0.499	0.647	0.658	0.621

## 4.2 Construct Equivalence

### 4.2.1 Exploratory Factor Analysis

Meritorious measures of sampling adequacy (over 0.90 in all groups) and significant Bartlett's test of sphericity statistics indicated that distribution of both TIMSS and PISA data were adequate for conducting factor analysis. In both groups mean communalities were about 0.50.

In Turkish TIMSS data, Principal Component Analysis (PCA) derived 11 components having eigenvalues greater than 1, which totally accounted for about 46% of the variance in the item data. The eigenvalue of the first component, 6.231, was about four times of the next, which may indicate unidimensionality.

<sup>1</sup> Mean of items' proportion correct indices

<sup>2</sup> Mean of item discriminations

In American TIMSS data, Principal Component Analysis (PCA) derived 7 components having eigenvalues greater than 1, which totally accounted for about 44% of the variance in the item data. The eigenvalue of the first component, 9.990, was about six times of the following, which may indicate unidimensionality (Zwick & Velicer, 1986).

Although these PCA analyses were similar in terms of variance accounted for and possible unidimensional structure, different number of components generated and some differences among the factor loadings would have been indicating different factor structures. For example, investigating the factor loadings given in Appendix C1, keeping in mind that the factors are not constrained to be in the same order across groups, it would be seen that most of the items have different loadings across groups. As it is only for illustration, only the first seven of the eleven components of Turkish data are given in the appendix.

In both Turkish and American PISA data, PCA derived 9 components having eigenvalues greater than 1. The cumulative variance accounted for by the nine components was about 53% and 52% for the Turkish and American data, respectively. In Turkish data, the eigenvalue of the first component, 7.759, was about four times of the following, and in American data, the eigenvalue of the first component, 7.480, was about five times of the next, both of which may again indicate unidimensionality.

However, investigating the factor loadings given in Appendix C2, it can be seen that there are some differences. For example, item m034q01t is under the fifth factor of the Turkish data with items m413q02, m474q01 and m520q03t, whereas these items are not all together in any of the American components.

To check whether the unidimensionality of the tests signed by the eigenvalues were justifiable, or differences seem to exist among the groups were large enough to warrant different factor structures, confirmatory factor analysis (CFA) was conducted.

#### 4.2.2 Confirmatory Factor Analysis

At the first step of CFA analyses, the unidimensional structure of the tests signed by the EFA analyses was tested. Tetrachoric (a special form of polychoric) correlation matrices and their asymptotic covariance matrices were calculated individually for each group through PRELIS program. Then, using this estimated matrix of polychoric correlations and corresponding asymptotic covariance matrix, selected booklets were tested through LISREL program, to check whether a single latent trait accounts for the variation among items. Weighted Least Squares (WLS) estimation method was used in the analyses. NC, GFI, AGFI, RMSEA, NFI, NNFI and SRMR indices were used in checking model-data fits.

The unidimensional model offered for Turkish and American groups in PISA and TIMSS data did not fit in any of the cases except American TIMSS data, which mean that differences that seem to exist in EFA were large enough to reject a unidimensional model in both groups. However, as the methods used in identifying items functioning differentially among groups require unidimensional data structures, subgroups of unidimensional items were selected in both groups.

To this reason EFA was conducted for TIMSS and PISA tests separately on the total sample by pooling the data from Turkish and American groups into one data file. As the dimensionality of the items emerges from the interaction between the abilities that items can detect and abilities that individuals possess, pooling the data is required in determining a common structure for both groups (Ackerman, 1992). Items selected for the further analyses that were in the first dimension of the rotated matrices from the pooled data are given in Appendix D.

Items selected through the process defined above were then tested individually through CFA, to check whether they were unidimensional. Goodness-of-fit indices are given in Table 4.2.

Table 4.2 Goodness-of-fit statistics for selected items.

STATISTICS	TIMSS		PISA	
	TUR (df = 189)	USA (df = 189)	TUR (df = 209)	USA (df = 209)
NC	2.7	2.5	2.7	1.85
GFI	0.98	0.99	0.96	0.97
AGFI	0.98	0.98	0.95	0.96
RMSEA	0.042	0.037	0.067	0.045
NNFI	0.96	0.97	0.92	0.95
CFI	0.96	0.97	0.93	0.95
RMR	0.11	0.15	0.28	0.22

All the indices but the RMR points a reasonable fit. As most of the indices provide an evidence of model data fit, it was decided to carry on the study despite the high RMR magnitudes. Estimated factor loadings and the error variances of the items selected to be unidimensional, can be seen in the path diagrams given Appendix E. Although the model fits the data in all the individual cases, item m022144 in Turkish TIMSS data had a low value of factor loading and high value of measurement error. But it was still kept in the analyses to be further investigated. In addition, item m124q03t in Turkish PISA data had a negative error variance.

Table 4.3 gives the scale statistics for the selected items of 7<sup>th</sup> booklet of TIMSS and 2<sup>nd</sup> booklet of PISA. The results indicate that, the selected items also resemble the performance difference between USA and Turkish students as it was in the whole tests. Although the selection procedure dropped the alpha levels to an extent, reliabilities were still at a reasonable level.

Table 4.3 Scale statistics for selected items of PISA and TIMSS

SCALE STATISTICS	TIMSS 1999 (21 Items)		PISA 2003 (22 Items)	
	TUR	USA	TUR	USA
# Examinees	980	1110	391	425
Mean	8.181	13.946	8.391	11.016
Std. Dev.	4.124	5.203	5.347	5.516
Skew	0.439	-0.487	0.414	-0.043
Kurtosis	-0.498	-0.865	-0.679	-1.016
Alpha	0.775	0.882	0.876	0.876
Mean PC <sup>3</sup>	0.390	0.664	0.381	0.501
Mean Biserial <sup>4</sup>	0.564	0.725	0.692	0.678

Selecting items assuring unidimensionality in all individual groups, whether the selected items did measure the same latent construct in Turkish and American groups were investigated in the second step of CFA. To this reason Multiple Group analyses through LISREL were conducted.

Multiple group analyses were conducted in two steps. In the first step, mean vector, the covariance matrix, and the asymptotic covariance matrix of the underlying variables were computed for each country. In these calculations, to specify a common metric, thresholds of the underlying variables in all individual analyses were fixed at the pre-calculated thresholds values of the pooled American and Turkish data in one data file.

The estimated means of the underlying variables are given in Table 4.4. The standard deviations of all the variables were fixed as 1.

---

<sup>3</sup> Mean of items' proportion correct indices

<sup>4</sup> Mean of item discriminations

Table 4.4. Estimated means of the underlying variables of the items

ITEM	PISA 2003 2 <sup>nd</sup> booklet (22 Items)		ITEM	TIMSS 1999 7 <sup>th</sup> booklet (21 Items)	
	TUR n=391	USA n=425		TUR n=980	USA n= 1110
m034q01t	-0.109	0.094	m012001	-0.282	0.243
m124q01	0.043	-0.040	m012002	-0.225	0.214
m124q03t	-0.152	0.135	m012003	-0.088	0.078
m145q01t	-0.069	0.065	m012007	-0.407	0.381
m150q01	-0.063	0.058	m012009	-0.229	0.208
m150q02t	-0.269	0.294	m012010	-0.822	0.488
m150q03t	-0.401	0.342	m012011	-0.360	0.313
m192q01t	-0.067	0.059	m012012	-0.430	0.411
m411q01	-0.333	0.274	m012021	-0.456	0.430
m411q02	-0.279	0.239	m012024	-0.272	0.274
m413q02	-0.216	0.209	m012043	-0.376	0.366
m413q03t	-0.342	0.261	m012044	-0.468	0.537
m438q02	-0.189	0.166	m012045	-0.391	0.534
m462q01t	0.128	-0.132	m012048	-0.510	0.619
m474q01	-0.220	0.212	m022135	-0.517	0.355
m520q01t	-0.160	0.151	m022144	-0.533	0.461
m520q02	-0.282	0.242	m022148	-0.380	0.311
m520q03t	-0.293	0.248	m022253	-0.224	0.194
m547q01t	0.032	-0.030	m022237	-1.351	0.592
m555q02t	-0.161	0.150	m022262a	-0.390	0.372
m702q01	-0.306	0.238	m022262b	-0.396	0.339
m806q01t	-0.075	0.069			

Table 4.4 shows that there are considerable differences in the means between countries. These estimated means of the variables must be provided to the LISREL program in order estimate the intercept values; otherwise all intercepts would be fixed to be zero.

In the second step of CFA analyses, multiple group factor analyses (MGFA) were carried on to check whether the items detected through the first step, measure the same latent variables in all countries. Invariance of factor loadings and intercepts were investigated further.

In this step, American and Turkish groups were compared starting with the strict invariance model forcing all parameters but the error variances to be reproduced in both countries. Syntax used in TIMSS data to this purpose is given in Appendix F. The syntax used in PISA analyses was also similar.

Fitting the model of strict invariance, in the sense of equal intercepts and equal factor loadings, to the USA and Turkey gave a chi-square of 3407.28 with 418 degrees of freedom and a RMSEA of 0.083 for TIMSS and a chi-square of 1434.32 with 460 degrees of freedom and a RMSEA of 0.072 for PISA. They both indicated that the models did not fit well.

Investigating the largest modification indices in TIMSS and PISA studies individually, the parameters of intercept or factor loading to be freely estimated were detected. AMI's were calculated in determining statistical significance of MI's. Then the new model allowing the parameter with the largest significant MI to be free in each country was tested. This process was carried on until no significant MI's were produced. Seven successive analyses for PISA and sixteen successive analyses for TIMSS were conducted. Results are given in Table 4.5 and Table 4.6. Tables also show the parameters allowed to be freely estimated at each step.

Table 4.5. Step-by-step Detection of Factorial Invariance for PISA

STEP #	$\chi^2$	df	ITEM	FREE PAR.	MI	AMI*
1	1434.32	460	m124q01	Intercept	46,90	15,52
2	1230,36	459	m462q01t	Intercept	30,80	11,76
3	1233,49	458	m474q01	Loading	32,70	12,45
4	1205,58	457	m145q01t	Loading	20,90	8,04
5	1186,56	456	m555q02t	Loading	22,00	8,60
6	1165,83	455	m520q02	Loading	21,60	8,57
7	1138,91	454	m150q03t	Loading	20,90	8,47

\* Significant at 1% level



Table 4.6. Step-by-step Detection of Factorial Invariance for TIMSS

STEP #	$\chi^2$	df	ITEM	FREE PAR.	MI	AMI*
1	3407,28	418	m022237	Intercept	316	42,63
2	2410,58	417	m022144	Loading	163,8	30,33
3	2108,94	416	m022262a	Intercept	131,4	27,58
4	2046,78	415	m022262b	Loading	82,3	17,34
5	1928,03	414	m012003	Intercept	73,5	16,37
6	1786,86	413	m022262a	Loading	88,4	21,44
7	1692,35	412	m022253	Intercept	76,1	19,35
8	1538,71	411	m012048	Intercept	51,7	14,25
9	1456,96	410	m012010	Intercept	36	10,36
10	1387,39	409	m012001	Intercept	29	8,71
11	1327,76	408	m022253	Loading	27,8	8,70
12	1289,81	407	m022148	Loading	25,3	8,12
13	1233,94	406	m012012	Loading	30,3	10,20
14	1186,28	405	m012002	Intercept	25,3	8,80
15	1164,62	404	m012024	Loading	22,9	8,08
16	1110,44	403	m012009	Intercept	18,9	6,96

The final multigroup models, allowing only the intercepts and loadings mentioned in the tables 4.5 and 4.6, to be different in the two countries, gave a chi-square of 1094.61 with 402 degrees of freedom and a RMSEA of 0.041 for TIMSS and a chi-square of 1112.72 with 453 degrees of freedom and a RMSEA of 0.060 for PISA. Despite the statistically significant chi-square values, Normed Chi-Square (NC) magnitudes of 2.72 and 2.46 for TIMSS and PISA, respectively, in addition to the reasonable RMSEA values, were considered to provide enough evidence for model-data-fit (Jöreskog, 2005)

---

\* Significant at 1% level

The result concerning the distribution of the latent variables is summarized in Table 4.7. Results are from the final MGFA models. Table gives the estimated means, and variances with their standard errors in parenthesis.

Table 4.7 USA vs Turkey Estimated Means and Variances

	TURKEY		USA	
	Mean	Variance	Mean	Variance
TIMSS	0.00	0.32 (0.02)	0.92 (0.03)	0.66 (0.03)
PISA	0.00	0.63 (0.03)	0.36 (0.01)	0.55 (0.03)

Dividing the estimations by corresponding standard errors gives t-values, which indicate that there is a significant mean difference between countries in both of the studies. In addition, variances indicated that Turkish students are more homogenous with regard to mathematics achievement, construct measured by TIMSS, whereas the situation is slightly reversed with regard to mathematical literacy, construct measured by PISA.

Estimations of the intercepts, factor loadings and error variances in the final models of PISA and TIMSS study are given in Appendix G.

### 4.3 Differential Item Functioning

This section includes results from differential item functioning analyses of the selected items in the previous section. Restricted Factor Analysis (RFA), M-H and Item Response Theory Likelihood Ratio (IRT-LR) methods were used in the DIF analyses. In addition, comparison of DIF results from different methodologies, the effect of using anchor items in IRT-LR, effect of purification in M-H are included in this section.

### 4.3.1 Restricted Factor Analysis

At the very first step of the RFA, American and Turkish data were pooled in a single data file before estimating polychoric correlations and their corresponding asymptotic covariance matrices.

In the LISREL step, main constructs PISA and TIMSS measures were named as “MathLit” indicating mathematics literacy as defined within the context of PISA and “MathAch”, indicating for mathematics achievement as defined within the context of TIMSS (OECD, 2005; Gonzalez & Miles, 1999). The PRELIS and LISREL syntaxes used to test the Null Model for the TIMSS data is given in Appendix H. The syntax used for PISA data was also similar with small modifications such as the name of the variables. Related part of the output is given in Table 4.8.

Table 4.8 Modification Indices and Expected Parameter Changes for PISA and TIMSS Items

PISA (N=816, Chi <sup>2</sup> =681, df=230)			TIMSS (N=2090, Chi <sup>2</sup> =1037, df=209)		
ITEM	MI	EPC	ITEM	MI	EPC
m124q01	15.5	-0.18	m012001	19.6	-0.76
m150q01	49.0	-0.35	m012003	85.2	-1.79
m150q03t	18.6	0.24	m012009	8.9	0.58
m192q01t	11.9	-0.20	m012010	8.1	0.46
m411q01	36.3	0.32	m012045	11.2	0.78
m411q02	26.1	0.29	m012048	19.4	0.82
m413q03t	18.6	0.24	m022144	19.3	0.83
m462q01t	88.6	-0.56	m022253	32.0	-0.94
m520q03t	12.5	0.21	m022237	40.1	1.08
m547q01t	40.2	-0.39			

To investigate the items functioning differentially between groups, the Null Model of no bias was fitted for each of the PISA and TIMSS data. Table 4.8 gives the resulting MI's and EPC's, offering to freely estimate the corresponding factor loading of the item on "Group" variable in PISA and TIMSS data.

Although it was possible to remove all the items with significant MI's at a time, this method is reported to produce slightly biased results such as failing to detect some DIF items or wrong detections of DIF items. For this reason step by step method was used in the analyses. That is, an item with largest AMI's was removed from the test and the Null Model was fitted on the reduced data set for the subsequent analysis. This process was repeated until all DIF items were removed from the tests. Table 4.9a and 4.9b give results from the step-by-step detection of items functioning differentially between American and Turkish groups in PISA and TIMSS, respectively.

Table 4.9a Step-by-step Detection of DIF in PISA

STEP #	# Items in Test	Chi Sq	df	ITEM	MI	AMI*	EPC
Step 0	22	681.42	230	m462q01t	88.6	34.23	-0.56
Step 1	21	554.88	209	m150q01	35.5	14.22	-0.27
Step 2	20	495.58	189	m124q01	34.9	14.24	-0.27
Step 3	19	406.46	170	m520q03t	25.5	11.31	0.28
Step 4	18	344.98	152	m150q03t	21.4	9.99	0.23
Step 5	17	266.06	135	m547q01t	22.1	12.14	-0.26
Step 6	16	229.06	119	m413q03t	14.7	8.09	0.20

\* Significant magnitudes at 1%

Table 4.9b Step-by-step Detection of DIF in TIMSS

STEP #	# Items in Test	Chi Sq	df	ITEM	MI	AMI*	EPC
Step 0	21	1037.63	209	m012003	85.2	18.61	-1.79
Step 1	20	911.73	189	m022253	43.2	9.35	-0.82
Step 2	19	777.31	170	m022237	47.4	10.97	0.81
Step 3	18	700.75	152	m012001	45.3	10.44	-0.52
Step 4	17	630.62	135	m022144	44.4	10.15	0.55

When looking at the tables, it is seen that RFA detected 7 items in PISA and 5 items in TIMSS functioning differentially between groups. In addition, positive EPC magnitudes indicate that 3 of 7 items in PISA and 2 of 5 items in TIMSS are more attractive to American students, whereas the rest of the items favor Turkish students.

#### 4.3.2 Item Response Theory Likelihood Ratio Analysis

IRTLRDIF program was used in conducting likelihood ratio (LR) tests. The program first evaluated whether there were overall differences in item parameters between USA and Turkish groups in TIMSS and PISA study. Upon significant difference between item parameters, the source of the differences were further investigated, i.e. it was investigated whether the source of the differences lies in guessing, slope or threshold parameters, respectively.

Table 4.10a and 4.10b give the results of the analyses for PISA and TIMSS indicating the hypothesis rejected. Benjamini and Hochberg (B-H) (1995) procedure was used to control the Type I error rate (Thissen, Steinberg & Kuang, 2002). The program outputs containing  $G^2$  values, parameter estimates, mean and standard deviation of focal group estimates are given in Appendix I1 and I2.

---

\* Significant magnitudes at 1%

In conducting the analyses, 3-parameter model for multiple choice items and 2-parameter model for coded response items were used in estimating the item parameters and loglikelihood magnitudes.

It should also be added that, only the items having at least one significant result are given in the tables 4.10a and 4.10b. In calculating the adjusted critical p-values for hypothesis tested, number of comparisons (or family size) for TIMSS and PISA was taken as 84 and 72, respectively. These were the all-possible number of hypotheses that can be tested within each analysis (Steinberg, 2001).

Table 4.10a Items Functioning Differentially in PISA. IRT-LR Results

ITEM	HYPOTHESES TESTED*			
	All Equal	c-Equal	a-Equal	b-Equal
m124q01	*	NA <sup>5</sup>		*
m145q01t				*
m150q01		NA		*
m150q02t	*	NA		*
m150q03t	*	NA		*
m411q01	*	NA		*
m411q02	*		*	
m413q03t	*	NA		*
m462q01t	*	NA		*
m520q02				*
m520q03t	*	NA		*
m547q01t	*	NA		*

\* Significant at 1% level according to B-H critical values

<sup>5</sup> NA (not applicable): These items were scaled with two-parameter model.

Table 4.10b Items Functioning Differentially in TIMSS. IRT-LR Results

ITEM	HYPOTHESES TESTED*			
	All Equal	c-Equal	a-Equal	b-Equal
m012001	*			*
m012002	*		*	*
m012003	*			*
m012009	*			*
m012010	*		*	
m012024	*		*	
m012044	*	*		
m012045				*
m012048	*	*		*
m022144	*		*	*
m022148	*	NA		*
m022253	*	NA	*	*
m022237	*	NA		*
m022262a		NA		*
m022262b	*	NA	*	*

In addition, as can also be seen in the Appendices I1 and I2, in each hypothesis tested the program estimated the mean and standard deviation of Turkey about  $-0.55$  and  $1.08$ , respectively, in PISA, and  $-1.21$  and  $0.81$ , respectively, in TIMSS while the mean and standard deviation of USA were fixed at 0 and 1 by default.

---

\* Significant at 1% level according to B-H critical values

### 4.3.3 Mantel-Haenszel and Anchor Items

EZDIF program was used to conduct DIF analyses using the Mantel-Haenszel procedure. To determine five score levels, Turkish and American data were pooled into a common file and then 20<sup>th</sup>, 40<sup>th</sup>, 60<sup>th</sup> and 80<sup>th</sup> percentiles of the total score was calculated, individually for TIMSS and PISA studies. These score levels were 0-4, 5-8, 9-11, 12-15, and 16-22 for PISA and 0-6, 7-9, 10-13, 14-17, and 18-21 for TIMSS. Introducing these levels to the EZDIF program, the odds ratio  $\alpha$ , MH D-DIF statistics, and effect size coding indicating negligible (A), moderate (B), or large (C) DIF were calculated by the program.

Table 4.11 gives the results of the MH analyses for PISA and TIMSS. These results were provided after a purification process conducted automatically by EZDIF program. To get purified results, DIF items detected at the first run were excluded from the matching criteria at the second step. Results of first step of EZDIF program are also given in Appendix I3. Items detected as significantly functioning differentially between Turkish and American groups with respect to the adjusted critical values calculated according to equation 5 at 1% level are given in bold in Table 4.11.

In the table it can be seen that, 7 PISA and 14 TIMSS items show DIF with respect to the significance test. However, level of A indicates that 2 of the 14 items show only negligible DIF in TIMSS.

According to the MH results, items showing a not significant A level DIF were further investigated to determine if they could work as anchor items in IRT-LR analysis. To this reason item difficulties, that is item proportion corrects, item discriminations, that is item total biserial correlations, and factor loadings of the items given in Appendix E, were investigated. These stated values are given in Table 4.12a and 4.12b for PISA and TIMSS, respectively.



Table 4.11 EZDIF Program Output of DIF Analysis

ITEM*	PISA 2003 2 <sup>nd</sup> booklet (22 Items)			TIMSS 1999 7 <sup>th</sup> booklet (21 Items)			
	DIF	$\alpha$	MH D-DIF	ITEM	DIF	$\alpha$	MH D-DIF
m034q01t	A	0,764	0,633	<b>m012001</b>	<b>A</b>	0,668	0,947
<b>m124q01</b>	CF	0,297	2,853	m012002	A	0,81	0,495
m124q03t	A	0,752	0,668	<b>m012003</b>	<b>CF</b>	0,436	1,952
m145q01t	B	0,583	1,267	<b>m012007</b>	<b>B</b>	1,648	-1,175
m150q01	B	0,589	1,244	m012009	A	0,8	0,526
<b>m150q02t</b>	B	1,797	-1,377	<b>m012010</b>	<b>CR</b>	3,527	-2,962
<b>m150q03t</b>	CR	2,328	-1,986	m012011	A	1,234	-0,494
m192q01t	B	0,643	1,039	<b>m012012</b>	<b>A</b>	1,498	-0,95
<b>m411q01</b>	B	1,806	-1,389	<b>m012021</b>	<b>B</b>	1,591	-1,092
m411q02	B	1,583	-1,08	m012024	A	1,198	-0,424
m413q02	A	1,12	-0,265	<b>m012043</b>	<b>B</b>	1,682	-1,221
m413q03t	B	1,752	-1,317	<b>m012044</b>	<b>CR</b>	2,383	-2,041
m438q02	A	1,059	-0,135	<b>m012045</b>	<b>CR</b>	2,536	-2,187
<b>m462q01t</b>	CF	0,243	3,321	<b>m012048</b>	<b>CR</b>	3,202	-2,735
m474q01	A	1,376	-0,751	<b>m022135</b>	<b>B</b>	1,603	-1,109
m520q01t	A	0,884	0,29	<b>m022144</b>	<b>CR</b>	2,467	-2,122
m520q02	B	1,539	-1,013	m022148	A	0,878	0,305
<b>m520q03t</b>	B	1,665	-1,198	<b>m022253</b>	<b>B</b>	0,512	1,571
<b>m547q01t</b>	B	0,479	1,732	<b>m022237</b>	<b>CR</b>	14,895	-6,347
m555q02t	A	0,892	0,269	m022262a	A	1,12	-0,266
m702q01	A	1,514	-0,975	m022262b	A	0,967	0,078
m806q01t	A	0,757	0,655				

Although the two items, m012001 and m012012, in TIMSS showed significant DIF, the effect size index of A shows that these were negligible differences. But still these items were not used as an anchor item in IRTLR analyses.

\* Bold items are showing DIF at 0.01

Table 4.12a Parameters of the Anchor Items in PISA

ITEMS	GROUP	DIFFIC.	DISCR.	LOADING	ERROR
m034q01t	TUR	0.230	0.711	0.84	0.29
	USA	0.296	0.604	0.65	0.58
m124q03t	TUR	0.335	0.819	1	0
	USA	0.445	0.832	0.93	0.14
m413q02	TUR	0.504	0.845	0.94	0.12
	USA	0.668	0.717	0.86	0.26
m438q02	TUR	0.327	0.631	0.79	0.38
	USA	0.464	0.688	0.74	0.45
m474q01	TUR	0.494	0.604	0.78	0.39
	USA	0.661	0.428	0.49	0.76
m520q01t	TUR	0.506	0.728	0.88	0.23
	USA	0.628	0.759	0.90	0.18
m555q02t	TUR	0.460	0.714	0.77	0.41
	USA	0.584	0.736	0.88	0.23
m702q01	TUR	0.182	0.776	0.90	0.18
	USA	0.358	0.754	0.84	0.30
m806q01t	TUR	0.527	0.574	0.69	0.52
	USA	0.584	0.592	0.68	0.54

The magnitudes in the Table 4.12a indicate that, items with a small and nonsignificant Mantel-Haenszel statistics also have a large item discrimination values and a wide range of item difficulties. In addition they have relatively high factor loadings. On the other hand, item m474q01 have a high error variance, but as it has a reasonable factor loading and discrimination value, it also was included as an anchor item.

Table 4.12b Parameters of the Anchor Items in TIMSS

ITEMS	GROUP	DIFFIC.	DISCR.	LOADING	ERROR
m012002	TUR	0.553	0.474	0.43	0.82
	USA	0.716	0.718	0.72	0.48
m012009	TUR	0.471	0.530	0.48	0.77
	USA	0.642	0.618	0.63	0.61
m012011	TUR	0.347	0.538	0.48	0.77
	USA	0.610	0.589	0.58	0.66
m012024	TUR	0.583	0.283	0.67	0.55
	USA	0.775	0.686	0.68	0.53
m022148	TUR	0.282	0.755	0.77	0.41
	USA	0.545	0.779	0.80	0.36
m022262a	TUR	0.424	0.686	0.95	0.10
	USA	0.716	0.822	0.99	0.03
m022262b	TUR	0.323	0.744	0.99	0.02
	USA	0.609	0.758	0.95	0.10

The magnitudes in the Table 4.12b indicate that, items with a small and nonsignificant Mantel-Haenszel statistics also have a large item discrimination values (except item m012024 in Turkish group) and a wide range of item difficulties. In addition they have relatively high factor loadings. Despite the low discrimination value of m012024, because of the reasonable factor loading this item was also included in the IR-TLR analyses with anchor items.

The mean differences between groups in anchor items were also investigated and compared with the mean differences in the entire set.

For PISA, the average score on the nine-item anchor test for all students combined was 4.149 (SD = 2.565); the means for Turkish and American students were 3.565 and 4.687, respectively. This means that American students scored 0.437 SDs above Turkish students on the nine-item anchor test. This difference was consistent with the mean American-Turkish difference on the entire 22-item test, in which American students scored 0.475 SDs above the Turkish students as can be seen from Table 4.3.

Additionally, reliability analysis produced an alpha of 0.77 for the 9-item anchor test for all students combined.

For TIMSS in the same manner, the average score on the seven-item anchor test for all students combined was 3.849 ( $SD = 2.074$ ); the means for Turkish and American student were 2.983 and 4.613, respectively. This means that American students scored 0.79 SDs above Turkish students on the seven-item anchor test. Although this difference was slightly less than the mean American-Turkish difference on the entire 21-item test, in which American students scored 1.04 SDs above the Turkish students, it was decided to be acceptable yet. Additionally, reliability analysis produced an alpha of 0.71 for the 7-item anchor test for all students combined.

It was decided that all these statistics provided a basis to use the specified items as anchors between USA and Turkey. To this reason, items in tables 4.12a and 4.12b were specified as anchor items in the program IRTLRDIF (Thissen, 2001). Then, the rest of the items, which are called candidate items, were investigated against DIF. In conducting the analyses, 3-parameter model for multiple choice items and 2-parameter model for coded response items were used in estimating the item parameters and loglikelihood magnitudes.

Among the candidate items, those having at least one significant result are given in the Table 4.13a and Table 4.13b for PISA and TIMSS, respectively. Benjamini-Hochberg (B-H) procedure was used in determining the significance of DIF levels.

Table 4.13a. Items Showing DIF in PISA. Anchored IRT-LR Results

ITEM	HYPOTHESES TESTED <sup>*</sup>			
	All Equal	c-Equal	a-Equal	b-Equal
m124q01	*	NA <sup>6</sup>		*
m145q01t				*
m150q01		NA		*
m150q03t	*	NA		*
m411q01		NA		*
m411q02	*		*	
m413q03t	*	NA		*
m462q01t	*	NA		*
m520q03t	*	NA		*
m547q01t	*	NA		*

According to the results, items m145q01t, m150q01, and m411q01 of PISA show b-DIF although they do not show an overall DIF. As this was an unexpected situation these items were treated as DIF-free items.

<sup>\*</sup> Significant at 1% level according to B-H critical values

<sup>6</sup> NA (not applicable): These items were scaled with two-parameter model.

Table 4.13b. Items Showing DIF in TIMSS. Anchored IRT-LR Results

ITEM	HYPOTHESES TESTED*			
	All Equal	c-Equal	a-Equal	b-Equal
m012003	*			*
m012007	*	*		*
m012010	*	*		*
m012012	*		*	*
m012021	*	*		*
m012043	*	*		
m012044	*	*		*
m012045	*	*		*
m012048	*	*	*	*
m022135	*	*		
m022144	*	*		
m022237	*	NA		*

IRT\_LR analyses investigates c, a, and b DIF in a hierarchical order. For example testing an a-DIF is based on the assumption that lower asymptote parameter c is equal between groups. To this reason, for an item only the very first significant result was considered, such as determining the item m012048 showing only c-DIF despite the significant a-DIF and b-DIF statistics.

---

\* Significant at 1% level according to B-H critical values

#### 4.3.4 Comparison of the Results of DIF Analyses

Table 4.14 combined the results for each of the 22 mathematics items of PISA from the procedures specified in previous sections. In comparing the item level analyses, RFA, M-H, and IRT-LR, it can be seen that among 22 items, 9 items were not flagged by all the three procedures and 6 items were flagged by all three procedures. MGFA is a construct level analysis; its results were used in interpreting the possible causes of DIF. In addition the effect of using anchor items was also discussed further in the current study.

In the same manner Table 4.15 presents the combined results for each of the 21 mathematics items of TIMSS from the procedures specified in previous sections. In comparison with the results from PISA, it can be seen that TIMSS has relatively high number of flagged items. The table indicates that, among 21 items, only 1 item was not flagged by all three procedures and 5 items were flagged by all the three procedures.

Table 4.14 Results of DIF Procedures in PISA

ITEMS	MGFA <sup>a</sup>	RFA <sup>b</sup>	M-H <sup>c</sup>	IRT-LR <sup>d</sup>	IRT-LR-Anchor
m034q01t	---	---	---	---	Anchor
m124q01	Intercept	*	CF	b-DIF	b-DIF
m124q03t	---	---	---	---	Anchor
m145q01t	Loading	---	BF	---	---
m150q01	---	*	BF	---	---
m150q02t	---	---	BR	b-DIF	---
m150q03t	Loading	*	CR	b-DIF	b-DIF
m192q01t	---	---	BF	---	---
m411q01	---	---	BR	b-DIF	
m411q02	---	---	BR	a-DIF	a-DIF
m413q02	---	---	---	---	Anchor
m413q03t	---	*	BR	b-DIF	b-DIF
m438q02	---	---	---	---	Anchor
m462q01t	Intercept	*	CF	b-DIF	b-DIF
m474q01	Loading	---	---	---	Anchor
m520q01t	---	---	---	---	Anchor
m520q02	Loading	---	BR	---	---
m520q03t	---	*	BR	b-DIF	b-DIF
m547q01t	---	*	BF	b-DIF	b-DIF
m555q02t	Loading	---	---	---	Anchor
m702q01	---	---	---	---	Anchor
m806q01t	---	---	---	---	Anchor

<sup>a</sup> MGFA: Multiple Group Factor Analysis, <sup>b</sup>RFA: Restricted Factor Analysis, <sup>c</sup>M-H: Mantel-Haenszel, <sup>d</sup>IRT-LR: Item Response Theory Likelihood Ratio Analysis



Table 4.15 Results of DIF Procedures in TIMSS

ITEMS	MGFA <sup>a</sup>	RFA <sup>b</sup>	M-H <sup>c</sup>	IRT-LR <sup>d</sup>	IRT-LR-Anchor
m012001	Intercept	*	AF	b-DIF	---
m012002	Intercept	---	---	a-DIF	Anchor
m012003	Intercept	*	CF	b-DIF	b-DIF
m012007	---	---	BR	---	c-DIF
m012009	Intercept	---	---	b-DIF	Anchor
m012010	Intercept	---	CR	a-DIF	c-DIF
m012011	---	---	---	---	Anchor
m012012	Loading	---	AR	---	a-DIF
m012021	---	---	BR	---	c-DIF
m012024	Loading	---	---	a-DIF	Anchor
m012043	---	---	BR	---	c-DIF
m012044	---	---	CR	c-DIF	c-DIF
m012045	---	---	CR	b-DIF	c-DIF
m012048	Intercept	---	CR	c-DIF	c-DIF
m022135	---	---	BR	---	c-DIF
m022144	Loading	*	CR	a-DIF	c-DIF
m022148	Loading	---	---	b-DIF	Anchor
m022253	Both	*	BF	a-DIF	---
m022237	Intercept	*	CR	b-DIF	b-DIF
m022262a	Both	---	---	b-DIF	Anchor
m022262b	Loading	---	---	a-DIF	Anchor

<sup>a</sup> MGFA: Multiple Group Factor Analysis, <sup>b</sup>RFA: Restricted Factor Analysis, <sup>c</sup>M-H: Mantel-Haenszel, <sup>d</sup>IRT-LR: Item Response Theory Likelihood Ratio Analysis

In order to examine the consistency between any of the three DIF procedures, the percentage of agreements, i.e. the rate of the items showing either DIF or no DIF in both analyses, among the three procedures were also computed. In PISA, agreement rate between the methods RFA and M-H, RFA and IRT-LR, and IRT-LR and M-H were 73%, 82%, and 82% respectively. In TIMSS, agreement rate between the methods RFA and M-H, RFA and IRT-LR, and IRT-LR and M-H were 57%, 52%, and 48% respectively. It is interesting to note that the agreement rates drop seriously in TIMSS with respect to PISA.

On the other hand, to determine the details of the agreement rates specified above, Table 4.16s and Table 4.17s are given.

Table 4.16a Agreement Between RFA and M-H Procedures in PISA

RESULTS FROM RFA	RESULTS FROM M-H		TOTAL
	# NON-DIF ITEMS	# DIF ITEMS	
# NON-DIF ITEMS	9	6	15
# DIF ITEMS	0	7	7
TOTAL	9	13	22

Table 4.16b Agreement Between RFA and IRT-LR Procedures in PISA

RESULTS FROM RFA	RESULTS FROM IRT-LR		TOTAL
	# NON-DIF ITEMS	# DIF ITEMS	
# NON-DIF ITEMS	12	3	15
# DIF ITEMS	1	6	7
TOTAL	13	9	22

Table 4.16c Agreement Between M-H and IRT-LR Procedures in PISA

RESULTS FROM M-H	RESULTS FROM IRT-LR		TOTAL
	# NON-DIF ITEMS	# DIF ITEMS	
# NON-DIF ITEMS	9	0	9
# DIF ITEMS	4	9	13
TOTAL	13	9	22

Table 4.17a Agreement Between RFA and M-H Procedures in TIMSS

RESULTS FROM RFA	RESULTS FROM M-H		TOTAL
	# NON-DIF ITEMS	# DIF ITEMS	
# NON-DIF ITEMS	7	9	16
# DIF ITEMS	0	5	5
TOTAL	7	14	21

Table 4.17b Agreement Between RFA and IRT-LR Procedures in TIMSS

RESULTS FROM RFA	RESULTS FROM IRT-LR		TOTAL
	# NON-DIF ITEMS	# DIF ITEMS	
# NON-DIF ITEMS	6	10	16
# DIF ITEMS	0	5	5
TOTAL	6	15	21

Table 4.17c Agreement Between M-H and IRT-LR Procedures in TIMSS

RESULTS FROM M-H	RESULTS FROM IRT-LR		TOTAL
	# NON-DIF ITEMS	# DIF ITEMS	
# NON-DIF ITEMS	1	6	7
# DIF ITEMS	5	9	14
TOTAL	6	15	21

#### 4.4 Sources of DIF

In Table 4.18, relative distribution of DIF items by subject area was examined to search evidence supporting curricular differences as explanation for DIF. The table gives the number of items favoring the corresponding countries in different content areas. Only items showing DIF in at least two of the three DIF procedures were considered. The area of Geometry in TIMSS is not included in the table, because there was only one geometry item.

Table 4.18 The Relative Distribution of DIF Items by Subject Area

PISA			TIMSS		
ITEM	USA	TUR	ITEM	USA	TUR
Space and Shape (5 items)	---	2	Fractions and Number Sense (9 items)	5	1
Change and Relationships (6 items)	2	2	Algebra (6 items)	1	1
Uncertainty (3 items)	1	---	Measurement (2 items)	---	1
Quantity (8 items)	3	---	Data Representation, Analysis, and Probability (3 items)	---	---

In the same manner, relative distribution of TIMSS items by cognitive expectations specified in the TIMSS publications is given in Table 4.19. The expectations of Recall and Predicting in TIMSS were not included in the table, because there was only one item at each of these levels. Additionally, a same table for PISA is not given because information about the cognitive expectations for 3 of 10 DIF items were not specified in PISA publications. However, it is worth adding that among 4 DIF items requiring reproduction, 3 were favoring Turkey. In addition, it is worth adding that items in Reproduction level are relatively easy items.

Table 4.19 The Relative Distribution of DIF Items by Cognitive Expectations

TIMSS		
ITEM	USA	TUR
Representing (5 items)	2	1
Solving (2 items)	---	---
Using more complex procedures (9 items)	2	1
Performing routine procedures (3 items)	1	1

For the subjective analyses of the items showing DIF with respect to the criteria given in Table 3.5, only the items showing DIF in at least two of the three methodologies were selected. In PISA, there were 10 DIF items 7 of which were released, and in TIMSS, there were 8 items 5 of which were released.

The Turkish and English versions of these released PISA and TIMSS items are given in Appendix J, Appendix K, Appendix L and Appendix M, respectively. The results from subjective analyses of the possible sources of DIF in these items are given in the next chapter.

## CHAPTER V

### CONCLUSION

In this chapter the results of this study are summarized and discussed in three main sections: (1) Construct Equivalence, (2) Item Level Analyses, (3) Sources of DIF. In addition, results from the comparisons of DIF methodologies, correspondence between the item and scale level analyses, the effect of purification on MH results, and the effect of using anchor items in IRT-LR analysis was discussed within the second section. Limitations of the study and future directions are also given at the end of the chapter.

#### 5.1 Construct Equivalence

Results from principal component analysis (PCA) with varimax rotation failed to provide evidence to support unidimensionality and equal factor structures. Although, PCA results in PISA indicated nine factors for both countries, investigating the rotated factor loadings revealed slight differences. For example, although the items m034q01, m124q01, m124q03t, m145q01t, and m150q01 loaded on the same factor in USA, they were distributed to three factors in Turkey.

On the other hand, comparison of the factor eigenvalues showed that, the eigenvalue for the first factor in Turkish TIMSS data was relatively lower (the difference between the eigenvalues was 3.759) than that of USA. But the eigenvalue for the first factor in Turkish PISA data was slightly bigger (the difference between the eigenvalues was 0.279) than that of USA. This means that, especially in TIMSS, the proportion of variances accounted for by the first factors were different in Turkey and American groups. Thus it was concluded that similarity of the factor structures in TIMSS was highly questionable.

Although they used Promax rotation, this conclusion is also in line with that of Arim and Ercikan (2005), who have reported that factor structure of the American and Turkish versions of TIMSS tests were non-equivalent.

These results indicated two problems. First of all construct equivalence is a prerequisite to carry on item level analysis (Sireci, 1997; Hui & Triandis, 1985). In addition, unidimensionality of the tests are required to determine a valid matching variable in DIF analyses (Shepard, 1982).

As the eigenvalues for the first factors across the groups were larger than the eigenvalues for the second factors it would have been concluded that a single trait underlined the test performance. To provide a statistical check of this assumption, confirmatory factor analyses (CFA) using polychoric correlations were conducted. Polychoric correlations were used to provide the ordinal data a metric Jöreskog (2005). However, except American data of TIMSS mathematics test, none of the other forms fit to a unidimensional model.

From all these analyses it was concluded that factor structure of both TIMSS and PISA were neither equivalent across Turkish and American groups, nor unidimensional except American TIMSS data. To continue the item level analysis, it was investigated whether items loaded on the first factor of the American and Turkish combined data, with respect to the PCA results, formed a unidimensional subtest.

Results from CFA for the selected items supported the unidimensionality assumption across groups for both TIMSS and PISA studies. Investigating unidimensionality, NC, GFI, AGFI, RMSEA, NNFI, and CFI indices provided reasonable values, whereas RMR values were relatively high. This contradicts with the claim of Sireci, Bastari and Allalouf (1998). They offered the use of RMR in CFA analysis, and concluded that GFI index was not reliable. However in this study, in addition to GFI all other fit indices signed a reasonable fit. In addition, not only in selected items, RMR statistic was also indicating misfit in all CFA analyses. Therefore it was decided that beyond model-data fit there were additional factors affecting RMR value, and high value RMR values were not taken as enough evidences of misfit.

Although the selected items fitted a unidimensional model in both groups individually, the results from multiple-group CFA suggested that some item parameters in unidimensional model were not equivalent across groups. This means that, Turkish and American groups had a comparable factor structures in TIMSS and PISA studies but not comparable factor loadings or intercepts for some items. Thus, group comparisons must be made with caution.

Items with different factor loadings or intercepts were further investigated. In PISA, 2 items had higher intercept values for Turkish group, and 4 items had lower factor loadings for Turkish group. Only 1 item had lower factor loading for American group.

In TIMSS, 7 items had different intercept values, 5 items had different factor loadings, and 2 items had both different intercept and factor loadings across groups. 5 of the factor loadings and 6 of the intercepts in Turkish group were larger than that of American. Other items had equal intercept and factor loadings in TIMSS and PISA.

For the items with different intercept values, it might be argued that there were differences between the mean vectors of the underlying variables of these items between the two countries that cannot be fully accounted by the mean differences in the abilities. In addition as a result of the differences in factor loadings, it can be concluded that these items had differential relations with the abilities intended to be measured by the tests, across groups.

Finally group means and variances in TIMSS and PISA studies were estimated. The means were larger in the USA both in TIMSS and PISA. USA was ahead of Turkey on mathematics achievement and mathematics literacy. This finding is in line with the TIMSS and PISA results (OECD, 2005; Gonzalez & Miles, 2001). Additionally, looking at the estimates of variances, in TIMSS Turkish students were found to be more homogenous with regard to the mathematics achievement, whereas in PISA American students were found to be more homogenous with regard to the mathematics literacy.



## 5.2 Item Level Analyses

### 5.2.1 RFA versus MH

In PISA, 7 items (32%) and in TIMSS, 5 items (24%) were flagged by RFA. On the other hand, 13 items (60%) in PISA and 14 items (67%) in TIMSS were flagged by MH. In both tests, all the items flagged by RFA were also flagged by MH as well. In PISA, RFA flagged all the high-DIF items with respect to MH results. However, in TIMSS, three items indicated by MH as showing high DIF were not flagged by RFA. The agreement rate between RFA and MH was, in sense of flagging the same items as showing DIF or not showing DIF across groups, 73% in PISA and 57% in TIMSS. It seemed that the larger the group differences and the number of problematic items, the divergent the results from RFA and MH.

The results indicated that MH detected all the items that RFA can detect. There may be various reasons for this: First of all RFA only applies to linear relations. However, the relation between dichotomous items and the trait measured by the test may be nonlinear. This can prevent RFA detecting DIF due to nonlinear fluctuations. These findings are in line with that of Benito and Ara (2000) as well. They also have reported that MH's Type II error rate was zero in a simulation study. That is MH did not fail in detecting any DIF item, although RFA did.

In addition it was easier to conduct MH than RFA. It required only a single run, whereas RFA required multiple runs with respect to AMI values. Therefore it was concluded that using RFA in addition to MH did not serve to reveal any additional information.

### 5.2.2 MH versus IRT-LR

IRT-LR flagged 9 items (41%) in PISA and 15 items (71%) in TIMSS as showing DIF. On the other hand with respect to the MH results, 13 items (60%) in PISA and 14 items (67%) in TIMSS were showing DIF. The agreement rate between MH and IRT-LR was 82% in PISA, but it seriously dropped to 48% in TIMSS. Thissen et al. (1988) also have reported similarity of results from MH and IRT-LR.

A closer look to the results revealed that, in PISA, all items flagged by IRT-LR were also detected by MH, however this was not the case in TIMSS. This issue was further investigated in terms of effect size measures, and guessing and discrimination indices.

Zwick and Ercikan (1989) have stated that absolute values of b-differences between the estimated values from reference and focal groups can be used as an effect size measure. They determined that, the absolute difference values from 0.5 to 1 indicate moderate DIF, and values greater than 1 indicate large DIF. In PISA, 7 of the 9 items flagged by IRT-LR were showing moderate to large b-differences. One of the rest two items was showing a-DIF, and only one flagged item had low b-difference. In addition to these 9 items, MH detected 4 more items as showing DIF. All these items had low b-differences, about 0.35. From these findings it was concluded that MH was more sensitive to b-differences than LRT-LR.

However, there was additional finding when the results from TIMSS were investigated. Investigating the low agreement rate between MH and IRT-LR in TIMSS, it seemed that the decrease in the agreement rate was mostly due to the disagreement on detecting the non-DIF items. Only one item was detected as non-DIF by both methods. On the other hand 6 items were detected to have a-DIF, and 2 items were detected to show c-DIF by IRT-LR. This was different from the PISA results, in which only one item was detected as showing a-DIF. In addition USA and Turkish group difference on total test score in TIMSS was larger than in PISA, and the difference between group homogeneities in TIMSS was also larger than that of PISA. Considering these findings, it was concluded that all these factors increased the potential of MH in flagging items incorrectly. MH results flagging items having only a little b-difference, for example 0.09 in item m022135, was also regarded as an evidence of this claim.

This finding is line with Penny and Johnson (1999). They have claimed that MH provided very powerful and unbiased test of DIF when items in the test could be characterized by 1-parameter IRT model. They also have reported that as items drifted from the 1-parameter model and more accurately characterized by 2 or 3-parameter models, MH provided some erroneous results. This was the case when especially group differences were large.

This result can be explained to a certain degree by the characteristic of common odds ratio, or alpha, statistic in MH. As Holland and Thayer (1988) have stated, this value is an average of the odds ratios comparing performances of individuals at each ability level determined by matching scores. In calculating alpha, MH test has an assumption that odds ratio is constant across ability levels. However, the more characteristics of groups, such as performances, or homogeneities, differ the more this can treat this assumption of MH to inflate error rates.

Finally, this study provided empirical evidence that MH and IRT-LR results were highly convergent when IRT-LR flagged only b-differences, and groups were similar in terms of performance and homogeneity on the test. However, corruption in these conditions diverged the agreeing results.

### 5.2.3 RFA versus IRT-LR

RFA is a modest model with respect to IRT-LR in the sense that, it flagged fewer items both in PISA and TIMSS. The agreement rate between these methodologies was 82% in PISA, but it considerably dropped to 52% in TIMSS. It was interesting to note that all the items flagged by RFA were also flagged by the two other models, MH and IRT-LR. However, the reverse was not true. That is, items flagged by MH and IRT-LR need not to be flagged by RFA as well.

The results were investigated in terms of b-differences as well. In TIMSS, the range of b-differences of items flagged by RFA changed from 0.50 to 1.73. On the other hand, in PISA b-differences fluctuate between 0.36 and 1.07. In addition, RFA was not able to detect items flagged by IRT-LR as showing a-DIF or c-DIF unless the items also had large b-differences. However this does not mean that RFA can always detect items with large b-differences, for example item m012009 in PISA with b-difference of 0.7 was not detected by RFA.

From these results it was concluded that, RFA produced similar results with IRT-LR when only b-DIF was reported by IRT-LR. When there were a-DIF and c-DIF with respect to IRT-LR results, the agreement rate between RFA and IRT-LR decreased in the sense that RFA could not able to detect these fluctuations. This

decrease in the agreement rate resembles the relation between MH and IRT-LR, however from an opposite direction.

That is, when items showed more complex parametric differences, such as differences in discrimination and guessing parameters, and group differences were large, MH flagged items even with very little differences. However, RFA was not sensitive to these differences. From these results, it might be argued that more complex parameters across groups increases the potential of Type I error in MH, and Type II error in RFA. But additional studies are required for further investigations.

Finally it was concluded that using RFA, MH, and IRT-LR in complimentary fashion counted to the DIF analyses. IRT-LR could detect a-DIF and c-DIF in addition to b-DIF. On the other hand MH had an outstanding power in detecting moderate and small fluctuations across groups. In addition RFA could control the possible inflation in Type I error of MH. A strict condition to determine an item as functioning differentially across groups would be to check whether the item is flagged by all these three methodologies.

#### 5.2.4 Scale Level Analysis versus Item Level Analysis

Use of different DIF methodologies for item level analysis produced somewhat divergent results. Because, different DIF methodologies can be affected by sample properties, such as ability distribution, or other procedures, such as computer algorithms, differently. So, the pattern of agreement of the procedures may produce more reliable results about the DIF items. In this context, items flagged by all three DIF procedures were also investigated with respect to CFA results.

In PISA, 3 of 6 items flagged by all three DIF procedures also had different parameters across groups with respect to CFA results. However, it is worth adding that all three items were high-DIF items. On the other hand, 4 items flagged to have different factor loadings with respect to CFA, were not flagged by RFA and IRT-LR, and only two of them were flagged by MH as moderate-DIF.

In TIMSS, 5 items flagged by all three DIF procedures also had different parameters across groups with respect to CFA. However, 6 items, two of which had different intercept values, flagged by CFA were not flagged by at least two of the DIF procedures.

As a conclusion, it was not possible to claim that items not flagged by CFA were free of bias as well. This finding is in line with that of Zumbo (2003). On the other hand it was concluded that items flagged as having different intercepts across groups are also candidates to be flagged by DIF methodologies as well. This finding is in line with the interpretation of different intercepts in CFA provided by Jöreskog (2005), who claimed that different intercepts point differences that can not be entirely accounted by corresponding differences in latent traits.

But this finding was slightly overcastted in TIMSS. This may be due to the relative complexity of TIMSS. It had more DIF items, there was a bigger ability difference between the groups, and groups' homogeneities were also very different. In addition, as Reise, Widaman and Pugh (1993) have reported, this may be due to that CFA cannot deal with non-linear differences yet.

#### 5.2.5 The Effects of Purifying Matching Criterion on MH Results

Comparing the MH analyses results of PISA before and after purification, which were given in first and second steps of EZDIF program respectively, indicated that there was no difference with respect to the effect size measures, except that item m192q01t showed B-DIF in the second step when it was A-DIF in the first step. However, with respect to statistical significance at 0.01, 9 (41%) items in the first step and 7 items (32%) in the second step were significantly different across groups.

In TIMSS, one of the two items showing high DIF, or C-DIF, in the first step was flagged to show negligible DIF in the second step, while the other item flagged to show moderate DIF. On the other hand some items, such as m012009, showing B-DIF in the first step changed to show A-DIF in the second step, whereas some A-DIF items in the first step, such as m012021, changed to show B-DIF in the second step.

Additionally, with respect to statistical significance at 0.01, 13 (62%) items in the first step and 14 items (67%) in the second step were significantly different across groups.

In terms of effect size measures, the results indicated that two-step procedure, or purification of the matching criterion, produced equal, as in PISA, or superior, as in TIMSS, results than that of first-step procedure. This finding is in line with that of Clauser, Mazor and Hambleton (1993). However, in terms of statistical significance, although the purification process clarified the DIF results to a certain degree in PISA (9% drop), it did not contribute in TIMSS (5% increase).

From these findings it was concluded that purification of matching criterion for subsequent analysis did contribute the results, if effect size measures (i.e. MH D-DIF vales) were taken into consideration. However, in terms of statistical significance, purification can either have a negative affect as well. This finding contradicts with that of Clauser at.al. (1993). They claimed that the affects of purifying the matching criterion should be most evident when the greatest contamination is present. But, in TIMSS, in which contamination was relatively larger, purifying the matching criterion did not contribute the results. This may be due to the potential of items in showing a-DIF and c-DIF in addition to b-DIF, or to the performance or homogeneity differences across groups. Further analysis on this issue may reveal the factors to be considered in purifying the matching criterion.

With respect to the results of this study it was concluded that purifying the matching criterion contributed to clarify the MH results when MH D-DIF statistics were considered.

#### 5.2.6 The Effects of Using Anchor Items on IRT-LR Results

In fact even when no items are specified as anchor items, IRTLRDIF program uses all but the studied item as anchor items, which is named as all-other method in the literature (Wang, Yeh, & Yi, 2003). Thus, investigating the effects of using anchor items in IRTLRDIF is comparing it with the all-other method.

In this context, when PISA items were considered, using anchor items did not lead a significant change in the results. However, using anchor items in TIMSS, it was observed that the power of analysis in detecting c-DIF increased noticeably. This finding supports the claim of Wang et al (2003) indicating all-other method works well in the reasonable tests in the sense that having few number of DIF items whose contamination is balanced between groups.

This study in addition concluded that using anchor items was seem to produce similar results to that of all-other method when tests were having few number of DIF items whose contamination were balanced between groups. However, when these conditions were broken using all-other method would decrease the power of IRT-LR analysis to detect c-DIF. In addition it should also be considered that this conclusion assumed that the process of detecting anchor items as defined in this study was reasonable.

### 5.3 Possible Sources of DIF

In determining the degree to which DIF may be due to curricular differences, Table 4.18 showing the relative distribution of DIF items by content area was examined. In TIMSS, DIF items were clustered in four area topics, namely fraction and number sense, algebra, measurement, and data representation. Six of nine (67%) of the Fraction and Number Sense items were identified as DIF, five of which were in favor of the USA.

Also in PISA, although it is not a curriculum-based study as TIMSS, a similar pattern in Quantity items was identified. The quantity items in PISA were those requiring an understanding of relative size, recognition of numerical patterns, and the use of numbers. Three of eight items (38%) of the Quantity items were identified as DIF, all of which were in favor of the USA.

It was concluded that these two findings provided support for the interpretation that DIF in the items requiring a number sense might be due to curricular differences. It should also be specified that curricular differences must be regarded in a broader sense to include instructional practices of the teachers as well.

In this context, relative failure of Turkish students in items requiring a number sense with respect to matched USA students would be attributed to the ineffectiveness of the curriculum and instructional practices in Turkey. On the other hand, the relative distribution of DIF items in TIMSS by cognitive expectation did not lead any interpretable results.

Qualitative reviewers also managed to reach some consensus about the characteristics of the items favoring a specific group, which revealed the following hypotheses.

Investigating the two released items, namely m124q01 and m547q01t, in PISA, functioning in favor of Turkey with respect to all three of the DIF methodologies, it was concluded that both of the items were relatively simple items with respect to the cognitive processes required to get the item correct. Item m124q01 was a single step question requiring a correct manipulation of expressions containing symbols, and item m547q01t was also a single step item requiring an interpretation of a simple picture and conducting a simple division by two-digit number. With respect to these cognitive activities both items were located in the reproduction cluster. In addition, item m150q01 was another PISA item favoring Turkish students with respect to RFA and M-H. This item was also located in reproduction cluster, requiring carrying out a simple subtraction.

Reviewers also labeled these items as curriculum-like and task oriented. It is also worth specifying that these items were the first items within the questions related with a single stem. PISA ordered the items related with a single question stem from relatively simple ones, usually at reproduction level, to relatively complex ones, at reflection or connection levels.

From all these findings it was concluded that items requiring competencies of reproduction of practiced knowledge, knowledge of facts, performance of routine procedures, application of technical skills are less likely to be biased against Turkish students with respect to American students at the same ability level.

On the other hand, an in depth analysis of released DIF items favoring USA students, namely items m150q03t, m413q03t, and m520q03t, revealed that these items were relatively more complex than the items favoring Turkish students.



Item m150q03 required students to interpret the given graph provide an explanation in support of the given proposition. In the same manner, item m413q03t also had a demand of conclusion and reasoning.

Although the item m520q03t did not require students to communicate mathematically, it required exploring possibilities to decide on which was the best, and interpret the results.

Considering these findings it was concluded that items requiring students to communicate mathematically, such as by providing explanations and reasoning, items where various results must be compared, and items that have real-world context are less likely to be in favor of Turkish students with respect to American students at the same ability level.

Reviewers also agreed that the translations of the items m150q02t and m150q03t changed the content in the sense that the translations did not preserve the quantitative language. The term “on average” were translated into Turkish as “ortalama olarak”, however it was argued that this term would have stimulated the Turkish students to perform an operation to calculate the arithmetic mean, although the items did not require any operation. So, it was also argued that DIF in these items might be due to this adaptation problem in addition to the possible sources specified above.

Unfortunately the reviewers would not reached a consensus on their arguments related with the sources of DIF in TIMSS items.

#### 5.4 Limitations of the Study

Three limitations for study can be noted. First, as the sample sizes were limited it was not possible to conduct cross-validation studies as offered by Camilli and Shepard (1994). Second, none of the DIF methodologies used in the study was able to detect non-uniform DIF.

Finally, the reviewers used in the study might not have been qualified enough to assess the possible sources of the DIF in the items.

In addition, in identifying the sources of DIF, interpretations would be speculative as the reviewers had knowledge about which items were functioning differentially functioning.

In addition it should also mentioned that, as only a limited number of items were released it were not possible to conduct a detailed review of all the items.

## 5.5 Future Directions

The results of this study suggest that future research should focus on the development of statistical methods for testing DIF, especially in tests having multifarious aspects, such as a considerable value of flagging items, suitable to be represented by 3-parameter model etc. This current study focused on item level DIF, future research can deal with the same data at the test level as well. Especially for the TIMSS data, as it seemed to be more problematic than PISA, it would be interesting to detect in what ways the results of item and test level analyses differ.

Future research would also focus on generating guidelines to adapt items into Turkish. Confirmatory approaches, as suggested by Gierl and Khalig (2000), would be conducted to develop confirmed hypotheses, which may lead to a better understanding of DIF in mathematics items. This study was an initial step in assessing the Turkish translation of math items used in international studies. Problematic items identified by both statistical and qualitative methods would be examined more thoroughly to determine any other potential sources that were not found in this study. Findings from various studies would help to achieve a better understanding of the cultural differences in international assessments. In this process it should also be considered that using more than one DIF method would lead better understandings because multiple methodologies would compensate each other's defects.

Developing systematic guidelines on reviewing mathematics items in international assessments would be invaluable contributions of future research. One of the most comprehensive guidelines is that of Allalouf et al.(1999). However it is not certain to what degree these guidelines can be applied to mathematics items as

well. Future research should not only investigate this appropriateness, but try to develop guidelines specific to mathematics items as well.

Finally, the relation between the equivalence of test structure and results from DIF analyses should also be further investigated. Some simulation studies should also be conducted to detect their reciprocal associations.

## REFERENCES

- Ackerman, T.A. (1992). A Didactic Explanation Of Item Bias, Item Impact, and Item Validity From a Multidimensional Perspective. *Journal of Educational Measurement*. 29(1), 67 – 91.
- Adams, R.J., & Gonzalez, E.J. (1996). The TIMSS Test Design. In M.O. Martin & D.L. Kelly (Eds.). *Third International Mathematics and Science Study technical report volume I: Design and development*. Chestnut Hill, MA: Boston College.
- Allalouf, A., Hambleton, R.K.& Sireci, S.G. (1999). Identifying the Causes of DIF in Translated Verbal Items. *Journal of Educational Measurement*. 36(3), 185 – 198.
- Angoff, W.H. & Ford, S.F. (1973). Item-Race Interaction on a Test of Scholastic Aptitude. *Journal of Educational Measurement*, 10, 95 – 105.
- Arim, R.G. & Ercikan, K. (2005) *Comparability Between The US and Turkish Versions of The Third International Mathematics and Science Study's Mathematics Test Results*. Paper presented at NCME April 12-14 Montreal, Canada
- Beller, M.&Gafni, N.(1996). The 1991 International Assessment of Educational Progress In Mathematics and Sciences. The Gender Differences Perspective. *Journal of educational psychology*, 88(2), 365-377.
- Benito J.G. & Ara M.J.N. (2000). A Comparison of  $\chi^2$  , RFA and IRT Based Procedures in the Detection of DIF, *Quality & Quantity*, v.34, 17-31.
- Benjamini, Y. & Hochberg, Y. (1995). Controlling The False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society, Series B*, 57, 289 – 300.

- Bentler, P.M. and Bonett, D.G. (1980), Significance Tests and Goodness of Fit in the Analysis of Covariance Structures. *Psychological Bulletin*, 88, 588 -606.
- Berberoğlu, G. (1995). Differential Item Functioning Analysis of Computation, Word Problem and Geometry Questions Across Gender and SES Groups. *Studies in Educational Evaluation*, 21, 439 – 456.
- Bock, D.R. & Aitkin, M. (1981) Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM Algorithm. *Psychometrika* 46(4), 443 – 459.
- Bontempo, R.(1993). Translation Fidelity of Psychological Scales. An Item Response Theory Analysis of an Individualism-Collectivism Scale. *Journal of Cross-Cultural Psychology*. 24(2), 149 – 166.
- Borsboom, D., Mellenbergh, G.J. & van Heerden, J. (2002). Different Kinds of DIF: A Distinction Between Absolute and Relative Forms of Measurement Invariance and Bias. *Applied Psychological Measurement*, 26(4), 433 – 450
- Camilli, G. & Shepard, L.A. (1994). *Methods for Identifying Biased Test Items*. Sage Publications, California.
- Clauser B., Mazor K. & Hambleton R.K. (1993). The Effects of Purification of The Matching Criterion on the Identification of DIF Using The Mantel- Haenszel Procedure. *Applied measurement in education*, 6(4), 269-279.
- Crocker, L. & Algina, J. (1986). *Introduction To Classical And Modern Test Theory*. New York: Holt, Rinehart and Winston.
- Dede, Y.& Argün, Z. (2003). Cebir, Öğrencilere Niçin Zor Gelmektedir? *Hacettepe Ün. Eğitim Fak. Dergisi* 24, 180-185.
- Donoghue, J.R. & Allen, N.L. (1993) Thin Versus Thick Matching In The Mantel-Haenszel Procedure For Detecting DIF. *Journal of Educational Statistics*. 18(2), 131 –154.

- Doolittle, A. E. & Cleary, T.A. (1987). Gender-Based Differential Item Performance In Mathematics Achievement Items. *Journal of Educational Measurement*, 24, 157 – 166.
- Dorans, N.J. & Holland, P.W.,(1993). DIF Detection And Description: Mantel-Haenzsel And Standardization. In P.W.Holland & H. Wainer (Eds.) *Differential item functioning: Theory and practice* (pp. 137 - 166) Hillsdale, NJ: Erlbaum.
- Drasgow, F. (1984). Scrutinizing Psychological Tests: Measurement Equivalence And Equivalent Relations With External Variables Are The Central Issues. *Psychological Bulletin*. 95(1), 134 – 135
- Du Toit, M. (2003). *IRT from SSI*, Scientific Software International, Inc,USA.
- EARGED, (2003) *Öğrenci Başarısının Belirlenmesi Durum Raporu*, EARGED, Ankara.
- Ellis, B.B. (1989). Differential Item Functioning: Implication For Test Translation. *Journal of Applied Psychology*. 74, 912 – 921
- Ellis, B.B., Becker, P. & Kimmel H.D.(1993). An Item Response Theory Evaluation Of An English Version Of The Trier Personality Inventory (TPI). *Journal of Cross-Cultural Psychology*. 24(2), 133 – 148.
- Engelhard, G.(1990). Gender Differences In Performance On Mathematics Items: Evidence From The United States And Thailand. *Contemporary Educational Psychology*. 15, 13-26.
- Ercikan, K.(1998). Translation Effects In International Assessments. *International Journal Of Educational Research*. 29(6), 543-553.
- Ercikan, K.(2002). Disentangling Sources Of Differential Item Functioning In Multilanguage Assessments. *International Journal Of Testing*. 2(3&4), 199-215.

- Ersoy, Y. & Erbas, A. K. (2000). Cebir öğretiminde öğrencilerin güçlükleri Yanlışlarla ilgili öğretmen görüşleri [Students' difficulties in Algebra-II: Teacher views about students' errors]. *IV.Fen Bilimleri Eğitimi Kongresi* (s. 625-629). Ankara, Türkiye: Milli Eğitim Bakanlığı Yay.
- Foy, P. and Joncas, M. (2000). Implementation of the Sample Design in Martin, M.O., Gregory, K.D. and Stemler, S.E. (Eds.), *TIMSS 1999 technical report: IEA's repeat of the Third International Mathematics and Science Study at the eighth grade*. Chestnut Hill, MA: Boston College.
- Gao L.& Wang C.(2005). *Using Five Procedures To Detect Dif With Passage-Based Testlets*. A paper prepared for the poster presentation at the graduate student poster session at the annual meeting of the national council of measurement in education, Montreal, Quebec.
- George, D. & Mallery, P. (2003). *SPSS for Windows Step By Step*, Pearson Education, Inc, USA.
- Gierl, M., Jodoin, M. & Ackerman T. (2000). *Performance Of Mantel-Haenszel, Simultaneous Item Bias Test, And Logistic Regression When The Proportion Of DIF Items Is Large*. Paper Presented at the Annual Meeting of the American Educational Research Association (AERA). New Orleans, Louisiana, USA
- Gierl, M.J (2005). Using Dimensionality-Based Dif Analysis To Identify And Interpret Constructs That Elicit Group Differences. *Educational Measurement Issues and practice*, 24(1), 3-13.
- Gierl, M.J. & Khaliq S.N. (2000). *Identifying Sources Of DIF On Translated Achievement Tests: A Confirmatory Analysis*. Paper Presented at the Annual Meeting of the National Council on Measurement in Education (NCME). New Orleans, Louisiana, USA
- Gierl, M.J. (2004). *Using A Multidimensionality- Based Framework To Identify And Interpret The Construct Related Dimensions That Elicit Group Differences*, Paper presented at the annual meeting of the American educational research association (AERA), San Diego, California, USA

- Gonzalez, E.J. & Miles, J.A. (2001). *TIMSS 1999 User Guide for the International Database*, IEA, Boston College, USA.
- Gulliksen, H. (1950). *Theory Of Mental Tests*. New York: John Wiley.
- Hambleton, R & Kanjee, A., (1995), Increasing The Validity Of Cross-Cultural Assessments: Use Of Improved Methods For Test Adaptations. *European journal of psychological assessment*, 11(3). Pp. 147-157.
- Hambleton, R.K. & Patsula, L. (2000) Adapting Tests For Use In Multiple Languages And Cultures. (*ERIC Document Reproduction Service, No: ED 459 207*)
- Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991). *Fundamentals of Item Response Theory*. Sage Publications, California.
- Harris, A. M. & Carlton, S. T. (1993). Patterns Of Gender Differences On Mathematics Items On The Scholastic Aptitude Test. *Applied Measurement in Education*, 6, 137 – 151.
- Holland, P.& Thayer, D., (1988) Differential Item Functioning And Mantel-Haenzsel Procedure. In H. Wainer & H.I. Braun (Eds.), *Test Validity* (pp. 129 – 145), Hillsdale, NJ: Lawrence Erlbaum.
- Hui, C.H. & Triandis, H.C. (1983). Multistrategy Approach To Cross-Cultural Research. The Case Of Locus Control. *Journal of Cross-Cultural Psychology*. 14(1), 65 – 83.
- Hui, C.H. & Triandis, H.C. (1985). Measurement In Cross-Cultural Psychology. A Review And Comparison Of Strategies. *Journal of Cross-Cultural Psychology*. 16(2), 131 –152
- Hui, C.H. & Triandis, H.C. (1989). Effects Of Culture And Response Format On Extreme Response Style. *Journal of Cross-Cultural Psychology*. 20(3), 296 – 309



- Hulin, C. & Mayer, L.(1986). Psychometric Equivalence Of A Translation Of The Job Descriptive Index Into Hebrew. *Journal of Applied Psychology* 71(1) pp. 83-94.
- Hulin, C.L. (1987). A Psychometric Theory Of Evaluations Of Item And Scale Translations: Fidelity Across Languages. *Journal of Cross-Cultural Psychology*. 18(2), 115 – 142.
- Hulin, C.L., Drasgow, F. & Komocar, J. (1982). Applications Of Item Response Theory To Analysis Of Attitude Scale Translations. *Journal of Applied Psychology*. 67(6), 818 – 825
- Jöreskog K.G. (1971)., Simultaneous Factor Analysis In Several Populations. *Psychometrika*, 36(4), 409-426
- Jöreskog, K., & Sörbom, D. (1993). *Structural Equation Modeling with the SIMPLIS Command Language*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Jöreskog, K., & Sörbom, D. (2001). *LISREL 8: User's Reference Guide*. Chicago: Scientific Software International Inc, USA.
- Jöreskog, K., & Sörbom, D. (2002). *PRELIS 2:User's Reference Guide*. Chicago: Scientific Software International Inc, USA
- Jöreskog, K.G. (2005) *Structural Equation Modeling With Ordinal Variables Using LISREL*, Retrieved from <http://www.ssicentral.com/lisrel/techdocs/ordinal.pdf>
- Kelloway, E. K. (1998). *Using LISREL for Structural Equation Modeling*. London, New Delhi: Sage Publications.
- King, J.P., (1998) *Matematik Sanatı*, TÜBİTAK Popüler Bilim Kitapları, Ankara.
- Klieme, E. & Baumert, J.(2001). Identifying National Cultures Of Mathematics Education: Analysis Of Cognitive Demands And Differential Item Functioning In TIMSS. *European Journal of Psychology of Education*, 15(3), 385 – 402.

- Li, Y., Cohen, A.S., & Ibarra, R. A. (2004). Characteristics of Mathematics Items Associated With Gender DIF, *International Journal of Testing*, 4(2), 115 – 136.
- Lord F.M. (1980). *Applications Of Item Response Theory To Practical Testing Problems*. Hilldale, NJ: Lawrence Erlbaum.
- Lord, F. M., & Novick, M.R., with Birnbaum, A. (1968). *Statistical Theories Of Mental Test Scores*. Reading, MA: Addison-Wesley.
- McKnight, C.C. & Valverde, G.A. (1999). Explaining TIMSS Mathematics Achievement. In *International Comparisons in Mathematics Education*, G. Kaiser, E. Luna & L. Huntley. (Eds).p.48-67., Falmer Press, London.
- Meara, K., Robin, F. & Sireci, S.G. (2000) Using Multidimensional Scaling To Assess The Dimensionality Of Dichotomous Item Data. *Multivariate Behavioral Research*, 35 (2), 229 –259.
- NCEE, (1983) *A Nation at Risk: The Imperative for Educational Reform* A Report to the Nation and the Secretary of Education United States Department of Education by The National Commission on Excellence in Education. Retrieved from <http://www.ed.gov/pubs/NatAtRisk/title.html>
- NCTM, (1996), *Curriculum and Evaluation Standards for School Mathematics*, The National Council of Teachers of Mathematics, Inc., USA.
- OECD (2003a). *The PISA 2003 Assessment Framework*, OECD Publishing.
- OECD (2005) *PISA 2003 Technical Report*, OECD Publishing.
- Oort, F.J. (1992) Using Restricted Factor Analysis To Detect Item Bias. *Methodika*, 6, 150 – 166.
- Osterlind, S.J. (1983). *Test Item Bias* Sage Publications, California

- Penny, J. & Johnson, R.L. (1999). How Group Differences In Matching Criterion Distribution And IRT Item Difficulty Can Influence The Magnitude Of The Mantel- Haenszel Chi-Square Dif Index. *Journal of experimental education*, 67(4), 343 – 366.
- Poortinga Y.H. & Van de Vijver, F.J. (1987). Explaining Cross-Cultural Differences. Bias Analysis And Beyond. *Journal of Cross-Cultural Psychology* .18(3), 259 – 282
- Poortinga, Y.H. (1989). Equivalence Of Cross-Cultural Data: An Overview Of Basic Issues. *International Journal of Psychology*. 24, 737 – 756.
- Reise, S.P., Widaman,K.F. & Pugh R.H.(1993). Confirmatory Factor Analysis And Item Response Theory: Two Approaches For Exploring Measurement Invariance. *Psychological Bulletin*. 114(3), 552 – 566.
- Robin, F., Sireci, S.G.& Hambleton, R.K.(2003). Evaluating The Equivalence Of Different Language Versions Of A Credentialing Exam. *International journal of testing*. 3(1), 1-20.
- Robitaille, D.F. & Beaton, A.E. (2002). A Brief Overview Of The Study. In D.F. Robitaille & A. E. Beaton. (Eds.). *Secondary Analysis of the TIMSS Data*. p. 11 –18. Kluwer Academic Publishers, Netherlands.
- Rogers, J. & Swaminathan, H., (1993). A Comparison Of Logistic Regression And Mantel-Haenszel Procedures For Detecting Differential Item Functioning. *Applied psychological measurement*. 17(2). Pp. 105-116.
- Roznowski, M. & Reith, J. (1999). Examining The Measurement Quality Of Tests Containing Differentially Functioning Items: Do Biased Items Result In Poor Measurement? *Educational and Psychological Measurement*. 59(2), 248 – 269.
- Scheuneman J.D. & Grima A. (1997). Characteristics Of Quantitative Word Items Associated With Differential Performance For Female And Black Examinees. *Applied measurement in education*, 10 (4), 299-319.

- Schumacker, R. E., & Lomax, R. G. (1996). *A Beginner's Guide to Structural Equation Modeling*. Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Shealy, R. & Stout, W. (1993). A Model Based Standardization Approach That Separates True Bias/DIF From Group Ability Differences And Detects Test Bias/DTF As Well As Item Bias/DIF. *Psychometrika*, 58(2), 159 – 194.
- Shepard, L.A. (1982). Definitions Of Bias. In R.A. Berk (Eds), *Handbook of methods for detecting test bias*. Baltimore: Johns Hopkins University Press.
- Sireci, S. & Geisinger, K., 1995. Using Subject Matter Experts To Assess Content Representation: An MDS Analysis. *Applied psychological measurement*. 19(3). Pp. 241-255.
- Sireci, S.G. & Allalouf, A. (2003) Appraising Item Equivalence Across Multiple Languages And Cultures. *Language Testing*, 20(2), 148 – 166.
- Sireci, S.G. (1997). Problems And Issues In Linking Assessment Across Languages. *Educational Measurement: Issues and Practice*. 16(1), 12 – 19.
- Sireci, S.G. & Berberoğlu, G. (2000). Using Bilingual Respondents To Evaluate Translated-Adapted Items. *Applied Measurement in Education*, 13(3), 229 – 248.
- Sireci, S.G., Bastari, B. & Allalouf, A. (1998) Evaluating Construct Equivalence Across Adapted Tests. Paper presented at APA August 14, San Francisco, CA.
- Steinberg, L. (2001) The Consequences Of Pairing Questions: Context Effects In Personality Measurement. *Journal of Personality and Social Psychology*. 81(2), 332 – 342.
- Swaminathan H. & Rogers, J.H. (1990). Detecting Differential Item Functioning Using Logistic Regression Procedures. *Journal of educational measurement*. 27(4), 361-370.

- Thissen, D. (2001) *IRTLRDIF v.2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning*. Retrieved from <http://www.unc.edu/~dthissen/dl.html>
- Thissen, D., Steinberg, L. & Gerrard, M. (1986) Beyond Group Mean Difference: The Concept Of Item Bias. *Psychological Bulletin* 99(1), 118 – 128.
- Thissen, D., Steinberg, L. & Kuang, D. (2002). Quick And Easy Implementation Of The Benjamini-Hochberg Procedure For Controlling The False Positive Rate In Multiple Comparisons. *Journal of Educational and Behavioral Statistics* 27(1), 77 – 83.
- Thissen, D., Steinberg, L. & Wainer, H. (1988) Use Of Item Response Theory In The Study Of Group Differences In Trace Lines. In H. Wainer & H. Braun (Eds.), *Test Validity*, (pp. 147 – 169) Hillsdale, NJ: Erlbaum.
- Thissen, D., Steinberg, L. & Wainer, H. (1993) Detection Of Differential Item Functioning Using The Parameters Of Item Response Models. In P.W.Holland & H. Wainer (Eds.) *Differential item functioning: Theory and practice* (pp. 67 – 113) Hillsdale, NJ: Erlbaum.
- Ülger,A., (2003) Matematğin Kısa Bir Tarihi. *Matematik Dünyası*, 2, 49 – 53.
- Van de Vijver, F. & Tanzer, N.K.(1997). Bias And Equivalence In Cross-Cultural Assessment: An Overview. *European Review of Applied Psychology*. 47(4), 263 – 279.
- Van de Vijver, F.J. & Poortinga Y.H.(1982). Cross-Cultural Generalization And Universality. *Journal of Cross-Cultural Psychology*. 13(4), 387 – 408.
- Waller, N.G. (2005) EZDIF: *A Computer Program For Detecting Uniform And Nonuniform Differential Item Functioning With The Mantel-Haenszel And Logistic Regression Procedures*. Retrieved from [http://peabody.vanderbilt.edu/depts/psych\\_and\\_hd/faculty/wallern/](http://peabody.vanderbilt.edu/depts/psych_and_hd/faculty/wallern/)
- Wang, W.C., Yeh, Y.L. & Yi, C. (2003). Effects Of Anchor Item Methods On Differential Item Functioning Detection With The Likelihood Ratio Test. *Applied Psychological Measurement*. 27(6), 479 – 498.

- Williams, V.S.L. (1997). The “Unbiased” Anchor: Bridging The Gap Between DIF And Item Bias. *Applied Measurement in Education*, 10, 353-267.
- Williams, V.S.L. (1997). The “Unbiased” Anchor: Bridging The Gap Between DIF And Item Bias. *Applied Measurement in Education*, 10, 353-267.
- Williams, V.S.L., Jones, L.V. & Tukey, J.W. (1999). Controlling Error In Multiple Comparisons, With Examples From State-To-State Differences In Educational Achievement. *Journal of Educational and Behavioral Statistics*. 24(1), 42 – 69.
- Wolf, R.M.(1998). Validity Issues In International Assessments. *International journal of educational research*. 29(6), 491-501.
- Yurdug l, H. & A kar, P. (2004a) Orta ğretim Kurumları  ğrenci Se me Ve Yerle tirme Sınavının  ğrencilerin Yerle im Yerlerine G re Diferansiyel Madde Fonksiyonu A ısından Incelenmesi *Hacettepe  niversitesi Eğitim Fak ltesi Dergisi*, 27, 268-275.
- Yurdug l, H & A kar, P. (2004b) Orta ğretim Kurumları  ğrenci Se me Ve Yerle tirme Sınavının Cinsiyete G re Madde Yanlılığı A ısından Incelenmesi. *Eğitim Bilimleri ve Uygulama Dergisi*, 3(5), 3-20.
- Zumbo, B.D. (2003) Does Item-Level DIF Manifest Itself In Scale-Level Analyses? Implications For Translating Language Tests. *Language Testing*, 20(2), 136 – 147.
- Zwick W.R.& Velicer W.F.(1986). Comparison Of Five Rules For Determining The Number Of Components To Retain. *Psychological Bulletin*, 99(3), 432-442.
- Zwick, R. & Ercikan, K., (1989). Analysis Of Differential Item Functioning In The NAEP History Assessment. *Journal of Educational Measurement*, 26(1), 55-66.

## APPENDIX A. PISA 2003 BOOKLET 2 PERCENTAGE OF RECODED ITEMS

Item No	Type	Item Scales	# Missing (TUR)	% (TUR)	# Missing (USA)	% (USA)
m034q01t	CR	Space and Shape	21	0,05	9	0,02
<b>m124q01</b>	CR	Change and Relationships	52	0,13	29	0,07
<b>m124q03t</b>	PCR	Change and Relationships	132	0,34	89	0,21
<b>m145q01t</b>	CMC	Space and Shape	20	0,05	14	0,03
<b>m150q01</b>	CR	Change and Relationships	44	0,11	27	0,06
<b>m150q02t</b>	PCR	Change and Relationships	74	0,19	20	0,05
<b>m150q03t</b>	CR	Change and Relationships	117	0,30	37	0,09
m192q01t	CMC	Change and Relationships	38	0,10	5	0,01
m305q01	MC	Space and Shape	19	0,05	11	0,03
m406q01	CR	Space and Shape	127	0,32	60	0,14
m406q02	CR	Space and Shape	211	0,54	135	0,32
m406q03	CR	Space and Shape	133	0,34	108	0,25
m408q01t	CMC	Uncertainty	5	0,01	1	0,00
m411q01	CR	Quantity	54	0,14	14	0,03
m411q02	MC	Uncertainty	42	0,11	18	0,04
<b>m413q01</b>	CR	Quantity	51	0,13	***	***
<b>m413q02</b>	CR	Quantity	68	0,17	27	0,06
<b>m413q03t</b>	CR	Quantity	124	0,32	61	0,14
m423q01	MC	Uncertainty	2	0,01	6	0,01
<b>m438q01</b>	CR	Uncertainty	63	0,16	***	***
<b>m438q02</b>	MC	Uncertainty	44	0,11	17	0,04
m446q01	CR	Change and Relationships	20	0,05	5	0,01
m446q02	CR	Change and Relationships	126	0,32	19	0,04
m462q01t	PCR	Space and Shape	62	0,16	44	0,10
m474q01	CR	Quantity	5	0,01	6	0,01
<b>m505q01</b>	CR	Uncertainty	103	0,26	***	***
<b>m510q01t</b>	CR	Quantity	34	0,09	15	0,04
<b>m520q01t</b>	PCR	Quantity	44	0,11	42	0,10
<b>m520q02</b>	MC	Quantity	20	0,05	6	0,01
<b>m520q03t</b>	CR	Quantity	31	0,08	7	0,02
<b>m547q01t</b>	CR	Space and Shape	56	0,14	65	0,15
<b>m555q02t</b>	CMC	Space and Shape	7	0,02	4	0,01
m598q01	CR	Space and Shape	36	0,09	28	0,07
<b>m702q01</b>	CR	Uncertainty	106	0,27	36	0,08
m710q01	MC	Uncertainty	31	0,08	20	0,05
<b>m806q01t</b>	CR	Quantity	5	0,01	13	0,03

(Items not responded although it was expected to be, non-reached items and items in which more than one alternative selected, were coded as missing in the study. Released items are given in **bold**.)

CR : Coded Response

PCR : Coded Response (with partial credit score)

MC : Multiple Choice

CMC : Complex Multiple Choice

\*\*\* : Items all coded as incorrect

## A2. TIMSS 1999 Booklet 7 Percentage of Recoded Items

Item No	Type	Item Scales	# Missing	%	# Missing	%
			(TUR)	(TUR)	(USA)	(USA)
m012001	MC	Fractions and Number Sense	12	0,01	11	0,01
m012002	MC	Algebra	11	0,01	9	0,01
m012003	MC	Measurement	7	0,01	14	0,01
m012004	MC	Fractions and Number Sense	11	0,01	13	0,01
m012005	MC	Geometry	23	0,02	13	0,01
m012006	MC	Data Rep. & Prob.	2	0,00	14	0,01
<b>m012007</b>	MC	Data Rep. & Prob.	15	0,02	15	0,01
<b>m012008</b>	MC	Fractions and Number Sense	14	0,01	19	0,02
<b>m012009</b>	MC	Fractions and Number Sense	88	0,09	33	0,03
<b>m012010</b>	MC	Fractions and Number Sense	15	0,02	24	0,02
<b>m012011</b>	MC	Geometry	19	0,02	14	0,01
<b>m012012</b>	MC	Algebra	25	0,03	16	0,01
<b>m012019</b>	MC	Geometry	23	0,02	26	0,02
<b>m012020</b>	MC	Algebra	25	0,03	34	0,03
<b>m012021</b>	MC	Fractions and Number Sense	5	0,01	25	0,02
<b>m012022</b>	MC	Algebra	33	0,03	29	0,03
<b>m012023</b>	MC	Measurement	4	0,00	31	0,03
<b>m012024</b>	MC	Fractions and Number Sense	4	0,00	26	0,02
<b>m012043</b>	MC	Data Rep. & Prob.	30	0,03	11	0,01
<b>m012044</b>	MC	Fractions and Number Sense	17	0,02	11	0,01
<b>m012045</b>	MC	Fractions and Number Sense	3	0,00	8	0,01
<b>m012046</b>	MC	Algebra	70	0,07	18	0,02
<b>m012047</b>	MC	Data Rep. & Prob.	16	0,02	13	0,01
<b>m012048</b>	MC	Algebra	22	0,02	10	0,01
m022135	MC	Data Rep. & Prob.	22	0,02	24	0,02
m022139	MC	Fractions and Number Sense	6	0,01	25	0,02
m022142	MC	Geometry	39	0,04	37	0,03
m022144	MC	Fractions and Number Sense	33	0,03	27	0,02
m022146	MC	Data Rep. & Prob.	15	0,02	28	0,03
m022148	CR	Measurement	140	0,14	57	0,05
m022253	CR	Algebra	191	0,19	81	0,07
m022154	MC	Geometry	13	0,01	25	0,02
m022156	CR	Fractions and Number Sense	110	0,11	113	0,10
<b>m022237</b>	CR	Fractions and Number Sense	307	0,31	98	0,09
<b>m022256</b>	PCR	Data Rep. & Prob.	209	0,21	110	0,10
<b>m022241</b>	MC	Fractions and Number Sense	61	0,06	23	0,02
<b>m022262a</b>	CR	Algebra	209	0,21	76	0,07
<b>m022262b</b>	CR	Algebra	208	0,21	62	0,06
<b>m022262c</b>	PCR	Algebra	520	0,53	213	0,19

Released items are given in **bold**)

CR : Coded Response / PCR : Coded Response (with partial credit score)

MC : Multiple Choice / MC : Complex Multiple Choice



## APPENDIX B. PROPORTION CORRECTS OF THE PISA & TIMSS ITEMS

PISA 2003 2 <sup>nd</sup> booklet			TIMSS 1999 7 <sup>th</sup> booklet		
Item	TUR n=391	USA n=425	Item	TUR n=980	USA n= 1110
m034q01t	0,2302	0,2965	m012001	0,3571	0,5631
m124q01	0,3453	0,3153	m012002	0,5531	0,7162
m124q03t	0,335	0,4447	m012003	0,5173	0,5829
m145q01t	0,5959	0,6471	m012004	0,4969	0,4577
m150q01	0,4885	0,5365	m012005	0,5112	0,5207
m150q02t	0,6292	0,8141	m012006	0,7439	0,7063
m150q03t	0,2788	0,5624	m012007	0,4	0,7036
m192q01t	0,2762	0,32	m012008	0,6745	0,6459
m305q01	0,4271	0,5224	m012009	0,4714	0,6423
m406q01	0,0818	0,1388	m012010	0,0878	0,482
m406q02	0,0409	0,0894	m012011	0,3469	0,6099
m406q03	0,0997	0,1176	m012012	0,4082	0,7288
m408q01t	0,4092	0,3553	m012019	0,3765	0,5
m411q01	0,2481	0,4706	m012020	0,599	0,7514
m411q02	0,289	0,4847	m012021	0,3847	0,7234
m413q02	0,5038	0,6682	m012022	0,4235	0,491
m413q03t	0,1714	0,3647	m012023	0,8153	0,6189
m423q01	0,8772	0,7694	m012024	0,5827	0,7748
m438q02	0,3274	0,4635	m012043	0,4571	0,7369
m446q01	0,6138	0,3318	m012044	0,5163	0,8523
m446q02	0,0486	0,0518	m012045	0,7041	0,9279
m462q01t	0,2225	0,1529	m012046	0,3071	0,5315
m474q01	0,4936	0,6612	m012047	0,5286	0,5459
m510q01t	0,2864	0,4447	m012048	0,5153	0,8784
m520q01t	0,5064	0,6282	m022135	0,1459	0,4279
m520q02	0,289	0,4871	m022139	0,2276	0,3856
m520q03t	0,2813	0,4847	m022142	0,3041	0,3432
m547q01t	0,711	0,6894	m022144	0,2908	0,6712
m555q02t	0,4604	0,5835	m022146	0,3867	0,5595
m598q01	0,5703	0,5835	m022148	0,2816	0,545
m702q01	0,1816	0,3576	m022253	0,3857	0,5505
m710q01	0,266	0,2965	m022154	0,3204	0,4441
m806q01t	0,5269	0,5835	m022156	0,4276	0,4432
			m022237	0,0265	0,5036
			m022256	0,4684	0,4712
			m022241	0,3286	0,5081
			m022262a	0,4245	0,7162
			m022262b	0,3235	0,609
			M022262c	0,1429	0,3856

## APPENDIX C. ROTATED FACTOR LOADINGS FOR TURKEY AND USA TIMSS DATA

Item	Scale <sup>1</sup>	TR COMPONENTS							USA COMPONENTS						
		TR1	TR2	TR3	TR4	TR5	TR6	TR7	US1	US2	US3	US4	US5	US6	US7
m012001	F&N	0,403			0,268				0,609						
m012002	AL								0,322				0,350		
m012003	MS				0,445				0,466				0,263		
m012004	F&N					0,262			0,537						
m012005	GM				0,640					0,290		0,368		0,398	
m012006	PR			0,437	0,413								0,269	0,262	
m012007	PR			0,614									0,488		
m012008	F&N			0,452									0,707		
m012009	F&N			0,428					0,418	0,255					0,314
m012010	F&N	0,392							0,533			0,267			
m012011	GM	0,376	0,301										0,283		0,491
m012012	AL	0,255			0,338			0,324	0,392	0,375	0,258		0,251		
m012019	GM	0,574							0,281		0,303				0,388
m012020	AL			0,306							0,596				
m012021	F&N	0,396		0,273			0,308		0,570		0,325				
m012022	AL							0,752	0,363						
m012023	MS						0,721				0,281	0,387			
m012024	F&N								0,252		0,524				
m012043	PR					0,629					0,425	0,272	0,310		
m012044	F&N			0,421									0,279	0,414	0,277
m012045	F&N			0,525										0,697	
m012046	AL								0,307			0,391	0,285		
m012047	PR				0,286				0,283			0,492			
m012048	AL			0,288	0,293	0,276							0,473	0,256	
m022135	PR								0,532						
m022139	F&N	0,485				0,266						0,565			
m022142	GM	0,568										0,580			
m022144	F&N										0,577				
m022146	PR	0,259						0,265	0,580			0,274			
m022148	MS	0,412							0,612						
m022253	AL	0,253	0,272	0,392	0,365				0,420	0,257	0,373				
m022154	GM		0,262			0,355	0,386				0,324	0,329			0,410
m022156	F&N						0,415		0,616		0,275	0,271			
m022237	F&N								0,571						
m022256	PR					0,263			0,482				0,272		
m022241	F&N	0,430							0,430						
M022262a	AL		0,840							0,770					
M022262b	AL		0,842							0,783					
M022262c	AL		0,618							,574					

<sup>1</sup> F&N: Fraction and Number Sense, AL: Algebra, PR: Data Representation and Probability, MS: Measurement Loadings less than 0.25 are omitted

## C2. Rotated Factor Loadings for Turkey and USA PISA Data

Item	Scale <sup>2</sup>	TR COMPONENTS							USA COMPONENTS						
		TR1	TR2	TR3	TR4	TR5	TR6	TR7	US1	US2	US3	US4	US5	US6	US7
m034q01t	S&P					0,633			0,446						
m124q01	C&R				0,562				0,531	0,275					
m124q03t	C&R			0,319	0,540		0,262		0,547	0,338					
m145q01t	S&P		0,627						0,305	0,536					
m150q01	C&R			0,468					0,283			0,446		0,262	
m150q02t	C&R			0,669						0,476		0,303			
m150q03t	C&R			0,677						0,479		0,408			
m192q01t	C&R						0,335	0,438	0,458						
m305q01	S&P							0,729				0,715			
m406q01	S&P	0,710									0,737				
m406q02	S&P	0,737									0,751				
m406q03	S&P	0,705									0,647				
m408q01t	U		0,462				0,401						0,762		
m411q01	Q			0,276	0,460				0,391			0,337			
m411q02	U	0,333						0,488	0,480						
m413q02	Q			0,451	0,296	0,368				0,399		0,287		0,377	0,277
m413q03t	Q						0,660		0,433			0,307			
m423q01	U		0,669										0,638		
m438q02	U						0,496		0,464		0,298				
m446q01	C&R		0,520		0,327									0,768	
m446q02	C&R	0,556			0,518				0,256					0,607	
m462q01t	S&P					0,321	0,396		0,619						
m474q01	Q				0,606									0,784	
m510q01t	Q														
m520q01t	Q		0,444			0,327			0,336	0,522					
m520q02	Q	0,359			0,296	0,315		0,292	0,380					0,340	
m520q03t	Q				0,455	0,281		0,349		0,726					
m547q01t	S&P		0,252	0,345	0,361					0,265				0,571	
m555q02t	S&P		0,489		0,266	0,339			0,422	0,439					
m598q01	S&P											0,496			
m702q01	U								0,524			0,271			
m710q01	U														
m806q01t	Q					0,553			0,506				0,315		

Loadings less than 0.25 are omitted

<sup>2</sup> S&P: Space and Shape, U: Uncertainty, Q: Quantity, C&R: Change and Relationships

APPENDIX D. ROTATED COMPONENT MATRIX OF PISA FROM THE  
TOTAL SAMPLE OF USA AND TURKEY.

	Component					
	1	2	3	4	5	6
M555Q02T	,637					
M520Q01T	,589					
M520Q03T	,587					
M124Q03T	,578					
M145Q01T	,561		,330			
M413Q02	,544			,277		
M124Q01	,485	,281	,323			
M150Q02T	,485					,394
M150Q03T	,450			,307		,306
M520Q02	,430	,294				
M806Q01T	,428					,303
M438Q02	,411	,343			,254	
M702Q01	,409				,308	
M462Q01T	,401	,318				
M034Q01T	,387	,285				
M411Q02	,371	,284		,260		
M411Q01	,370			,314	,310	
M192Q01T	,366	,319				
M413Q03T	,355	,262				
M547Q01T	,334		,262			
M474Q01	,318					
M406Q02		,755				
M406Q01		,725				
M406Q03		,650				
M446Q02		,489	,262		,308	
M446Q01			,664			
M423Q01			,599			,283
M408Q01T			,511			
M305Q01				,782		
M150Q01	,349			,405		
M710Q01					,782	
M510Q01T					,403	
M598Q01						,737

Extraction Method: Principal Component Analysis.  
Rotation Method: Varimax with Kaiser Normalization.

Loadings less than 0.25 are omitted

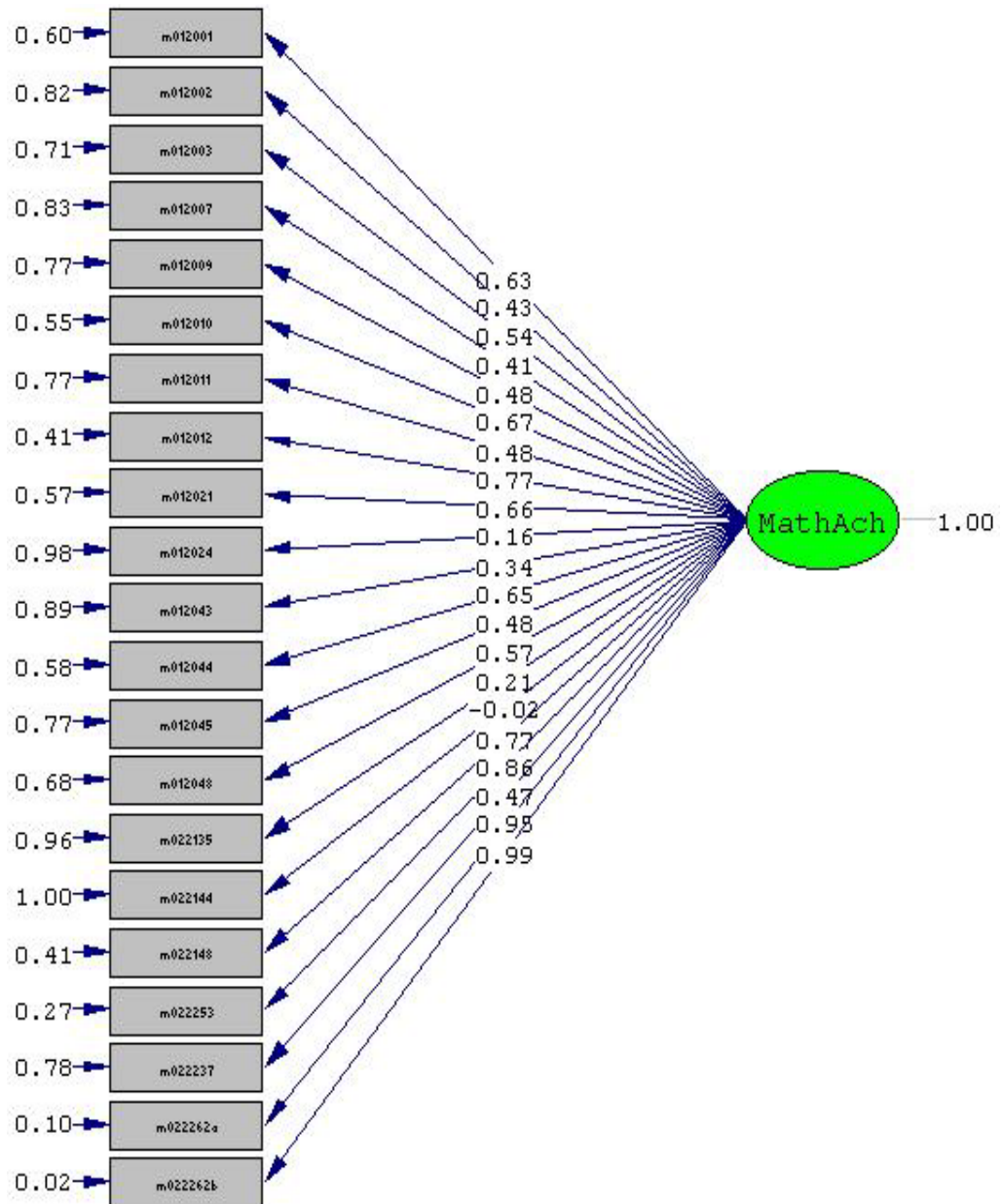
D2. Rotated Component Matrix of TIMSS from the total sample of USA and Turkey.

	Component				
	1	2	3	4	5
M012044	,627				
M012045	,601				
M012048	,581				
M012021	,492	,282			
M012007	,466				
M012012	,427		,273		,341
M012010	,422	,370	,372		
M012001	,409	,286		,295	
M012043	,393				
M012002	,381				
M022148	,378	,328	,273	,289	
M012011	,326	,307			,285
M012009	,314				
M022139		,595			
M022142		,564			
M012019		,530			
M022154		,422			
M022241		,304	,297		
M012020		,294			
M012022			,651		
M012024	,266		,470		
M022237	,458	,299	,458		
M022146		,292	,416	,310	
M022144	,372		,416		
M012046			,381		
M022135	,254	,341	,374		
M012006				,521	
M012023				,485	
M012004		,251		,475	
M012047		,259		,456	
M022156		,325	,341	,423	
M012008				,418	,302
M012003	,310			,395	
M022256			,320	,384	
M022253	,329			,351	,301
M012005				,346	
M022262A	,259				,782
M022262B	,250				,781
M022262C		,309			,553

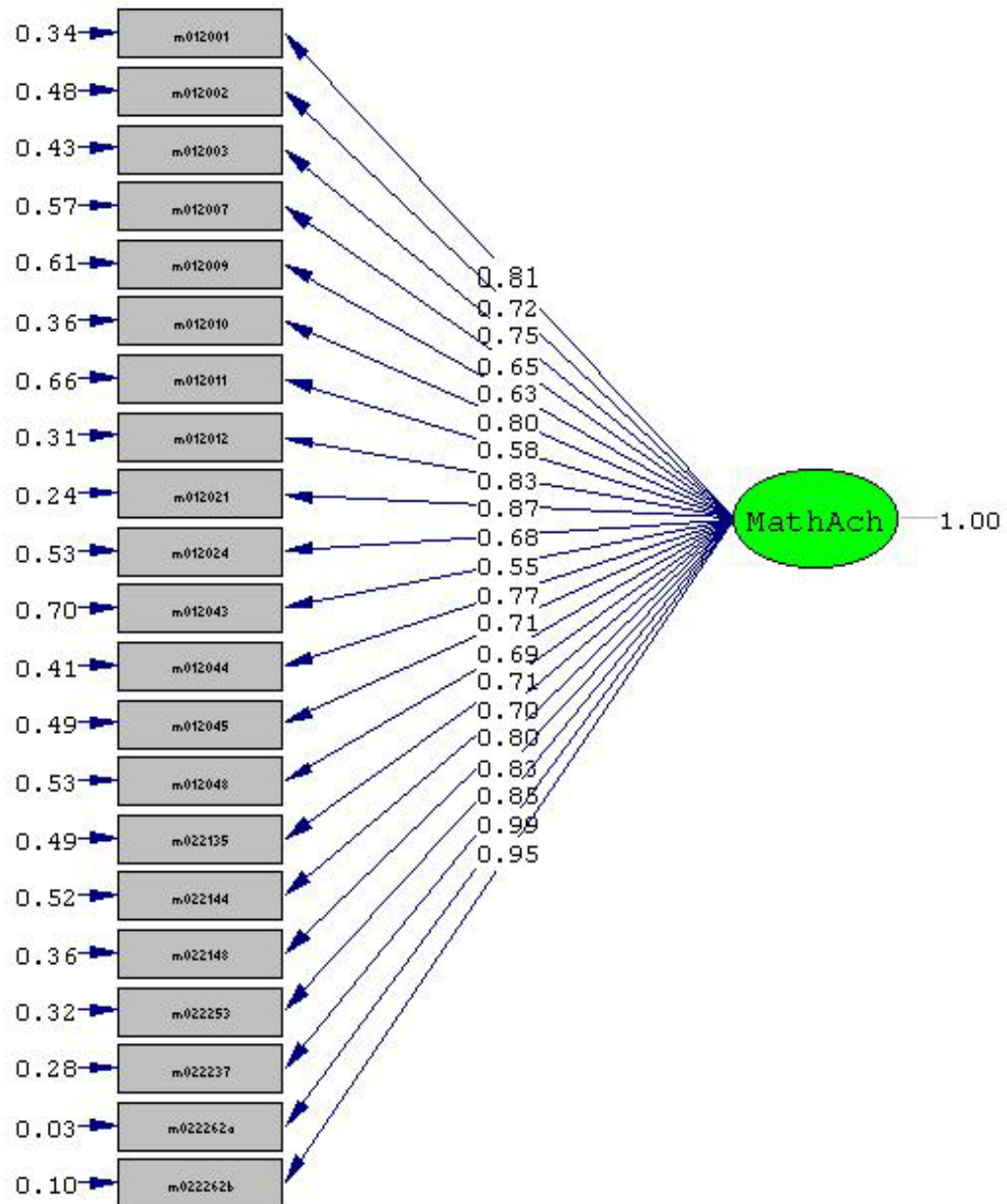
Extraction Method: Principal Component Analysis.  
Rotation Method: Varimax with Kaiser Normalization.

Loadings less than 0.25 are omitted

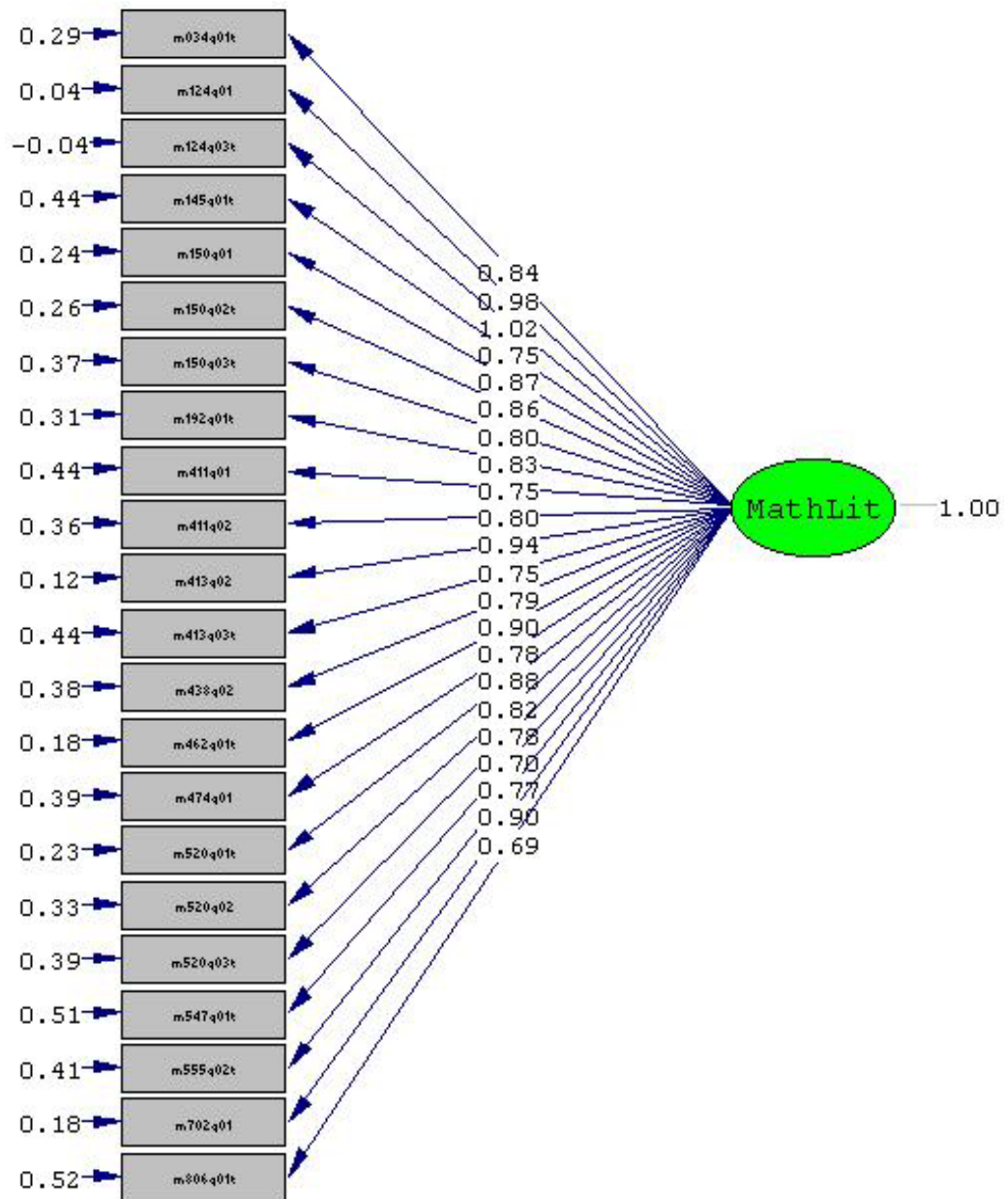
# APPENDIX E. FACTOR LOADINGS AND ERROR VARIANCES OF SELECTED ITEMS OF TIMSS TURKISH DATA



## E2. Factor Loadings and Error Variances of Selected Items of TIMSS American Data

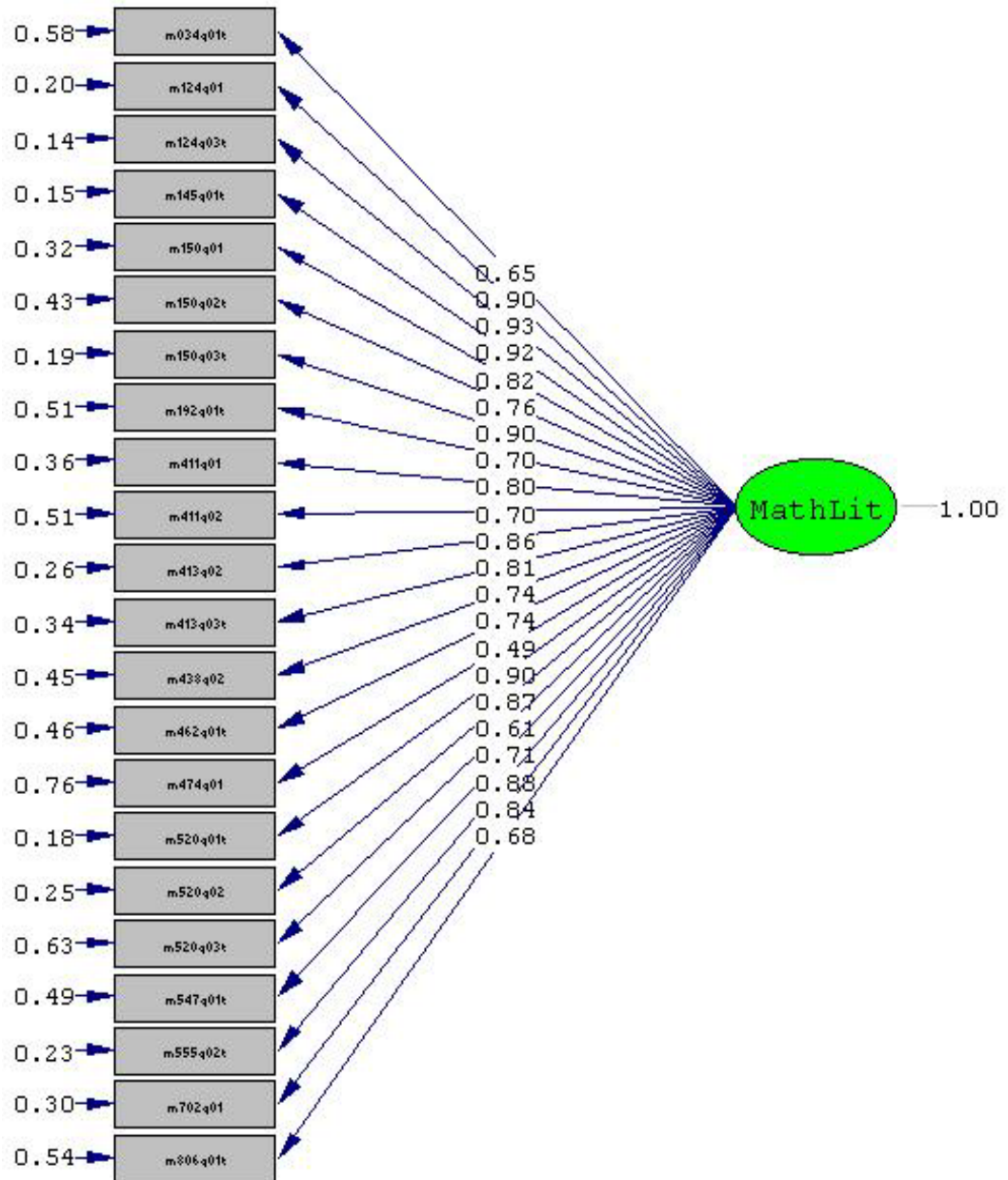


### E3. Factor Loadings and Error Variances of Selected Items of PISA Turkish Data





#### E4. Factor Loadings and Error Variances of Selected Items of PISA American Data



APPENDIX F. SYNTAX IN SIMPLIS COMMAND LANGUAGE USED  
TO TEST STRICT INVARIANCE MODEL IN TIMSS

*Group TUR*

*Observed Variables:*

*m012001 m012002 m012003 m012007 m012009  
m012010 m012011 m012012 m012021 m012024  
m012043 m012044 m012045 m012048 m022135  
m022144 m022148 m022253 m022237 m022262a m022262b*

*Means from File Tims\_tr.ME*

*Covariance Matrix from File Tims\_tr.CM*

*Asymptotic Covariance Matrix from File Tims\_tr.ACC*

*Sample Size: 980*

*Latent Variables: MathAch*

*Relationships:*

*m012001 = CONST 1\*MathAch*

*m012002 - m022262b = CONST MathAch*

*Group USA*

*Observed Variables:*

*m012001 m012002 m012003 m012007 m012009  
m012010 m012011 m012012 m012021 m012024  
m012043 m012044 m012045 m012048 m022135  
m022144 m022148 m022253 m022237 m022262a m022262b*

*Means from File Tims\_usa.ME*

*Covariance Matrix from File Tims\_usa.CM*

*Asymptotic Covariance Matrix from File Tims\_usa.ACC*

*Sample Size: 1110*

*Latent Variables: MathAch*

*Relationships:*

*MathAch = CONST*

*Set the error variances of m012001 - m022262b free*

*Set the variances of MathAch free*

*Method of Estimation: Weighted Least Squares*

*Path Diagram*

*End of Problem*

APPENDIX G. ESTIMATIONS OF THE INTERCEPTS, FACTOR  
LOADINGS AND ERROR VARIANCES IN THE FINAL MODELS OF  
PISA

	INTERCEPTS		FACTOR LOAD.		ERROR VAR.	
	TUR	USA	TUR	USA	TUR	USA
m034q01t	- 0.22	- 0.22	1.00	1.00	0.38	0.45
m124q01	0.25	- 0.53	1.20	1.20	0.13	0.21
m124q03t	- 0.25	- 0.25	1.26	1.26	0.008	0.13
m145q01t	- 0.083	- 0.083	0.95	1.21	0.43	0.28
m150q01	- 0.21	- 0.21	1.07	1.07	0.30	0.38
m150q02t	- 0.15	- 0.15	1.07	1.07	0.29	0.37
m150q03t	- 0.26	- 0.26	0.98	1.21	0.41	0.22
m192q01t	- 0.17	- 0.17	0.96	0.96	0.43	0.51
m411q01	- 0.22	- 0.22	1.03	1.03	0.34	0.43
m411q02	- 0.17	- 0.17	0.95	0.95	0.44	0.51
m413q02	- 0.21	- 0.21	1.16	1.16	0.15	0.26
m413q03t	- 0.23	- 0.23	1.06	1.06	0.30	0.39
m438q02	- 0.19	- 0.19	1.00	1.00	0.37	0.45
m462q01t	0.069	- 0.52	1.07	1.07	0.28	0.37
m474q01	- 0.12	- 0.12	0.97	0.69	0.42	0.75
m520q01t	- 0.21	- 0.21	1.13	1.13	0.19	0.30
m520q02	- 0.29	- 0.29	0.95	1.16	0.42	0.27
m520q03t	- 0.17	- 0.17	0.89	0.89	0.51	0.57
m547q01t	- 0.16	- 0.16	0.89	0.89	0.54	0.60
m555q02t	- 0.27	- 0.27	0.95	1.19	0.44	0.22
m702q01	- 0.33	- 0.33	1.14	1.14	0.18	0.31
m806q01t	- 0.14	- 0.14	0.86	0.86	0.54	0.60

G2. Estimations of The Intercepts, Factor Loadings and Error Variances in  
The Final Models of TIMSS

	INTERCEPTS		FACTOR LOAD.		ERROR VAR.	
	TUR	USA	TUR	USA	TUR	USA
m012001	- 0.29	- 0.69	1.00	1.00	0.68	0.34
m012002	- 0.21	- 0.55	0.87	0.87	0.76	0.51
m012003	- 0.11	- 0.78	0.94	0.94	0.72	0.42
m012007	- 0.38	- 0.38	0.79	0.79	0.80	0.59
m012009	- 0.23	- 0.50	0.79	0.79	0.80	0.59
m012010	- 0.79	- 0.42	1.01	1.01	0.68	0.33
m012011	- 0.36	- 0.36	0.74	0.74	0.82	0.63
m012012	- 0.45	- 0.45	1.35	1.01	0.42	0.33
m012021	- 0.48	- 0.48	1.08	1.08	0.63	0.24
m012024	- 0.35	- 0.35	0.33	0.79	0.97	0.60
m012043	- 0.30	- 0.30	0.64	0.64	0.88	0.74
m012044	- 0.42	- 0.42	0.97	0.97	0.70	0.38
m012045	- 0.35	- 0.35	0.91	0.91	0.74	0.46
m012048	- 0.55	- 0.15	0.88	0.88	0.76	0.49
m022135	- 0.46	- 0.46	0.81	0.81	0.79	0.57
m022144	- 0.47	- 0.47	- 0.045	0.88	1.00	0.50
m022148	- 0.51	- 0.51	1.39	0.96	0.41	0.40
m022253	- 0.17	- 0.75	1.45	1.01	0.34	0.33
m022237	- 1.31	- 0.37	1.02	1.02	0.67	0.31
m022262a	- 0.29	- 0.89	1.67	1.19	0.13	0.100
m022262b	- 0.53	- 0.53	1.72	1.14	0.072	0.17

APPENDIX H. PRELIS SYNTAX USED TO CALCULATE  
CORRELATION MATRIX AND ASYMPTOTIC COVARIANCE  
MATRIX OF TIMSS POOLED DATA

*'USA&TUR\_TIMSS PRELIS Run for RFA*  
*'Computing Tetrachoric correlation to be used in RFA*  
*Data Ninputvariables = 22*  
*Labels*  
*Country m012001 m012002 m012003 m012007*  
*m012009 m012010 m012011 m012012 m012021*  
*m012024 m012043 m012044 m012045 m012048*  
*m022135 m022144 m022148 m022253 m022237*  
*m022262a m022262b*  
*Rawdata=Timss\_USA&TUR.dat*  
*Output BT MA=PM PM=Timss.PM AC=Timss.ACP TH=Timss.THR*

## H2. LISREL Syntax Used to Test the Null Model in RFA of TIMSS Data. Latent Variables Uncorrelated.

*'TIMSS RFA*

*'Country coded 0 for Turkey and 1 Usa*

*Observed Variables:*

*Country m012001 m012002 m012003 m012007*

*m012009 m012010 m012011 m012012 m012021*

*m012024 m012043 m012044 m012045 m012048*

*m022135 m022144 m022148 m022253 m022237*

*m022262a m022262b*

*Correlation Matrix from File Timss.PM*

*Asymptotic Covariance Matrix from File Timss.ACP*

*Sample Size: 2090*

*Latent Variables: MathAch Group*

*Relationships:*

*Country = 1\*Group*

*m012001 - m022262b = MathAch*

*Set the Error Variance of Country equal to 0*

*Set the Correlations of MathAch - Country to 0*

*Method of Estimation: Weighted Least Squares*

*Path Diagram*

*End of Problem*

# APPENDIX I. IRTL RDIF OUTPUT FOR PISA ITEMS

Item	Test	G <sup>2</sup>	d.f.	Reference Group			Focal Group			Focal	
				a	b	c	a	b	c	Mean	s.d.
m145q01t	c	0.00	1	1.27	-0.34	0.1	0.83	-0.75	0.1	-0.56	1.09
m520q02	c	0.00	1	1	0.22	0.09	1.11	0.56	0.09	-0.54	1.09
m520q02	a	0.00	1	1.04	0.22	0.09	1.04	0.59	0.09	-0.54	1.09
m547q01t	a	0.00	1	1.16	-0.86		1.16	-1.56		-0.56	1.09
m555q02t	All	0.00	3	1.21	0	0.16	1.01	-0.17	0.11	-0.55	1.08
m702q01	a	0.00	1	1.79	0.52		1.79	0.78		-0.54	1.09
m150q01	a	0.10	1	1.5	-0.11		1.5	-0.49		-0.56	1.08
m438q02	b	0.10	1	1.87	0.52	0.18	1.87	0.52	0.18	-0.55	1.08
m520q01t	All	0.90	2	1.65	-0.45		1.46	-0.55		-0.55	1.08
m411q02	c	0.30	1	1.05	0.47	0.19	2.39	0.78	0.19	-0.54	1.09
m413q02	b	0.30	1	1.86	-0.55		1.86	-0.55		-0.55	1.08
m192q01t	All	2.00	3	0.93	0.99	0.1	1.06	0.63	0.1	-0.55	1.08
m124q01	a	0.70	1	2.15	0.65		2.15	-0.01		-0.57	1.07
m411q01	a	0.70	1	1.5	0.12		1.5	0.54		-0.54	1.08
m034q01t	All	2.50	2	1.11	0.97		1.4	0.66		-0.55	1.08
m806q01t	All	2.70	2	1.02	-0.39		0.94	-0.68		-0.55	1.08
m124q03t	All	3.60	2	2.3	0.21		2.05	0.06		-0.56	1.08
m462q01t	a	2.00	1	1.78	1.43		1.78	0.54		-0.56	1.07
m438q02	a	2.20	1	1.91	0.52	0.19	1.91	0.55	0.19	-0.55	1.08
m438q02	All	6.10	3	2.79	0.54	0.22	1.2	0.54	0.15	-0.55	1.09
m702q01	All	5.30	2	1.81	0.51		1.76	0.79		-0.54	1.09
m411q02	b	3.30	1	1.41	0.63	0.19	1.41	0.63	0.19	-0.55	1.08
m520q02	All	7.20	3	1.03	0.25	0.11	1.26	0.61	0.12	-0.54	1.09
m150q02t	a	3.70	1	1.31	-1.49		1.31	-1.07		-0.53	1.07
m438q02	c	3.70	1	2.18	0.52	0.2	1.54	0.59	0.2	-0.55	1.09
m150q03t	a	4.00	1	1.3	-0.26		1.3	0.47		-0.53	1.08
m474q01	a	4.40	1	0.79	-0.96		0.79	-0.5		-0.54	1.08
m413q02	All	6.70	2	1.53	-0.65		2.35	-0.51		-0.54	1.06
m145q01t	a	4.60	1	1.01	-0.41	0.09	1.01	-0.72	0.09	-0.55	1.08
m702q01	b	5.30	1	1.84	0.61		1.84	0.61		-0.55	1.08
m474q01	b	5.50	1	0.83	-0.71		0.83	-0.71		-0.55	1.08
m413q03t	a	5.60	1	1.42	0.54		1.42	1		-0.54	1.09
m520q03t	a	6.10	1	0.98	0.08		0.98	0.64		-0.54	1.08
m413q02	a	6.40	1	1.84	-0.58		1.84	-0.53		-0.55	1.08
m145q01t	All	11.30	3	1.29	-0.32	0.12	0.88	-0.64	0.15	-0.56	1.09
m150q01	All	9.80	2	1.46	-0.12		1.55	-0.49		-0.56	1.08
m474q01	All	9.90	2	0.61	-1.2		0.99	-0.49		-0.54	1.07
<b>m150q02t</b>	b	8.00	1	1.37	-1.24		1.37	-1.24		-0.55	1.08
<b>m145q01t</b>	b	8.30	1	1.03	-0.48	0.13	1.03	-0.48	0.13	-0.55	1.08
<b>m520q02</b>	b	8.70	1	1.08	0.37	0.09	1.08	0.37	0.09	-0.55	1.08
<b>m150q02t</b>	All	11.70	2	1.08	-1.69		1.57	-1.01		-0.53	1.06
<b>m150q01</b>	b	9.70	1	1.4	-0.3		1.4	-0.3		-0.55	1.08
<b>m411q01</b>	All	12.70	2	1.6	0.12		1.38	0.59		-0.54	1.09
<b>m411q02</b>	All	15.10	3	1.02	0.45	0.18	2.74	0.78	0.19	-0.54	1.09
<b>m520q03t</b>	b	11.10	1	1.04	0.3		1.04	0.3		-0.55	1.08
<b>m413q03t</b>	b	11.20	1	1.49	0.7		1.49	0.7		-0.55	1.08
<b>m411q02</b>	a	11.60	1	1.55	0.63	0.23	1.55	0.79	0.23	-0.55	1.09
<b>m411q01</b>	b	12.00	1	1.56	0.29		1.56	0.29		-0.55	1.08
<b>m413q03t</b>	All	16.70	2	1.73	0.49		1.08	1.28		-0.54	1.1
<b>m520q03t</b>	All	17.10	2	0.77	0.09		1.3	0.45		-0.54	1.07
<b>m547q01t</b>	All	19.50	2	1.16	-0.85		1.15	-1.57		-0.56	1.09
<b>m547q01t</b>	b	19.50	1	1.01	-1.29		1.01	-1.29		-0.55	1.08
<b>m150q03t</b>	b	30.10	1	1.37	0.06		1.37	0.06		-0.55	1.08
<b>m150q03t</b>	All	34.10	2	1.53	-0.23		1.05	0.61		-0.53	1.09
<b>m124q01</b>	All	44.70	2	2.32	0.64		1.98	0.01		-0.57	1.07
<b>m462q01t</b>	All	48.20	2	1.53	1.55		2.05	0.48		-0.56	1.06
<b>m124q01</b>	b	44.00	1	1.76	0.4		1.76	0.4		-0.55	1.08
<b>m462q01t</b>	b	46.20	1	1.4	1.19		1.4	1.19		-0.55	1.08

## 12. IRTL RDIF Output for TIMSS Items

Item	Test	G <sup>2</sup>	d.f.	Reference Group			Focal Group			Focal	
				a	b	c	a	b	c	Mean	s.d.
m012001	a	0.00	1	1.49	0.01	0.11	1.49	-0.57	0.11	-1.23	0.79
m012003	c	0.00	1	1.07	-0.11	0.09	0.82	-1.11	0.09	-1.24	0.81
m012010	c	0.00	1	1.13	0.15	0.04	2.59	0.16	0.04	-1.21	0.81
m012010	b	0.00	1	1.42	0.21	0.03	1.42	0.21	0.03	-1.21	0.81
m012011	b	0.00	1	0.76	-0.2	0.13	0.76	-0.2	0.13	-1.21	0.81
m012024	c	0.00	1	0.94	-0.75	0.25	0.24	-0.66	0.25	-1.22	0.81
m012045	c	0.00	1	0.88	-2.14	0.16	0.85	-1.74	0.16	-1.21	0.8
m012045	a	0.00	1	0.86	-2.16	0.16	0.86	-1.73	0.16	-1.21	0.8
m022144	c	0.00	1	1.11	-0.13	0.28	4.41	0.9	0.28	-1.21	0.83
m012007	c	0.10	1	0.79	-0.61	0.15	0.74	-0.36	0.15	-1.21	0.81
m022135	a	0.10	1	1.19	0.45	0.1	1.19	0.71	0.1	-1.21	0.81
m022262a	a	0.30	1	1.89	-0.75		1.89	-0.98		-1.23	0.81
m012009	a	0.40	1	0.91	-0.1	0.24	0.91	-0.57	0.24	-1.22	0.81
m012043	a	0.40	1	0.95	-0.22	0.36	0.95	-0.1	0.36	-1.22	0.81
m012021	a	0.60	1	1.53	-0.53	0.14	1.53	-0.6	0.14	-1.22	0.81
m012007	b	0.90	1	0.82	-0.53	0.14	0.82	-0.53	0.14	-1.21	0.81
m012002	c	1.20	1	0.88	-0.67	0.13	0.57	-1.14	0.13	-1.22	0.81
m012011	All	3.90	3	0.81	0	0.21	1.02	-0.29	0.15	-1.21	0.8
m012021	b	1.40	1	1.48	-0.57	0.14	1.48	-0.57	0.14	-1.21	0.81
m012007	All	4.30	3	0.8	-0.58	0.17	0.72	-0.38	0.14	-1.21	0.81
m012011	a	1.50	1	0.76	-0.2	0.13	0.76	-0.2	0.13	-1.22	0.81
m012024	b	1.60	1	1.27	-0.17	0.49	1.27	-0.17	0.49	-1.21	0.81
m012044	b	1.70	1	1.14	-1.16	0.11	1.14	-1.16	0.11	-1.21	0.81
m012009	c	2.30	1	0.92	-0.07	0.25	1	-0.56	0.25	-1.22	0.81
m012012	b	2.30	1	1.38	-0.61	0.15	1.38	-0.61	0.15	-1.21	0.81
m012011	c	2.40	1	0.78	-0.08	0.18	1.03	-0.19	0.18	-1.21	0.8
m012044	a	2.70	1	1.19	-1.13	0.17	1.19	-1.04	0.17	-1.21	0.81
m022148	a	2.70	1	1.91	-0.12		1.91	-0.52		-1.23	0.8
m022135	c	2.90	1	1.18	0.46	0.1	1.37	0.6	0.1	-1.21	0.81
m022237	a	3.00	1	2	-0.01		2	1.17		-1.2	0.84
m012021	All	6.70	3	1.89	-0.39	0.22	1.32	-0.65	0.11	-1.22	0.81
m012003	a	3.10	1	0.94	-0.16	0.08	0.94	-1.14	0.08	-1.23	0.8
m012012	c	3.10	1	1.3	-0.58	0.15	1.9	-0.67	0.15	-1.22	0.8
m012007	a	3.20	1	0.86	-0.47	0.19	0.86	-0.37	0.19	-1.22	0.81
m012043	b	4.00	1	0.86	-0.34	0.29	0.86	-0.34	0.29	-1.21	0.81
m022135	All	8.50	3	1.12	0.44	0.09	1.91	0.53	0.12	-1.21	0.81
m012021	c	4.70	1	1.55	-0.53	0.14	1.44	-0.61	0.14	-1.22	0.81
m012048	a	5.50	1	1.45	-0.95	0.33	1.45	-0.78	0.33	-1.21	0.81
m022135	b	5.50	1	1.28	0.49	0.1	1.28	0.49	0.1	-1.21	0.81
m012043	All	11.20	3	0.77	-0.48	0.3	0.49	-0.32	0.18	-1.21	0.81
m012043	c	6.90	1	0.93	-0.19	0.38	1.18	-0.1	0.38	-1.22	0.81
m012012	a	7.00	1	1.45	-0.55	0.15	1.45	-0.65	0.15	-1.22	0.81
m012001	c	7.50	1	1.43	0	0.11	1.53	-0.58	0.11	-1.23	0.79
m012012	All	12.40	3	1.3	-0.58	0.15	2.56	-0.62	0.2	-1.22	0.8
<b>m022262a</b>	All	10.50	2	1.94	-0.74		1.82	-0.98		-1.23	0.81
<b>m012045</b>	All	12.90	3	0.88	-2.12	0.18	0.87	-1.69	0.18	-1.21	0.8
<b>m012002</b>	a	8.50	1	0.72	-0.83	0.09	0.72	-1.23	0.09	-1.23	0.8
<b>m012044</b>	All	13.60	3	0.97	-1.34	0.17	1.23	-1.03	0.13	-1.2	0.8
<b>m012024</b>	All	13.70	3	0.93	-0.78	0.23	0.22	-0.96	0.2	-1.22	0.81
<b>m012044</b>	c	9.20	1	1.12	-1.15	0.21	1.37	-0.97	0.21	-1.21	0.8
<b>m022262b</b>	a	9.70	1	1.83	-0.35		1.83	-0.65		-1.23	0.81
<b>m022262a</b>	b	10.20	1	1.73	-0.88		1.73	-0.88		-1.21	0.81
<b>m012002</b>	b	11.00	1	0.79	-0.66	0.25	0.79	-0.66	0.25	-1.21	0.81
<b>m012024</b>	a	12.20	1	1.07	-0.35	0.46	1.07	-0.06	0.46	-1.21	0.81
<b>m012045</b>	b	13.20	1	0.96	-1.81	0.16	0.96	-1.81	0.16	-1.21	0.81
<b>m012048</b>	b	13.30	1	1.31	-1.02	0.22	1.31	-1.02	0.22	-1.21	0.81
<b>m012002</b>	All	20.70	3	0.9	-0.66	0.13	0.71	-0.85	0.24	-1.22	0.81
<b>m012009</b>	All	21.00	3	1.01	-0.01	0.27	0.86	-0.71	0.18	-1.22	0.8



Continued

Item	Test	G <sup>2</sup>	d.f.	Reference Group			Focal Group			Focal	
				a	b	c	a	b	c	Mean	s.d.
<b>m022262b</b>	b	16.20	1	1.62	-0.49		1.62	-0.49		-1.21	0.81
<b>m012009</b>	b	18.30	1	0.72	-0.39	0.19	0.72	-0.39	0.19	-1.21	0.81
<b>m012010</b>	All	24.60	3	1.19	0.19	0.07	2.45	0.17	0.03	-1.21	0.81
<b>m012048</b>	c	21.30	1	1.31	-0.98	0.37	1.85	-0.71	0.37	-1.21	0.8
<b>m022262b</b>	All	25.90	2	1.61	-0.38		2.3	-0.71		-1.22	0.79
<b>m022253</b>	a	25.70	1	1.91	-0.13		1.91	-0.88		-1.24	0.81
<b>m012010</b>	a	26.10	1	1.44	0.2	0.05	1.44	0.31	0.05	-1.21	0.81
<b>m022148</b>	All	32.20	2	1.79	-0.13		2.15	-0.56		-1.23	0.79
<b>m022144</b>	a	28.70	1	1.42	0.07	0.37	1.42	17.03	0.37	-1.21	0.83
<b>m022148</b>	b	29.50	1	1.61	-0.28		1.61	-0.28		-1.21	0.81
<b>m012048</b>	All	40.10	3	1	-1.48	0.21	1.32	-0.81	0.23	-1.2	0.8
<b>m022144</b>	b	55.00	1	1.74	0.19	0.3	1.74	0.19	0.3	-1.21	0.81
<b>m012001</b>	All	66.90	3	1.46	0.01	0.11	2.59	-0.49	0.18	-1.23	0.79
<b>m012001</b>	b	59.60	1	1.17	-0.16	0.14	1.17	-0.16	0.14	-1.21	0.81
<b>m022144</b>	All	83.50	3	1.04	-0.2	0.25	4.48	0.87	0.29	-1.21	0.83
<b>m022253</b>	b	100.60	1	1.35	-0.47		1.35	-0.47		-1.21	0.81
<b>m012003</b>	All	118.10	3	1.09	-0.09	0.1	0.96	-0.94	0.18	-1.23	0.81
<b>m012003</b>	b	115.90	1	0.71	-0.38	0.18	0.71	-0.38	0.18	-1.21	0.81
<b>m022253</b>	All	126.30	2	1.55	-0.16		2.77	-0.92		-1.23	0.78
<b>m022237</b>	All	141.60	2	2.1	0		1.45	1.73		-1.2	0.85
<b>m022237</b>	b	138.60	1	2.62	0.19		2.62	0.19		-1.21	0.81

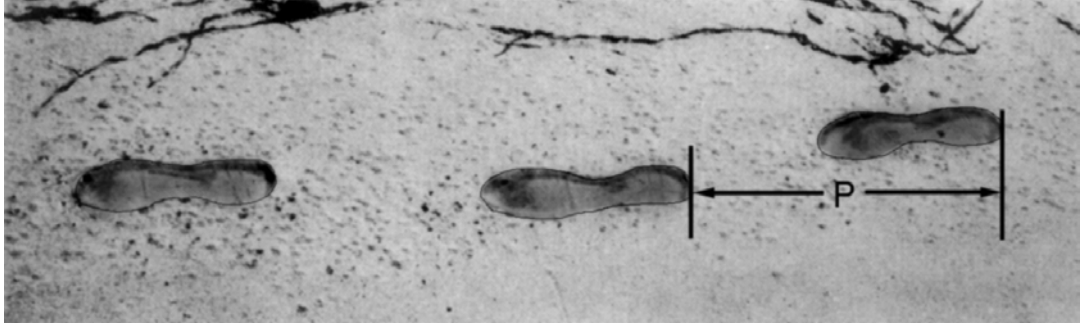
### 13.EZDIF Program MH Results Before Purification

PISA 2003 2 <sup>nd</sup> booklet (22 Items)				TIMSS 1999 7 <sup>th</sup> booklet (21 Items)			
ITEM	DIF	$\alpha$	MH D-DIF	ITEM	DIF	$\alpha$	MH D-DIF
m034q01t	A	0.83	0.437	<b>m012001</b>	CF	0.435	1.955
<b>m124q01</b>	CF	0.306	2.783	<b>m012002</b>	B	0.638	1.057
m124q03t	A	0.782	0.579	<b>m012003</b>	CF	0.281	2.984
m145q01t	B	0.603	1.187	m012007	A	1.349	-0.703
m150q01	B	0.599	1.203	<b>m012009</b>	B	0.649	1.016
<b>m150q02t</b>	B	1.858	-1.456	<b>m012010</b>	CR	2.657	-2.297
<b>m150q03t</b>	CR	2.592	-2.239	m012011	A	1.064	-0.146
m192q01t	A	0.705	0.821	m012012	A	1.064	-0.146
<b>m411q01</b>	B	1.932	-1.548	m012021	A	1.144	-0.317
<b>m411q02</b>	B	1.72	-1.274	m012024	A	1.031	-0.071
m413q02	A	1.173	-0.374	<b>m012043</b>	A	1.43	-0.84
<b>m413q03t</b>	B	1.929	-1.544	<b>m012044</b>	CR	1.966	-1.588
m438q02	A	1.135	-0.297	<b>m012045</b>	CR	2.175	-1.826
<b>m462q01t</b>	CF	0.241	3.343	<b>m012048</b>	CR	2.651	-2.291
m474q01	A	1.458	-0.887	m022135	A	1.315	-0.643
m520q01t	A	0.919	0.197	<b>m022144</b>	CR	2.24	-1.895
m520q02	B	1.635	-1.156	<b>m022148</b>	B	0.606	1.176
<b>m520q03t</b>	B	1.743	-1.306	<b>m022253</b>	CF	0.345	2.503
<b>m547q01t</b>	B	0.486	1.694	<b>m022237</b>	CR	11.56	-5.752
m555q02t	A	0.927	0.178	m022262a	A	0.85	0.382
m702q01	B	1.644	-1.168	m022262b	A	0.771	0.61
m806q01t	A	0.793	0.545				

## APPENDIX J. RELEASED TURKISH DIF ITEMS IN PISA

Item No: M124q01

### YÜRÜYÜŞ



Resim, yürüyen bir erkeğin ayak izlerini gösteriyor. Adım uzunluğu  $P$ . ardışık iki ayak izinin topukları arasındaki mesafedir.

Erkekler için,  $n$  ile  $P$  arasındaki ilişki yaklaşık olarak  $\frac{n}{P} = 140$  formülü ile gösterilmektedir.

Burada;

$n$  = bir dakikadaki adım sayısı

$P$  = metre cinsinden adım uzunluğunu göstermektedir.

#### Soru 1: YÜRÜYÜŞ

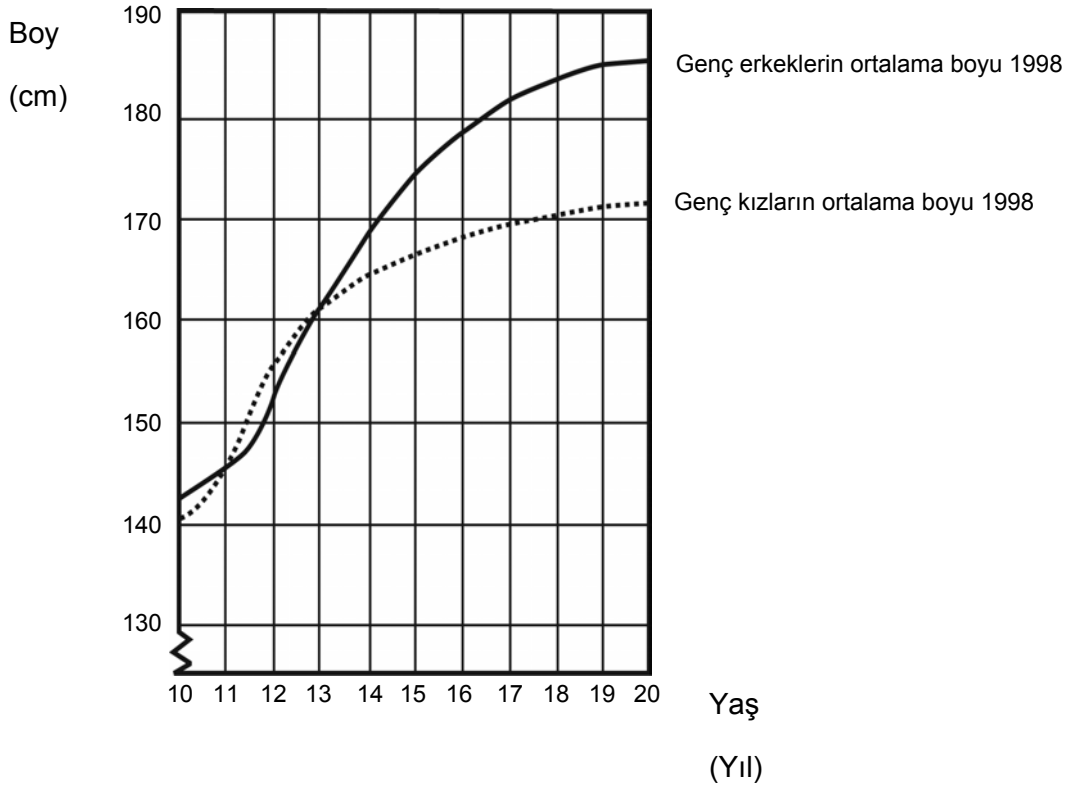
M124Q01- 0 1 2 9

Dakikada 70 adım atarak yürüyen Hakkı'ya bu formül uygulandığında, Hakkı'nın bir adım uzunluğu ne olur? İşleminizi gösteriniz.

## BÜYÜME

### YENİ KUŞAK GENÇLERİN BOYU DAHA UZUN OLUYOR

1998 yılında. Hollanda'daki hem genç erkeklerin hem de genç kızların ortalama boyları aşağıdaki grafikte gösterilmiştir.



**Soru 1: BÜYÜME***M150Q01- 0 1 9*

1980'den bu yana. 20 yaşındaki kızların ortalama boyu 2.3 cm artmış ve 170.6 cm'ye ulaşmıştır. 20 yaşındaki kızların 1980 yılındaki ortalama boyu kaç cm. idi?

Yanıt: .....cm

**Soru 3: BÜYÜME***M150Q03- 01 02 11 12 13 99*

12 yaşından sonra ortalama olarak kızların büyüme hızlarındaki yavaşlamayı grafiğin nasıl gösterdiğini açıklayınız.

**Soru 2: BÜYÜME***M150Q02- 00 11 21 22 99*

Bu grafiğe göre. ortalama olarak. yaşamlarının hangi döneminde kızlar aynı yaştaki erkeklerden daha uzundur?

J3. Item No: M413q03

### DÖVİZ KURU

Singapur'dan Mei-Ling karşılıklı değişim öğrencisi olarak 3 ay süreyle Güney Afrika'ya gitmek için hazırlık yapıyordu. Onun. bir miktar Singapur dolarını (SGD) Güney Afrika para birimi olan randa (GAR) çevirmesi gerekti

#### Soru 3: DÖVİZ KURU

*M413Q03 - 01 02 11 99*

Bu 3 ay süresince döviz kuru oranı bir SGD için 4.2'den 4.0 GAR'a değişmiştir.

Mei-Ling Güney Afrika randını yeniden Singapur dolarına çevirdiğinde. döviz kurunun 4.2 GAR yerine 4.0 GAR olması Mei-Ling'in yararına mı olmuştur? Yanıtınızı destekleyecek bir açıklama yazınız

J4. Item No: M520q03

## KAYKAY

Ercan koyu bir kaykay meraklısıdır. O. bazı fiyatları öğrenmek için

KAYKAYCILAR adlı mağazaya gidiyor

Bu mağazada bütün halde bir kaykay satın alabilirsiniz. Ya da bir kaykay tahtası.

bir tane 4'lü tekerlek seti. bir 2'li tekerlek mili seti ve bir kaykay birleştirme setini

satın alabilir ve bunları birleştirerek kendi kaykayınızı yapabilirsiniz

Mağazanın ürün fiyatları şöyledir:

Ürün	Zed cinsi fiyat	
Bütün olarak bir kaykay	82 ya da 84	
Kaykay Tahtası	40. 60 ya da 65	
Bir tane 4'lü tekerlek seti	14 ya da 36	
Bir tane 2'li tekerlek mili seti	16	
Bir tane kaykay birleştirme seti (mil yatakları. lastik destek gereçleri. civatalar ve vida somunları)	10 ya da 20	

**Soru 3: KAYKAY**

M520Q03

Ercan'ın harcayabileceği 120 zed'i var ve elindeki parayla alabileceği en pahalı kaykayı satın almak istiyor.

Ercan. 4 parçanın her birine ne kadar para harcayabilir? Yanıtlarınızı aşağıdaki çizelgeye yazınız.

Parça	Miktar (zed)
Kaykay Tahtası	
Tekerlekler	
tekerlek Milleri	
Kaykay Birleştirme Gereçleri	

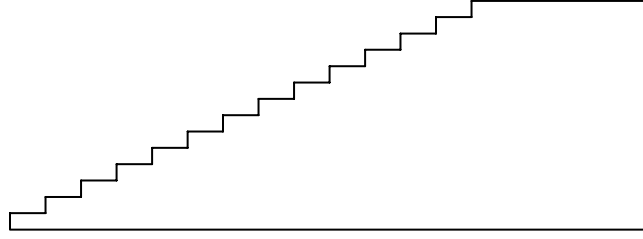


J5. Item No: M547q01

## MERDİVEN

### Soru 1 MERDİVEN

M547Q01



Toplam yükseklik 252 cm

Toplam genişlik 400 cm

14 basamağın her birinin yüksekliği nedir?

Yükseklik: .....cm.

APPENDIX K. RELEASED TURKISH DIF ITEMS IN TIMSS

K1. Item No: M012010

**Alan:** Kesirler ve sayıların anlamı

**Madde Tanımı:** En küçük ondalık kesir

**Madde:** Aşağıdakilerden hangisi en küçük sayıdır?

A. 0.625

B. 0.25

C. 0.375

D. 0.5

☒ E. 0.125

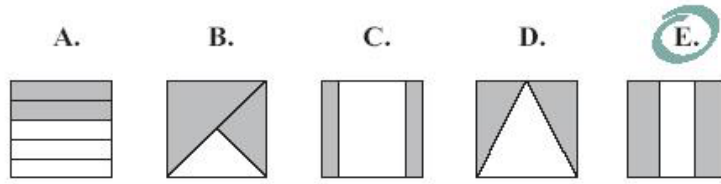
K2. Item No: M012044

**Alan:** Kesirler ve sayıların anlamı

**Madde Tanımı:** Verilen bir kesrin şekilsel gösterimi

**Madde:**

Hangisi taranmış karenin  $\frac{2}{3}$  ünü göstermektedir?



K3. Item No: M012045

**Alan:** Kesirler ve sayıların anlamı

**Madde Tanımı:** 691+208 toplamına en yakın toplama işlemi

**Madde:** 691 + 208 toplamı şu toplamlardan hangisine en yakındır ?

- A. 600 + 200  
B. 700 + 200  
C. 700 + 300  
D. 900 + 200

K4. Item No: M012048

Alan: Cebir

Madde Tanımı: Dergilerin sembolik doğrusal denklemi

**Madde:**  $\square$  Leman'ın her hafta okuduğu dergilerin sayısını göstermektedir. Aşağıdakilerden hangisi Leman'ın 6 haftada okuduğu dergilerin toplam sayısını göstermektedir?

- A.  $6 + \square$
- ☒ B.  $6 \times \square$
- C.  $\square + 6$
- D.  $(\square + \square) \times 6$

K5. Item No: M012237

Alan: Kesirler ve sayıların anlamı

Madde Tanımı: Çocuğun gerçek boy uzunluğuyla ilgili iki olasılığı yazma

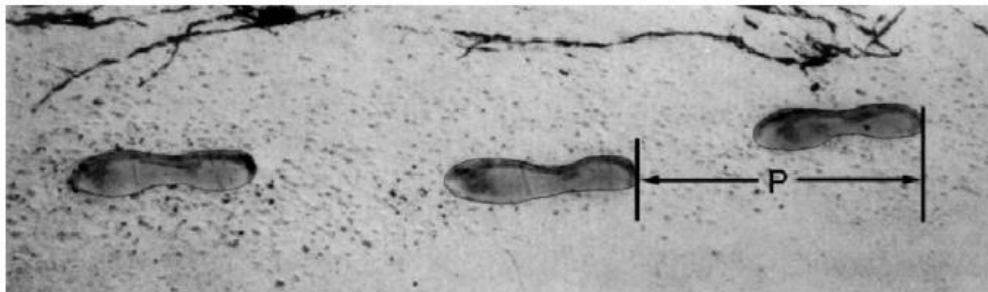
**Madde:** Bir çocuğun boyu 140 cm olarak bildirildi. Boy en yakın 10 cm'e yuvarlanmıştır. Çocuğun gerçek boyu ile ilgili iki olasılık nedir ?

Cevap: \_\_\_\_\_ cm ve \_\_\_\_\_ cm

## APPENDIX L RELEASED ENGLISH DIF ITEMS IN PISA

Item No: M124q01

### WALKING



The picture shows the footprints of a man walking. The pacelength  $P$  is the distance between the rear of two consecutive footprints.

For men, the formula,  $\frac{n}{P} = 140$ , gives an approximate relationship between  $n$  and  $P$  where,

$n$  = number of steps per minute, and

$P$  = pacelength in metres.

#### Question 1: WALKING

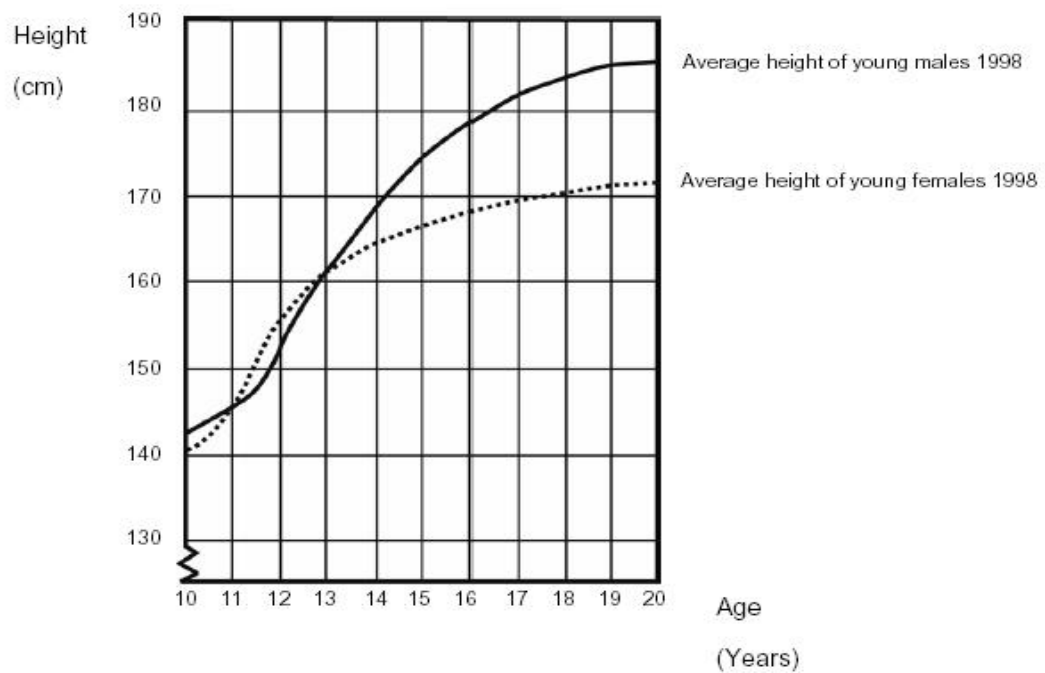
M124Q01- 0 1 2 9

If the formula applies to Heiko's walking and Heiko takes 70 steps per minute, what is Heiko's pacelength? Show your work.

## GROWING UP

### YOUTH GROWS TALLER

In 1998 the average height of both young males and young females in the Netherlands is represented in this graph.



### Question 3: GROWING UP

M150Q03- 01 02 11 12 13 99

Explain how the graph shows that on average the growth rate for girls slows down after 12 years of age.

.....

.....

.....

Item No: M150q01, M150q02

**Question 1: GROWING UP**

M150Q01- 0 1 9

Since 1980 the average height of 20-year-old females has increased by 2.3 cm, to 170.6 cm. What was the average height of a 20-year-old female in 1980?

Answer: .....cm

**Question 2: GROWING UP**

M150Q02- 00 11 21 22 99

According to this graph, on average, during which period in their life are females taller than males of the same age?

.....

.....

Item No: M413q03

## EXCHANGE RATE

Mei-Ling from Singapore was preparing to go to South Africa for 3 months as an exchange student. She needed to change some Singapore dollars (SGD) into South African rand (ZAR).

### Question 3: EXCHANGE RATE

M413Q03 - 01 02 11 99

During these 3 months the exchange rate had changed from 4.2 to 4.0 ZAR per SGD.

Was it in Mei-Ling's favour that the exchange rate now was 4.0 ZAR instead of 4.2 ZAR, when she changed her South African rand back to Singapore dollars? Give an explanation to support your answer.




Item No: M520q03

## SKATEBOARD

Eric is a great skateboard fan. He visits a shop named SKATERS to check some prices.

At this shop you can buy a complete board. Or you can buy a deck, a set of 4 wheels, a set of 2 trucks and a set of hardware, and assemble your own board.

The prices for the shop's products are:

Product	Price in zeds	
Complete skateboard	82 or 84	
Deck	40, 60 or 65	
One set of 4 Wheels	14 or 36	
One set of 2 Trucks	16	
One set of hardware (bearings, rubber pads, bolts and nuts)	10 or 20	

### Question 3: SKATEBOARD

M520Q03

Eric has 120 zeds to spend and wants to buy the most expensive skateboard he can afford.

How much money can Eric afford to spend on each of the 4 parts? Put your answer in the table below.

Part	Amount (zeds)
Deck	
Wheels	
Trucks	
Hardware	

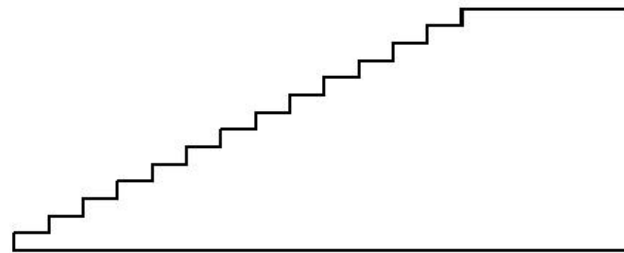
Item No: M547q01

## STAIRCASE

### Question 1: STAIRCASE

M547Q01

The diagram below illustrates a staircase with 14 steps and a total height of 252 cm:



Total height 252 cm

Total depth 400 cm

What is the height of each of the 14 steps?

Height: ..... cm.

## APPENDIX M RELEASED ENGLISH DIF ITEMS IN TIMSS

Item No: M012010

Smallest decimal fraction					B10
Content Category	Performance Expectation	Item Key	Score Points	International Average Percentage of 8th Grade Students Responding Correctly	Used in 1995
Fractions and Number Sense	Knowing	E	1	46	Y

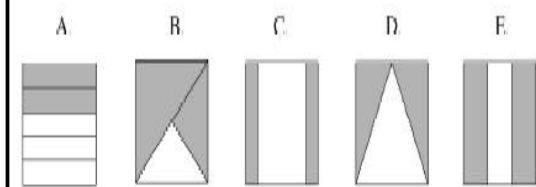
Which of these is the smallest number?

- A. 0.625
- B. 0.25
- C. 0.375
- D. 0.5
- E. 0.125

Item No: M012044

Figure showing fraction of shaded square					H08
Content Category	Performance Expectation	Item Key	Score Points	International Average Percentage of 8th Grade Students Responding Correctly	Used in 1995
Fractions and Number Sense	Knowing	E	1	68	Y

Which shows  $\frac{2}{3}$  of the square shaded?



Item No: M012045

Sum closest to $691 + 208$					H09
Content Category	Performance Expectation	Item Key	Score Points	International Average Percentage of 8th Grade Students Responding Correctly	Used in 1995
Fractions and Number Sense	Using Complex Procedures	B	1	80	Y

The sum  $691 + 208$  is closest to the sum

- A.  $600 + 200$
- B.  $700 + 200$
- C.  $700 + 300$
- D.  $900 + 200$

Item No: M012048

Symbolic linear equation of magazines					H12
Content Category	Performance Expectation	Item Key	Score Points	International Average Percentage of 8th Grade Students Responding Correctly	Used in 1995
Algebra	Knowing	B	1	72	Y

$\square$  represents the number of magazines that Lina reads each week. Which of these represents the total number of magazines that Lina reads in 6 weeks?

- A.  $6 + \square$
- B.  $6 \times \square$
- C.  $\square + 6$
- D.  $(\square + \square) \times 6$

Item No: M012237

Two possibilities for actual height					V01
Content Category	Performance Expectation	Item Key	Score Points	International Average Percentage of 8th Grade Students Responding Correctly	Used in 1995
Fractions and Number Sense	Using Complex Procedures	Rubric	1	44	N

<p>The height of a boy was reported as 140 cm. The height had been rounded to the nearest 10 cm. What are two possibilities for the boy's actual height.</p> <p>Answer: _____ cm and _____ cm</p>	Code	Response	Item: M022237
		Correct Response	
	10	One answer 140 and the other answer is in an acceptable range, $135 \leq x < 140$ or $140 < x \leq 145$ .	
	11	Neither answer is 140 cm but both are in the acceptable range $135 \leq x < 140$ and/or $140 < x \leq 145$ .	
		Incorrect Response	
	70	Both answers within the intervals $145 \leq x \leq 150$ and/or $130 \leq x < 135$	
	71	130 AND 150	
	79	Other incorrect (including crossed out/erased, stray marks, illegible, or off task)	
		Nonresponse	
	99	BLANK	

## CURRICULUM VITAE

### PERSONAL INFORMATION

Surname, Name: Yıldırım, Hüseyin Hüsnü  
Nationality: Turkish (T.C.)  
Date and Place of Birth: 13 April 1974, Kars  
Marital Status: Single  
Phone: +90 312 210 36 59  
Fax: +90 312 210 12 57  
email: yhuseyin@metu.edu.tr

### EDUCATION

Degree	Institution	Year of Graduation
MS	Marmara Univ Mathmetics Teaching	2000
BS	Marmara Univ. Mathmetics Teaching	1997
High School	H. Avni Sözen Anatolian High School, İstanbul	1992

### WORK EXPERIENCE

Year	Place	Enrollment
2001- Present	METU Department of SSME	Research Assistant
2000 - 2001	MEB İlköğretim Okulu, İstanbul	Mathematics Teacher
1997 - 2000	Özel Kalamış Lisesi	Mathematics Teacher

### PUBLICATIONS

1.Yıldırım, H.H., Çömlekoğlu, G. & Berberoğlu, G. (2003). The fit of Ministry of National Education Private School Examination Data to Item Response Theory Models. *Hacettepe Journal of Education*, 24, pp. 159-168

2. Yıldırım, H.H. & Berberoğlu, G. (2005) Comparison of L-R and Mantel Haenszel Methods in Evaluating Mathematics Items of TIMSS and PISA projects across Turkish and English languages. Paper presented at the annual meeting of ECER, Dublin.

### HOBBIES

Music, Play Bağlama