A MULTIVARIATE ANALYSIS IN DETECTING DIFFERENTIALLY FUNCTIONING ITEMS THROUGH THE USE OF PROGRAMME FOR INTERNATIONAL STUDENT ASSESSMENT (PISA) 2003 MATHEMATICS LITERACY ITEMS

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

BY

SELDA ÇET

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: Selda, ÇET

Signature :

ABSTRACT

A MULTIVARIATE ANALYSIS IN DETECTING DIFFERENTIALLY FUNCTIONING ITEMS THROUGH THE USE OF PROGRAMME FOR INTERNATIONAL STUDENT ASSESSMENT (PISA) 2003 MATHEMATICS LITERACY ITEMS

Çet, Selda
Ph.D., Department of Secondary Science and Mathematics Education
Supervisor: Prof. Dr. Giray BERBEROĞLU

April 2006, 146 pages

Differential item functioning (DIF) analyses investigates whether individuals with same ability in different groups also show similar performance on an item. In matching the individuals of the same ability, most of the methodologies use total scores of the tests which are usually constructed to be unidimensional. The purpose of the present study is evaluating the PISA 2003 mathematical literacy items trough the use of DIF methodology which uses a multidimensional approach in matching students instead of a single total test score, improve the matching for DIF analyses.

In the study, factor structures of the tests will be determined via both exploratory and confirmatory analyses in a complimentary fashion. Then DIF analyses conducted using Logistic regression (LR) and Mantel-Haenszel methods. Analyses showed that the matching criterion improved when multivariate analyses were used. The number of DIF items was decreased when the matching criterion is defined based on multiple criterion scores such as mathematical literacy and problem solving scores or two different mathematical subtest score.

In addition, qualitative reviews and examination of the distribution of DIF items by content categories, cognitive demands, item types, item text, visual-spatial factors and linguistic properties of items were analyzed to explain the differential performance. Curriculum, cultural and translation differences were the main criteria for the qualitative analyses of DIF items. The results imply that curriculum and translation differences in items might be causing the DIF across Turkish and English versions of the tests.

Keywords: Differential Item Functioning, Multivariate Analysis, Logistic Regression Method, Mantel-Haenszel Method, Programme for International Student Assessment (PISA)

ÖZ

PISA 2003 MATEMATİK MADDELERİ KULLANILARAK YANLI ÇALIŞAN MADDELERİN TESPİTİNDE ÇOK BOYUTLU EŞLEŞTİRME ANALİZİ

Çet, Selda
Doktora., Ortaöğretim Fen ve Matematik Alanları Eğitimi Bölümü
Tez Yöneticisi: Prof. Dr. Giray BERBEROĞLU

Nisan 2006, 146 sayfa

Madde yanlılığı analizleri aynı yetenekteki fakat farklı gruplardaki kişilerin farklı performans gösterip göstermediklerini araştırır. Aynı yetenekteki kişilerin tespitinde sadece toplam test puanını kullanan tek boyutlu analizler çoğunluktadır. Bu çalışma aynı yetenekteki kişileri tespit etmede, farklı faktör puanlarını aynı anda kullanarak yapılan analizlerin, tek bir toplam test puanını kullanarak yapılan analizlere göre daha etkili olduğunu savunmaktadır. Bu çalışmanın amacı PISA 2003 Matematik sorularının Türkçe ve İngilizce formları arasındaki madde yanlılığını araştırmaktır. Bunun için yayınlanmış madde sayısının çoğunlukta olduğu iki kitapçık seçilmiştir. Bu çalışmada testlerin faktör yapıları faktör çözümlemesi yöntemleri ile tespit edildikten sonra seçilen maddeler analiz edilmiştir.

Tek boyutlu DIF analizleri ile çok boyutlu eşleştirme analizlerinin sonuçları Mantel-Haenszel (M-H) ve Logistic Regression (LR) metotları kullanılarak karşılaştırılmıştır. Bu karşılaştırma sonucunda çok boyutlu eşleştirme yöntemleri ile yapılan analizlerde madde yanlılığı gösteren maddelerde her iki kitapçıkta da bir farklılık görülmüştür.

Yanlı çalıştığı tespit edilen maddelerde ölçtükleri matematiksel beceriler ve bilişsel yeterlilikler, madde türü, madde kökü ve diğer görsel ve uzamsal unsurlar dikkate alınarak madde yanlılığının kaynağının tespit edilmesi için niteliksel analizler yapılmıştır. Bu analizler müfredat farklılıkları, kültürel farklılıklar ve çeviriden kaynaklanan farklılıklar olarak üç ana başlık altında yapılmıştır. Türkiye ve Amerika'daki öğrencilerden eşit yeteneklerde olanların niye bazı maddelere doğru cevap verme olasılıklarının farklı olduğu araştırıldığında bunun matematik programlarının farklılığından kaynaklanabileceği ya da İngilizce'den Türkçe'ye çeviri yapılırken matematik maddelerindeki bazı nicelik bildiren kelimelerin anlamlarının değişebileceği görülmüştür.

Anahtar Kelimeler: Madde Yanlılığı, Çok Boyutlu Eşleştirme, Lojistik Regresyon Analizi, Mantel-Haenszel Analizi, Uluslararası Öğrenci Başarısını Belirleme Programı (PISA)

ACKNOWLEDGEMENTS

Appreciation is expressed to my advisor, Prof. Dr. Giray BERBEROĞLU, for his guidance and directions.

Also thanks to Dr. Michael JODOIN for his help in Logistic Regression Syntax.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

Mathematics is one of the most important components of a fundamental education system in different societies. Today's society, in an information age, requires mathematically literate individuals to become informed citizens. In this context, one of the characteristics of an informed citizen is the knowledge and understanding of technology.

Because of the recent advancements in the use of technology, a greater understanding and using of mathematical ideas and procedures are necessary. This latest pace in the socioeconomic field also has its reflection in educational arrangements as well. For example, National Council of Teachers of Mathematics (NCTM) has stated that there is a shift from an industrial to an information society (NCTM, 1996). In the 21st century, students must understand mathematical models, structures and simulations applicable in a variety of situations. It is important for students to become mathematically literate, i.e. they should be equipped with a capacity to analyze, reason, and communicate mathematical ideas effectively and to formulate, solve and interpret mathematical problems in many disciplines (OECD, 2003).

Monitoring the effectiveness of the educational procedures in developing students having competencies in line with the needs of the society, research studies, both national and international, provide deep insights in understanding whether the students develop a sense of mathematical concepts, symbols, and procedures. These studies supply invaluable information to both the parents, the students, the public and those who manage education systems in deciding whether students are able to analyze, reason and communicate ideas effectively, whether they are well

prepared for the future and have the capacity of continue learning throughout their lives (OECD, 2002).

The results from these studies are the measure of student success and are often used for various purposes. For example, state officials and the public determine the state of students' and schools' performances, and policy makers use these results in setting up educational policies.

Like many other countries, Turkey also gives increasing attention to the quality of its education and assessment of students' academic performance. At the national level, studies of EARGED determine the performance of students (EARGED, 2003). In addition, Turkey also participated in various international studies such as Program for International Student Assessment (PISA) and Third International Mathematics and Science Study (TIMSS).

One of the most up-to-date issues in the context of the international assessments specified above is developing fair instruments among different countries. As achievement tests and questionnaires in international assessments are to be administered in various languages, the main version of the tests, developed usually in English, are to be translated to the native languages of the countries. However, many studies in literature showed that even the most cautious translations, or adaptations, do not quarantine the equivalence of tests (Ellis, 1989; Ercikan, 1998; Ercikan, 2002). For example, there can be linguistic differences such as changing the difficulty of words and sentences or cultural differences such as unfamiliar content related cultural relevance that the test developers should consider.

Although the details are further discussed in the next section, to give some examples of studies dealing with translation fidelity, the study of Sireci and Berberoğlu (2000) investigating the method of using bilingual test takers in evaluating fidelity of translated items, the study of Huang, Church and Katigbak (1997) analyzing how well the items in English-language version function where English is a second language, and the study of Sireci, Yang and Bhola (2003)

examining the structural equivalence of an employee attitude survey form large international corporation can be specified. In all these studies researchers found items functioning differentially across groups.

The growing interest in cross-cultural assessments also requires stringent procedures to assure that the translated, or adapted, versions of a test is fair, reliable and valid in corresponding cultures because, as specified above, translation of even a valid and reliable test can contain some distorting effects to cause bias across cultures. Three kinds of bias can occur during the process of adaptation, administration, and making use of results of cross-cultural instruments; construct bias, method bias and item bias. Construct bias is the non-negligible differences across cultures in the construct being measured, method bias is the different conditions in testing administration across cultures and item bias is anomalies at item level, such as poor wording, incorrect translations etc. (Van de Vijver & Hambleton, 1996).

In this context, adapting a test should aim to create the best possible measures in terms of validity and quality to prevent undesirable differences for various group of interest (Roznowski & Reith, 1999). . In any assessment, if a test is loaded with items that are appropriate for only various groups of students but not for others, it may cause the inadequacy of the test in comparing the individuals from different groups (Beaton, 1998). That's why Hambleton & Rodgers (1995) have indicated that when important decisions are to be made based on test scores, factors that unfairly affect examinees' scores must be avoided.

A biased item functions differently for groups. Statistical procedures that are currently used by test publishers to identify items that function differently across, for example, gender, language or racial/ethnic groups are known as differential item function (DIF) analyses. DIF analyses are useful for flagging items that may need to be eliminated or, at least, submitted to additional review.

In developing a large-scale assessment, researchers should conduct DIF analysis to investigate possible bias at the item level. The DIF analysis is based on

the principle of comparing the performance of focal groups (e.g., female, African Americans, or Hispanic examinees) on an item with the performance of reference group (e.g., male or White examinees), by controlling overall knowledge of the subject tested.

The measure of overall knowledge of the subject is usually the total test score and called the "matching criterion" (Linn, 1993). However, how well the total test score can match the students from different groups is still one of the most important issues in DIF.

From this perspective, before DIF analyses at the item level, the issue of finding the dimensional structure of a test was emerged with the problem of improving matching criterion. A DIF item means it does not seem to measure the same construct as the total test. In other words violations from unidimensionality are causes of DIF. This definition of DIF requires a univariate matching criterion (Dorans & Holland, 1993). But univariate matching criterion may be insufficient in specifying same ability individuals from different groups and this may lead to errors in identification of DIF, if individual items measures more than one ability or if all items in a test measure different abilities (Hambleton, Clauser, Mazor & Jones, 1993).

In the same manner, Ackerman (1992) also explained differential performance from a multidimensional perspective. According to his explanation, a test composed of two or more items is hardly unidimensional. If there are multidimensional abilities as he stated, and only a unidimensional criterion is used in matching individuals, unaccounted abilities can cause cultural, language differences across groups. This approach requires multidimensional DIF analyses methods in examining the equivalence of test items.

However, it should also be added that determining the dimensional structure of a test is itself another challenge to deal with. For example, Gierl (2005) have stated that it is difficult to determine the dimensional structure of the test. Dimensionality of a test may yield a simple or complex structure and outcomes of

dimensionality assessments affect the interpretation of matching and studied subtests.

Identifying the dimensionality requires complex analyses with numerous decisions and consequences resulting from these decisions. But despite these difficulties, there are studies providing considerable perspectives to deal with the dimensionality of the tests.

One of these studies is that of Shealy and Stout's (1993) multidimensional model for DIF which combines the substantive and statistical analyses. In this approach they suggested to conduct substantive analysis to generate hypothesis before the statistical analyses, and then to test these generated hypotheses indicating potential DIF items. The confirmed hypotheses then can help to develop guidelines and test construction principles for reducing DIF on translated tests.

Another solution of this dimension dilemma is using multivariate matching criteria. Related studies (Clauser, Nungester & Swaminathan, 1996; Clauser, Nungester, Mazor & Ripkey, 1996) have concluded that when tests have a dimensionally complex structure, finding an appropriate matching criterion for this structure is an unavoidable procedure.

What is common among the studies dealing with DIF from a dimensional perspective is using multivariate matching (Zwick & Ercikan, 1989; Williams, 1997; Clauser, Nungester & Swaminathan, 1996). Multivariate matching is using more than one variable determined with respect to the factor structure of the data such as, using factor scores in matching individuals from different groups (Hamilton & Snow, 1998). In addition, an external variable, such as educational background variable can also be used in addition to the internal matching criterion, such as subtest scores (Clauser, Nungester, Mazor & Ripkey, 1996).

The studies investigating the effect of using multivariate matching, have demonstrated that multivariate matching can substantially reduce the number of items as exhibiting DIF by enhancing the matching criterion. This means that true group differences in multiple dimensions accounted by multivariate matching,

reduces the probability of finding items as differential functioning although they are not.

On the other hand, matching on only total score instead of multiple valid dimensions may cause multidimensional item impact to be identified as DIF.

Finally, it is worth specifying that not only identifying items showing DIF but also disentangling possible sources of DIF is also required within the studies investigating translation fidelity.

DIF studies can serve to determine culture specific aspects of psychological constructs such as mathematics literacy as defined by PISA (Ercikan, Gierl, McCreith, Puhan & Koh, 2004).

In addition, information provided through DIF studies can also lead developing, or adapting, more valid cross-cultural assessment instruments in future studies. Unfortunately in the DIF literature, there are few studies dealing with the sources of translation DIF, possibly because of the difficulty in interpreting sources of DIF in statistically flagged items (Van de Vijver, 1998). One of the possible causes of this difficulty may be that, the samples of individuals are in many cases different from each other and their differences are not stable and easy to describe. So identification of DIF related factors is more complex process than identifying DIF items.

As an example, one of the most promising studies investigating the possible sources of DIF may be that of Allauf, Hambleton and Sireci‘s (1999). They have indicated that there was a little research exploring why some translated items function differentially across languages. They found in their study that the main reasons of these differences were the changes in word difficulty, item format and content and differences in cultural relevance.

Beaton (1998) has also indicated that when mathematics items were contextualized to make them more realistic, it causes to introduce differences in complexity attributable to national variances. It must be analyzed in large- scale assessment in cross- cultural studies by using different DIF detection procedures.

In this context, this present study aimed at assessing cross-cultural and translation equivalence of English and Turkish versions of mathematics items of PISA 2003through different matching strategies. It is also aimed to disentangle possible sources of DIF-related factors in the mathematics items.

2.1 Purpose of the Study

The purpose of the present study is to evaluate the PISA mathematics literacy items across English and Turkish language versions of the test via univariate and multivariate matching criteria used within M-H and LR approaches. Thus, for this purpose,

1) Total test score on mathematics literacy test

2) Simultaneous use of problem solving and mathematical literacy test scores

3) Subtest scores determined through factor analysis, are used as matching criteria in identifying DIF in the mathematics literacy items of PISA 2003

M-H and LR approaches are used in detecting the DIF items. Thus, items detected as DIF across these methodologies will be compared as well, on the basis of three different matching criteria. In this comparison possible sources of DIF will be evaluated in the curricular, cultural and translation differences between Turkey and USA.

The research questions of the study are:

1) Is there any difference for the items flagged as DIF across LR and MH methods when the matching criterion is mathematical literacy standard score?
2) Is there any difference in items flagged as DIF in LR method when the matching criterion is multivariate such as using two ability scores?
3) What possible factors caused items to be flagged as DIF in the PISA 2003 mathematical literacy test forms across languages and cultures?

2.2 Definition of Terms

Followings are the definitions of the terms that were used in the study:

Item impact: The significant group difference, i.e. when one group has a higher proportion of examinees answering an item correctly than other group, is called item impact. In other words, true group differences in proficiency are the reason of item impact (Sireci & Allalouf, 2003).

Differential Item Functioning (DIF): An item functions differentially between groups if individuals with the same ability level but from different groups do not have equal probability of answering the item correctly (Li & Stout, 1996). Differential item functioning analyses enable to understand whether the reason of item impact is irrelevant to the construct being measured by the test after controlling for ability.

Matching variable (criterion): Some measure of test performance to specify the students of the same ability in different groups to assess DIF. In present study total test score and subtest scores were used as the matching variable.

Reference group: The group of examinees who are used to compare performance of the focal group. USA is the reference group of the present study.

Focal group: The group of examinees whose test performance are of primary interest and believed to be disadvantaged. Turkey is the focal group of the present study.

Item Bias: A DIF item is considered biased when believed that this item measures some irrelevant construct that function at a disadvantage of one group of examinees DIF is required, but not sufficient, for item bias, because the cause of DIF can also be the item impact.

Mathematical Literacy: Mathematical literacy is defined by OECD/PISA (2003) as; “an individual’s capacity to identify and understand the role that mathematics plays in the world, to make well-founded judgments and to use and

engage with mathematics in ways that meet the needs of that individual's life as a constructive, concerned and reflective citizen" p.24.

Problem solving: In PISA 2003 study defined the problem solving as " an individual's capacity to use cognitive processes to confront and resolve real, cross-disciplinary situations where the solution path is not immediately obvious and where the literacy domains of curricular areas that might be applicable are not within a single domain of mathematics, science or reading"p.156 (PISA, 2003).

2.3 Significance of the Study

DIF analysis is an approach, which may be effective in understanding variables that might be related to significant differences among students in mathematical achievement. Emphasis in mathematics assessment research has lead to many new DIF analyses in different subgroups of examinees and DIF detection strategies; however studies are needed to examine the DIF strategies using different matching criteria. The changes in mathematics assessment are affected by the results of such studies.

Most recently accepted approaches for identifying differentially functioning test items compare performance across groups after matching examinees on the ability of interest, generally total test score. The optimal matching criterion is the single total test score when the test is approximately unidimensional. However, when the test is dimensionally complex, matching through the use of single total score may result in an inflated Type I error rate (Ackerman, 1992). Multivariate matching may provide an attractive and practical alternative to the using single total test score for the study of differential item functioning. Conditioning multiple valid dimensions influencing item responses may eliminate Type I error. In other words this type of matching may decrease the likelihood that item impact is misinterpreted as DIF (Gierl, 2005; Clauser, Nungester, Mazor & Ripkey, 1996).

Multivariate matching is not only useful for identifying items with DIF but also for explaining the appearance of DIF. Multivariate analysis of DIF offers a more complete approach to DIF thus enhances our understanding of the nature of DIF. Trying to identify DIF using multivariate perspectives offers clues as to the causes of DIF which may not be evident through analysis with univariate matching. The ability of using multivariate analysis, that is more than one matching criteria in logistic regression procedure, may have important usefulness for the researchers. DIF analysis may be more difficult in multivariate case but it can provide deep understanding of the factors influencing the performance of DIF items (Clauser, Nungester & Swaminathan, 1996).

DIF detection plays an important role in the test adaptation process as well. Conducting a DIF study has become an essential part of test development and test evaluation (Allauf, Hambleton & Sireci, 1999). It seems necessary for test developers, to determine the features of items with DIF for different groups of examinees gains importance because of the increase in use of large-scale assessments. Identifying the mathematics items exhibiting statistical DIF and examining the characteristics of these items provide suggestions to improve mathematics items to revise assessment tasks. The origin of rationale is that, removing of modification of biased items will increase the validity of a test, and in combination with more direct assessments of validity, will result a fair test for all groups of examinees (Camilli & Congdon, 1999).

Using DIF analyses helps to increase confidence in making decisions that are based on test data. Zieky (1993) noted in the statement "The use of DIF procedures has caused us to focus more clearly on exactly what knowledge, skills, and abilities we are trying to measure. In the long run, the continued use of DIF statistics will result in more valid as well as fairer tests" p.346.

Finally, becoming aware of patterns of differential item functioning has implications for teachers, curriculum specialists, state boards of education, and test developers. The results of the study can give an idea to test developers, users who

interested fair and unbiased testing in mathematics assessment. In some cases the causes of item level difference can be the curriculum, teaching methods etc. so these findings can alert the curriculum specialists to need of curriculum change. Instruction and assessment strategies can be included into the classroom to give more opportunities to the students who are at a disadvantage with the knowledge and skills addressed by the mathematics items.

This study also aimed to provide suggestions to improve items in translation procedures and instructional strategies in order to help to reduce the presence of DIF in cross culture studies so researchers and also national governments will have some benefits of using the results in mathematics education.

CHAPTER II

LITERATURE REVIEW

In the literature review, the bias and equivalence issues in translated tests and the detailed description of the differential item functioning and item bias were presented. Then the differences in univariate and multivariate analysis and the explanations of DIF methods were given. In the last section, possible sources of DIF specified in the literature were discussed.

2.1 Bias and Equivalence Issues in Translated Tests

Many international tests were administered in multiple languages. There is substantial evidence that the different language versions of tests are not equivalent (Hambleton & Patsula, 2000). Interpretations of the results are inappropriate if this equivalence does not exist. The equivalence of the different versions of tests should be established for valid comparisons across cultural and ethnic groups (Robin, Sireci & Hambleton, 2003). Test translation is an important topic because the validity of scores on any translated test depends on the accuracy of the test adaptation. The original and translated versions of items must display equal probabilities of a correct response from individuals and assess the same amounts of traits to obtain the equivalence of the original and the translated item (Hui & Triandis, 1985; Hulin, 1987; Van de Vijver & Leung, 1997).

In cross-cultural research in psychology, establishing equivalence is viewed as key in making valid cross-cultural comparisons (Poortinga 1989). Van de Vijver (1998) defined a three hierarchical order for equivalence issue depending on the type of cultural comparison.

In the first level same construct is measured in each cultural group which is defined as construct equivalence. If the construct equivalence is present in each group, constructs can be compared. A theoretical representation of the underlying trait, concept, attribute, processes or structures the test is designed to measure is labeled as "construct". After assuring that the same construct measured in each group, second level requires the same scale or measurement unit in each group. If there is the same scale in each group, the differences between scores can be compared. And finally, assuring that there exists the same scale with same origin in each group then full comparability is obtained between groups. This equivalence is called scalar equivalence. The scalar equivalence provides the comparison of scores between groups.

Equivalence of measurements in the social sciences across cultural groups can be threatened in various ways. In the literature, a distinction is made between construct bias, method bias, and item bias as threats to equivalence. Construct inequivalence is the non negligible differences in the construct being measured and also labeled as construct bias (Van de Vijver & Tanzer, 1997; Van de Vijver, 1998). In construct bias the measured construct or the behaviors from which items are sampled are not identical across cultures. Method bias is presence of nuisance variables due to the methodology related factors such as sample, instrument and administration procedures. And item bias can be due to the appropriateness of the item content, inadequate item formulation or inadequate translation (Van de Vijver, 1998).

This study related with the item bias. Construct equivalence is the prerequisite for the item level bias analysis. According to hierarchy of the equivalence levels explained above, construct equivalence is to be achieved first before going into item bias analysis. If the construct is determined unidimensional, it must be demonstrated that the same unidimensional construct is measured by both language versions of test. Additionally, if the construct is multidimensional

the common dimensions across languages must be identified (Sireci & Swaminathan, 1996; Allaouf, Hambleton & Sireci, 1999). Construct equivalence was evaluated in the literature by means of structure oriented approaches. Exploratory factor analysis, multidimensional scaling and confirmatory factor analysis is one of the techniques that the researchers used to evaluate construct equivalence (Berberoğlu & Hei, 2003; Robin, et all. 2003; Gierl, Rogers & Klinger, 1999; Gierl, 2000; Hui & Triandis, 1985).

In exploratory factor analysis (EFA), separate factor analyses are performed for each group and then results are compared. If there exist similar patterns of factor loadings across groups, evidence of construct equivalence obtained. But to evaluate construct equivalence using exploratory factor analyses separately for each group makes difficult to understand the common factor structure and there is no any statistical test to determine the degree of testing construct equivalence.

Less used to determine construct equivalence are multidimensional scaling techniques, because most multidimensional scaling techniques do not provide the statistical structural equivalence. Generally descriptive fit indices are used for model data fit. Due to the limitations of EFA and MDC, confirmatory factor analysis (CFA) is more popular technique to evaluate the construct equivalence because the statistical test and descriptive indices of model fit are available and one can manage the multi-group analyses (Sireci, Bastari & Allalouf, 1998).

Obtaining construct equivalence does not ensure the item equivalence. The studies of Zumbo (2003) showed that the results of construct equivalence analysis should ensure construct equivalence before investigating item bias. He reported also that the construct equivalence does not guarantee that there is no DIF or bias in tests.

Investigating construct equivalence was the prerequisite step for the DIF analysis in this study. That is whether the mathematical literacy which was defined in PISA 2003 means the same thing for American and Turkish students. The main step was the investigating DIF. Methods in investigating DIF and item bias were given in following sections.

2.2 Differential Item Functioning (DIF) and Item Bias

In the beginning of 1900's, it was recognized that some items were measured the effects of cultural training instead of mental capacity in IQ tests. Then group differences in IQ tests have been the major research area for researchers. In 1950's, the original purpose of item bias research was to make free IQ tests of group differences that resulted from unequal learning (Camilli & Sheapard, 1994). Then studies of item bias gained importance in achievement tests to create culture fair tests when test results were used for making important selection and placement decisions.

In the past, DIF has been called item bias but this mean of item bias was more evaluative than descriptive. Therefore, it has been replaced by the term DIF (Thissen, Steinberg, Wainer, 1988). The term biased is used if one is emphasizing cause, the term DIF is used, if one is emphasizing effect. Analyses of DIF are statistical. On the other hand analyses of item bias are qualitative (Sireci & Allouf, 2003; Camilli, 1993).

Item bias implies a qualitative review, item may have different meaning or may be measuring an unwanted nuisance factor for one group as compared to another. For example in a mathematics achievement test, mathematics knowledge is a primary dimension and test wise-ness and verbal ability are irrelevant secondary dimensions. If there exists an irrelevant factor that is the item measuring, test authors conclude that the item is biased against members of the affected group (Gierl, 2005; Camilli & Sheapard, 1994; Clauser & Mazor, 1998). Removal or modification of these items will improve the validity of a test and this test will be fair to all groups of examinees (Camilli & Congdon, 1999).

If the source of the differential item functioning of the item is relevant to the purpose of the test such as critical thinking in mathematics achievement test, there exists item impact not item bias. Individuals from groups of interest may

actually differ in ability and this difference in performance is expected. That is item impact is the real differences on the underlying ability between the groups (Gierl, 2005; Camilli & Sheapard, 1994; Clauser & Mazor, 1998).

In the literature there are two approaches for qualitative analysis of DIF items. Traditionally, content reviews (substantive methods) are implemented after statistical analyses for identifying sources of DIF.

With this approach each item is tested statistically using DIF detection methods. This approach can lead to inflated Type I errors because a large number of DIF hypothesis are tested. So a non-DIF item can be thought to be a DIF item. Due to the misidentified items there is a little progress in substantive methods identifying the causes DIF items (Camilli & Sheapard, 1994; Gierl, 2005). An alternative approach was suggested by many researchers who are studying DIF in multidimensional perspective. In this approach substantial analysis conducted in the first stage of DIF analysis to generate DIF hypotheses. With this approach testing fewer DIF hypotheses using statistical analysis provides the better understanding of causes of DIF items (Gierl 2005; Gierl & Khalig, 2000; Shealy& Stout, 1993).

With the recognition of various irrelevant factors that the item measures in qualitative analysis of items, researchers classified item bias in different ways. Gierl (2005) indicated that bias could be content related or response related. For example test-wiseness is content related and verbal ability is response related irrelevant factors. Another classification was made by Hambleton and Rodgers (1995). They classified the item bias as content bias, language bias, item structure and format bias. For example, if there is a content bias item contains content that unfamiliar to focal or reference groups. In language bias, item contains words that have different or unfamiliar meanings or item has difficult vocabulary or group specific language for focal or reference group. In item structure and format bias, there are clues in the item that facilitates the performance of one group over another. If item is biased, test item stem, keyed response or distracters may not be

adequate and clear. Explanation concerning the nature of the task required to successfully complete the item may tend to differentially confuse the studied groups.

2.3 Multivariate Analysis in DIF Detection Procedures

In univariate matching, when substantial analysis is conducted after the statistical analysis, it is difficult to interpret the causes of DIF. The reason of the differential performance in the item may be the item impact that is relevant ability. Multivariate matching is an important approach to control this relevant ability before the DIF analysis in matching process.

If a test is designed to measure a single trait, DIF analysis using total score is appropriate to identify items that measuring irrelevant factors. But if a test is designed to measure a complex skill, DIF analysis using total score may not be appropriate. Because when there exists more than one relevant ability, items measuring multiple relevant dimensions may be identified as displaying DIF (Clauser, Nungester, Mazor & Ripkey 1996; Mazor, Kanjee & Clauser 1995; Ackerman, 1992; Camilli & Sheapard, 1994; Gierl & Khalig, 2000).

In unequal multidimensional ability distributions between groups, the interpretation of total score is difficult across the ability range of groups. Mazor, Kanjee and Clauser (1995) introduced the term multidimensional item impact to refer the case in which the cause of DIF is uncontrolled between-group ability differences on at least one of relevant abilities.

When these relevant abilities are controlled by conditioning on all relevant abilities, matching will be more accurate and the number of the DIF items will reduce, because the probability of the multidimensional impact will be reduced.

Matching criteria are valid if all irrelevant dimensions are not included or all relevant dimensions are included in the matching. In a simulation study Ackerman (1992) have reported the effects of choice of matching criterion when

irrelevant dimensions exist. In this simulation study first he used total score in a two dimensional data to identify DIF items and then used a valid unidimensional subtest score. Using the valid subtest score reduced the number of items flagged as DIF. Secondary dimension in this study was irrelevant with the purpose of testing and most of DIF items were loaded on the second dimension. So he showed that when irrelevant secondary dimension was included in the matching score, matching procedure was violated.

On the contrary of Ackerman (1992) study, when relevant secondary abilities are not included in the matching score, matching procedure is violated. Studies based on using multiple relevant abilities have examined the improvement in matching criterion for differential item functioning analyses (Zwick & Ercikan, 1989; Mazor, Kanjee & Clauser, 1995; Clauser, Nungester & Swaminathan, 1996; Hamilton & Snow, 1998 and Clauser, Nungester, Mazor & Ripkey, 1996).

Various matching strategies are available using external or internal variables in multivariate matching. The choice of these relevant variables should be considered according to the purpose of studied test (Mazor, Kanjee & Clauser, 1995). Zwick and Ercikan (1989) used a background variable relevant to history education in addition to total score with the Mantel-Haenszel statistics. The hypothesis of their study was the between group differences in the studied historical periods were related to the history achievement. But adding background variable did not result in reduction in the number of items identified as DIF. They reported that the choice of this background variable or limitation in the M-H statistics might be caused this result.

Similar study was conducted by Clauser, Nungester and Swaminathan (1996) to improve matching with categorical background variable. In this study, researchers used logistic regression procedure. They hypothesized that addition of categorical variable representing educational background might improve the matching and cause the reduction in the number of DIF items. In contrast to findings of Zwick and Ercikan (1996), when educational background variable was

used, there was a reduction in the number of DIF items in their study. They concluded that educational background variable used in their study might be more appropriate than the educational variable that used in the study of Zwick and Ercikan (1996). Also using LR statistics could be more effective than M-H statistics.

In the study of Mazor, Kanjee and Clauser (1995), two continuous achievement external variables were used with M-H and LR statistics for the comparison of the results. They showed that conditioning on two relevant abilities provides more accurate matching than the conditioning on single ability. Also they showed that M-H and LR results were similar identifying items as DIF.

External ability estimates are not always available. Clauser, Nungester, Mazor and Ripkey (1996) analyzed the usefulness of interval ability estimates by comparing the results of M-H and LR. They used both real and simulated data and their findings supported that items identified with the total test score as the matching criterion but not identified using the subtest score and multiple subtest scores are more likely to represent Type I error. M-H and LR statistics produced similar results for identifying uniform DIF. This finding is consistent with the study of Mazor, Kanjee and Clauser (1995).

Another study using the internal multiple criteria was conducted by Hamilton and Snow (1998). They used science achievement data to identify DIF items with M-H and LR procedures. In LR they used subtest scores which were identified with factor analysis. Their study also showed that taking into account multiple constructs eliminated some items identified as DIF.

2.4 Comparison of LR and M-H Methods

The most common used chi-square methods are the Mantel Haenszel (M-H) DIF detection procedure, which was adapted and extended by Holland and Thayer in 1986 (Hambleton, Rogers, 1989) and the Logistic Regression (LR) which was adapted by Swaminathan and Rogers (1990).

Chi-square methods include the contingency tables. The strategy of chi-square techniques is to eliminate the differential functioning item from the dependency on the groups x items interaction. Chi-square approach is sensitive to within-groups item discrimination and the differences among groups in item difficulty levels.

To examine the degree of difference between the score interval proportions total test score is divided into the different number of categories. The chi-square statistics is comparatively simple to calculate and it is appropriate also for small sample sizes (Osterlind, 1983). Multivariate analysis conditioning multiple ability estimates are another advantage of these two methods. With these advantages chi-square methods are widely used to investigate item bias. There are similarities and differences in univariate and multivariate analysis of M-H and LR methods in the literature.

In Univariate Analysis

There are studies in the literature which compares the detection of number of DIF items in M-H and LR procedures. One of the advantages of M-H and LR DIF methods is that, these procedures obtain valid results with relatively small numbers of examinees. But some simulation studies showed that when 500 or fewer examinees were retained in each group more than 50% of the differentially functioning items especially which were the most difficult or with a small difference in item difficulty between the two groups or poorly discriminating items were missed (Mazor, Clauser & Hambleton, 1992). But in their simulation study Gierl, Jodoin and Ackerman (2000), showed that when the proportion of DIF items is large, M-H and LR methods obtained good type error protection by manipulating the amount of DIF, sample size, and ability distributions between groups. In the literature there are studies suggesting that the MH procedure may be a good choice when sample sizes of between 100 and 300 (Hills, 1990). But Mazor, Clauser and Hambleton (1994), found that the results of the MH procedure is questionable at small sample sizes. Sample sizes of 200 in a group may be adequate if one needs to

identify only most noticeable DIF items. Sample size of 500 gives more accurate results than the sample size of 200. If groups with different ability distributions are compared it is advisable to use sample sizes of more than 1000. But large sample also may fail to identify DIF items if the compared groups have the unequal ability distribution.

Benito and Ara (2000) have proposed several IRT and non-IRT DIF methods in their study. All DIF detection techniques tend to over identify items with DIF except LR. They found that the tendency of over identifying DIF items is slightly reversed in the LR procedure. Their simulation study showed that the DIF technique that appears to do the best job was the Mantel Haenszel statistic. On the contrary, Hidalgo and Pina (2004) compared the MH and LR methods in their efficacy for detecting DIF. They compared the effect size measures and manipulated the conditions of item difficulty and discrimination. In this simulation study, their results have suggested that LR analysis generally detected more DIF items then M-H analysis.

Both M-H and LR can be used to detect non-uniform DIF with some modifications in M-H analysis. Swaminathan and Rogers (1990) showed through simulation studies that the LR procedure was more powerful than the M-H procedure in detection of non-uniform DIF but as powerful in detection of uniform DIF. Although M-H is not powerful as LR in detecting non-uniform DIF, Mazor, Clauser and Hambleton (1994) studied the detection of non-uniform DIF using Mantel Haenszel DIF method. They split examinees into two samples by breaking the full sample at approximately the middle of the test score distribution. Then they reanalyzed the tests across these low and high performing samples. This procedure improved the detection rate of non-uniform DIF items especially items having largest differences in discrimination and difficulty parameters without increasing the Type I error rate.

Rogers and Swaminathan (1993) concluded that the M-H procedure was quick and inexpensive to implement. Only cell frequencies were needed for

calculation of M-H. But LR procedure was iterative and so more expensive in terms of computer time.

In Multivariate Analysis

When it is possible to account another variables that are related the studied variables with a sampling design, this may be preferable to detect DIF items more accurate and increase the power of DIF detection analyses.

This matching could be used also in MH analysis but this variation in MH analysis requires large sample sizes. With small sample sizes, each cell may not contain members of both reference and focal groups and both 0 and 1 scores on the studied item in contingency table (Clauser, Nungester & Swaminathan, 1996). Another advantages using LR procedure over M-H is the potential of accommodating more than two ability estimates. It is also possible to use more than two ability variable in M-H, but in this case M-H procedure is inconvenient and interpretations of the additional variables are difficult (Mazor, Kanjee & Clauser, 1995).

2.5 Other DIF Detection Methods

In the literature many DIF methods have been described and classified based on different properties.

Methods for DIF detection is divided into two groups; observed score methods and latent score methods. They are also called non-IRT (non-parametric) and IRT (parametric) methods, respectively (Camilli & Sheapard, 1994). The most commonly used observed score methods are based on classical test theory. In classical test theory, ability is defined as the expected value of observed performance on the test. Also probability of getting the item right is defined as the proportion correct scores, so this probability depends on the ability of examinees taking the test (Crocker & Algina, 1986).

Hambleton et all. (1993) classified DIF methods as methods using classical test theory, using item response theory and involving Chi-square analysis. Methods which utilize classical test theory include analysis of variance, correlational methods, transformed item difficulty or delta plot which is based on the difference between the difficulty parameter estimates obtained in each group. These methods use observed scores as a criterion and compare the classical item difficulty values for the studied groups. Being sample dependent is an important disadvantage of classical theory methods.

DIF results can be different according to selection of the samples of the groups. So results can not be generalized to the population.

By using IRT, a researcher can place the item response curves from each test on the same scale. In contrast to classical test theory DIF detection methods, IRT methods are not sample dependent. Item characteristic curves are independent of the groups, and estimated ability is independent of test difficulty. In IRT terms, DIF exists if individuals having identical levels of the latent trait from different groups have unequal probabilities of correctly answering an item. Matching criterion is the estimate of latent ability rather than the observed score. Between group differences in the item parameters is used to identify DIF. Item parameters are estimated separately for the focal and reference groups and then these parameters are placed on the same scale for comparison. So an item displays DIF if the item characteristic curves (ICC) or item parameters are not the identical across two different groups of examinees (Hambleton, Swaminathan, & Rogers, 1991). There are different statistical procedures to compare ICC's across different examine groups (Thissen, Steinberg, Wainer, 1988; Lim, Drasgow, 1990).

One of these statistical procedures is the Likelihood Ratio Test (LRT). Thissen et al. (1988) have applied the Likelihood Ratio Test (LRT) to detect DIF in IRT by testing the improvement in fit for the model, comparing fit with and without separate group parameter estimates. The limitations of the IRT DIF detection methods are the need of unidimensionality assumption, large sample sizes

for accurate parameter estimation especially for two or three parameter model and more complicated calculations.

Another DIF detection procedure is the simultaneous item bias testing SIBTEST. Matching variable is latent score as in IRT. This is an iterative procedure that all items are used in the matching variable. Items that are flagged as DIF are removed from the analyses until no-DIF items are found. SIBTEST is a non-parametric procedure and calculates the size of DIF in multidimensional IRT model based approach. This procedure is appropriate either to detect item bias or DIF or to detect test bias or DTF (differential test functioning) (Shealy & Stout, 1993).

In addition to classical test theory, chi-square and item response theory based DIF methods, Benito and Ara (2000) indicated the forth classification as factor analysis FA-based methods. There exist two FA-based method; Unrestricted FA and Restricted FA methods. The logic of these analyses comparing the factor solutions obtained when factoring the data matrices in the different groups with constraints or without constraints.

Differences among DIF methods can be characterized according to whether they are parametric or non-parametric; are based on latent or observed variables; can model multiple traits; can detect uniform and non-uniform DIF; can examine polytomous responses; can include covariates in the model etc. There are advantages and disadvantages of DIF detection methods according to each other. They produced different results in different conditions.

Identification of items as DIF depends on which DIF detection method is used. The question of which statistical method is most adequate for DIF detection has not an exact answer. In high stakes testing situations the most adequate solution for choosing appropriate method is using more than one method. Using more than one method provides easy controlling of the type I error rate (Hambleton et all., 1993).

2.6 Possible Sources of DIF in Achievement Tests

Differential item functioning (DIF) analyses control ability levels of individuals belonging to different groups. This makes it more dependable to claim that the results are not reflections of ability but group differences. In addition, as DIF analyses are item level analyses it is also possible to disentangle the characteristics of items functioning differentially across groups.

Judgmental methods which items are judged subjectively according to linguistic and psychological characteristics may provide understanding of possible causes of DIF. But in the literature results of judgmental reviews and empirical DIF methods show little agreement.

Empirical studies showed that item bias was not clearly understood and item bias conclusions was not consistent across instruments and samples (Van de Vijver, 1998). Even there exist inconsistent results, conducting judgmental analyses as a substantial analysis before or after in DIF detection procedure may help researchers to combine results of each analysis. Many studies in DIF literature was done in this manner. Judgmental studies to understand the sources of DIF are less common in large scale assessments especially in mathematics. Some researchers found that the reason of differential performance might be due to the characteristics of mathematics items such as cognitive complexity (Engelhard, 1990; Tatsuoka, Linn, Tatsuoka & Yamamato, 1988), content or item format (Gamer & Engelhard, 1999; Scheuneman & Grima, 1997; Harris & Carlton, 1993), curriculum differences and adaptation or translation differences (Ercikan, Gierl, McCreith, Puhan & Koh 2004).

For example, Engelhard (1990) investigated the relationship between gender and performance on a set of test items, which vary in both level of cognitive complexity and content. Nationally representative samples of 13 years old students of US and Thailand who were participated TIMSS were analyzed. Mantel-Haenszel procedure was used to detect DIF. And then repeated measures of ANOVA was

designed to examine the observed gender differences on these items were related to cognitive level (computation, comprehension, analysis) and content category (algebra, arithmetic, geometry). They showed that both level of cognitive complexity and content category are related to gender differences. Gender differences tend to become more favorable toward boys as the level of cognitive complexity increases and also as the content changes from arithmetic through algebra to geometry.

Similar result obtained in the study of Harris and Carlton (1993). They examined the patterns of gender differences on mathematics items of SAT. They used M-H procedure to investigate differential item functioning and one-way of analyses of variance (ANOVA) techniques to identify categories of item characteristics that resulted in significant differences between male and female students. Items were analyzed in item format and item content categories. The results of the study showed that male and female students who achieved the same score did not arrive at that score with the same pattern of responses. Male students performed relatively better than female students in geometry and geometry/arithmetic items. Female students performed better than male students in arithmetic/algebra items. These results indicated that the female students were good at in abstract item and male students good at items that are related in real life situations.

Socio economic status also is another effect on mathematics performance in addition to gender effect (Yurdugül & Aşkar, 2004a; Yurdugül & Aşkar, 2004b). Also Berberoğlu (1995) studied DIF, by comparing ICC across gender and socio-economic status (SES) groups to provide evidence on whether one of the groups had an advantage in solving mathematics questions in the content areas of computation, word problem and geometry in the mathematics subtest of the University Entrance Examination in Turkey. On the contrary the study of Engelhard (1990), the findings showed that most of the computation items favored

males. In the word problems and geometry items males had disadvantage in solving questions. All of the items of the word problem type favored the high SES group. In the geometry and computation parts about the half of the items favored high SES group.

Another study dealing with item format and content is the study of Gamer and Engelhard (1999). They examined gender differences in performance on multiple-choice and constructed response items in mathematics. A random sample of 3 952 eleventh graders who took the 1994 Georgia High School Graduation Test was used for the analysis. The mathematics portion consists of 60 multiple-choice items and eight constructed response items. Mean performance on subtests was compared for the two groups, and DIF was explored using the many-faceted Rasch measurement model (FACETS). In both mean scores ($p < .001$) and DIF indexes-the constructed response items exhibited less DIF than the multiple-choice items.

Women showed a statistically significant and consistent advantage over men on multiple-choice items involving algebra, whereas men showed a less consistent advantage on items involving geometry and measurement, number and computation, data analysis, and proportional reasoning. Mean scores were significantly higher for men than for women on 2 out of 8 constructed response items. However, when men and women were statistically matched according to ability, the only significant difference in performance on constructed response items was in favor of women. It was concluded that gender differences in mathematics might well be linked to content and item format.

Scheuneman and Grima (1997) examined the characteristics of quantitative word items in GRE by considering sources of group performance differences. They classified the factors that may be the causes of DIF for quantitative item properties as the cognitive nature of the task, mathematical content, and the surface properties of item such as item format or key position. Verbal properties of items are also classified as readability, semantic content verbal structure and quantitative language. They analyzed DIF for female-male and Black-White groups. They

found that verbal properties of items were found to be associated with differential performance of women and men but not of Black and White examinees. Black examinees showed differential difficulty on data-interpretation. Items with one or more diagrams, and real setting tended to be relatively more difficult for black examinees. Another finding of the study, key position was related to DIF. Both female and black examinees showed poor performance on items with A or B keys. Their performances were better in D or E keys. These findings concerning the response style differences of examinees.

Zenisky, Hambleton, and Robin (2003) conducted an example of such studies in science items. The purpose of the study was to identify gender DIF and try to understand DIF due to the content, cognitive demands, item type, item text and visual-spatial or reference factors. Elementary, middle and high school levels were used with approximately 360.000 students. Multiple choice and open response the item types were used in each test with 32 to 42 items. They searched the possible patterns of items which related content category, visual-spatial and reference component and item type in each level. Their findings are the indicative of possible sources of DIF and can be used by item writers as guidelines. They found differences in content category, visual–spatial component and item type dimensions.

Educational systems in different countries produce different patterns of outcomes (Beaton, 1998; Klieme & Baumert 2001). The study of Beaton (1998) has implications for the teaching and encouragement of mathematics. Addressing the question how fair the TIMSS tests, he used different subtests of mathematics and science items using test curriculum matching analysis (TCMA). Countries curriculum may vary in different subject and teaching methods. This may limit the international achievement studies comparison. He found that allowing countries to select the items that they are scored not substantially affect the overall picture on their international standings. Countries performances in sub areas were highly correlated. For example, fractions and proportionality overlapped but they were not

the same thing as Geometry. Students who got high score in one sub area also got high scores in other sub areas.

Klieme and Baumert (2001) used DIF to identify proficiency profiles in TIMSS study using data from the advanced mathematics test for the upper secondary sample. To find the country specific the strengths and weaknesses of the advanced mathematics they examined other countries compared to Germany using IRT approach. They found some main differences in examined countries such as while US curriculum focuses declarative and procedural knowledge, Germany is weak on advanced knowledge and understanding, but has strengths in the use of visual and graphical representations.

The effect of language and culture differences on mathematics performance gained importance since the achievement tests were used in different ethnic groups and international assessments. In the study of Gierl and Khalig (2000) the test development and analyses committee identified language and cultural differences might affect the performance of one group.

These sources were omissions of additions that affect meaning, differences in the words, expressions, or sentence structure inherent and not inherent to language and culture and differences in item structure. They used the data eight different English and French student samples from the 1997 administration of Mathematics and Social Studies Achievement Test at grade 6 and grade 9. They found that the outcomes in social studies were more complex and less interpretable than the outcomes of mathematics. In mathematics the translators predicted correctly seven of the eight items and only one bundle consist of two items was incorrectly predicted. The majority of the DIF related factors were associated with the differences in the words, expressions, or sentence structure of items that are not inherent to the language and culture. Two items from the grade 6 could not be interpreted by the translators. The result of SIBTEST indicated that these items produced a systematic effect that favored the French examinees. Factors that identified in substantial analyses were less effective in social studies test than in mathematics test.

The reason of the DIF items expected to be multiple factors. So many strategies were used in different studies to find the sources of DIF. Ercikan, Gierl, McCreith, Puhan and Koh (2004) indicated that the sources of DIF depend on the type of the test. If a test is an achievement test, expected performance differences related to curricular and instructional factors. But in licensure tests curricular differences less affects the performance. The identification of DIF is more complex in licensure tests than the achievement tests. In multilanguage versions of tests researchers focus on the comparability of item format, content, translation and adaptation. Ercikan et all. (2004) examined the degree of comparability of bilingual versions of assessment of English and French versions of reading, mathematics, and science tests that were administered in Canada. They also examined the sources of incomparability due to the adaptation effects and curricular differences. Sources of DIF items considered as belonging to adaptation effects more than curricular differences.

Similar result found also in the study of Ercikan (2002), adaptation effects and curricular differences of DIF discussed using the data Third International Mathematics and Science Study (TIMSS) assessment of USA, English ad French tests. She used the DIF identification procedure described by Linn and Harnisch using IRT based approach. As a result in mathematics 27% of the DIF items related the adaptation effects and 23% of the DIF items related curricular differences. In science items 37% of the DIF items related to adaptation effects and 13% of the DIF items related to curricular differences.

To assess the possible causes of DIF in translated verbal items Allalouf, Hambleton and Sireci (1999) used the types of items which were most likely to display DIF when translated form one language to another. Analyses of DIF detection and analyses of translators showed that changes in difficulty of words and sentences, changes in content, changes in format and differences in cultural relevance were the possible causes for DIF.

To understand the causes of DIF, researchers used the Gallagher's and Ibarra's classifications in differential performance between groups (Li, Cohen & Ibarra, 2004; Gierl, Bizans & Li, 2004). They developed coding schemes; Gallagher's cognitive structure analyses and Ibarra's Multicontext Theory. According to Multicontext Theory, culture is the set of learned patterns and these patterns play an important role in people's learning, thinking and communication. The item type and item format form the culture context. Gallagher's method is based on cognitive factors which favor females or males. Social and cultural domain of items, real world application, spatial reasoning, definition-based and indefinite answer questions are some factors that Li, Cohen and Ibarra (2004) explored these two approaches to explain why gender DIF occurs. They were set up coding categories in cognitive and cultural structures. Their study suggested that Multicontext Theory was more effective than Gallagher's method in predicting gender DIF.

Gierl, Bisanz and Li (2004) used Gallagher taxonomy to generate hypothesis in gender differences to identify specific content areas and cognitive skills. Then these hypotheses were tested using SIBTEST using data from the grade 9 mathematics achievement test administered in Canadian province of Alberta. They obtained inconsistent results between the statistical and substantial analyses. They indicated that current cognitive theories might not be a good substantive basis for generating DIF hypothesis or statistical DIF analyses might not be appropriate testing cognitive theory based hypotheses.

2.7 Translation Fidelity in Multilingual Comparisons

Possible sources of DIF in achievement tests which explained the previous section showed that the adequacy of translation can be threatened by various sources of bias. Some studies reported poor translation for the sources of DIF in the literature (Allalouf, Hambleton & Sireci, 1999; Ercikan, Gierl, McCreith, Puhan & Koh, 2004).

The findings related poor translation in the sources of DIF in multilingual studies alerts the researchers for the translation inequivalency of the instruments. Ellis (1989) has reported this issue, saying that when cultural differences and similarities are under investigation, language differences become a serious problem to obtain valid inferences from the results, because language is a defining characteristic of a culture. Also, Bontempo (1993) stated for an instrument that are developed in one language and translated into another, to produce comparable scores, it is necessary to demonstrate the translation fidelity. Test translation is a difficult task because it requires all of the psychological, linguistic, and cultural considerations.

To take into account these considerations in translation there are different procedures. In most multilingual assessments e.g. PISA 2003, instruments are developed in a single language and cultural setting instead of using simultaneous translation.

There are three options to translate instrument from one language to another: applied, adapted and assembly. Application option is the case; a literal translation is used linguistically and psychologically appropriate. In adapted option, also there is a change in wording and contents of other items. Assembly option is appropriate in the case the original instrument is assumed to be inadequate in new context. So a new instrument developed in the new cultural context (Van de Vijver & Leung, 1997).

There are cross cultural studies in the literature that show these translation options may not be appropriate (Van de Vijver & Leung, 1997). For example, Robin, Sireci and Hambleton (2003) found that the adapted forms were less reliable and new dimensions were necessary for the structure of all the response data. They conducted first descriptive analyses to evaluate the psychometric properties and second dimensionality analyses to assess the equivalence. Then to indicate the potential translation problems or other sources of item bias DIF analysis were conducted. They concluded that impact was large across the different versions of credentialing exams.

2.8 Summary of the Literature

In translated tests, bias and equivalence are important issues to investigate the validity of comparisons of different cultures through the DIF analysis. Although most of the statistical procedures in DIF analysis require a unidimensional data, finding unidimensional tests are difficult due to the items measuring complex abilities. Therefore, investigating DIF using univariate DIF analysis in a multidimensional test is not appropriate; because, if done so, multiple relevant dimensions are measured by the items may be identified as displaying DIF. In this case detecting a unidimensional subset of items and using scores of these items in matching process have advantages over using the total score as a matching criterion. In this study a unidimensional set of items were selected and the score of these items were used as a matching criterion in univariate DIF analysis.

But even in a unidimensional test, some or all items may measure more than one relevant ability. Reckase, Ackerman, and Carlson (1988) showed that a unidimensional test might be consisting of multidimensional items. Nandakumar (1991) have also reported that there could be minor dimensions in a unidimensional test and when the effect of the minor dimensions increased, the unidimensionality of the test might be violated.

In many studies in the literature DIF analysis were conducted in an approximately unidimensional data, but the effects of the minor dimensions in the DIF analysis results were not considered. One of the aims of the present study was to investigate the possible differences when these minor dimensions were used in the matching process. To this purpose mathematical literacy subtest scores were used in examining the differences in DIF results between the multivariate and the univariate DIF analyses.

In the literature, there are studies indicating using external test scores or other variables that are related to primary dimension measured by the test is an effective way in improving the matching of the same ability individuals (Mazor, Kanjee & Clauser, 1995; Clauser, Nungester & Swaminathan, 1996). However, it is worth adding that identification of the meaningful and relevant dimensions with the main dimension measured by the test is difficult, if not impossible. Within the context of this current study, problem solving scores were determined to be used as an additional matching dimension in investigating the items of the mathematics literacy test of PISA 2003 study through DIF methodologies. Results from this analysis using both problem solving and mathematics literacy test scores as matching variables were compared with the results of the analysis using only mathematics literacy test scores in determining the same ability students.

Another purpose of this study was investigating the possible sources of DIF. There are a few studies in the literature mentioning the possible sources of DIF, most of which have inconsistent results with each other. Investigating DIF in different contexts may produce a set of consistent results which may lead disentangling the sources of DIF in future studies.

CHAPTER III

METHOD

3.1. Population and Sample

The target population of PISA 2003 was international 15 year-olds students attending educational institutions located in each country. PISA 2003 survey was conducted in 41 countries and national target population was, "all students born in 1987 who were attending a school or any educational institution". Accessible population was more than a quarter of a million students, representing almost 30 million 15 year-olds students. The sample design for PISA 2003 was a two stage stratified sampling in most countries. In a few countries three-stage design was used. According to variables such that school type (public/private), school size, geographical area and language, the formulation of the minimum number of schools and students were developed and used in each country. As a result of these sampling designs, minimum 150 schools and 4500 students were selected in each participating country. With this formulation, 4855 Turkish and 5456 American students were sampled (OECD, 2005).

Students who answered the 3rd and the 13th booklets of the study were selected in this study. Because these booklets include the maximum number of released items and students answered both mathematical literacy and problem solving items. Demographic information of these students is given in Table 3.1 and Table 3.2.

Table 3.1 Demographic information of students in the 3rd Booklet

	3rd Booklet		
	Female	Male	Total
Turkish	167 (44%)	212 (56%)	379 (100%)
American	202 (48%)	218 (52%)	420 (100%)

Table 3.2 Demographic information of students in the 13th Booklet

	13th Booklet		
	Female	Male	Total
Turkish	164 (46%)	196 (54%)	360 (100%)
American	188 (45%)	228 (56%)	416 (100%)

3.2. Instruments

PISA 2003 survey covered reading, mathematical and scientific literacy, and problem solving. Data from the mathematical literacy section was the focus of the analysis in this study. In PISA 2003 double translation (i.e. two independent translations from the source language with reconciliation by a third person) from two different languages was used and tests were administered in 33 languages. Items used in Turkey were double translated from the English versions. Experts' form participating countries ensured that instruments were valid and took into account the cultural and educational contexts of the member of the OECD countries.

PISA 2003 mathematics literacy items consisted of an introduction part, then the actual question. Items selected for the mathematics instrument represent four situation types i.e. personal, educational or occupational, public and scientific.

Item contexts were related real life situations with four overarching ideas: space & shape, change & relation, uncertainty and quantity. Item solutions require reproduction, connection and reflection processes (OECD, 2003).

PISA 2003 problem solving items include three types of problem; decision-making, system analysis and design, and trouble shouting (OECD, 2003).

For the mathematical literacy domain, 85 items were selected for use in the study of PISA 2003. For the problem solving minor domain, 19 items were selected for use in the study of PISA 2003. Item types, in mathematics literacy and problem-solving tests, were open constructed and closed constructed and multiple-choice type (OECD, 2005).

In each booklet there was different number of mathematical literacy and problem solving items. There were 34 mathematical literacy items and 9 problem solving items in the 3rd booklet and there were 23 mathematical literacy items and 9 problem solving items in the 13th booklet. Examples of items that were used in the mathematics literacy and problem solving tests were given in Appendix F1-F2.

3.3. Test Design

Student achievement in mathematics was assessed using 85 test items representing approximately 210 minutes testing time. Problem solving assessment consisted of 19 items representing approximately 60 minutes of testing time. The 167 main study items were allocated to 13 clusters (seven mathematics clusters, and two clusters in each of the other domains). Each cluster represented 30 minutes of test time. There were 13 test booklets and there were 4 clusters in each booklet according to rotation design. Each cluster appeared in each of the four possible positions within a booklet exactly ones. Each test item therefore appeared in four of the test booklets. Students were randomly assigned one of the booklets. A special one-hour booklet was prepared for students with special needs. The two-hour test booklets were administered in two one-hour parts and there was short break

between administrations of these two part test booklets. But there were longer break between the administration of the test and the questionnaire (OECD, 2005).

3.4. Analysis of Data

To obtain consistency and reliability a detailed coding scheme was developed to code the student responses. Double-digit code was used to distinguish cognitive processes and knowledge for items requiring constructed responses. First digit indicates the score (degree of correctness for the constructed response) and second digit indicates the approach or method, which used by the student to get the correct answer. One digit code was used for multiple choice and some open constructed items (OECD, 2003).

3.4.1. Dimensionality

To determine the dimensionality of PISA 2003 mathematical literacy test, both exploratory and confirmatory factor analyses (EFA and CFA) were conducted through the use of SPSS 13 and LISREL 8.72 program (Jöreskog & Sörbom, 2001).

It is worth restating that to conduct item level analyses within the context of cross-cultural studies evaluating equivalence, the tests under investigation should possess a common structure, or in other words the tests should have an equivalent construct. To this purpose, before carrying on the DIF analyses, structure of the constructs measured by the different language forms of the tests was investigated through EFA and CFA.

3.4.2. Construct Equivalence

In this study multi-group factor analysis (MGFA) was used to investigate the construct equivalence by structural equation models. Construct equivalence of the groups was tested through the use of PRELIS 2.72 and LISREL 8.72 programs (Jöreskog & Sörbom, 2001; Jöreskog & Sörbom, 2002).

MGFA offers a specific structural equation model and investigates whether the model can be reproduced in both of the groups. In this current study, it was investigated whether a unidimensional model was reproduced in USA and Turkish groups. In other words, it was investigated whether both English and Turkish versions of the tests were unidimensional.

However, a common scale is required to form a basis for the comparison of constructs of different versions of the tests. In addition, as ordinal variables do not have a unit or an origin, a continuous variable to define a metric for the corresponding ordinal variable is required (Joreskog & Sörbom, 2001)

Fortunately, PRELIS and LISREL programs offer solutions to overcome this metric and common scale challenges. In this current study, an underlying continuous variable (threshold) for each ordinal variable (items) was estimated through PRELIS program, using the pooled data of American and Turkish groups in one data file. Then to estimate variable means, these common threshold values were used in each of the American and Turkish groups. Using these means, factor loadings and measurement errors for each item were estimated through the use of LISREL program (Jöreskog & Sörbom, 1993). Factor loadings indicate the relationship between the observed and latent variables. Factor loadings can also be considered as the validity coefficients and the measurement errors are the basis of the reliability coefficients.

However, an additional issue to be specified, which also accounts for the reason of conducting multivariate analysis in this study, is that; beyond a model which fits a data an alternative model may also fit the data as well. Determining the fidelity of the model is usually a context issue, i.e. the best model to be selected

among the models fitting a data can be specified with respect to the purpose of the study. That is why Joreskog says that, "a model is need not to be true to be useful" (Jöreskog, 2005).

In this context, it was investigated whether a two-dimensional model also fits both English and Turkish data.

The model data fit was evaluated through the goodness of fit indices provided in the output of LISREL 8.72 program. There are different fit indices and recommendations about interpretation of these indices in evaluating model-data fit. The fit indices used in the study were as follows:

Chi-square (χ^2): It measures the difference between the sample covariance (correlation) matrix and the fitted covariance (correlation) matrix. A small (zero) chi-square indicates good (perfect) fit and a large chi-square indicates bad fit. It is depend on sample size (Jöreskog & Sörbom, 1993).

Normed Chi-square: Adjusted chi-square that is the ratio of χ^2 and its degrees of freedom. χ^2 / df value less than 5 indicates good fit. If this ratio is less than 2, model over fits the data (Kelloway, 1998).

Root-Mean-Squared Error of Approximation (RMSEA): It is a measure of discrepancy per degree of freedom. The value of 0.05, and smaller, for RMSEA means a close fit and the value of 0.08 acceptable with reasonable errors of approximation in the population (Jöreskog & Sörbom, 1993).

Goodness of Fit Index (GFI): It does not depend on sample size and measures how much better the model fits as compared to no model. The range of the GFI is from 0 to 1. The values exceeding 0.9 indicates a good fit to the data (Jöreskog & Sörbom, 1993; Kelloway, 1998).

Adjusted goodness of fit index (AGFI): It is an adjusted goodness of fit measures. This index has a range from 0 to 1. 0.90, and higher, indicates a goof fit to the data (Kelloway, 1998).

Comparative fit index (CFI): CFI have been recommended by Bentler (1980). CFI supposed to lie between 0 and 1 and the value of 0.90 and higher indicates a good fit (Jöreskog & Sörbom, 1993).

Non-Normed fit index (NNFI): NNFI measures how much better fits as compared to a baseline model usually the independence model (Jöreskog & Sörbom, 1993). NNFI is the adjusted NFI which is based on percentage improvement in fit over the baseline independence model (Bentler & Bonet, 1980). Underestimation of the fit of the model with small samples is the disadvantage of NFI. NFI take control of this disadvantage. Higher values of NNFI of 0.90 indicate a good fit (Kelloway, 1998).

Root-Mean-Square Residual (RMR): The last index that was used in the study is the square root of the mean of the squared differences between the implied and observed covariance matrices which is called root- mean-square residual (RMR). Low values of standardized RMR values in LISREL indicate good fit with a lower bound of 0 and upper bound of 1. The RMR values which is less than 0.05 generally accepted values for the good fit (Kelloway, 1998).

If the model does not fit the data, one should consider how the model can be modified to fit the data better. For this purpose, fitted and standardized residuals and modification indices (MI) are useful. MI values determine the estimated decrease in chi-square value when a corresponding parameter is set to be freely estimated. The distribution of MI is approximately chi-square with one degree of freedom. However, as MI values are influenced by sample size, in determining the significance of MI's an adjustment procedure as suggested by Oort (1992) was used.

Adjusted modification indexes (AMI) were calculated by the formula:

$$\mathrm{AMI} = \frac{(\mathrm{df} - 1)}{\left(\chi^2 - \mathrm{MI}\right)} \times \mathrm{MI}$$

χ^2, df and MI values are the estimated values in LISREL output. The corresponding parameter specified by the largest modification index was set to be freely estimated prior to re-running the LISREL program. This process continued until there were no significant MI indices, i.e. the values greater than the critical value of 3.841 at the 0.05 level of significance.

3.4.3. Matching Criterion

In DIF analysis, matching individuals having same ability from reference and focal groups is an important issue. Because an item functioning differentially across groups is defined as an item affected by additional dimensions to that of specified by the matching criterion, matching criterion should be an adequate representation of all the dimensions required to respond an item correctly. To this purpose not only the univariate matching criterion was used to specify the students of the same ability but the affect of using multivariate matching criterion to the result of DIF analyses was investigated in this study. In the univariate analysis students matched on the total test scores. On the other hand two different perspectives were used in the multivariate analysis.

First, to determine whether differences that were found in the univariate analysis were depending on an additional ability, problem solving scores in addition to the mathematical literacy scores were used simultaneously in multivariate LR analysis. It was hypothesized that this external matching variable may provide an additional contribution to distinguish item impact from DIF (Williams, 1997; Clauser, Nungester & Swaminathan, 1996). Pearson correlation between the problem solving and mathematics literacy scores provided evidence that problem solving scores were related to mathematics literacy scores. Pearson correlation of problem solving and mathematics literacy scores was 0.77 and 0.75 in the 3rd and in the 13th booklets, respectively.

In the second perspective an internal matching criterion was used. To this purpose the total test score was divided into subdimension scores with respect to EFA results (Hamilton & Snow, 1998). Then, instead of a single total score these subtest scores were used to match students in the same ability.

It was hypothesized that, by matching on a more refined criterion, fewer items would be revealed as showing DIF, because item performance would be compared for groups of students whose ability levels are presumable more similar than those matched on total score alone.

3.4.4. Purification of Matching Criterion

As stated before, comparing the individuals within the context of DIF analysis requires identification of the best matching variable. Total test score may not always be a perfect matching criterion (Zieky, 1993). For example, when a large number of DIF items are present, the appropriateness of the conditioning on total test score can be questionable. Because total test score may be distorted by the large number of DIF items. This problem is known as circularity problem in DIF literature. To overcome this problem, purification of matching criteria was used in both univariate and multivariate analysis in the present study (Dorans & Holland, 1993; Donoghue, Holland & Thayer, 1993; Zenisky, Hambleton & Robin, 2003; Camilli & Shepard, 1994).

To purify the matching criterion, the items determined as showing high-DIF in the first run of the programs were not included in calculating the matching variable scores in the subsequent analyses. However, the item under investigation is always included in the matching variables as argued by Zumbo (1999).

Another reason of using purification strategy in this study was that, purification strategy has been shown to work empirically in LR analysis (Zumbo, 1999) and purification strategy has been reported to be equal or superior to the single step M-H analysis with equal and unequal ability distributions of groups in M-H analysis (Clauser, Mazor & Hambleton, 1993).

3.4.5. DIF Methods

In this study, two nonparametric DIF methods; Mantel Haenszel (M-H) and Logistic Regression (LR) were used for the item level analysis. This section gives descriptions of the (M-H) and (LR) DIF methods.

3.4.5.1. Mantel-Haenszel Method

M-H DIF method assumes that if individuals know approximately the same amount according to test score, then they should perform in approximately the same way on an individual test item regardless of group membership.

With this assumption, M-H DIF method tests the hypothesis that there is no relation between group membership and test performance on the item after controlling for ability.

To test this null hypothesis, for the comparison of studied groups M-H DIF method creates 2x2 contingency tables consisting group by item success for each item.

Table 3.3 Contingency Table for M-H Statistics

	Score on studied item		
Group	**Right**	**Wrong**	**Total**
Focal	R_{fm}	W_{fm}	N_{fm}
Reference	R_{rm}	W_{rm}	N_{rm}
Total	R_{tm}	W_{tm}	N_{tm}

The contingency table of score level (m), studied item (i), reference group (r) and focal group (f) can be displayed as in Table 3.3. This table indicates the

number of individuals in reference and focal groups having right (R) and wrong (W) answers to the studied item.

With respect to the indices given in the 2x2 contingency table, the null hypothesis to be tested is (Holand & Wainer, 1993):

$$H_0: \left[R_{rm}/W_{rm}\right]/\left[R_{fm}/W_{fm}\right]=1 \; m=1, 2, \ldots M$$

The measure of item performance is obtained by dividing the number of correct answers by the number of incorrect answers. This ratio is called an odds ratio of the right to wrong answers. It is formed for each group at each score categories.

These score categories can be obtained in two different ways. First way is using total score as the matching variable which is called thin matching. In this matching type each total score determines a score category indicating individuals of the same ability. Second way is forming the matching variable by pooling the total score levels which is called thick matching (Donoghue & Allen, 1993). The main difference between these matching strategies is the number of score categories. The score categories of thick matching are less than the score categories of thin matching.

Each score category must include both correct and incorrect responses of reference and focal groups in M-H DIF analysis. If the number of examinees is small, thin matching may not satisfy this requirement. If this condition is not satisfied, using thick matching is suggested in the literature to obtain a good power of M-H statistics. Thick matching can increase stability and so decrease the variability of M-H statistics. Thick matching can improve the performance of the MH procedure. For short tests (5 or 10 items), thin matching gives the poor results. But for long tests thin matching is the best solution. For shorter tests (20 items or fewer) thick matching is better than thin matching. When $MH_{\chi 2}$ is used for DIF, pooling approximately equal numbers of examinees (percent total) or equal numbers of focal group members (percent focal) yields the best results (Donoghue & Allen, 1993).

In the present study there were 19 and 23 selected items in the 13th and in the 3rd booklets respectively, so total percent thick matching strategy was used in MH analysis. To find the categories for thick matching, 20th, 40th, 60th and 80th percentiles of the total score of the pooled data were calculated.

To get a common value to represent all the odds ratios of each score category constant odds ratio is calculated. The estimate of the constant odds ratio is,

$$\alpha_{MH} = \frac{\sum_{m} R_{rm} W_{fm} / N_{tm}}{\sum_{m} R_{fm} W_{rm} / N_{tm}}$$

This formula is also an estimate of DIF effect size and its metric ranges from 0 to ∞ with a value of 1 indicating no-DIF. Under the null hypothesis, α_{MH} is equal to one and it means that focal and reference group perform equally on studied item. If α_{MH} is greater than 1, studied item favors reference group. If α_{MH} is less than 1, studied item favors focal group (Dorans & Holland, 1993; Donoghue, Holland & Thayer, 1993).

The M-H DIF method yields chi-square test which is distributed with one degree of freedom. In this study this chi-square value was calculated at 5% level in determining the significance of the statistics. With reference to Table 3.3, M-H statistics is calculated as follows:

$$MH\chi^2 = \frac{\left[\left|\sum_{m} R_{rm} - \sum_{m} E(R_{rm})\right| - 0.5\right]^2}{\sum_{m} Var(R_{rm})} \text{ where,}$$

$$E(R_{rm}) = E(R_{rm} | \alpha = 1) = N_{rm} R_{tm} / N_{tm}$$

$$Var(R_{rm}) = Var(R_{rm} | \alpha = 1)$$
$$= \left[N_{rm} R_{tm} N_{fm} W_{tm}\right] / \left[N_{tm}^2 (N_{tm} - 1)\right]$$

In the expression, –0.5 serves as a continuity correction to improve the accuracy of the chi-square percentage (Hambleton & Rogers, 1989).

Odds are converted to log odds to interpret easier due to the symmetrical property around zero i.e. MH D-DIF= $-2.35\log_e(\alpha_{MH})$, as M-H measure of DIF. The negative sign in equation is to make its value negative if item is more difficult for the members of the focal group than the reference group. This value presents the average degree of increased difficulty that members of one group found the item than did comparable members of the other group.

By the classification of M-H D-DIF value into three categories as explained below, an effect size measure was provided by Educational Testing Service (ETS) for the M-H DIF method.

These three categories are;

1) Negligible DIF (A). $|\text{MH D - DIF}| < 1$. This is interpreted as item does not show DIF.

2) Moderate DIF (B). $1 \leq |\text{MH D - DIF}| \leq 1.5$. Revision is recommended for this item.

3) Large DIF (C). $|\text{MH D - DIF}| \geq 1.5$. Substantive revision or elimination or this item should be performed (Dorans & Holland, 1993; Gierl, Jodoin & Ackerman, 2000).

In the present study M-H DIF analysis was conducted using EZDIF program developed by Niels Waller (Waller, 2005).

3.4.5.2. Logistic Regression Method

LR DIF method is a contingency table approach and has the capability of using both continuous and multiple ability estimates as well as the dichotomous ability estimates. LR DIF procedure can be used with both polytomous and dichotomous items (Agresti, 2002).

In logistic regression DIF procedure, an item shows DIF if individuals with same ability but from different groups do not have the same probability of getting an answer correct (Zumbo, 1999; Swaminathan & Rogers, 1990).

The LR model for ordinal response format is

$$P(u=1)=\frac{e^{z}}{(1+e^{z})} \quad \text{where} \quad z=\tau_0+\tau_1\theta+\tau_2 g+\tau_3(\theta g).$$

In this model, u is the response of the individual 1 indicating right answer, 0 indicating wrong answer, P is the probability of individuals getting an answer correct and θ is the observed ability of an individual. g represents group membership which is defined as;

$$g=\begin{cases}1 \text{ if individual is a member of reference group}\\ 0 \text{ if individual is a member of focal group}\end{cases}$$

The term θg is the product of two independent variables, observed ability of individuals θ and group membership g. The parameters $\tau_0,\tau_1,\tau_2,\tau_3$ correspond to the intercept and weights for the ability, group difference and interaction between group and ability, respectively.

In the LR DIF model, the null hypothesis is $H_0=\tau_2=\tau_3=0$. If $\tau_2\neq 0$ and $\tau_3=0$ an item shows uniform DIF. The uniform DIF favors reference and focal groups if $\tau_2>0$ and $\tau_2<0$, respectively. An item shows non-uniform DIF if $\tau_3\neq 0$ (whether or not $\tau_2=0$). The item favors higher ability members of the reference group and the lower ability members of the focal group if $\tau_3>0$. The item favors lover ability members of the reference group and the higher ability members of the focal group if $\tau_3<0$ (Jodoin & Gierl, 2001; Rogers & Swaminathan, 1993).

The estimate of ability most often used in the LR model is total score. Also there exists a great flexibility using other estimates of ability, concomitant variables, or some combination (Camilli & Shepard, 1994). In the present study LR DIF procedure was conducted first using only total score in a univariate analysis, and using combination of two subtest scores in a multivariate analysis.

Logistic regression DIF procedure is based on a model strategy comparison. To calculate and interpret model comparison statistics, instead of the likelihood function, i.e. $(P(u=1)=\frac{e^{z}}{(1+e^{z})})$, logged likelihood function $(z=\tau_0+\tau_1\theta+\tau_2 g+\tau_3(\theta g))$ is used. The likelihood values range from 0 to 1 where log likelihood values range from negative infinity to zero. Reversing this range as from 0 to positive infinity by multiplying -2 provides the same interpretation with regression models.

This model strategy comparison is made by adding the ability, group and interaction terms into the model in a hierarchical order as shown in model1, model2 and model3 below. The univariate LR DIF model which was used in the present study was,

$$model1: Z=\tau_0+\tau_1\theta$$
$$model2: Z=\tau_0+\tau_1\theta+\tau_2 g$$
$$model3: Z=\tau_0+\tau_1\theta+\tau_2 g+\tau_3(\theta * g)$$

Model3 is the full model $(\tau_0+\tau_1\theta+\tau_2 g+\tau_3(\theta g))$ in which ability, group and interaction terms is included. Model2 is the second model $(\tau_0+\tau_1\theta+\tau_2 g)$ in which interaction term is removed from the full model. And the last model, model1, $(\tau_0+\tau_1\theta)$ in which the group variable is removed from the second model.

As a result of this formulation, the larger the difference between the models, the larger the improvement in the model due to the ability and group variables (Pampel, 2000). The improvement in the model can be analyzed in two ways. First testing uniform and non uniform DIF simultaneously, and second

testing uniform and non uniform DIF separately. In first analysis, the change between the model1 ($\tau_0 + \tau_1\theta$) and the model3 ($\tau_0 + \tau_1\theta + \tau_2 g + \tau_3(\theta g)$) is tested with a chi-square statistics with two degrees of freedom. This analysis provides testing uniform and non-uniform DIF simultaneously. When the test statistics exceeds $\chi^2_{\alpha:2}$, the hypothesis of there is no DIF is rejected. To measure the magnitude of DIF, corresponding effect size is ΔR^2, a weighted least squares measure (Swaminathan & Rogers, 1990; Zumbo, 1999). If an item shows DIF, by comparing the ΔR^2 values of model2 and model3, one can determine whether this item shows uniform or non-uniform DIF.

Zumbo (1999) have recommended ΔR^2 values below 0.13 to be regarded as negligible DIF, values between 0.13 and 0.26 as moderate DIF and values above 0.26 as large DIF. With this model1 and model3 comparison, the interaction term may decrease the power of the LR procedure when only uniform DIF is present because one degree of freedom is lost necessarily.

Although non-uniform DIF occurs with substantially lower frequency than uniform DIF (Camilli & Shepard, 1994), it is reasonable to modify the two-degrees of freedom chi-square test into separate two one-degree of freedom tests. These alternative comparisons can be made between model3 and model2, and between model2 and model1. The change between models is tested separately using chi-square statistics with one degree of freedom. Jodoin and Gierl (2001) developed an effect size criterion for this comparison and their simulation study suggested that this effect size criterion was accurate. This effect size criterion of this alternative comparison is,

Type A items: negligible DIF, $\Delta R^2 < 0.0035$

Type B items: moderate DIF, $0.0035 \leq \Delta R^2 \leq 0.070$

Type C items: large DIF, $\Delta R^2 > 0.070$.

They also found that this criterion is more powerful for detecting moderate DIF items than Zumbo's (1999) ΔR^2, a classification criterion for moderate DIF items. So, the effect size criterion of Jodoin and Gierl (2001) was used to classify DIF items in the present study. LR analysis was computed by using SPSS 13 software.

Same model comparison was used in the multivariate LR DIF model in the present study. With two ability estimates in the LR DIF procedure the exponent for each model was,

$$model1: Z = \tau_0 + \tau_1\theta_1 + \tau_2\theta_2 + \tau_3(\theta_1 * \theta_2)$$
$$model2: Z = \tau_0 + \tau_1\theta_1 + \tau_2\theta_2 + \tau_3(\theta_1 * \theta_2) + \tau_4 g$$
$$model3: Z = \tau_0 + \tau_1\theta_1 + \tau_2\theta_2 + \tau_3(\theta_1 * \theta_2) + \tau_4 g + \tau_5(\theta_1 * g) + \tau_6(\theta_2 * g) + \tau_7(\theta_1 * \theta_2 * g)$$

where model1 matches on both test scores individually and the covariance of the test scores. Model2 adds a term for uniform DIF analogous to above univariate case and model3 adds the various non-uniform DIF terms in dimensions θ_1 and θ_2.

3.5. Qualitative Review of DIF Items

After identifying B- and C-level DIF items in statistical analyses, two reviewers were tried to identify the causes of DIF items subjectively for the presence or absence of any characteristic or feature. Qualitative reviews of items refined by using the findings of the earlier studies by the reviewers to construct item review criteria for potential bias.

The reviewers were qualified to evaluate student performance because of their teaching experiences, university education, and mathematics background. The one of the reviewer was research assistant in secondary science and mathematics education department in Middle East Technical University and the other was the mathematics teacher in a collage and also author of this thesis.

Both of them had an experience in teaching mathematics in secondary and primary schools and skilled in understanding the cognitive strategies typically used by the students to solve mathematics items. They are familiar with the mathematics curriculum in Turkey, and English is their foreign language. The reviewers tried to specify the possible ways in which DIF items could differ by considering the DIF related factors that have found in the literature. First each reviewer worked independently then they discussed their findings and differences on DIF related factors resolved through their discussion.

Qualitative Review Criteria

In the literature findings showed that the sources of DIF could be classified in three groups as given in the following paragraphs (Allalouf, Hambleton & Sireci, 1999; Ercikan, 1998; Gierl & Khalig, 2000). This classification provided a basis of the criteria to disentangle the sources of DIF in the qualitative analysis of items in the present study.

Translation and Adaptation Differences

The poor translations or adaptations can affect the meaning of the content of the items and skills measured by the items. The changes in difficulty of words or sentences, content and format have been reported in the literature as the causes of DIF due to the incorrect translations. For example, the changes in difficulty of word may be a cause particularly in analogy items. A very difficult word may be translated into very easy word (Allalouf, Hambleton & Sireci, 1999).

On the other hand, changes in content of translated and the original items can change meanings of words, sentences or passages in compared groups in favor of one group. For example, grammatical structures in original language may have different equivalent forms or may not have an equivalent form in translated language. Some words that are easy in one language may not be in another language. Ercikan (1998) have found that the word “work” in English has meanings in real life and physics contexts, but translation of this word in French do not have the same meaning in physics context.

Gierl and Khalig (2000) have indicated that these differences in meanings in the words and expressions could be inherent or not inherent to language. As an inherent difference in language, they showed that some words in English have no expression that is directly parallel in French. As a not inherent language differences, they have indicated that a word in English translated into French incorrectly had alternative words and these alternatives produce items that were closer in meaning across the languages.

Punctuation, capitalization, item structure, typeface etc. are the formatting usages that may affect the performance of individuals on an item (Gierl & Khalig, 2000; Allalouf, Hambleton & Sireci, 1999). For example, in a translation, a sentence completion item may turn into a four alternative responses item. Allalouf et all. (1999) indicated that due to the constraints of Russian language, translating an English item in this way was unavoidable.

In this study it was examined whether frequency, difficulty or commonness of a vocabulary, length or complexity of sentences, contextual meaning of vocabulary and item format could be the sources of DIF due to the poor translation in mathematics literacy items.

Sources of DIF may not always be due to the poor translation or adaptation, culture differences and curriculum-related differences may also be the sources of DIF (Ercikan et all., 2004).

Cultural Differences

Gierl and Khalig (2000) have indicated that inherent characteristics of cultures may be causes of the differential performance of the individuals. The familiarity in content or context of items can change according to these inherent interests in different cultures. An example they found for cultural difference is an English item with a 12-hour clock using AM and PM while the French translation uses 24-hour clock. Differences in cultural relevance as a source of DIF also have been reported in the study of Allalouf, Hambleton and Sireci (1999). In their study, they have claimed that content of a reading comprehension passage or content of a sentence completion item might be more relevant or familiar to one of the groups.

Within this context, in the current study the content of DIF items were investigated to understand whether there were a cultural familiarity in DIF items.

Curriculum-related Differences

In this study it was investigated that whether DIF in an item was due to the curricular differences. Performance in international achievement tests mostly depends on curricular coverage of the countries. This means that, any difference among countries in topics such as algebra, data handling, number sense, or any difference in order of these topics can influence the relative performance of countries (Gierl & Khalig, 2000).

In addition to differences explained above, the DIF related factors that have been investigated by Scheuneman and Grima (1997) were also considered in this study. They have classified DIF related factors as the cognitive requirements, mathematical content and task presentation variables (notations, variables, figures, etc.) of the quantitative items. They have also studied verbal properties of items such as readability (e.g. sentence length), semantic content (e.g. adjectives, adverbs or propositions) and quantitative language (average, area, product, sum, etc.) to understand whether the differences in these properties could be causes of DIF.

CHAPTER IV

RESULTS

This chapter is divided into six sections. Each section consists of the results of the relevant statistical analyses for each booklet. In the first section, descriptive statistics of mathematics literacy and problem solving items are reported. In the second section dimensionality and construct equivalence analyses and then in the third section, selected unidimensional and two-dimensional mathematics literacy items' descriptive statistics are presented. In the forth section, combined results of M-H, LR and multivariate LR analyses and in the fifth section the comparison of univariate and multivariate DIF analyses are given. Finally the sixth section presents the qualitative reviews of the items to determine the possible sources of DIF.

In this study, all items from selected booklets were recoded according to the following criteria: if the response is fully or partially correct it was recoded as 1, and other responses recoded as 0. The item m438q01 in the 3rd booklet was not answered in the U.S. group, so this item was not included in the analyses of the 3rd booklet.

4.1 Descriptive Statistics of All Items in Booklets

Table 4.1 presents the descriptive statistics of mathematical literacy items for the 3rd and the 13th booklets of PISA 2003. The results showed that USA students performed better than Turkish students in each booklet. It is apparent that the groups have the unequal test score distributions. The difference between the skewness values of Turkey and USA indicates that Turkey has more scores than USA toward the lower end of scale.

Table 4.1 Descriptive Statistics of Mathematics Literacy Items

	3rd Booklet		13th Booklet	
STATISTICS	**TURKEY**	**USA**	**TURKEY**	**USA**
N of examines	379	420	360	416
N of items	34	34	23	23
Mean	13.10	16.68	7.67	11.21
S.D.	7.66	7.57	5.12	5.13
Skewness	0.648	0.011	0.891	0.098
Kurtosis	-0.375	-0.824	0.086	-0.747
Alpha	0.91	0.90	0.86	0.85
Mean PC	0.39	0.49	0.33	0.49
Mean Biserial	0.65	0.62	0.69	0.63

Table 4.2 indicates similar results with Table 4.1. In problem solving items USA students performed better than Turkish students in each booklet as in mathematics literacy items. There exist unequal test score distributions between groups. The difference between the skewness values of Turkey and USA indicates that Turkey has more scores than USA toward the lower end of scale.

Mathematics literacy and problem solving items were more difficult for Turkey than for USA, however discrimination values were similar for both groups in each booklet.

Table 4.2 Descriptive Statistics of Problem Solving Items

	3rd Booklet		13th Booklet	
STATISTICS	**TURKEY**	**USA**	**TURKEY**	**USA**
N of examines	379	420	360	416
N of items	10	10	9	9
Mean	3.34	4.63	2.75	4.04
S.D.	2.30	2.62	2.06	2.43
Skewness	0.609	0.166	0.763	0.145
Kurtosis	-0.230	-0.882	0.007	-0.952
Alpha	0.70	0.75	0.69	0.76
Mean PC	0.33	0.46	0.31	0.45
Mean Biserial	0.68	0.72	0.77	0.76

Difficulties and discriminations of all mathematics and problem solving items in studied booklets were given in appendices A1, A2, A3 and A4. Mean of proportion corrects of items (Mean PC) and mean of item discriminations (Mean Biserial) were calculated through the use of ITEMAN program. Other descriptive statistics were calculated using SPSS 13 (George & Mallery, 2003).

4.2 Dimensionality and Construct Equivalence

According to the rationale cited in Chapter III, the first step is to test the dimensionality of the data set. Before DIF analyses, an exploratory factor analysis was run on the entire sample of students of Turkey and USA to determine the factor structure for the whole group. Scree plots indicated a dominant factor structure in each booklet for mathematics items. It seems to suggest that factors come much closer to satisfy the unidimensionality assumption (Appendices B3 and B4).

The test of all problem solving items fit to the unidimensional model in CFA, but this is not the case for the test of mathematics literacy items. EFA produced 3 factors and 6 factors in the 3rd and in the 13th booklets, respectively (Appendices B1 and B2). To get a unidimensional test for DIF analysis, mathematics items which had highest loadings in first and second factors in each booklet were selected. Then these items tested for unidimensional and two dimensional models in CFA whether the same factor structure was present in groups. Each model gave the acceptable fit to the data. This finding was consistent with the literature. Reckase, Ackerman and Carlson (1988) showed that a unidimensional test might consist of multidimensional items that measure more than one ability to obtain a correct answer.

In CFA analyses the error variances of m421q01, m496q01t and m704q02 items in the 3rd booklet and error variances of m810q03t, m464q01t and m462q01t items in 13th booklet were negative. So these items were not included in the further analyses. Finally, 19 and 23 mathematics items were selected from the 13th and the 3rd booklets, respectively.

Then as a preliminary of DIF analyses, the construct equivalence of the selected mathematics literacy items and problem solving items was investigated via multi-group CFA. Followings are the details of the analyses.

4.2.1 Dimensionality

Table 4.3 Fit Indices of Dimensionality in the 3rd Booklet Mathematics Literacy Items

	3rd Booklet			
	Unidimensional		Two-dimensional	
STATISTICS	**TURKEY**	**USA**	**TURKEY**	**USA**
Chi-square	571.07	617.64	549.63	568.58
P values	0.0	0.0	0.0	0.0
d.f.	230	230	229	229
CFI	0.96	0.88	0.96	0.90
GFI	0.97	0.95	0.97	0.96
AGFI	0.96	0.94	0.96	0.95
NFI	0.94	0.83	0.94	0.84
NNFI	0.96	0.87	0.96	0.89
RMSEA	0.063	0.063	0.061	0.059
RMR	0.29	0.26	0.28	0.24

The dimensionality fit indices of selected mathematics literacy items are given in Table 4.3 and Table 4.4.

In the 3rd and the 13th booklets, although P values of chi-square was zero, the ratio of χ^2 / df is less than 5. This normed chi- square indicated good fit for one and two factor models in each booklet. RMSEA values indicated that the degree of approximation in the population was acceptable and the models fit to the data. CFI, GFI, and AGFI values also showed that the models fit in each group.

It is worth specifying that the NFI and NNFI values of USA were less than the 0.90 in the 3rd booklet which indicated that the model fit in Turkey better than the model fit in USA. Model fit indices demonstrated an acceptable fit except RMR values. RMR values were greater than generally accepted value of 0.05.

Table 4.4 Fit Indices of Dimensionality in the 13th Booklet Mathematics Literacy Items

	13th Booklet			
	Unidimensional		Two-dimensional	
STATISTICS	**TURKEY**	**USA**	**TURKEY**	**USA**
Chi-square	254.54	398.67	251.48	348.47
P values	0.00	0.00	0.00	0.00
d.f.	152	152	151	151
CFI	0.94	0.85	0.94	0.88
GFI	0.97	0.96	0.97	0.96
AGFI	0.96	0.95	0.96	0.96
NFI	0.86	0.79	0.87	0.81
NNFI	0.93	0.84	0.93	0.87
RMSEA	0.043	0.063	0.043	0.056
RMR	0.20	0.18	0.19	0.17

NNFI, CFI and NFI values also revealed that the models fit the data for Turkey better than for USA in the 13th booklet.

Table 4.5 Fit Indices of Dimensionality in Problem Solving Items

	3rd Booklet		13th Booklet	
STATISTICS	**TURKEY**	**USA**	**TURKEY**	**USA**
Chi-square	35.27	41.80	59.52	49.33
P value	0.46	0.20	0.000	0.005
d.f.	35	35	27	27
CFI	1.00	0.99	0.91	0.96
GFI	0.99	0.99	0.98	0.99
AGFI	0.99	0.99	0.97	0.98
NFI	0.91	0.92	0.85	0.91
NNFI	1.00	0.98	0.88	0.95
RMSEA	0.0046	0.022	0.058	0.045
RMR	0.073	0.063	0.11	0.097

The dimensionality fit indices of selected problem solving items are given in Table 4.5. The chi-square value of exact fit was less than 5 times of degrees of freedom, and the RMSEA values were approximately around the recommended value of 0.05 in the 13th booklet.

In the 3rd booklet fit is better than the 13th booklet. Chi-square values indicated exact fit for the model. It was seen that the RMSEA values were below the recommended value of 0.05. It means that the degree of approximation in the population was too large and the models fit well. RMR values were also small, indicating good fit for the model.

The degree of approximation in the population is large and the model fit well. CFI, GFI and AGFI values were above the recommended value. These values also showed that the model fit well. NFI and NNFI values indicate better fit in USA than in Turkey.

4.2.2 Construct Equivalence

In the 3rd booklet, according to suggestions of Oort (1992), intercepts and loadings of some items calculated in each group in one-factor and two-factor multi-group analysis to assess whether less constraints improved the model fit according to modification indices of LISREL output. Modification indices were interpreted at alpha level of 0.05 and only significant indices were interpreted. An example of LISREL syntax for multi-group analysis is given in appendix G1.

In CFA, selected items fitted a unidimensional model in both groups but multi-group CFA analysis indicated that some item parameters in unidimensional model were not equivalent across groups. To determine possible sources of the lack of fit, factor intercepts of items m124q01, m421q03, m438q02 and m155q02t and factor loadings of m124q03t, m547q01t and m571q01 were allowed to be different for the two groups in the 3rd booklet. Although both unidimensional and two-dimensional models fit the data of each groups, it seems that the two-dimensional model fits better than the unidimensional model in the 3rd booklet.

Table 4.6 Multi-group Fit Indices of Mathematics Items

	3rd Booklet		13th Booklet	
	Unidimensional	**Two-dimensional**	**Unidimensional**	**Two-dimensional**
STATISTICS	**Global goodness of fit**		**Global goodness of fit**	
Chi-square	1569.00	1484.29	899.78	830.31
P values	0.0	0.0	0.00	0.00
d.f.	521	524	340	356
CFI	0.91	0.92	0.84	0.86
RMSEA	0.071	0.068	0.065	0.059
NFI	0.87	0.88	0.76	0.78
NNFI	0.91	0.92	0.83	0.87

Multi-group analyses fit indices of mathematics literacy items were given in Table 4.6. Although P values of chi-square is zero, the ratio of χ^2 / df is less than 5. This normed chi-square indicated fit of unidimensional and two-dimensional models. In addition, RMSEA values indicated that the degree of approximation in the population acceptable.

Multi-group fit indices of problem solving items are given in Table 4.7. In multi-group analysis of problem solving items, factor intercept of the item x412q01 was allowed to be different for the two groups to provide the better fit in the 3rd booklet. Although the chi-square value of exact fit was rejecting the model, since P values were very small, the ratio of χ^2 / df was less than 5. It was seen that the RMSEA value exceeded the recommended value of 0.05 but the values indicated a reasonable error of approximation in the population. NNFI values indicated good fit but NFI values less than the recommended value.

Table 4.7 Multi-group Fit Indices of Problem Solving Items

	3rd Booklet	**13th Booklet**
STATISTICS	**Global goodness of fit**	**Global goodness of fit**
Chi-square	158.34	158.23
P values	0.00	0.00
d.f.	97	62
CFI	0.92	0.89
RMSEA	0.059	0.063
NFI	0.83	0.82
NNFI	0.93	0.90

The path diagrams which present the estimated factor loadings and the error variances of the selected unidimensional and two-dimensional items were given in appendices C1-C6.

The goodness of fit statistics given above indicated that an acceptable equivalent construct was present in selected items for original and translated tests for unidimensional and two-dimensional structures. This result was consistent with the literature i.e. there is not a single solution to the construct equivalence of translated achievement tests. Several methods that deal with different kinds of equivalence should be used (Jöreskog & Sörbom, 1993). The evidence of this study

suggested that although the tests have unidimensional equivalent structure, there might be a multidimensional equivalent structure among groups.

Investigating dimensional structures, NC, GFI, AGFI, RMSEA, NFI, NNFI and CFI indices generally provided acceptable values for the model fit for mathematical literacy and problem solving items but RMR values were relatively high especially in mathematical literacy test. Although in the literature, Sireci, Bastari and Allalouf (1998) suggested the use of RMR for model fitting, this value did not indicate the model fit to the data.

4.3 Descriptive Statistics of Selected Items in Booklets

Descriptive statistics of unidimensional mathematics literacy test scores on the 3rd and on the 13th booklets is given in Table 4.8. The differences were consistent with the difference of the all items in each booklet.

Table 4.8 Descriptive Statistics of Unidimensional Mathematics Items

	3rd Booklet		**13th Booklet**	
STATISTICS	**TURKEY**	**USA**	**TURKEY**	**USA**
N of examines	379	420	360	416
N of items	23	23	19	19
Mean	9.36	12	6.88	10.06
S.D.	5.74	5.45	4.50	4.45
Skewness	0.515	-0.119	0.705	-0.077
Kurtosis	-0.718	-0.818	-0.354	-0.816
Alpha	0.88	0.86	0.85	0.83
Mean PC	0.41	0.52	0.36	0.53
Mean Biserial	0.68	0.65	0.69	0.64

A 15- item set and an 8- item set were selected for first and second matching criteria respectively in the 3rd Booklet. Psychometric characteristics of these items are summarized in Table 4.9.

Table 4.9 Descriptive Statistics of Two-dimensional Mathematics Literacy Items in the 3rd Booklet

	3rd Booklet			
STATISTICS	**TURKEY**		**USA**	
	Factor		Factor	
	1	**2**	1	**2**
N of examines	399		420	
N of items	15	**8**	15	**8**
Mean	6.30	**3.06**	8.80	**3.20**
S.D.	3.81	**2.32**	3.90	**2.08**
Skewness	0.438	**0.513**	-0.349	**0.361**
Kurtosis	-0.738	**-0.850**	-0.853	**-0.653**
Alpha	0.81	**0.77**	0.82	**0.69**
Mean PC	0.42	**0.38**	0.59	**0.40**
Mean Biserial	0.68	**0.80**	0.69	**0.73**

A 10- item set and a 9- item set were selected for first and second matching criteria respectively in the 13th Booklet. Psychometric characteristics of these items are summarized in Table 4.10.

Table 4.10 Descriptive Statistics of Two- dimensional Mathematics Literacy Items in the13th Booklet

	13th Booklet			
STATISTICS	**TURKEY**		**USA**	
	Factor		Factor	
	1	**2**	1	**2**
N of examines	360		416	
N of items	10	**9**	10	**9**
Mean	2.89	**4,14**	4.10	**6,21**
S.D.	2.51	**2.53**	2.68	**2.40**
Skewness	1.046	**0,327**	0.349	**-0.560**
Kurtosis	0.236	**-0.808**	-0.861	**-0.287**
Alpha	0.77	**0.71**	0.75	**0.70**
Mean PC	0.29	**0.44**	0.41	**0.66**
Mean Biserial	0.78	**0.71**	0.71	**0.72**

Exploratory factor analysis of all items gives the item labels and loadings in each factor (Appendix B1 and B2). In both dimensions USA students perform better than Turkey students. For the selected mathematics literacy items, it is seen that the groups have the unequal test score distributions in first and second dimensions.

In a multidimensional case, for shorter subtests matching examinees using subtest scores may not be appropriate if these subtests are not reliable (Donoghue, Holland & Thayer, 1993). In the present study although the subtests were consisted

of 8 to15 items, alpha values of these subsets were approximately 0.70, which was reasonable.

4.4 Analyses of Differential Item Functioning

To examine the differences in results of using different matching criteria, four analyses were conducted. The first and second analyses were the matching total score using M-H and LR methods. The second and third analyses were multivariate matching of two different score using LR, matching on mathematics literacy and problem solving total scores and matching two factor subtest score. The studied item was included in forming the matching criteria and iterative purification was used in all of the DIF analyses. All test statistics were interpreted at an alpha level of 0.05. In all comparisons described below, items with B- or C-level rating were considered DIF items whereas those with an A-level rating were not. Detailed outputs of univariate M-H and LR and multivariate LR analyses were given in the appendices D1, D2 and E1-E6. Also syntax for the LR DIF analysis is given in appendix G2.

The combined results of the MH, LR and multivariate LR analyses are given in Table 4.11 and Table 4.12 in the 3rd and the 13th booklets, respectively. All items flagged as DIF showed uniform DIF in LR analyses.

Table 4.11 Results of M-H, LR and Multivariate LR analyses in the 3rd Booklet

Item Name	p-value Turkey	p-value USA	MH	LR	Multivariate LR	
					LR_{M1}	LR_{M2}
m124q01	0.37	0.25	CF	CF	CF	CF
m124q03t	0.37	0.43	BF	A	A	A
m144q03	0.60	0.76	A	A	A	A
m155q01	0.49	0.66	A	A	A	A
m155q02t	0.35	0.73	CR	CR	CR	CR
m155q04t	0.34	0.52	A	A	A	A
m420q01t	0.34	0.57	BR	A	A	A
m421q03	0.39	0.29	CF	CF	CF	A
m438q02	0.40	0.44	BF	A	BF	A
m442q02	0.23	0.34	A	A	A	A
m447q01	0.47	0.63	A	A	A	A
m462q01t	0.21	0.25	A	A	A	A
m468q01t	0.37	0.55	A	A	A	A
m474q01	0.49	0.68	BR	A	A	A
m484q01t	0.37	0.56	A	A	A	A
m496q02	0.48	0.58	A	A	A	A
m505q01	0.29	0.36	A	A	A	A
m509q01	0.36	0.54	A	A	A	A
m510q01t	0.25	0.41	A	A	A	A
m547q01t	0.67	0.69	BF	A	A	BF
m559q01	0.49	0.57	A	A	A	A
m571q01	0.36	0.41	A	A	A	A
m704q01t	0.67	0.77	A	A	A	A

LR_{M1}: analyses based on problem solving and mathematics scores
LR_{M2}: analyses based on two- mathematics subtest scores
CR and BR: favoring reference group
CF and BF: favoring focal group

Table 4.11 indicates that in unidimensional case, M-H and LR statistics produced same results for C-level DIF items but MH results showed additional B-level DIF items although univariate LR method did not find them. 8 of 23 items (35 %) displayed DIF in MH and 3 of 23 items (13 %) displayed DIF in LR. Results showed that MH and LR analyses identified same items as C-level DIF in univariate case. Using problem solving scores additional to mathematics literacy scores did not change the result. Only one more B-level item identified as DIF. Same C-level DIF items were found as DIF items in both univariate LR and multivariate LR based on using problem solving scores and mathematics literacy scores.

Using two-factor mathematics literacy subtest scores to match the students did not reduce the number of DIF items. 3 of 23 items (23 %) displayed DIF matching on mathematics factor subtest scores. But the item m547q01t showed B-level DIF in the matching using two-mathematics literacy scores although this item did not showed DIF in other LR analyses and the item m421q03 showed no-DIF in the matching using two-mathematics literacy scores although this item showed C-level focal DIF in all other analyses in the study.

Table 4.12 Results of M-H, LR and Multivariate LR Analyses in the 13th Booklet

Item Name	p-value Turkey	p-value USA	MH	LR	Multivariate LR	
					LR_{M1}	LR_{M2}
m033q01	0.52	0.76	BR	A	A	A
m124q01	0.28	0.31	BF	BF	BF	A
m124q03t	0.32	0.45	A	A	A	A
m179q01t	0.19	0.50	CR	BR	BR	CR
m402q01	0.41	0.46	A	A	A	A
m402q02	0.12	0.32	BR	A	A	CR
m438q01	0.50	0.84	CR	CR	CR	CR
m438q02	0.30	0.39	A	A	A	A
m467q01	0.34	0.56	A	A	A	A
m474q01	0.48	0.65	A	A	A	A
m505q01	0.24	0.45	A	A	A	A
m510q01t	0.22	0.49	CR	BR	BR	A
m547q01t	0.64	0.72	A	A	A	BF
m564q01	0.41	0.47	A	A	A	A
m564q02	0.37	0.38	BF	BF	BF	A
m806q01t	0.43	0.63	A	A	A	A
m810q01t	0.44	0.73	CR	A	A	A
m810q02t	0.52	0.69	A	A	A	A
m833q01t	0.15	0.25	A	A	A	A

LR_{M1}: analysis based on problem solving and mathematics scores
LR_{M2}: analysis based on two- mathematics subtest scores
CR and BR: favoring reference group
CF and BF: favoring focal group

Table 4.12 indicates that M-H and LR procedures produced similar results in matching total score. 8 of 19, or 42 % of items tested in MH analysis and 6 of 19, or 32 % of items tested in LR analysis displayed DIF.

The items m124q01 and m564q02 showed significant B-level DIF, which favors focal group Turkey, in M-H, LR and multivariate LR analysis based on problem solving and mathematics literacy scores. But these items showed no DIF in multivariate LR analysis based on two mathematical literacy subtest scores.

The items m402q02 showed C-level reference and m547q01t showed B-level focal DIF in LR analysis based on two factor subtest score although they were not DIF items in other LR analyses. The items m179q01t and m438q01 showed significant DIF, which favor reference group USA, in all analyses. Three items (m124q01, m510q01t and m564q02) did not flagged as DIF in LR analysis based on two mathematical literacy subtest scores although they showed DIF in univariate LR analysis and multivariate LR analysis based on problem solving and mathematical literacy scores.

4.5 Comparison of the DIF Procedures

The M-H and LR comparison in univariate analyses in the 3rd and in the 13th booklets is given in Table 4.13. The agreement is 78% in the 3rd booklet and 84% in the 13th booklet. M-H analysis identified more items than LR analysis.

Table 4.13 M-H and LR Comparison

	3rd Booklet			**13th Booklet**		
	DIF	**No-DIF**	**Total**	**DIF**	**No-DIF**	**Total**
M-H	8	15	23	8	11	19
LR	3	20	23	5	14	19
Agreement	3	15	18 (78%)	5	11	16 (84%)

Common item comparison between booklets for M-H and LR analyses is given in Table 4.14.

Table 4.14 Common Item Comparison in Univariate Analysis

	M-H			**LR**		
	DIF	**No-DIF**	**Total**	**DIF**	**No-DIF**	**Total**
3rd Booklet	5	2	7	1	6	7
13th Booklet	2	5	7	2	5	7
Agreement	1	2	3 (43%)	1	5	6 (86%)

The comparison of the univariate LR and the multivariate LR (LR_{M1}) based on problem solving and mathematical literacy scores, in the 3rd and in the 13th booklets is given in Table 4.15. The agreement was high between analyses. The number of items identified as DIF and no-DIF was same in each booklet. Also the univariate LR and the multivariate LR (LR_{M1}), based on problem solving and mathematical literacy scores, identified the same items as DIF and no-DIF in each booklet.

Table 4.15 LR and LR_{M1} Comparison

	3rd Booklet			**13th Booklet**		
	DIF	**No-DIF**	**Total**	**DIF**	**No-DIF**	**Total**
LR	3	20	23	5	14	19
LR_{M1}	4	19	23	5	14	19
Agreement	3	19	22 (96%)	5	14	19 (100%)

LR_{M1}: analysis based on problem solving and mathematics scores

The comparison of the univariate LR and the multivariate LR (LR_{M2}), based on two- mathematics literacy subtest scores, in the 3rd and in the 13th booklets is given in Table 4.16. Although there was a high consistency between the univariate LR and the multivariate LR (LR_{M1}) based on problem solving and mathematical literacy scores, this consistency was decreased between the univariate LR and the multivariate LR (LR_{M2}) based on two-mathematical literacy subtest scores, especially in the 13th booklet.

Table 4.16 LR and LR_{M2} Comparison

	3rd Booklet			**13th Booklet**		
	DIF	**No-DIF**	**Total**	**DIF**	**No-DIF**	**Total**
LR	3	20	23	5	14	19
LR_{M2}	3	20	23	4	15	19
Agreement	2	19	21 (91%)	2	12	14 (74%)

LR_{M2}: analysis based on two- mathematics subtest scores

4.6 Qualitative Analyses of Released DIF Items

When two booklets and all analyses were considered, significant findings indicate that when Turkey and USA were matched on different scores, American students perform relatively better than Turkish students on items m179q01, m402q02, m438q01 and m155q02t and Turkish students perform relatively better than American students on items m547q01t, m124q01.

Followings are the some characteristics of items flagged as exhibiting DIF in this study. This discussion is hunted to the released items in the PISA 2003 study. Explanations and findings were limited because all items were not released and described in PISA 2003 technical report.

There was not consistency according to item content category in DIF or non-DIF items. Two of four items favoring USA were in uncertainty content and other two were in change and relation content. One of two items favoring Turkey was in space and shape content and the other was in change and relation content.

The four items favoring USA and two items favoring Turkey were primarily coded- response items.

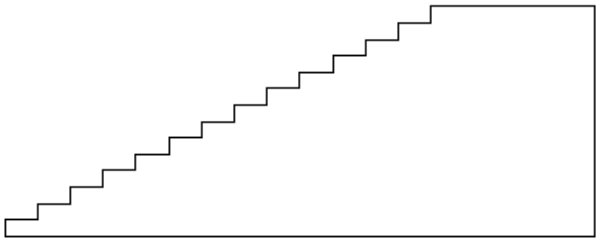
An interesting finding was performance on the item demands and clusters and context, which defined in PISA 2003 study. The followings are the items and the item demands of these items given in PISA 2003 technical report.

Item m547q01 requires interpreting simple and familiar picture, simple calculation (division by two- digit number). It is in reproduction competency and educational / occupational context

STAIRCASE

Question 1: STAIRCASE *M547Q01*

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



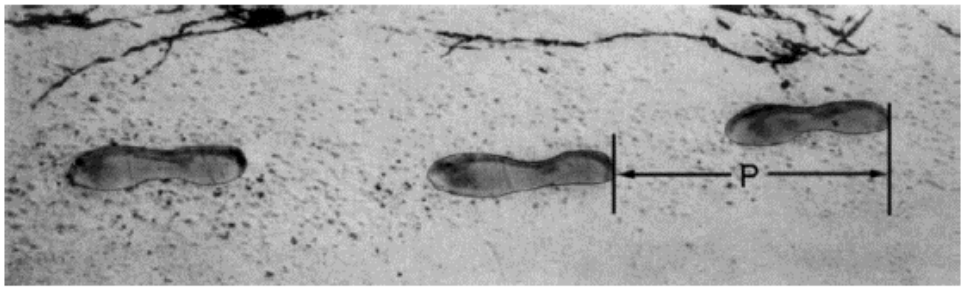
What is the height of each of the 14 steps?

Height: ..... cm.

Turkish version was given in appendix F2.

Item m124q01 needs to interpret and link picture, text and algebra; algebraic substitution, solve basic equation, single step, correct manipulation of expressions containing symbols. It is in reproduction competency and personal context.

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.

Turkish version was given in appendix F2.

Item m179q01 needs to interpret a graphical representation, construct a partially correct explanation of a mathematical concept, mathematical argumentation skills based on use of data. It is in connections competency cluster and public context.

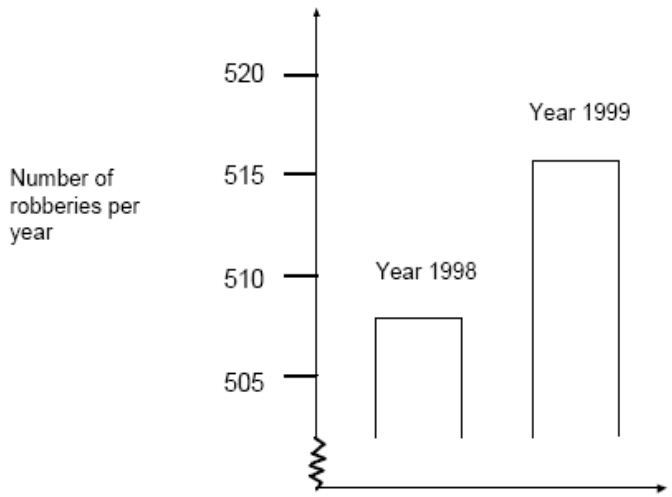
ROBBERIES

Question 1: ROBBERIES

M179Q01- 01 02 03 04 11 12 21 22 23 99

A TV reporter showed this graph and said:

"The graph shows that there is a huge increase in the number of robberies from 1998 to 1999."



Do you consider the reporter's statement to be a reasonable interpretation of the graph? Give an explanation to support your answer.

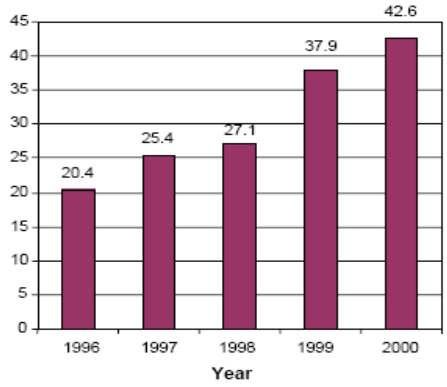
Turkish version was given in appendix F2.

Item m438q01 needs to link representations (text and graphic); identify relevant information, read value directly from a bar graph. It is in reproduction competency cluster and public context.

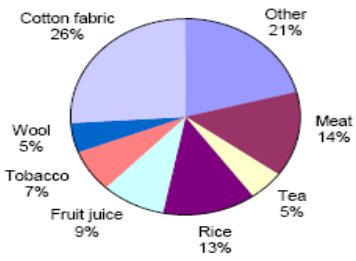
EXPORTS

The graphics below show information about exports from Zedland, a country that uses zeds as its currency.

Total annual exports from Zedland in millions of zeds, 1996-2000



Distribution of exports from Zedland in 2000



Question 1: EXPORTS

M438Q01 - 0 1 9

What was the total value (in millions of zeds) of exports from Zedland in 1998?

Answer: .....

Question 2: EXPORTS

M438Q02

What was the value of fruit juice exported from Zedland in 2000?

A 1.8 million zeds.
B 2.3 million zeds.
C 2.4 million zeds.
D 3.4 million zeds.
E 3.8 million zeds.

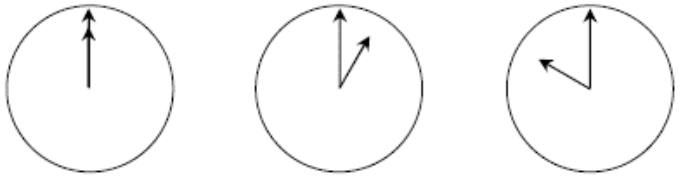
Turkish version was given in appendix F2.

The cognitive demands of items m155q02t and m402q02 were not given in technical report, and the item m155q02t was not a released item. Followings are the items m402q01 and m402q02.

INTERNET RELAY CHAT

Mark (from Sydney, Australia) and Hans (from Berlin, Germany) often communicate with each other using "chat" on the Internet. They have to log on to the Internet at the same time to be able to chat.

To find a suitable time to chat, Mark looked up a chart of world times and found the following:



Question 1: INTERNET RELAY CHAT

M402Q01 - 0 1 9

At 7:00 PM in Sydney, what time is it in Berlin?

Answer: .....

Question 2: INTERNET RELAY CHAT

M402Q02 - 0 1 9

Mark and Hans are not able to chat between 9:00 AM and 4:30 PM their local time, as they have to go to school. Also, from 11:00 PM till 7:00 AM their local time they won't be able to chat because they will be sleeping.

When would be a good time for Mark and Hans to chat? Write the local times in the table.

Place	Time
Sydney	
Berlin	

Turkish version was given in appendix F2.

CHAPTER V

CONCLUSION AND DISCUSSION

The purpose of this chapter is to summarize and discuss the main findings from the results chapter. In this chapter findings from univariate and multivariate DIF analyses of the original (English) and translated (Turkish) PISA mathematics and problem solving tests are compared, possible sources of DIF in flagged items are also discussed. In addition, limitations of the study and implications and future directions are also given in the last section of this chapter.

In the study, as a prerequisite of item level DIF analyses, construct equivalence between translated and original items was investigated through multi-group factor analysis. Investigating the booklets, it was determined that mathematics literacy was not the same thing for Turkish and American students, i.e. the constructs measured by Turkish and English versions of the tests were not equivalent. So, Turkish and American groups could not be compared using original and translated tests. It was concluded that the one should be careful in comparison of American and Turkish cultures with respect to mathematics literacy as measured by PISA 2003, because translated and original tests may not be measuring the same construct in these cultures.

Despite this construct inequivalence in the booklets, as stated in the previous chapter, to carry on DIF analysis, a subtest of items assuring construct equivalence was selected through EFA and CFA. Then, both univariate and multivariate DIF analyses were conducted. However, it is worth specifying that because of the small sample size, M-H procedure were not used in the multivariate analyses. Followings are the discussions of the findings from the DIF analyses.

5.1. M-H versus LR in Univariate Analysis

With respect to the results of DIF analyses; M-H flagged more items than did LR. However, there were an agreement of 78% and 84% in the 3rd and in the 13th booklets, respectively, with respect to the items flagged or not flagged by both methodologies. Specifically, items flagged as showing high-DIF in M-H were also flagged as high-DIF item in LR. But the agreement, both of the methods flagging an item as showing DIF or not, between M-H and LR decreased considerably when moderate DIF items were considered. M-H detected all the items flagged by LR, however the reverse was not true. That can probably be a result of higher Type I error rate in M-H analysis. This result is also in line with that of Benito and Ara (2000) reporting the tendency of M-H in detecting more items as showing DIF than LR.

Type I error rate deserves an additional discussion. Jodoin and Huff (2001) have argued that, the difference between ability distributions of the groups might inflate the Type I error rate of DIF detection procedure. Also they have concluded that using effect size measure in LR as an indicator of DIF items reduced the probability of making Type I error rate when there were unequal ability distributions between groups. In this context, it can be concluded that M-H is more vulnerable than LR in case of groups having unequal ability distributions. However it is worth specifying that, the high Type I error rate have mostly affected the results of moderate DIF level items.

In this study, the results of univariate DIF analyses via M-H and LR methodologies were also compared with respect to the common items of third and thirteenth booklets. It was investigated whether DIF methodologies within themselves detected same items as showing DIF in different booklets. Agreement rates of M-H and LR within themselves was 43% and 86%, respectively. In fact, as DIF is not an intrinsic property of items but mostly determined with respect to items relative function in the test, it was not expected that both M-H and LR would produce same results for the common items of the 3rd and the 13th booklets.

Also in the literature, empirical studies have reported that item bias (DIF) conclusions for the same item were not consistent across instruments and samples (Van de Vijver, 1998). However, as LR produced the same results for 86% of the common items, that is twice as much than did M-H, M-H might be more context dependent than the LR analysis. It is highly possible that factors like ability distributions of the groups or difficulty levels of the items etc. are more effective on the M-H results than that of LR.

In this context, it was concluded that, when total test score is the only matching variable M-H and LR produces strictly similar results in detecting high-DIF items. However, M-H methodology is more open to distorting effects, like different ability distributions, than LR, in the sense of Type I error rates.

5.2. Univariate LR versus Multivariate LR

In the present study, in addition to total test score, problem-solving scores that was figured to be relevant to the mathematical literacy scores was also used in matching the students of the same ability. It was expected that there would be a reduction in the number of items detected as DIF, because the additional variable used in matching would account for an additional dimension that was neglected in the univariate matching. Also, findings in the literature have indicated that in multivariate matching, there were substantially fewer items detected as showing DIF than univariate analysis. For example; in the study of Mazor, Kanjee and Clauser (1995), two continuous achievement variables, total achievement score and SAT-Verbal score, were used as matching variables in M-H and LR methodologies. They have found that conditioning on two relevant abilities provides more accurate matching than conditioning on a single ability. But in this current study including the problem solving scores as an additional matching criterion did not reduce the number of items identified as displaying DIF. It is highly probable that problem solving dimension did not contribute to get a more precise matching of students. There are similar studies in the literature as well. For

example, Zwick and Ercikan (1989) also did not find any reduction in the number of items identified as DIF when they used a matching background variable relevant to history education in addition to total score with the Mantel-Haenszel statistics.

On the other hand, there was an unexpected result in this current study in the sense that an additional item, item m438q02 in favor of Turkey, was flagged as DIF in multivariate LR analyses although it was not in univariate LR analysis. Simpson's Paradox (Dorans & Holland, 1993) may provide an explanation to this result; the difference in mathematical literacy or problem solving could be cancelled on matching with respect to only mathematical literacy score. With concerning the problem solving ability in matching criteria, difference between the groups performance could be obvious. There are additional studies in the literature claiming that adding matching variables may not only decrease the appearance of DIF for an item but also increase it (Clauser, Nungester & Swaminathan, 1996).

However, the reason of the increment in this current study might also be the Type I error due to the small sample size and unequal ability distributions between groups. Because, when students were matched according to the mathematics literacy and problem solving abilities, in most of the ability levels there was only one student from each group, and only one individual may not provide sufficient information in determining the characteristics of the corresponding ability level.

Similar result was found when two mathematics literacy subtest scores were used as a matching variable instead of a single test score. Item m402q02, in the 13th booklet and item m547q01t in the 13th and in the 3rd booklets not showing DIF in univariate LR analysis were flagged in multivariate analysis. Item m402q02 was identified as showing a C-level DIF and m547q01t was identified as showing a B-level DIF in multivariate matching, with two mathematical literacy subtest scores. Items identified as DIF by multivariate matching that were not identified using the total score alone may be explained with Simpson's Paradox (Dorans & Holland, 1993). In this current study there might be a difference between the groups performance that were related to the second dimension. In the 13th booklet, there

was a difference between the groups performance according to first and second factors separately. The difference in the first or second factor could be cancelled on matching with respect to a single score. So, without concerning the second dimension in matching criteria, difference between the groups performance could be masked and when including this dimension in the analysis this difference in performances could be obvious.

On the other hand, using two subtest scores in matching the individuals also reduced some items showing DIF in the univariate case. In the 3rd booklet item m421q03 and in the 13th booklet items m124q01, m510q01t and m564q02 were not flagged as DIF in the multivariate case, although they did in all other analysis used in this study. This finding supports the results of other LR analyses using multivariate matching. Hamilton (1998) showed the shift in status of the DIF items using science test scores with different conditioning variables and concluded that when more than one dimension was included as conditioning variable, LR procedure identified fewer items as showing DIF. In the present study, the second factor loadings of items m124q01, m510q01t and m421q03 were bigger than 0.30. So, it can be concluded that item impact due to this second factor might be reduced by taking into account the second factor in these items. However, item m564q02 has second factor loading less than 0.10, although it was not flagged as DIF in subtest scores matching. To conclude that there might be the second factor impact for this item can not be reasonable. But the rotation in EFA analysis may be inappropriate in the study. In this current study varimax rotation was used in EFA to determine the factor loadings of the items. Other rotations (promax or oblique) may be more appropriate than varimax rotation to explain the second factor effects in the items.

In the booklets considered in the present study, the results suggested that there were not a substantial reduction in number of items identified as DIF when subtest scores was used. But using subtest scores changed the DIF results. Using multiple subtest scores simultaneously in matching may be appropriate when single

test items require more than one ability (Clauser, Mazor, Nungester & Ripkey, 1996). In this study rotated component matrix showed that some items have loaded on more than one factor. Subtest scores, which were determined via factor analysis, were also correlated. The Pearson correlation between the subtest scores was 0.66 in the 3rd booklet and 0.90 in the 13th booklet. The correlation between the factor subtest scores in the 13th booklet was more than the correlation between the factor subtest scores in the 3rd booklet. This can be an explanation of why shift in status of identified items as DIF in the 13th booklet was more than the 3rd booklet. The speculation that the shift in the status of items identified as DIF might be a result due to the improved matching (reduction of item impact).

Finally, it may be argued that in a unidimensional test where subdimensions revealed in EFA do not threat the unidimensionality with respect to CFA, using total test score as a matching criterion may have a distorting effect on DIF results. Namely, DIF in an item may be due to the impact of any of the subdimensions. So it was concluded that using multiple subdimension scores instead of a single total test score in matching the individuals, might control the potential of impact on DIF items due to the differences of individuals with respect to their positions in the subdimensions. This result supports the study of Nandakumar (1991). Nandakumar (1991) have reported that there could be minor dimensions in a unidimensional test and when the effect of minor dimensions increased, the unidimensionality of the test was violated.

5.3. Causes of DIF

To put the statistical information to best use, two reviewers analyzed the DIF items, to disentangle the possible sources of DIF, with respect to adaptation and translation differences including possible cultural and curriculum related discrepancies as well. Several findings about curriculum-related differences and poor translation emerged from the study may be useful for test users.

Followings are the results from the qualitative review of released items flagged as showing DIF. Attempt to assign causes of DIF identifying patterns that were described in the Chapter III gave some interesting results.

Items m124q01 and m547q01, favoring Turkish students, were curriculum-like items requiring an interpretation of a simple and familiar picture, text and algebra, and a simple calculation containing symbols. On the other hand, items m179q01 and m402q02, favoring American students, were real life problems requiring an interpretation of data and reasoning mathematically.

Although there are a restricted number of items, these properties of the items may be indicating some curricular differences between the two countries. For example, it is possible that while Turkish curriculum has a focus on algebra and simple calculation, USA curriculum may be focusing on data interpretation and mathematical reasoning. There are also DIF studies in the literature concerning the country specific strengths and weaknesses in the context of countries' curriculum. For example, studies of Ercikan, (2002), Klieme and Baumert, (2001) and Beaton, (1998) have claimed that differences in countries curriculum may cause some items to function differentially among groups.

So according to the results of the study it might be possible to argue that the probability of Turkish students to perform better than the same ability American students in algebra and items requiring simple calculation is higher, whereas this situation is reversed in items requiring data interpretation.

Also it might be reasonable to argue that the probability of Turkish students to perform better than the same ability American students in curriculum-like problems is higher, whereas the probability of American students to perform better than the same ability Turkish students in real life problems is higher.

In the same manner, item m179q01 and item m402q02 are open constructed items and they require a supporting explanation of the answer. In Turkey curriculum, this type of questions in lessons is rare. American students might be more familiar than Turkish students with this type of questions due to the coverage

of their curriculum. It was concluded that items requiring a supporting–explanation of the answer might have a potential to function differentially against Turkish students compared to matched American students.

An important suggestion this current study provides is about the translation and adaptation process. Beyond curriculum differences, there may be translation differences causing some items to function differentially across groups. Two reviewers suggested that some items might have been incomparable in meaning due to the translation problems between the English (original) and Turkish (translated) version of the tests. In the items stated below, due to the translation problems, translated Turkish item might not convey the same meaning as the English item.

Item m438q01 is a DIF item which favors American students and requires reading value on a bar graph. On the other hand, although item m438q02 also requires reading value on a graph and using this value in a simple calculation, it did not show DIF. Items m438q01 and m438q02 have the same content and these items require the same cognitive ability as well. So, it was figured out that there might be another factor affecting Turkish students' responses on the item m438q01. The discussion on this item focused on the fact that there might be a translation problem for this item. The DIF exhibited between this English item and its Turkish counterpart suggested that poor translation might affect performance on this item. The item "What was the total value (in millions of zeds) of exports from Zedland in 1998?" have been translated as "1998 yılında Zed ülkesinden yapılan dışsatımın toplam değeri (milyon zed olarak) nedir?" In this translation the word "total" in English and "toplam" in Turkish might have different meanings. This word was not used in second question m438q02, so there was not any difference on performances of matched Turkish and American students in the item m438q02. This word might be misunderstood in Turkish version of the test and students might use this word in meaning of "sum or add". Turkish students probably added all exports values in given years in the graph instead of reading the value of the export of the year 1998.

This misunderstanding might be also due to the place of this word in the sentence. In English version of the item, the word "total" is at the beginning of the sentence, but it is at the end of the sentence in Turkish version due to the grammatical structures of this language. Understanding the meaning of the sentence in item m438q01 might be more difficult for Turkish students. This finding was consistent with the previous studies which have reported that the translation problem in meanings in the words and expressions was the cause of DIF in achievement tests, such as verbal tests (Allalouf et al., 1999) and social studies and mathematics achievement tests (Gierl et al, 1999).

Another explanation of differential performance in this item might be explained with the study of Scheuneman and Grima (1997). They have stated that quantitative words like add, circle, equal, sum, product…etc. indicate the operations that are critical for expressing the conditions. They considered that verbal properties such as quantitative language of items might be associated with differential performance. The reviewers concluded that if the word "toplam" did not use in the Turkish version of the test, Turkish students might not perform relatively worse with respect to matched American students.

Although some additional research is needed, in this current study the combination of statistical and qualitative analyses have provided some hypotheses concerning the sources of translation between English and Turkish versions of the test, including some curriculum differences between Turkish and American Education.

5.4. Limitations

The reviewers used for this study might not have been qualified enough to assess the cultural relevance factors for the DIF items. In addition, qualitative analysis of the reviewers would be speculative because they know which items were flagged as DIF before the qualitative analysis.

In the study sample size was limited. Although M-H and LR DIF methods are appropriate using with small samples, moderate to high sample size could increase the power of the study.

There were also a few numbers of released items in addition to small sample size. These limitations decreased the probability of getting more generalizable results.

5.5. Implications

The results of the dimensionality analyses alert the researchers to find optimal matching criteria in the study of investigating differential performance. This study showed that the information about the construct equivalence of translated tests should be used carefully.

In the study it seemed that in some cases, the causes of item level difference in performances could be the curriculum, teaching methods etc. These findings can be useful in understanding the need of curriculum change for the curriculum developers.

This study also provided an evidence of difficulty in translation of quantitative language in cross-cultural assessments. This finding may also be useful for test development and test adaptation processes.

5.6. Future Directions

In this study some items were polytomous and these items were dichotomized. The effect of this process on DIF items may be examined with another research.

Differential performance can be specific to cultures. Another research is needed to develop a systematic translation and adaptation guidelines from English to Turkish. More researches are needed to identify sources of group performance differences on quantitative language issue in linguistic translations.

The methods in finding more appropriate criterion are an important issue to investigate. This study may be a step to identify an appropriate matching criterion. However, much more research is needed to identify the most appropriate matching criterion for translated achievement tests.

This study only focused on DIF at the item level, but differential performance can also be assessed at the test level. There exist also other statistical methods such as IRT methods. Different DIF methods use different strategies for identifying item performance and matching criterion. The effectiveness of the other DIF methods in translation and culture DIF research can be examined.

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.

Agresti, A. (2002). Categorical Data Analysis. Second edition. John Wiley & Sons, Inc., Hoboken, New Jersey.

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.

Beaton, A.E. (1998). Comparing Cross-National Student Performance On TIMSS Using Different Test Items. *International Journal of Educational Research* 29, 529-542

Benito J.G. & Ara M.J.N. (2000). A Comparison of X^2 , RFA and IRT Based Procedures in the Detection of DIF, *Quality & Quantity*, v.34, 17-31.

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. & Hei, L.M. (2003). A comparison of university students' approaches to learning across Taiwan and Turkey. *International Journal of Testing.* 3(2), 173 – 187.

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.

Bontempo, R. (1993). Translation Fidelity of Psychological Scales. An Item Response Theory Analysis of an Individualism-Collectivism Scale. Journal of cross cultural Psychology, vol.24. no.2,149-166.

Camilli, G. & Congdon, P. (1999). Application of A Method of Estimating DIF for Polytomous Test Items. *Journal of Educational and Behavioral Statistics,* v.24, no.4, 323-341.

Camilli, G. & Shepard, L.A. (1994). *Methods for Identifying Biased Test Items.* Sage Publications, California.

Camilli, G. (1993). The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues? In P.W.Holland & H. Wainer (Eds.). *Differential item functioning: Theory and practice* (pp. 397 – 417) Hillsdale, NJ: Erlbaum.

Carlton, S. & Harris, A.M. (1992). *Characteristics Associated with Differential Item Functioning on the Scholastic Aptitude Test: Gender and Majority/Minority Group Comparisons.* Reports-Research/Technical. Educational Testing Service, Princeton, N.J.

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.

Clauser, B.E. & Mazor, K.M (1998). Using Statistical Procedures to Identify Differentially Functioning Items. *Educational Measurement Issues and Practice*, spring, 34-44.

Clauser, B.E., Nungester R.J., Mazor, K. & Ripkey, D. (1996). A Comparison of Alternative Matching Strategies for DIF Detection in Tests That are Multidimensional. *Journal of Educational Measurement*, vol.33, no.2, pp.202-214.

Clauser, B.E., Nungester, R.J. & Swaminathan, H. (1996) Improving the matching for DIF analysis by conditioning on both test score and an educational background variable. *Journal of Educational Measuremen,.* 33(4), 453-464.

Crocker, L. & Algina, J. (1986). *Introduction to Classical and Modern Test Theory.* New York: Holt, Rinehart and Winston.

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.

Donoghue, J.R., Holland, P.W. & Thayer, D.T. (1993) A Monte Carlo Study of Factors That Affect the Mantel-Haenszel and Standardization Measures of Differential Item Functioning. In P.W.Holland & H. Wainer (Eds.) *Differential item functioning: Theory and practice* (pp. 137 - 166) Hillsdale, NJ: Erlbaum.

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.

Ellis, B.B. (1989). Differential Item Functioning: Implication for Test Translation. *Journal of Applied Psychology.* 74, 912 – 921.

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.

Ercikan, K., Gierl, M.J., Mccreith, T., Puhan, G., Koh, K. (2004). Comparability of Bilingual Versions of Assessments: Sources of Incomparability of English and French Versions of Canada's National Achievement in Tests. *Applied Measurement in Education,* 17(3), 301-321.

Gamer, M. & Engelhard, J., G. (1999). Gender Differences In Performance On Multiple- Choice And Constructed Response Mathematics Items. *Applied Measurement in Education,* Vol. 12, Issue 1

George, D. & Mallery, P. (2003). *SPSS for Windows Step By Step,* Pearson Education, Inc, USA.

Gierl M.J., Rogers W.T. & Klinger D. (1999). *Using Statistical and Judgmental Reviews to Identify and Interpret Translation Dif*. Paper Presented At the Annual Meeting of the National Council on Measurement in Education (NCME) At the Symposium Entitled "Translation Dif: Advances and Applications" Canada

Gierl, M. J., 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., (2000). Construct Equivalence on Translated Achievement Tests. Canadian Journal of Education 25 (4), 280-296.

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., Bisanz, J. & Li, Y. (S) Y. (2004). *Using The Multidimensionaliy-Based DIF Analysis Paradigm To Study Cognitive Skills That Elicit Group Differences: A Critique.* Paper Presented At the Annual Meeting of the National Council on Measurement in Education (NCME), San Diego, California, USA

Hambleton, R., Clauser, B., Mazor,K. & Jones, R. (1993). Advances in the Detection of Differentially Functioning Test Items. *European Journal of Psychological Assessment*, 9(1), pp. 1-18.

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.& Rogers, H.J. (1989). Detecting Potentially Biased Test Items: Comparison of IRT Area and Mantel-Haenszel Methods. *Applied Measurement in Education,* 2(4), 313 –334

Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991). *Fundamentals of Item Response Theory*. Sage Publications, California.

Hambleton, Ronald & Rodgers, Jane (1995). Item Bias Review. *Practical Assessment, Research & Evaluation*, 4(6)

Hamilton, L. S. & Snow, R. E. (1998). Exploring Differential Item Functioning on Science Achievement test. CSE Technical Report 483.

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.

Hidalgo, M.D. & Pina, S.A.L., (2004) Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement*. 64(6), 905 – 915.

Hills, J. (1990). Screening for Potentially Biased Items in Testing Programs. *Educational Measurement: Issues and Practice*, 8, 5-11.

Huang, C.D., Church, A.T. & Katigbak, M.S. (1997). Identifying Cultural Differences in Item Traits: Differential Item Functioning In the NEO Personality Inventory. *Journal of Cross-Cultural Psychology.* 28(2), 192 – 218.

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.

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.

Hyde, J.S., Fennema, E. & Lamon S.J. (1990). Gender Differences in Mathematics Performance: A Meta Analysis. *Psychological Bulletin*. V.107, no: 2, 139-155.

Jodoin, M., G. & Gierl, M., J. (2001). Evaluating Type I Error and Power Rates Using an Effect Size Measure with the Logistic Regression Procedure for DIF Detection. *Applied Measurement in Education*, 14(4), 329-349.

Jodoin, M.G. & Huff,K.L. (2001). Examining Type I Error Rates When Ability Distributiond are Unequal With the Logistic Regression Procedure for DIF Detection. Paper presented at the annual meeting of the National Council on measurement in education, Seattle, WA.

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

Kelloway, E. K. (1998). *Using LISREL for Structural Equation Modeling*. London, New Delhi: Sage Publications.

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, H. & Stout W. (1996). A New Procedure for Detection of Crossing DIF. *Psychometrica*, V61, n.4, p.647-677.

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.

Lim, R.G. & Drasgow, F. (1990). Evaluation of Two Methods for Estimating Item Response Theory Parameters When Assessing Differential Item Functioning. *Journal of Applied Psychology*, 75(2), pp. 164-174.

Linn, M., C. & Hyde, J., S. (1989). Gender, Mathematics, and Science. *Educational Researcher,* vol. 18, no.8 p.17-19, 22-27.

Linn, R.L. (1993). The use of differential item functioning statistics: A discussion of current practice and future implications. In P.W.Holland & H. Wainer (Eds.) *Differential item functioning: Theory and Practice* (pp. 67 – 113) Hillsdale, NJ: Erlbaum.

Mazor, K.M., Kanjee, A. & Clauser, B.E. (1995). Using Logistic Regression and Mantel Haenszel with Multiple Ability Estimates to Detect Differential Item Functioning. *Journal of Educational Measurement*, vol.32, no.2, 131-144.

Mazor, K.M., Clauser, B.E. & Hambleton R.K. (1992). The Effect of Sample Size on the Functioning of the Mantel-Haenszel Statistic. *Educational and Psychological Measurement,* vol.52, pp. 443-451.

Mazor, K.M., Clauser, B.E. & Hambleton R.K. (1994). Identification of Nonuniform Differential Item Functioning Using a Variation of the Mantel-Haenszel Procedure. *Educational and Psychological Measurement*, vol.54, no. 2, pp. 284-291.

Nandakumar, R. (1991). Traditional Dimensionality versus Essential Dimensionality. *Journal of Educational Measurement*, vol.28, no.2, 99-117.

NCTM, (1996). *Curriculum and Evaluation Standards for School Mathematics*, The National Council of Teachers of Mathematics, Inc., USA.

OECD (2002). Sample Tasks from the PISA 2000 Assessment of Reading, Mathematical and Scientific Literacy, Paris.

OECD (2003). *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.

Pampel, F.C. (2000). *Logistic Regression A Primer*. Sage Publications. Thousand Oaks. London, New Delhi

Poortinga, Y.H. (1989). Equivalence of Cross-Cultural Data: An Overview of Basic Issues. *International Journal of Psychology.* 24, 737 – 756.

Reckase, M.D., Ackerman, T.A. & Carlson J.E. (1988). Building a Unidimensional Test Using Multidimensional Items. *Journal of Educational Measurement,* vol.25, no.3, 193-203.

Robin, F., Sireci, S.G. & Hambleton, R.K. (2003). Evaluating the Equivalence of Different Language Versions of A Credentialing Exam. *International Journal of Ttesting*, 3(1), 1-20.

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.G. (1997). Problems And Issues In Linking Assessment Across Languages. *Educational Measurement: Issues and Practice.* 16(1), 12 – 19.

Sireci, S.G. & Allalouf, A. (2003). Appraising Item Equivalence across Multiple Languages and Cultures. *Language Testing*, 20(2), 148 – 166.

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.

Sireci, S.G., Harter, J., Yang, Y. & Bhola, D. (2003). Evaluating the Equivalence of an Employee Attitude Survey across Languages, Cultures, and Administration Formats. *International Journal of Testin,.* 3(2). 129 – 150.

Sireci, S.G. & Swaminathan, H. (1996). Evaluating Translation Equivalence: So What's the Big DIF?. Paper presented at the Annual Meeting of the Northeastern Educational Research Association, Ellenville, NY.

Swaminathan H.& Rogers, J.H.(1990). Detecting Differential Item Functioning Using Logistic Regression Procedures. *Journal of Educational Measuremen,.* 27(4), 361-370.

Tatsuoka, K.K., Linn, R.L., Tatsuoka, M.M. & Yamamoto, K. (1988). DIF Resulting From the Use of Different Solution Strategies. *Journal of Educational Measurement,* 25(4), 301 – 319.

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.

Van de Vijver, F (1998). Towards a Theory of Bias and Equivalance. *ZUMA-Nachrichten Spezial*, January, pp.41-65.

Van de Vijver, F. & Hambleton, R.K. (1996) Translating test: Some practical Guidelines. *European Psychologist*, 1(2), 89-99.

Van de Vijver, F. & Leung, K. (1997). *Methods and Data Analysis for Cross-Cultural Research.* Sage Publications, Thousand Oaks, London, New Delhi.

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.

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/

Williams, V.S.L. (1997). The "Unbiased" Anchor: Bridging the Gap between DIF and Item Bias. *Applied Measurement in Education*, 10, 353-267.

Yurdugül, H & Aşkar P. (2004b) Ortaöğretim Kurumları Öğrenci Seçme ve Yerleştirme Sınavının Cinsiyete Göre Madde Yanlılığı Açısından Incelenmesi. *Eğitim Bilimleri ve Uygulama Dergisi*, 3(5), 3-20.

Yurdugül, H. & Aşkar, P. (2004a) Ortaöğretim Kurumları Öğrenci Seçme ve Yerleştirme Sınavının Öğrencilerin Yerleşim Yerlerine Göre Diferansiyel Madde Fonksiyonu Açısından Incelenmesi *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 27, 268-275.

Zenisky, A.L., Hambleton, R.K. & Robin, F. (2003). Detection of DIF in Large-Scale State Assessments: A Study Evaluating a Two-Stage Approach. *Educational and Psychological Measurement.* 63(1), 51 – 64.

Zieky, M. (1993). Practical Questions in the Use of DIF Statistics in Test Development. In P.W.Holland & H. Wainer (Eds.) *Differential İtem Functioning: Theory And Practice* (pp. 337 – 347) Hillsdale, NJ: Erlbaum.

Zumbo, B.D. (1999). A Handbook of the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling As a Unitary Framework for Binary and Likert-Type (Ordinal) Item Score. Ottawa, Ontario, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

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, R. & Ercikan, K. (1989). Analysis of Differential Item Functioning In the NAEP History Assessment. *Journal of Educational Measurement*, 26(1), 55-66.

APPENDICES

APPENDIX A1

Mathematics Literacy Items Descriptions of Booklet 13

Item name	Item type	Item label	Item scale	Item Discrimination		PC		#Missing (%)			
				USA	Turkey	USA	Turkey	USA		Turkey	
m033q01	MC	View Room Q1	Space & Shape	0.54	0.64	0.76	0.52	4	(1%)	2	(0.6%)
m124q01*	CR	Walking Q1	Change & Relation	0.78	0.81	0.31	0.28	45	(11%)	69	(19%)
m124q03t*	CR	Walking Q3	Change & Relation	0.75	0.83	0.45	0.32	107	(26%)	152	(42%)
m179q01t*	CR	Robberies Q1	Uncertainty	0.74	0.82	0.50	0.19	33	(8%)	54	(15%)
m305q01	MC	Map Q1	Change & Relation	0.63	0.61	0.34	0.57	6	(2%)	26	(8%)
m402q01*	CR	Internet Q1	Change & Relation	0.37	0.10	0.46	0.41	9	(2%)	14	(4%)
m402q02*	CR	Internet Q2	Change & Relation	0.52	0.57	0.32	0.12	36	(9%)	62	(17%)
m438q01*	CR	Exports Q1	Uncertainty	0.67	0.79	0.84	0.50	14	(3.4%)	77	(22%)
m438q02*	MC	Exports Q2	Uncertainty	0.51	0.66	0.39	0.30	17	(4%)	41	(12%)
m462q01t	CR	Third Side Q1	Space & Shape	0.62	0.71	0.16	0.17	64	(15%)	87	(24%)
m464q01t	CR	Fence Q1	Space & Shape	0.77	0.99	0.56	0.34	21	(5%)	21	(6%)
m467q01*	MC	Coloured Candies Q1	Uncertainty	0.65	0.73	0.65	0.48	4	(1%)	4	(1%)
m474q01	CR	Running Time Q1	Quantity	0.63	0.82	0.45	0.24	6	(2%)	6	(2%)
m505q01*	CR	Litter Q1	Uncertainty	0.78	0.97	0.49	0.22	40	(10%)	120	(33%)
m510q01t*	CR	Choices Q1	Quantity	0.73	0.69	0.72	0.64	16	(4%)	19	(5%)
m547q01t*	CR	Staircase Q1	Space & Shape	0.59	0.58	0.47	0.41	23	(6%)	28	(8%)
m564q01	MC	Chair Lift Q1	Quantity	0.69	0.83	0.38	0.37	7	(2%)	6	(2%)
m564q02	MC	Chair Lift Q2	Uncertainty	0.57	0.56	0.63	0.43	9	(2%)	21	(6%)
m806q01t*	CR	Step Pattern Q1	Quantity	0.71	0.51	0.73	0.44	12	(3%)	7	(2%)
m810q01t	CR	Bicycles Q1	Quantity	0.58	0.59	0.69	0.52	13	(3%)	34	(10%)
m810q02t	CR	Bicycles Q2	Quantity	0.56	0.62	0.56	0.34	15	(4%)	20	(6%)
m810q03t	CR	Bicycles Q3	Change & Relation	0.51	0.71	0.17	0.23	64	(15%)	129	(36%)
m833q01t	CMC	Seeing The Tower Q1	Space & Shape	0.66	0.64	0.25	0.15	19	(5%)	44	(12%)

#missing (%): missing values before recoded
PC: proportion corrects after recoded
MC: Multiple choice
CR: Coded response
CMC: Complex Multiple choice
* indicates released items

APPENDIX A2

Problem Solving Items Descriptions of Booklet 13

Item name	Item type	Item label	Item discrimination		PC		#Missing (%)			
			USA	Turkey	USA	Turkey	USA		Turkey	
x402q01t	CR	Library System Q1	0.60	0.73	0.79	0.62	19	(5%)	11	(3%)
x402q02t	CR	Library System Q2	0.86	0.88	0.75	0.47	34	(8%)	65	(18%)
x414q01	CR	Course Design Q1	0.69	0.80	0.41	0.15	8	(2%)	56	(16%)
x415q01t	CR	Transit System Q1	0.68	0.63	0.22	0.17	16	(4%)	18	(5%)
x602q01	CR	Holiday Q1	0.78	0.68	0.42	0.30	8	(2%)	17	(5%)
x602q02	CR	Holiday Q2	0.85	1.00	0.39	0.11	44	(11%)	61	(17%)
x603q01	CR	Irrigation Q1	0.74	0.71	0.64	0.54	47	(11%)	68	(19%)
x603q02t	CMC	Irrigation Q2	0.82	0.77	0.46	0.29	15	(4%)	24	(7%)
x603q03	CR	Irrigation Q3	0.81	0.69	0.53	0.43	16	(4%)	29	(8%)

#missing (%): missing values before recoded
PC: proportion corrects after recoded
MC: Multiple choice
CR: Coded response
CMC: Complex Multiple choice

APPENDIX A3

Mathematics Literacy Items Descriptions of Booklet 3

Item name	Item type	Item label	Item scale	Item Discrimination		PC		#Missing (%)			
				USA	Turkey	USA	Turkey	USA		Turkey	
m124q01*	CR	Walking Q1	Change & Relation	0.68	0.82	0.25	0.37	22	(5%)	45	(12%)
m124q03t*	CR	Walking Q3	Change & Relation	0.62	0.80	0.43	0.37	74	(18%)	98	(26%)
m144q01t	CR	Cube Painting Q1	Space & Shape	0.56	0.60	0.47	0.42	9	(2%)	10	(3%)
m144q02t	CR	Cube Painting Q2	Space & Shape	0.51	0.54	0.19	0.11	20	(5%)	22	(6%)
m144q03	MC	Cube Painting Q3	Space & Shape	0.68	0.67	0.76	0.60	8	(2%)	15	(4%)
m144q04t	CMC	Cube Painting Q4	Space & Shape	0.65	0.73	0.32	0.24	39	(9%)	51	(13%)
m155q01	CR	Pop Pyramids Q1	Change & Relation	0.71	0.64	0.66	0.49	35	(8%)	77	(20%)
m155q02t	CR	Pop Pyramids Q2	Change & Relation	0.77	0.76	0.73	0.35	24	(6%)	139	(37%)
m155q03t	CR	Pop Pyramids Q3	Change & Relation	0.84	0.81	0.21	0.08	102	(24%)	215	(58%)
m155q04t	CMC	Pop Pyramids Q4	Change & Relation	0.52	0.52	0.52	0.34	17	(4%)	26	(7%)
m305q01	MC	Map Q1	Space & Shape	0.27	0.26	0.47	0.43	67	(16%)	63	(17%)
m420q01t	CMC	Transport Q1	Uncertainty	0.72	0.67	0.57	0.34	6	(1%)	10	(3%)
m421q01	CR	Height Q1	Uncertainty	0.81	0.89	0.66	0.42	30	(7%)	97	(27%)
m421q02t	CMC	Height Q2	Uncertainty	0.63	0.46	0.18	0.07	9	(2%)	9	(2%)
m421q03	MC	Height Q3	Uncertainty	0.50	0.67	0.29	0.39	18	(4%)	22	(6%)
m438q02*	MC	Exports Q2	Uncertainty	0.67	0.71	0.44	0.40	9	(2%)	27	(7%)
m442q02	CR	Braille Q2	Quantity	0.71	0.81	0.34	0.23	33	(8%)	81	(21%)
m447q01	MC	Tile Arrange1 Q1	Space & Shape	0.66	0.64	0.63	0.47	8	(2%)	8	(2%)
m462q01t	CR	Third Side Q1	Space & Shape	0.50	0.77	0.25	0.21	40	(10%)	49	(13%)
m468q01t*	CR	Science Tests Q1	Uncertainty	0.63	0.73	0.55	0.37	21	(5%)	20	(5%)
m474q01	CR	Running Time Q1	Quantity	0.38	0.50	0.68	0.49	2	(0.5%)	1	(0.3%)
m484q01t*	CR	Bookshelves Q1	Quantity	0.79	0.75	0.56	0.37	20	(5%)	20	(5%)

(Continued)

m496q01t	CMC	Cash Withdrawal Q1	Quantity	0.73	0.82	0.48	0.32	16	(4%)	13	(3%)
m496q02	CR	Cash Withdrawal Q2	Quantity	0.70	0.65	0.58	0.48	27	(6%)	47	(12%)
m505q01*	CR	Litter Q1	Uncertainty	0.55	0.70	0.36	0.29	18	(4%)	88	(23%)
m509q01*	MC	Earthquake Q1	Uncertainty	0.63	0.50	0.54	0.36	25	(6%)	13	(3%)
m510q01t*	CR	Choices Q1	Quantity	0.53	0.69	0.41	0.25	50	(12%)	31	(8%)
m547q01t*	CR	Staircase Q1	Space & Shape	0.66	0.45	0.69	0.67	31	(7%)	17	(5%)
m559q01	MC	Telephone Rates Q1	Quantity	0.56	0.68	0.57	0.49	4	(1%)	6	(2%)
m571q01	MC	Stop The Car Q1	Change & Relation	0.62	0.42	0.41	0.36	13	(3%)	23	(6%)
m704q01t*	CR	The Best Car Q1	Change & Relation	0.73	0.74	0.77	0.67	29	(7%)	12	(3%)
m704q02t*	CR	The Best Car Q2	Change & Relation	0.77	0.92	0.23	0.22	48	(11%)	46	(12%)
m800q01	MC	Computer Game Q1	Quantity	0.30	0.34	0.85	0.92	6	(1%)	3	(0.8%)
m806q01t*	CR	Step Pattern Q1	Quantity	0.65	0.55	0.61	0.51	15	(4%)	3	(0.8%)

#missing (%): missing values before recoded
PC: proportion corrects after recoded
MC: Multiple choice
CR: Coded response
CMC: Complex Multiple choice
* indicates released items

APPENDIX A4

Problem Solving Items Descriptions of Booklet 3

Item name	Item type	Item label	Item discrimination		PC		#Missing (%)			
			USA	Turkey	USA	Turkey	USA		Turkey	
x412q01	MC	Design by Numbers Q1	0.58	0.20	0.34	0.09	19	(5%)	17	(5%)
x412q02	MC	Design by Numbers Q2	0.62	0.57	0.41	0.33	19	(5%)	25	(7%)
x412q03	CR	Design by Numbers Q3	0.81	0.81	0.31	0.30	113	(27%)	122	(32%)
x417q01	CR	Children's Camp Q1	0.80	0.82	0.48	0.23	30	(7%)	59	(17%)
x423q01t	CMC	Freezer Q1	0.68	0.74	0.44	0.31	6	(1%)	18	(5%)
x423q02t	CMC	Freezer Q2	0.66	0.64	0.33	0.34	6	(1%)	15	(4%)
x430q01	CR	Energy Needs Q1	0.70	0.70	0.70	0.72	10	(2%)	27	(7%)
x430q02	CR	Energy Needs Q2	0.90	0.88	0.29	0.17	65	(16%)	158	(42%)
x601q01t	CMC	Cinema Outing Q1	0.78	0.74	0.72	0.48	6	(1%)	22	(6%)
x601q02	MC	Cinema Outing Q2	0.62	0.72	0.62	0.37	9	(2%)	13	(4%)

#missing (%): missing values before recoded
PC: proportion corrects after recoded
MC: Multiple choice
CR: Coded response
CMC: Complex Multiple choice

APPENDIX **B**1

Rotated Component Matrix of Booklet 13 for Over All Data of Turkey and USA

Rotated Component Matrix[a]

	Component		
	1	2	3
m810q03t	,655	,192	
m464q01t	,638	,222	-,158
m564q02	,548		,127
m124q01	,544	,300	
m438q02	,519	,234	
m564q01	,516		,180
m462q01t	,505	,157	
m124q03t	,504	,381	
m179q01t	,476	,404	,105
m467q01	,456	,429	-,115
m402q02	,444	,237	,332
m402q01	,353	,138	,309
m833q01t	,346	,209	,174
m510q01t		,637	
m810q01t		,618	,232
m810q02t	,131	,606	,211
m438q01	,183	,518	,105
m547q01t	,166	,511	
m806q01t	,305	,499	
m474q01	,209	,471	
m505q01	,428	,447	
m033q01	,257	,391	,230
m305q01			,823

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

a. Rotation converged in 5 iterations.

APPENDIX B2

Rotated Component Matrix of Booklet 3 for Over All Data of Turkey and USA

Rotated Component Matrix [a]

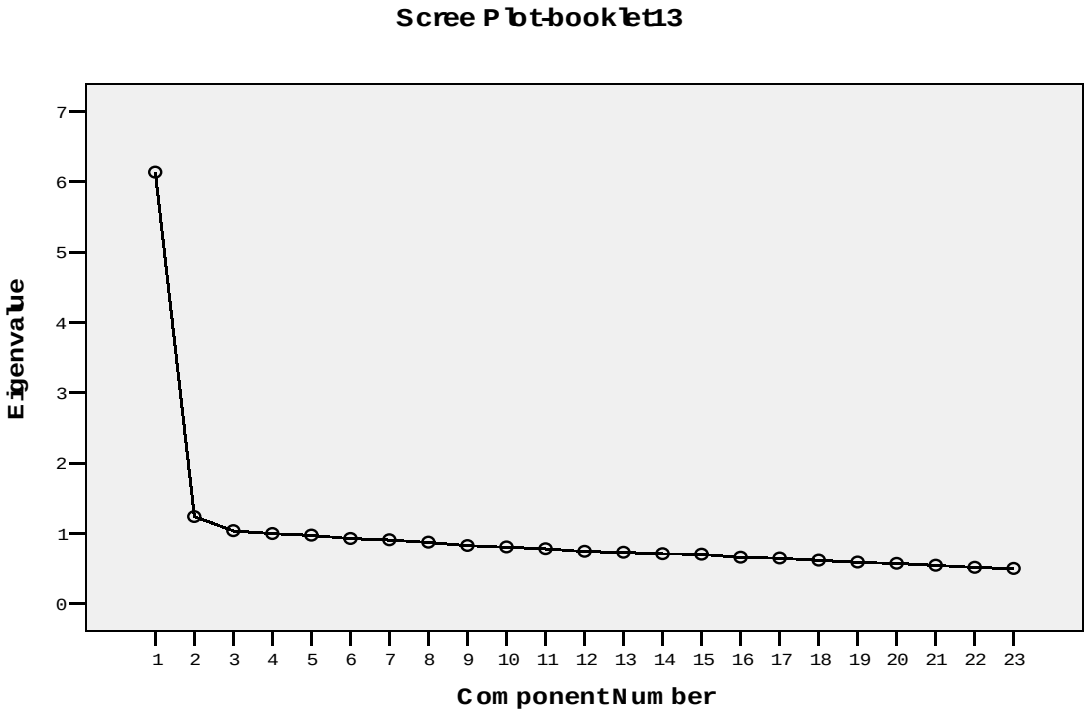
	Component					
	1	2	3	4	5	6
m155q01	,638		,203		,134	
m155q02t	,609	,135	,214	,208		
m155q04t	,559	,142				-,150
m420q01t	,534	,215	,157	,219		
m421q01	,529	,308	,203	,243	,167	
m704q01t	,529	,281			,256	
m474q01	,509	,101	-,227	,177		,200
m484q01t	,469	,169	,289	,255	,170	
m496q01t	,441		,278	,271	,283	,139
m468q01t	,429	,288	,230		,132	
m447q01	,385		,191	,321	,224	,109
m510q01t	,357	,105	,253	,166	,130	
m442q02	,333	,215	,257	,316	,202	
m559q01	,296	,237	,263	,112	,204	
m124q01	,121	,715		,153	,132	
m124q03t	,329	,650				
m421q03		,516	,198	,129	,263	
m438q02	,206	,461	,328	,103		
m462q01t	,200	,453	,144			,413
m704q02t	,193	,433	,369	,232	,127	,129
m505q01	,303	,365	,163			
m421q02t			,710	,120		,110
m155q03t	,203	,285	,555	,159		
m571q01	,173	,272	,456			
m509q01	,306		,389	,148	,138	-,141
m144q02t				,652		
m144q01t	,132	,104		,648	,143	
m144q04t	,120	,204	,179	,630	,101	
m144q03	,378			,518		
m800q01		,220	-,206		,667	-,157
m496q02	,444		,142		,486	,142
m806q01t	,227		,145	,212	,478	,169
m547q01t	,269		,209		,420	
m305q01					,116	,852

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

a. Rotation converged in 14 iterations.

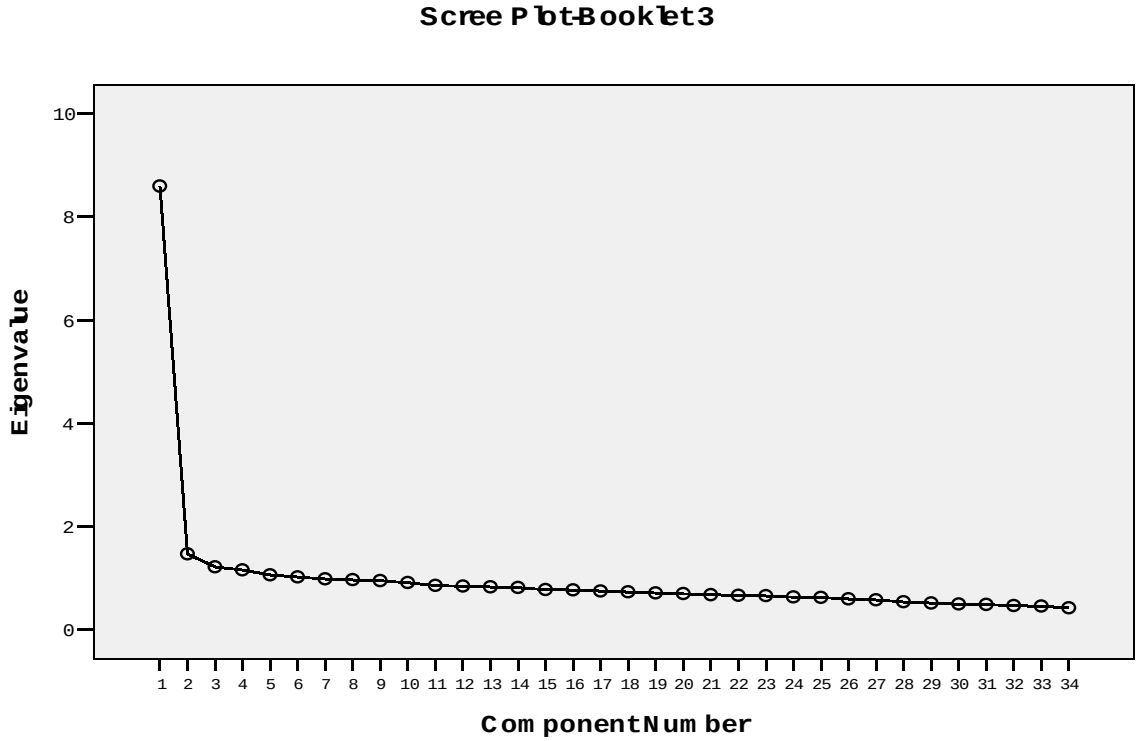
APPENDIX B3

Principal Component Scree Plot of Booklet 13



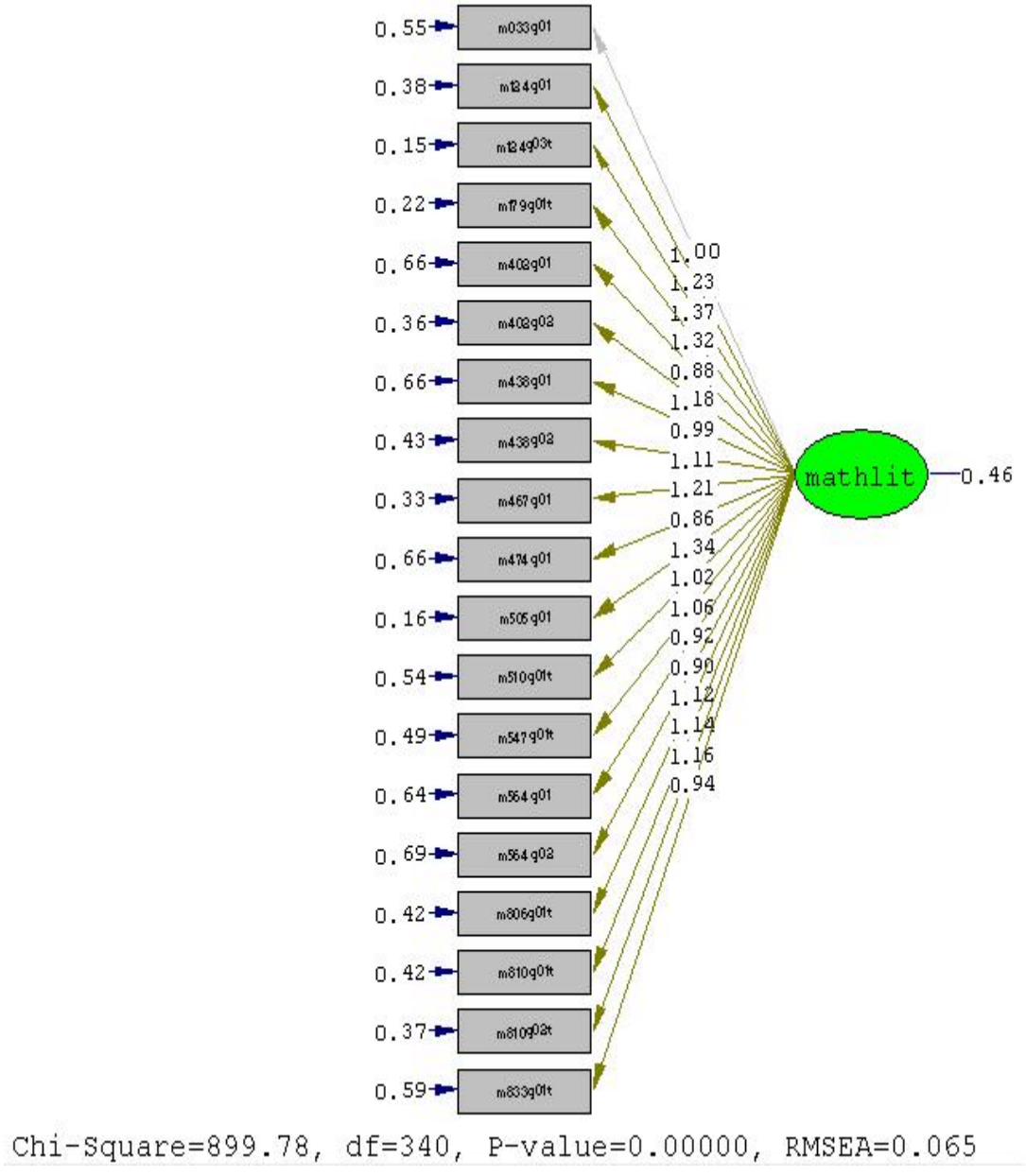
APPENDIX B4

Principal Component Scree Plot of Booklet 3



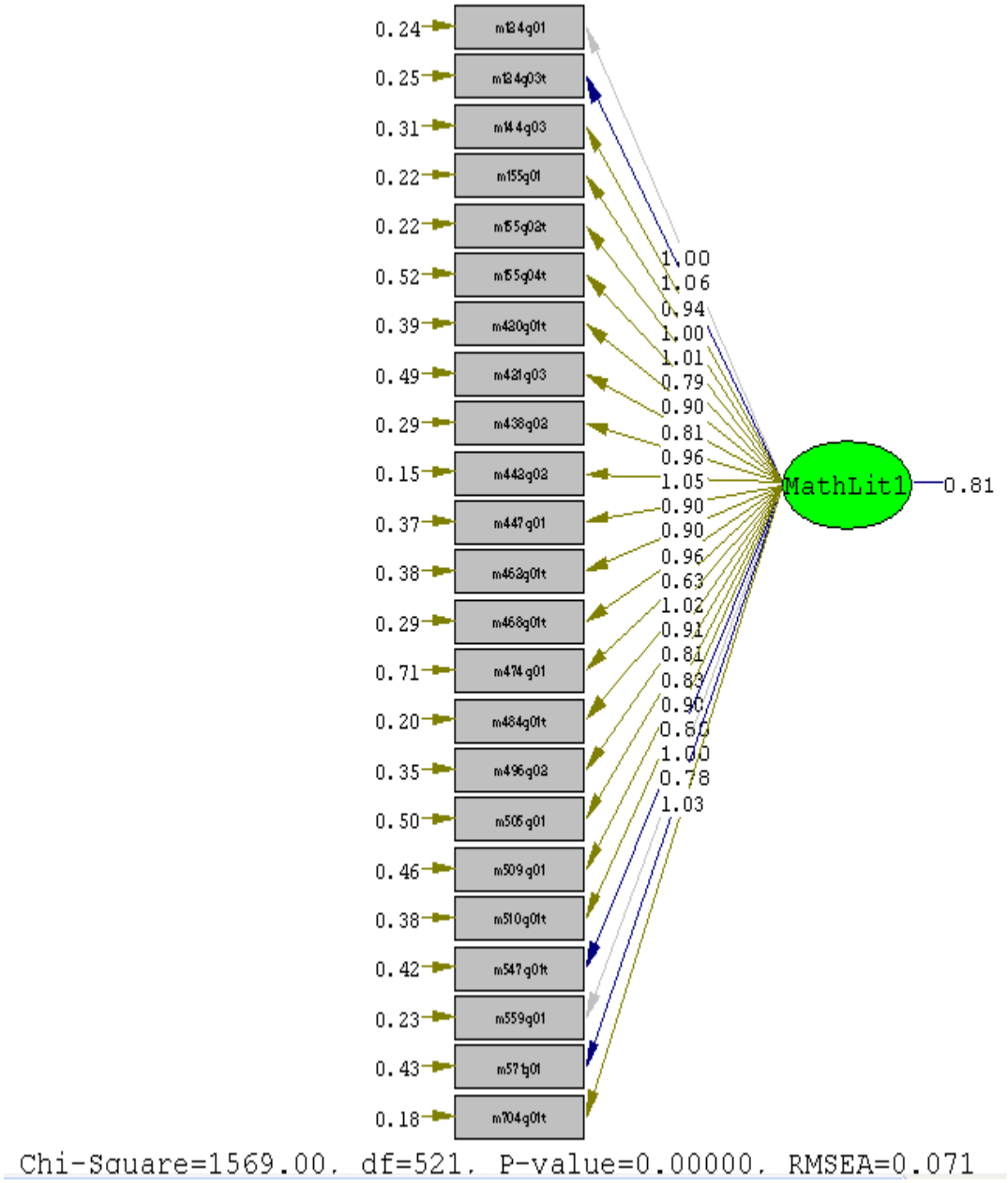
APPENDIX C1

Multi-group Analyses of Unidimensional Mathematics Items in Booklet 13



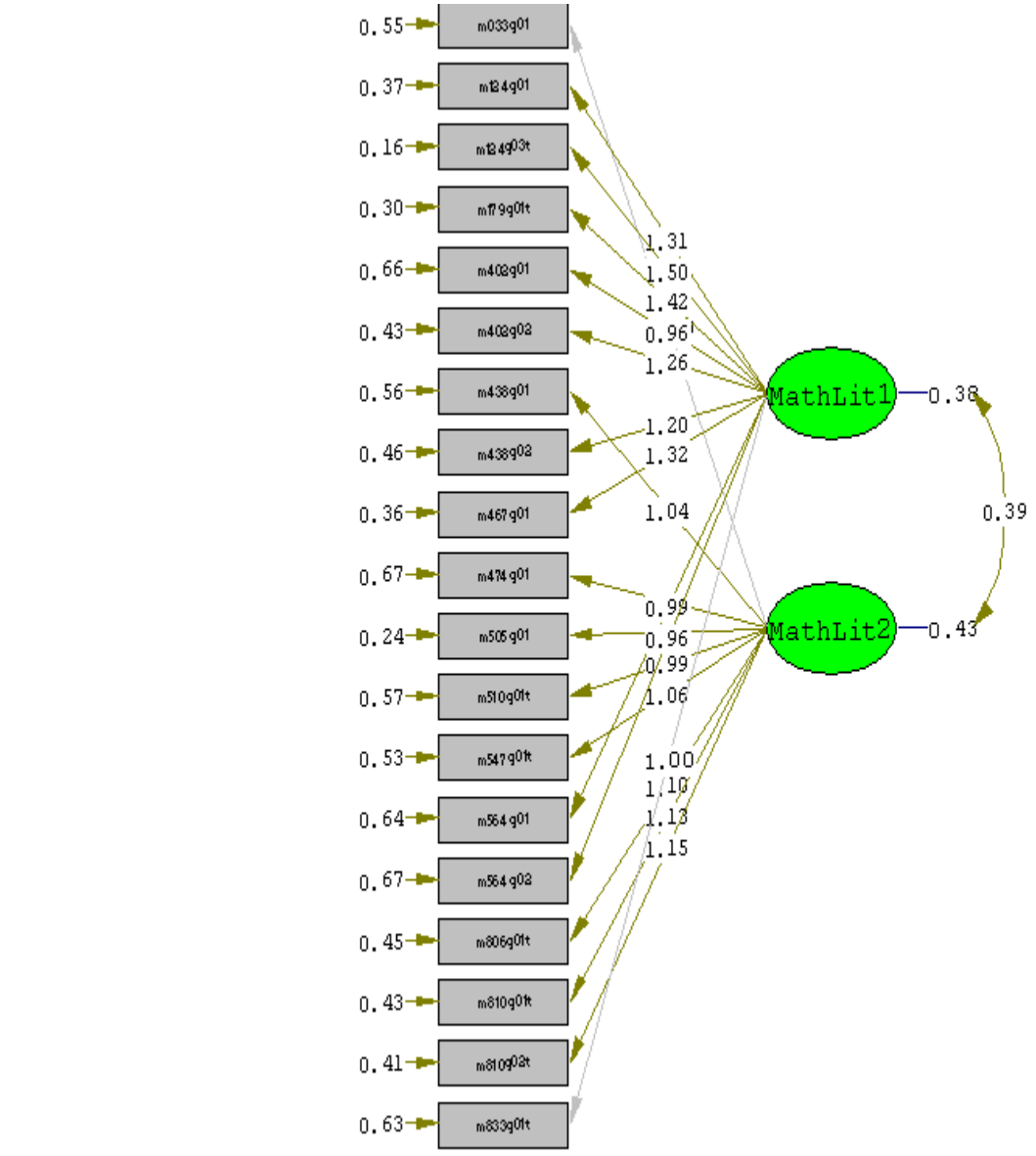
APPENDIX C2

Multi-group Analyses of Unidimensional Mathematics Items in Booklet 3



APPENDIX C3

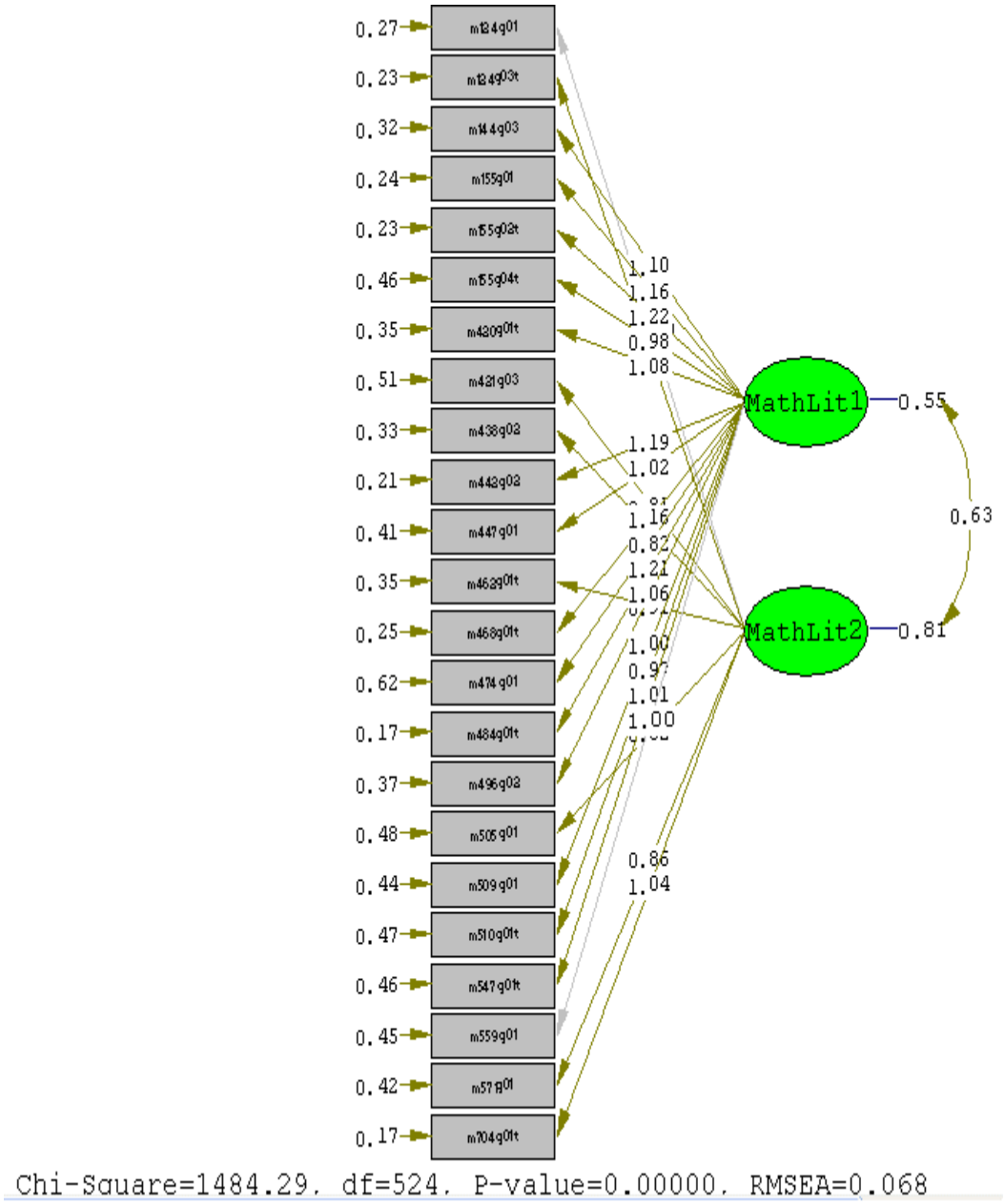
Multi-group Analyses of Two-dimensional Mathematics Items in Booklet 13



Chi-Square=830.31, df=356, P-value=0.00000, RMSEA=0.059

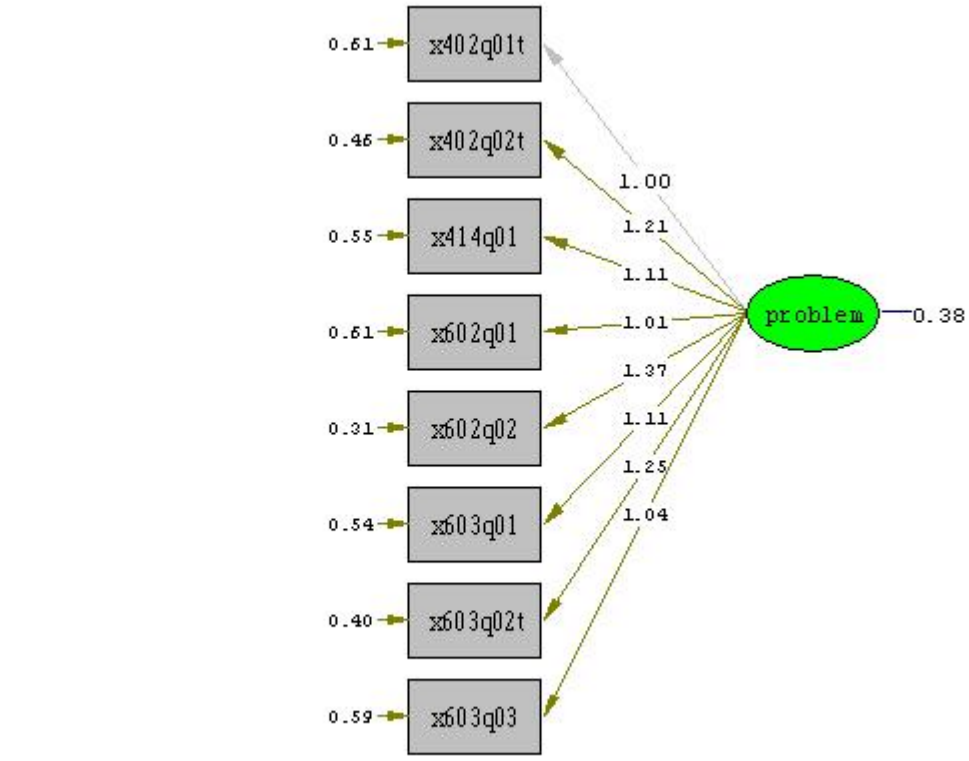
APPENDIX C4

Multi-group Analyses of Two-dimensional Mathematics Items in Booklet 13



APPENDIX C5

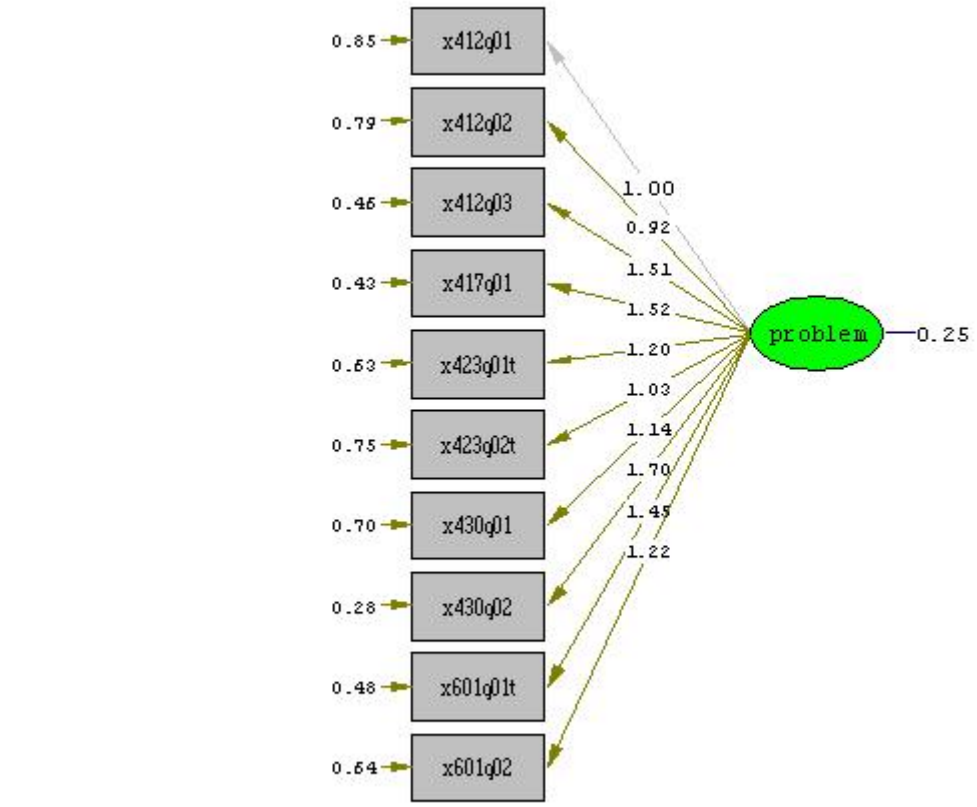
Multi-group Analyses of Problem Solving Items in Booklet 13



Chi-Square=158.23, df=62, P-value=0.00000, RMSEA=0.063

APPENDIX C6

Multi-group Analyses of Problem Solving Items in Booklet 3



APPENDIX D1

Results of M-H Analyses in Booklet 13

Item name	ES	Alpha	X^2	P-Value	MH D-DIF	SE (MH D-DIF)
m033q01	B	1.810	10.389	0.001	-1.395	0.421
m124q01	B	0.443	14.334	0.000	1.911	0.501
m124q03t	A	0.773	1.486	0.223	0.604	0.459
m179q01t	CR	2.653	23.159	0.000	-2.293	0.473
m402q01	A	0.681	4.771	0.029	0.904	0.402
m402q02	B	2.148	10.715	0.001	-1.796	0.537
m438q01	CR	3.301	39.097	0.000	-2.806	0.448
m438q02	A	0.718	2.758	0.097	0.779	0.448
m467q01	A	1.313	1.914	0.167	-0.639	0.435
m474q01	A	1.202	0.975	0.323	-0.433	0.403
m505q01	A	1.442	3.068	0.080	-0.861	0.463
m510q01t	CR	2.243	19.961	0.000	-1.898	0.423
m547q01t	A	0.697	3.437	0.064	0.848	0.438
m564q01	A	0.659	5.194	0.023	0.980	0.414
m564q02	B	0.540	10.726	0.001	1.449	0.432
m806q01t	A	1.211	0.971	0.325	-0.449	0.421
m810q01t	CR	2.186	18.278	0.000	-1.838	0.425
m810q02t	A	1.018	0.000	0.999	-0.041	0.437
m833q01t	A	1.045	0.011	0.918	-0.103	0.495

Alpha > 1.00 favors Reference Group; Alpha < 1.00 favors Focal Group
D-DIF < 0.00 favors Reference Group; D-DIF > 0.00 favors Focal Group

APPENDIX D2

Results of M-H Analyses in Booklet 3

Item name	ES	Alpha	X^2	P-Value	MH D-DIF	SE (MH D-DIF)
m124q01	CF	0.187	67.890	0.000	3.945	0.513
m124q03t	B	0.587	7.771	0.005	1.251	0.436
m144q03	A	1.278	1.533	0.216	-0.577	0.433
m155q01	A	1.191	0.806	0.369	-0.412	0.417
m155q02t	CR	4.104	56.109	0.000	-3.318	0.456
m155q04t	A	1.437	4.350	0.037	-0.852	0.391
m420q01t	B	1.669	8.047	0.005	-1.204	0.415
m421q03	CF	0.388	29.365	0.000	2.226	0.417
m438q02	B	0.579	8.610	0.003	1.285	0.428
m442q02	A	0.911	0.134	0.714	0.219	0.470
m447q01	A	1.191	0.869	0.351	-0.410	0.401
m462q01t	A	0.708	2.840	0.092	0.813	0.458
m468q01t	A	1.278	1.724	0.189	-0.577	0.408
m474q01	B	1.583	7.831	0.005	-1.080	0.377
m484q01t	A	1.290	1.609	0.205	-0.599	0.439
m496q02	A	0.813	1.226	0.268	0.486	0.407
m505q01	A	0.770	1.898	0.168	0.614	0.421
m509q01	A	1.461	4.788	0.029	-0.891	0.392
m510q01t	A	1.477	4.337	0.037	-0.916	0.420
m547q01t	B	0.619	7.194	0.007	1.129	0.414
m559q01	A	0.771	2.058	0.151	0.612	0.404
m571q01	A	0.816	1.324	0.250	0.478	0.389
m704q01t	A	0.730	1.908	0.167	0.739	0.496

Alpha > 1.00 favors Reference Group; Alpha < 1.00 favors Focal Group
D-DIF < 0.00 favors Reference Group; D-DIF > 0.00 favors Focal Group

APPENDIX E1

Results of LR Analysis in Booklet 13

item label	LLM1	CSR2M1	NR2M1	LLM2	CSR2M2	NR2M2	LLM3	CSR2M3	NR2M3	rsqm1	rsqm2	rsqm3	CHI21	P_CHI21	CHI32	P_CHI32
m033q01	766,807	0,262	0,362	761,46	0,267	0,369	758,581	0,27	0,372	0,899	0,914	0,922	5,35	0,02	2,88	0,09
m124q01	649,609	0,315	0,448	622,853	0,338	0,481	621,236	0,34	0,483	0,887	0,951	0,955	26,76	0	1,62	0,2
m124q03t	691,849	0,36	0,488	685,887	0,365	0,495	685,597	0,365	0,495	0,91	0,923	0,924	5,96	0,01	0,29	0,59
m179q01t	664,028	0,361	0,495	641,478	0,379	0,521	641,305	0,379	0,521	0,873	0,925	0,925	22,55	0	0,17	0,68
m402q01	908,618	0,18	0,241	899,763	0,189	0,253	899,49	0,189	0,254	0,868	0,898	0,899	8,86	0	0,27	0,6
m402q02	592,44	0,265	0,403	585,191	0,271	0,413	585,031	0,272	0,413	0,889	0,911	0,911	7,25	0,01	0,16	0,69
m438q01	728,557	0,268	0,376	695,275	0,299	0,419	692,992	0,301	0,422	0,819	0,907	0,913	33,28	0	2,28	0,13
m438q02	765,203	0,265	0,365	757,146	0,273	0,376	757,119	0,273	0,376	0,874	0,895	0,895	8,06	0	0,03	0,87
m467q01	756,219	0,333	0,445	756,054	0,333	0,445	754,656	0,334	0,447	0,903	0,904	0,907	0,17	0,68	1,4	0,24
m474q01	844,761	0,242	0,325	844,753	0,242	0,325	844,752	0,242	0,325	0,92	0,92	0,92	0,01	0,93	0	0,97
m505q01	688,505	0,336	0,463	687,535	0,337	0,464	687,218	0,337	0,464	0,946	0,949	0,949	0,97	0,32	0,32	0,57
m510q01t	818,831	0,228	0,311	803,03	0,243	0,333	802,007	0,244	0,334	0,844	0,89	0,893	15,8	0	1,02	0,31
m547q01t	763,814	0,234	0,328	754,376	0,243	0,341	750,316	0,247	0,347	0,829	0,858	0,87	9,44	0	4,06	0,04
m564q01	891,946	0,2	0,268	882,501	0,209	0,281	882,498	0,209	0,281	0,894	0,925	0,925	9,45	0	0	0,96
m564q02	872,157	0,181	0,246	853,044	0,201	0,273	852,896	0,201	0,274	0,798	0,859	0,859	19,11	0	0,15	0,7
m806q01t	797,849	0,296	0,396	797,788	0,296	0,396	797,186	0,297	0,397	0,89	0,89	0,892	0,06	0,8	0,6	0,44
m810q01t	794,372	0,279	0,376	781,657	0,291	0,392	780,841	0,291	0,393	0,829	0,862	0,864	12,72	0	0,82	0,37
m810q02t	761,867	0,3	0,407	761,106	0,301	0,408	757,554	0,304	0,412	0,839	0,84	0,849	0,76	0,38	3,55	0,06
m833q01t	637,535	0,178	0,279	637,472	0,178	0,279	635,951	0,18	0,282	0,9	0,9	0,906	0,06	0,8	1,52	0,22

Continued)

R2UN	R2NONUN	dif	ABC	B0	B1	B2	B3
0,02	0,01	1	A	0,974	1,387	0,137	-0,206
0,06	0	2	B	-1,42	1,965	-0,638	0,182
0,01	0	1	A	-0,648	1,898	-0,241	-0,074
0,05	0	2	B	-0,994	1,748	0,47	0,055
0,03	0	1	A	-0,29	1,096	-0,258	-0,051
0,02	0	1	A	-1,87	1,575	0,334	-0,055
0,09	0,01	3	C	1,206	1,372	0,47	-0,193
0,02	0	1	A	-0,84	1,482	-0,27	-0,019
0	0	1	A	-0,237	1,623	0,039	0,142
0	0	1	A	0,421	1,269	0,007	-0,004
0	0	1	A	-0,921	1,686	0,116	-0,071
0,05	0	2	B	-0,749	1,12	0,339	0,102
0,03	0,01	1	A	1,04	1,534	-0,185	0,26
0,03	0	1	A	-0,278	1,177	-0,27	-0,006
0,06	0	2	B	-0,609	1,176	-0,391	-0,039
0	0	1	A	0,263	1,484	0,038	0,09
0,03	0	1	A	0,573	1,369	0,345	0,103
0	0,01	1	A	0,782	1,639	-0,167	-0,245
0	0,01	1	A	-1,72	1,246	0,048	-0,147

APPENDIX E2

Results of LR_{M1} Analysis in Booklet 13

ITEM LABEL	LLM1	CSR2M1	NR2M1	LLM2	CSR2M2	NR2M2	LLM3	CSR2M3	NR2M3	rsqm1	rsqm2	rsqm3	CHI21	P_CHI21	CHI32	P_CHI32
m033q01	764,001	0,265	0,365	757,809	0,271	0,373	754,668	0,274	0,377	0,43	0,45	0,46	6,19	0,01	3,14	2,08
m124q01	648,867	0,316	0,449	626,404	0,335	0,477	625,004	0,336	0,478	0,46	0,52	0,52	22,46	0	1,4	2,24
m124q03t	710,692	0,344	0,467	705,985	0,348	0,472	701,395	0,352	0,477	0,52	0,53	0,54	4,71	0,03	4,59	2,03
m179q01t	656,054	0,367	0,504	634,896	0,384	0,528	634,46	0,385	0,528	0,5	0,56	0,56	21,16	0	0,44	2,51
m402q01	898,523	0,19	0,255	890,207	0,199	0,267	881,218	0,208	0,279	0,37	0,4	0,43	8,32	0	8,99	2
m402q02	581,515	0,275	0,418	573,416	0,282	0,43	571,721	0,284	0,432	0,45	0,47	0,48	8,1	0	1,7	2,19
m438q01	725,517	0,271	0,38	693,96	0,3	0,42	688,147	0,305	0,428	0,35	0,42	0,43	31,56	0	5,81	2,02
m438q02	757,157	0,273	0,376	749,562	0,28	0,385	747,396	0,282	0,388	0,43	0,45	0,46	7,6	0,01	2,17	2,14
m467q01	739,35	0,347	0,464	739,182	0,347	0,464	738,231	0,348	0,465	0,47	0,47	0,47	0,17	0,68	0,95	2,33
m474q01	840,456	0,246	0,33	840,427	0,246	0,33	840,185	0,246	0,331	0,44	0,44	0,44	0,03	0,86	0,24	2,62
m505q01	684,727	0,339	0,467	683,664	0,34	0,468	682,682	0,341	0,47	0,49	0,49	0,49	1,06	0,3	0,98	2,32
m510q01t	811,981	0,234	0,321	796,6	0,25	0,341	793,994	0,252	0,345	0,41	0,45	0,46	15,38	0	2,61	2,11
m547q01t	763,061	0,235	0,329	754,925	0,243	0,34	748,388	0,249	0,349	0,4	0,42	0,44	8,14	0	6,54	2,01
m564q01	874,666	0,217	0,291	866,3	0,226	0,302	864,603	0,227	0,305	0,43	0,46	0,46	8,37	0	1,7	2,19
m564q02	850,822	0,203	0,276	832,033	0,222	0,302	824,651	0,229	0,313	0,35	0,4	0,42	18,79	0	7,38	2,01
m806q01t	786,436	0,307	0,41	786,335	0,307	0,41	782,128	0,31	0,415	0,48	0,48	0,49	0,1	0,75	4,21	2,04
m810q01t	768,205	0,303	0,409	755,583	0,314	0,424	754,468	0,315	0,425	0,46	0,49	0,49	12,62	0	1,12	2,29
m810q02t	736,01	0,323	0,438	735,206	0,324	0,439	733,783	0,325	0,441	0,5	0,5	0,5	0,8	0,37	1,42	2,23
m833q01t	633,045	0,183	0,286	632,93	0,183	0,287	630,288	0,186	0,291	0,36	0,36	0,36	0,12	0,73	2,64	2,1

(Continued)

R2UII	R2HOHUII	diff	ABC	B0	B1	B2	B3	B4	B5	B6	B7
0,02	0,01	1	A	0,992	1,33	0,113	-0,008	0,143	-0,117	-0,12	0,004
0,06	0	2	B	-1,42	1,96	-0,153	0,118	-0,588	0,203	-0,149	0,048
0,01	0,01	1	A	-0,609	1,828	-0,074	-0,052	-0,258	-0,288	0,163	0,184
0,06	0	2	B	-0,954	1,696	0,19	-0,115	0,508	0,016	0,032	-0,093
0,03	0,03	1	A	-0,345	1,237	-0,264	0,099	-0,405	-0,148	-0,003	0,283
0,02	0,01	1	A	-2,11	1,716	-0,448	0,428	0,474	-0,202	0,108	-0,092
0,07	0,01	3	C	1,312	1,474	-0,133	-0,178	0,613	-0,235	0,081	-0,247
0,02	0,01	1	A	-0,994	1,424	-0,071	0,37	-0,249	-0,212	0,058	0,012
0	0	1	A	-0,326	1,599	0,041	0,249	0,029	-0,014	0,103	0,064
0	0	1	A	0,469	1,239	0,07	-0,06	0,007	-0,005	-0,048	-0,007
0	0	1	A	-0,957	1,825	-0,157	0,045	0,165	-0,133	0,088	-0,08
0,04	0,01	2	B	-0,709	1,149	0,047	-0,07	0,328	0,223	-0,174	0,004
0,02	0,02	1	A	1,146	1,448	0,054	-0,245	-0,122	0,108	0,201	-0,103
0,03	0	1	A	-0,371	1,496	-0,363	0,161	-0,223	0,054	-0,114	-0,096
0,05	0,02	2	B	-0,824	1,308	-0,204	0,454	-0,449	-0,029	-0,278	0,027
0	0,01	1	A	0,331	1,585	-0,115	-0,128	-0,049	0,028	0,07	0,204
0,03	0	1	A	0,808	1,361	0,14	-0,42	0,37	0,086	0,041	-0,046
0	0	1	A	1,004	1,547	0,174	-0,495	-0,148	-0,111	-0,05	0,053
0	0	1	A	-1,71	1,375	-0,067	-0,08	0,046	-0,282	0,188	0,006

APPENDIX E3

Results of LR_{M2} Analysis in Booklet 13

ITEM LABEL	LLM1	CSR2M1	NR2M1	LLM2	CSR2M2	NR2M2	LLM3	CSR2M3	NR2M3	rsqm1	rsqm2	rsqm3	CHI21	P_CHI21	CHI32	P_CHI32
m033q01	746,187	0,282	0,388	742,7	0,285	0,393	738,41	0,289	0,398	0,59	0,61	0,62	3,49	0,06	4,29	2,04
m124q01	616,245	0,344	0,489	609,923	0,349	0,496	607,198	0,351	0,5	0,62	0,64	0,65	6,32	0,01	2,73	2,1
m124q03t	648,419	0,395	0,535	648,129	0,395	0,536	644,018	0,398	0,54	0,64	0,64	0,65	0,29	0,59	4,11	2,04
m179q01t	629,234	0,389	0,534	578,756	0,427	0,587	575,062	0,43	0,59	0,48	0,61	0,62	50,48	0	3,69	2,05
m402q01	828,024	0,261	0,35	827,786	0,261	0,35	824,652	0,264	0,354	0,64	0,64	0,66	0,24	0,63	3,13	2,08
m402q02	566,094	0,289	0,44	539,609	0,313	0,476	537,363	0,315	0,479	0,55	0,64	0,64	26,49	0	2,25	2,13
m438q01	693,986	0,3	0,42	669,977	0,321	0,45	666,953	0,324	0,454	0,56	0,64	0,65	24,01	0	3,02	2,08
m438q02	718,319	0,308	0,424	718,229	0,308	0,425	710,329	0,315	0,434	0,62	0,62	0,64	0,09	0,76	7,9	2
m467q01	730,43	0,355	0,474	718,668	0,365	0,487	715,551	0,367	0,491	0,58	0,61	0,62	11,76	0	3,12	2,08
m474q01	788,985	0,294	0,395	786,349	0,297	0,398	784,302	0,299	0,401	0,58	0,59	0,59	2,64	0,1	2,05	2,15
m505q01	644,71	0,373	0,513	644,601	0,373	0,513	638,776	0,377	0,519	0,61	0,61	0,63	0,11	0,74	5,83	2,02
m510q01t	700,755	0,337	0,461	699,37	0,338	0,462	696,612	0,34	0,465	0,62	0,62	0,63	1,39	0,24	2,76	2,1
m547q01t	717,518	0,279	0,39	695,13	0,299	0,419	691,706	0,302	0,423	0,51	0,58	0,59	22,39	0	3,42	2,06
m564q01	794,174	0,294	0,394	793,832	0,295	0,395	789,345	0,299	0,4	0,64	0,64	0,66	0,34	0,56	4,49	2,03
m564q02	787,527	0,265	0,362	786,468	0,266	0,363	782,295	0,27	0,368	0,52	0,52	0,54	1,06	0,3	4,17	2,04
m806q01t	747,081	0,341	0,455	746,121	0,342	0,457	742,554	0,345	0,461	0,6	0,61	0,62	0,96	0,33	3,57	2,06
m810q01t	705,722	0,357	0,482	702,7	0,359	0,485	700,252	0,361	0,488	0,62	0,63	0,64	3,02	0,08	2,45	2,12
m810q02t	690,182	0,362	0,491	682,501	0,368	0,499	681,406	0,369	0,5	0,61	0,64	0,64	7,68	0,01	1,1	2,3
m833q01t	617,623	0,199	0,312	614,453	0,202	0,317	611,542	0,205	0,321	0,51	0,52	0,53	3,17	0,08	2,91	2,09

(Continued)

R2UII	R2NONUII	dif	ABC	B0	B1	B2	B3	B4	B5	B6	B7
0,02	0,01	1	A	0,957	0,348	1,21	0,078	0,113	0,06	-0,251	0,018
0,02	0,01	1	A	-1,27	1,489	0,623	-0,086	-0,321	0,252	-0,134	-0,037
0	0,01	1	A	-0,531	1,588	0,569	-0,087	0,048	0,045	-0,241	0,131
0,13	0,01	3	C	-0,759	2,095	0,223	-0,637	0,855	0,336	-0,117	-0,27
0	0,02	1	A	-0,119	1,465	-0,141	-0,34	-0,036	-0,086	-0,059	0,192
0,09	0	3	C	-2,01	1,726	-0,076	0,006	0,822	-0,205	0,167	-0,11
0,08	0,01	3	C	1,194	0,171	1,298	-0,037	0,444	-0,169	-0,059	-0,053
0	0,02	1	A	-0,989	1,338	0,181	0,276	0,026	-0,418	0,19	0,163
0,03	0,01	1	A	-0,207	1,479	0,389	-0,035	0,324	-0,138	0,223	0,097
0,01	0	1	A	0,325	-0,068	1,562	0,132	-0,229	-0,043	0,069	0,193
0	0,02	1	A	-1,3	0,116	1,912	0,326	-0,145	-0,285	0,355	0,112
0	0,01	1	A	-1,13	-0,518	2,002	0,31	0	-0,06	0,196	0,143
0,07	0,01	2	B	0,977	-0,033	1,811	0,172	-0,41	0,074	0,206	0,003
0	0,02	1	A	-0,073	1,738	-0,173	-0,333	0,148	0,27	-0,168	-0,237
0	0,02	1	A	-0,623	1,48	-0,082	0,066	-0,012	0,012	-0,104	-0,227
0,01	0,01	1	A	0,174	0,157	1,622	0,251	-0,127	0,139	-0,011	0,168
0,01	0,01	1	A	0,688	0,029	1,673	-0,355	0,162	-0,004	0,178	0,062
0,03	0	1	A	0,768	0,035	1,875	-0,295	-0,345	-0,101	0,041	0,124
0,01	0,01	1	A	-1,71	1,156	0,249	-0,125	0,225	-0,152	-0,145	0,176

APPENDIX E4

Results of LR Analysis in Booklet 3

item label	LLM1	CSR2M1	NR2M1	LLM2	CSR2M2	NR2M2	LLM3	CSR2M3	NR2M3	rsqm1	rsqm2	rsqm3	CHI21	P_CHI21	CHI32	P_CHI32
m124q01	737,722	0,266	0,376	659,808	0,334	0,472	658,568	0,335	0,473	0,726	0,907	0,91	77,91	0	1,24	0,27
m124q03t	793,815	0,297	0,402	785,02	0,305	0,412	779,584	0,31	0,419	0,877	0,896	0,909	8,8	0	5,44	0,02
m144q03	726,826	0,285	0,4	725,187	0,286	0,402	724,629	0,287	0,403	0,896	0,9	0,902	1,64	0,2	0,56	0,46
m155q01	760,247	0,335	0,451	759,624	0,335	0,451	757,519	0,337	0,454	0,883	0,884	0,889	0,62	0,43	2,11	0,15
m155q02t	704,378	0,39	0,522	649,737	0,43	0,576	649,691	0,43	0,576	0,807	0,921	0,921	54,64	0	0,05	0,83
m155q04t	900,004	0,215	0,288	894,637	0,22	0,295	893,927	0,221	0,296	0,841	0,857	0,859	5,37	0,02	0,71	0,4
m420q01t	779,158	0,333	0,445	771,584	0,34	0,454	768,978	0,342	0,456	0,884	0,901	0,906	7,57	0,01	2,61	0,11
m421q03	882,928	0,158	0,22	845,87	0,197	0,273	843,22	0,199	0,276	0,711	0,828	0,836	37,06	0	2,65	0,1
m438q02	823,755	0,281	0,378	811,971	0,292	0,392	811,851	0,292	0,392	0,868	0,896	0,896	11,78	0	0,12	0,73
m442q02	645,531	0,329	0,47	644,972	0,33	0,47	644,971	0,33	0,47	0,931	0,932	0,932	0,56	0,45	0	0,97
m447q01	842,116	0,275	0,368	841,51	0,276	0,369	841,504	0,276	0,369	0,92	0,922	0,922	0,61	0,44	0,01	0,94
m462q01t	692,388	0,192	0,29	688,442	0,196	0,296	684,618	0,199	0,302	0,817	0,83	0,842	3,95	0,05	3,82	0,05
m468q01t	820,221	0,298	0,398	818,707	0,299	0,4	817,832	0,3	0,401	0,905	0,908	0,91	1,51	0,22	0,88	0,35
m474q01	943,385	0,161	0,216	934,68	0,17	0,229	934,076	0,17	0,23	0,811	0,842	0,844	8,71	0	0,6	0,44
m484q01t	725,873	0,378	0,505	724,749	0,379	0,506	723,725	0,38	0,507	0,946	0,949	0,951	1,12	0,29	1,02	0,31
m496q02	839,345	0,282	0,376	837,374	0,284	0,379	835,997	0,285	0,38	0,895	0,9	0,904	1,97	0,16	1,38	0,24
m505q01	799,988	0,231	0,322	797,142	0,233	0,325	794,5	0,236	0,329	0,866	0,874	0,881	2,85	0,09	2,64	0,1
m509q01	890,489	0,232	0,31	886,577	0,235	0,315	883,25	0,239	0,319	0,886	0,898	0,907	3,91	0,05	3,33	0,07
m510q01t	800,279	0,239	0,331	796,434	0,242	0,336	794,974	0,244	0,338	0,889	0,9	0,904	3,85	0,05	1,46	0,23
m547q01t	829,328	0,195	0,272	820,624	0,203	0,284	817,001	0,207	0,289	0,844	0,872	0,885	8,7	0	3,62	0,06
m559q01	866,325	0,257	0,344	863,285	0,26	0,348	857,939	0,265	0,354	0,872	0,879	0,893	3,04	0,08	5,35	0,02
m571q01	911,076	0,177	0,241	908,621	0,18	0,244	901,833	0,187	0,254	0,813	0,82	0,842	2,46	0,12	6,79	0,01
m704q01t	607,402	0,341	0,493	605,229	0,343	0,495	602,374	0,345	0,499	0,894	0,899	0,906	2,17	0,14	2,86	0,09

(Continued)

R2UN	R2NONUN	dif	ABC	B0	B1	B2	B3
0,18	0	3	C	-1,2	1,833	-0,859	-0,153
0,02	0,01	1	A	-0,512	1,605	-0,239	-0,274
0	0	1	A	1,214	1,579	0,078	-0,094
0	0,01	1	A	0,531	1,689	0,112	0,178
0,11	0	3	C	0,4	1,896	0,717	0,029
0,02	0	1	A	-0,354	1,093	0,185	0,079
0,02	0,01	1	A	-0,236	1,592	0,239	0,185
0,12	0,01	3	C	-0,788	1,126	-0,5	-0,163
0,03	0	1	A	-0,433	1,51	-0,307	-0,039
0	0	1	A	-1,44	1,848	-0,078	0,004
0	0	1	A	0,303	1,368	0,068	0,008
0,01	0,01	1	A	-1,55	1,324	-0,08	-0,231
0	0	1	A	-0,193	1,443	0,111	-0,101
0,03	0	1	A	0,443	0,889	0,226	-0,07
0	0	1	A	-0,158	1,838	0,098	0,129
0,01	0	1	A	0,19	1,45	-0,108	0,128
0,01	0,01	1	A	-0,931	1,322	-0,102	-0,175
0,01	0,01	1	A	-0,259	1,168	0,156	0,175
0,01	0	1	A	-0,927	1,274	0,214	-0,128
0,03	0,01	1	A	0,944	1,255	-0,186	0,207
0,01	0,01	1	A	0,264	1,375	-0,188	-0,247
0,01	0,02	1	A	-0,623	1,072	-0,174	0,245
0,01	0,01	1	A	1,828	2,183	-0,332	-0,286

APPENDIX E5

Results of LR_{M1} Analysis in Booklet 3

ITEM LABEL	LLM1	CSR2M1	NR2M1	LLM2	CSR2M2	NR2M2	LLM3	CSR2M3	NR2M3	rsqm1	rsqm2	rsqm3	CHI21	P_CHI21
m124q01	733,445	0,27	0,381	658,508	0,335	0,473	651,245	0,341	0,482	0,33	0,5	0,52	74,94	0
m124q03t	793,633	0,298	0,402	784,899	0,305	0,412	777,592	0,311	0,421	0,45	0,47	0,49	8,73	0
m144q03	724,25	0,287	0,403	722,285	0,289	0,406	721,435	0,29	0,407	0,43	0,44	0,44	1,97	0,16
m155q01	750,17	0,343	0,462	749,616	0,344	0,462	745,626	0,347	0,467	0,48	0,48	0,49	0,55	0,46
m155q02t	704,215	0,39	0,522	649,539	0,431	0,576	649,383	0,431	0,576	0,43	0,56	0,56	54,68	0
m155q04t	889,628	0,225	0,302	884,03	0,23	0,309	878,007	0,236	0,317	0,37	0,39	0,4	5,6	0,02
m420q01t	777,948	0,334	0,447	770,478	0,34	0,455	766,644	0,344	0,459	0,48	0,5	0,51	7,47	0,01
m421q03	923,566	0,115	0,159	889,924	0,151	0,209	879,421	0,162	0,225	0,2	0,28	0,31	33,64	0
m438q02	812,86	0,291	0,391	800,177	0,302	0,406	796,721	0,305	0,41	0,45	0,49	0,5	12,68	0
m442q02	643,24	0,331	0,472	642,327	0,332	0,473	639,431	0,334	0,477	0,47	0,47	0,48	0,91	0,34
m447q01	842,01	0,275	0,369	841,352	0,276	0,369	836,026	0,281	0,376	0,47	0,47	0,49	0,66	0,42
m462q01t	689,904	0,194	0,294	686,307	0,198	0,299	680,371	0,204	0,308	0,36	0,37	0,39	3,6	0,06
m468q01t	817,841	0,3	0,401	816,629	0,301	0,402	808,376	0,308	0,412	0,48	0,49	0,51	1,21	0,27
m474q01	942,156	0,162	0,218	932,897	0,172	0,231	930,952	0,174	0,234	0,34	0,37	0,37	9,26	0
m484q01t	722,359	0,381	0,508	721,569	0,382	0,509	720,218	0,383	0,511	0,47	0,48	0,48	0,79	0,37
m496q02	834,664	0,286	0,382	833,031	0,287	0,384	830,213	0,29	0,387	0,47	0,48	0,48	1,63	0,2
m505q01	799,947	0,231	0,322	797,008	0,234	0,326	793,601	0,237	0,33	0,41	0,42	0,43	2,94	0,09
m509q01	880,319	0,241	0,323	876,46	0,245	0,328	873,24	0,248	0,332	0,48	0,49	0,51	3,86	0,05
m510q01t	798,281	0,24	0,334	794,206	0,244	0,339	788,284	0,25	0,347	0,42	0,43	0,45	4,08	0,04
m547q01t	828,94	0,195	0,273	820,15	0,204	0,285	815,489	0,208	0,292	0,37	0,39	0,41	8,79	0
m559q01	861,876	0,262	0,349	858,765	0,264	0,353	849,763	0,273	0,364	0,39	0,39	0,41	3,11	0,08
m571q01	904,807	0,184	0,25	902,692	0,186	0,252	895,59	0,193	0,262	0,36	0,36	0,38	2,12	0,15
m704q01t	600,135	0,347	0,501	598,155	0,348	0,504	594,557	0,351	0,508	0,43	0,43	0,44	1,98	0,16

(Continued)

CHI32	P_CHI32	R2UN	R2NONUN	dif	ABC	B0	B1	B2	B3	B4	B5	B6	B7
7,26	2,01	0,17	0,02	3	C	-1,26	1,806	-0,198	0,171	-0,913	-0,365	0,082	0,239
7,31	2,01	0,02	0,02	1	A	-0,569	1,606	0,008	0,17	-0,219	-0,277	-0,065	-0,085
0,85	2,36	0,01	0	1	A	1,262	1,682	-0,209	-0,138	0,079	-0,132	0,109	0,053
3,99	2,05	0	0,01	1	A	0,716	1,67	0,03	-0,42	0,082	0,214	0,019	0,043
0,16	2,69	0,13	0	3	C	0,424	1,919	-0,036	-0,059	0,718	0,065	-0,05	0,001
6,02	2,01	0,02	0,01	1	A	-0,197	1,215	-0,017	-0,344	0,224	-0,011	0,247	-0,095
3,83	2,05	0,02	0,01	1	A	-0,166	1,614	0,005	-0,143	0,203	0,261	-0,099	0,067
10,5	2	0,08	0,03	3	C	-0,917	0,673	0,252	0,364	-0,431	-0,281	-0,037	-0,095
3,46	2,06	0,04	0,01	2	B	-0,608	1,444	0,037	0,419	-0,382	-0,085	-0,101	0,114
2,9	2,09	0	0,01	1	A	-1,4	1,69	0,242	-0,051	-0,14	0,077	-0,262	0,184
5,33	2,02	0	0,02	1	A	0,316	1,453	-0,072	0,024	0,067	0,221	-0,293	-0,007
5,94	2,01	0,01	0,02	1	A	-1,67	1,353	-0,166	0,246	-0,047	-0,214	-0,059	-0,053
8,25	2	0,01	0,02	1	A	-0,262	1,367	0,145	0,117	0,197	-0,336	0,296	-0,195
1,95	2,16	0,03	0	1	A	0,484	0,982	-0,13	-0,068	0,176	0,011	-0,087	0,1
1,35	2,25	0,01	0	1	A	-0,175	1,717	0,199	0,085	0,107	0,15	-0,075	-0,078
2,82	2,09	0,01	0	1	A	0,285	1,608	-0,182	-0,202	-0,095	0,198	-0,014	-0,007
3,41	2,06	0,01	0,01	1	A	-0,932	1,274	0,037	0,008	-0,149	-0,18	-0,042	0,095
3,22	2,07	0,01	0,02	1	A	-0,431	1,201	-0,076	0,306	0,155	0,059	0,116	0,041
5,92	2,01	0,01	0,02	1	A	-1,04	1,335	-0,106	0,227	0,325	-0,095	-0,03	-0,207
4,66	2,03	0,02	0,02	1	A	0,996	1,209	0,024	-0,118	-0,192	0,157	0,099	0,016
9	2	0	0,02	1	A	0,126	1,401	0,01	0,298	-0,223	-0,371	0,072	0,017
7,1	2,01	0	0,02	1	A	-0,723	1,166	-0,151	0,142	-0,094	0,107	0,17	-0,092
3,6	2,06	0	0,01	1	A	1,933	2,182	-0,254	-0,392	-0,317	-0,037	-0,098	0,193

APPENDIX E6

Results of LR_{M2} Analysis in Booklet 3

ITEM LABEL	LLM1	CSR2M1	NR2M1	LLM2	CSR2M2	NR2M2	LLM3	CSR2M3	NR2M3	rsqm1	rsqm2	rsqm3	CHI21	P_CHI21
m124q01	570,64	0,405	0,571	544,555	0,424	0,598	541,712	0,426	0,601	0,47	0,54	0,55	26,09	0
m124q03t	649,523	0,413	0,559	649,497	0,413	0,559	644,33	0,417	0,564	0,6	0,6	0,61	0,03	0,87
m144q03	709,132	0,3	0,422	708,674	0,301	0,422	706,665	0,303	0,425	0,56	0,57	0,57	0,46	0,5
m155q01	734,854	0,356	0,479	734,822	0,356	0,479	731,754	0,358	0,482	0,56	0,56	0,57	0,03	0,86
m155q02t	683,449	0,406	0,543	638,68	0,438	0,586	634,564	0,441	0,59	0,46	0,58	0,59	44,77	0
m155q04t	874,988	0,239	0,321	873,399	0,241	0,323	872,715	0,241	0,324	0,48	0,49	0,49	1,59	0,21
m420q01t	759,639	0,349	0,467	756,097	0,352	0,471	753,578	0,354	0,473	0,56	0,57	0,58	3,54	0,06
m421q03	756,696	0,281	0,39	744,116	0,293	0,406	738,875	0,297	0,412	0,48	0,51	0,53	12,58	0
m438q02	678,229	0,401	0,539	678,133	0,401	0,539	677,585	0,401	0,54	0,56	0,56	0,56	0,1	0,76
m442q02	622,039	0,349	0,497	617,515	0,352	0,503	612,229	0,357	0,509	0,5	0,51	0,53	4,52	0,03
m447q01	806,03	0,307	0,411	805,962	0,307	0,411	805,53	0,308	0,412	0,57	0,57	0,57	0,07	0,79
m462q01t	594,557	0,285	0,431	593,033	0,286	0,433	589,918	0,289	0,438	0,52	0,52	0,53	1,52	0,22
m468q01t	811,345	0,306	0,409	810,194	0,307	0,41	808,717	0,308	0,412	0,55	0,55	0,55	1,15	0,28
m474q01	911,762	0,193	0,26	908,168	0,197	0,265	904,415	0,201	0,27	0,41	0,43	0,44	3,59	0,06
m484q01t	696,911	0,4	0,534	696,909	0,4	0,534	696,279	0,401	0,535	0,61	0,61	0,61	0	0,96
m496q02	820,374	0,299	0,399	815,059	0,303	0,405	809,408	0,308	0,412	0,5	0,52	0,53	5,32	0,02
m505q01	728,852	0,296	0,413	728,043	0,297	0,414	727,269	0,298	0,415	0,52	0,52	0,53	0,81	0,37
m509q01	869,806	0,251	0,336	867,132	0,254	0,339	862,941	0,258	0,344	0,49	0,49	0,51	2,67	0,1
m510q01t	773,49	0,264	0,366	772,857	0,264	0,367	768,05	0,269	0,373	0,48	0,48	0,5	0,63	0,43
m547q01t	807,951	0,216	0,302	793,283	0,23	0,322	789,398	0,234	0,327	0,44	0,48	0,5	14,67	0
m559q01	852,975	0,27	0,36	848,375	0,274	0,366	836,726	0,284	0,38	0,45	0,46	0,49	4,6	0,03
m571q01	800,029	0,284	0,386	797,866	0,286	0,388	779,342	0,303	0,41	0,47	0,47	0,52	2,16	0,14
m704q01t	541,888	0,393	0,568	541,85	0,393	0,568	540,745	0,394	0,569	0,49	0,49	0,5	0,04	0,85

(Continued)

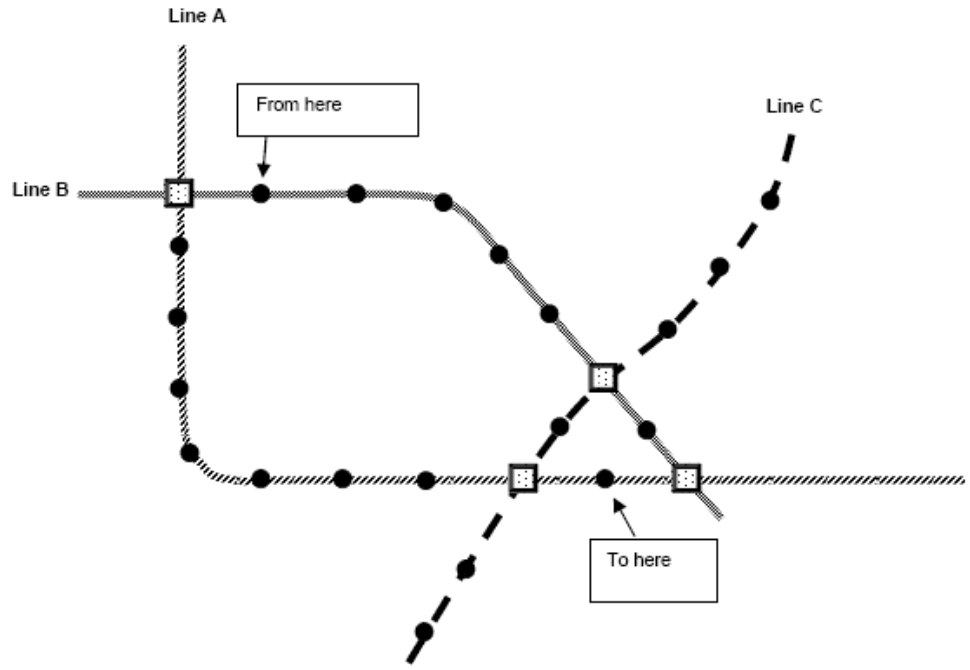
CHI32	P_CHI32	R2UN	R2NONUN	dif	ABC	B0	B1	B2	B3	B4	B5	B6	B7
2,84	2,09	0,07	0,01	3	C	-1,32	-0,029	2,227	-0,085	-0,62	-0,051	-0,208	0,279
5,17	2,02	0	0,01	1	A	-0,589	-0,012	2,262	-0,191	0,136	0,101	-0,304	-0,199
2,01	2,16	0,01	0	1	A	1,341	1,58	0,029	-0,215	0,015	-0,216	0,113	-0,026
3,07	2,08	0	0,01	1	A	0,644	1,803	-0,026	-0,163	-0,046	-0,071	0,192	0,086
4,12	2,04	0,12	0,01	3	C	0,454	1,827	0,303	0,055	0,624	-0,149	0,221	0,187
0,68	2,41	0,01	0	1	A	-0,285	1,223	0,003	-0,158	0,121	0,022	0,065	-0,042
2,52	2,11	0,01	0,01	1	A	-0,143	1,55	0,235	-0,247	0,175	0,053	0,14	-0,047
5,24	2,02	0,03	0,02	1	A	-1,05	-0,263	1,61	0,343	-0,243	-0,004	-0,24	-0,12
0,55	2,46	0	0	1	A	-0,571	-0,023	1,997	0,288	-0,03	0,095	-0,063	0,028
5,29	2,02	0,01	0,02	1	A	-1,65	2,106	-0,195	0,153	-0,369	0,232	-0,359	0,254
0,43	2,51	0	0	1	A	0,256	1,615	-0,145	0,142	-0,025	-0,04	-0,036	-0,024
3,12	2,08	0	0,01	1	A	-1,82	-0,225	1,821	0,01	0,214	-0,105	-0,199	0,114
1,48	2,22	0	0	1	A	-0,265	1,231	0,307	0,162	0,104	-0,044	-0,091	-0,021
3,75	2,05	0,02	0,01	1	A	0,434	1,197	-0,312	0,027	0,091	0,031	-0,162	0,098
0,63	2,43	0	0	1	A	-0,223	1,937	0,029	0,134	0,013	0,081	0,032	-0,023
5,65	2,02	0,02	0,01	1	A	0,201	1,549	-0,03	-0,056	-0,291	0,211	-0,116	0,211
0,77	2,38	0	0,01	1	A	-0,946	0,137	1,498	-0,145	0,091	-0,112	0,034	0,013
4,19	2,04	0	0,02	1	A	-0,386	1,165	0,063	0,325	0,134	-0,001	0,18	0,036
4,81	2,03	0	0,02	1	A	-1,04	1,429	-0,109	0,106	0,178	0,028	-0,134	-0,128
3,89	2,05	0,04	0,02	2	B	1,096	1,428	-0,141	-0,197	-0,358	0,075	0,128	0,092
11,65	2	0,01	0,03	1	A	0,2	1,454	0,051	0,278	-0,29	-0,405	0,047	0,034
18,52	2	0	0,05	1	A	-0,677	-0,291	1,746	-0,031	0,103	0,157	0,371	-0,078
1,11	2,29	0	0,01	1	A	2,157	0,546	1,958	-0,491	-0,053	-0,039	0,091	0,155

APPENDIX F1

An Example of Problem Solving Item

TRANSIT SYSTEM

The following diagram shows part of the transport system of a city in Zedland, with three railway lines. It shows where you are at present, and where you have to go.



● means a station on a railway line.

⊡ means a station that is a junction where you can change from one railway line to another (Lines A, B or C).

Question 1: TRANSIT SYSTEM

X415Q01 – 01 02 11 12 13 21 22 99

The diagram indicates a station where you are currently at ("From here"), and the station where you want to go ("To here"). **Mark on the diagram** the best route in terms of cost and time, and indicate below the fare you have to pay, and the approximate time for the journey.

Fare: ..... zeds.

Approximate time for journey: ..... minutes.

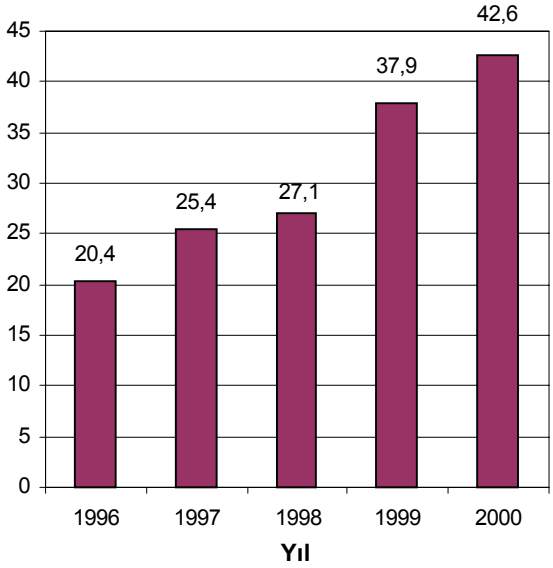
APPENDIX F2

Examples of Released Turkish DIF Items

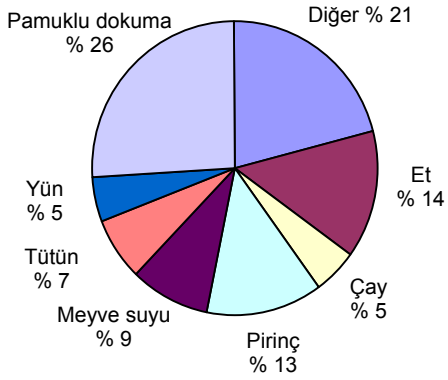
DIŞSATIM

Aşağıdaki grafikler, para birimi olarak zed kullanan, Zed ülkesinden yapılan dışsatımla ilgili bilgileri göstermektedir.

1996-2000 yılları arasında Zed ülkesinden milyon zed olarak toplam yıllık dışsatımı



2000 yılında Zed ülkesinden dışsatımın dağılımı



M438q01

Soru 1: DIŞSATIM

1998 yılında Zed ülkesinden yapılan dışsatımın toplam değeri (milyon zed olarak) nedir?

M438q02

Soru 2: DIŞSATIM

2000 yılında Zed ülkesinden dışarıya satılan meyve suyunun değeri ne idi?

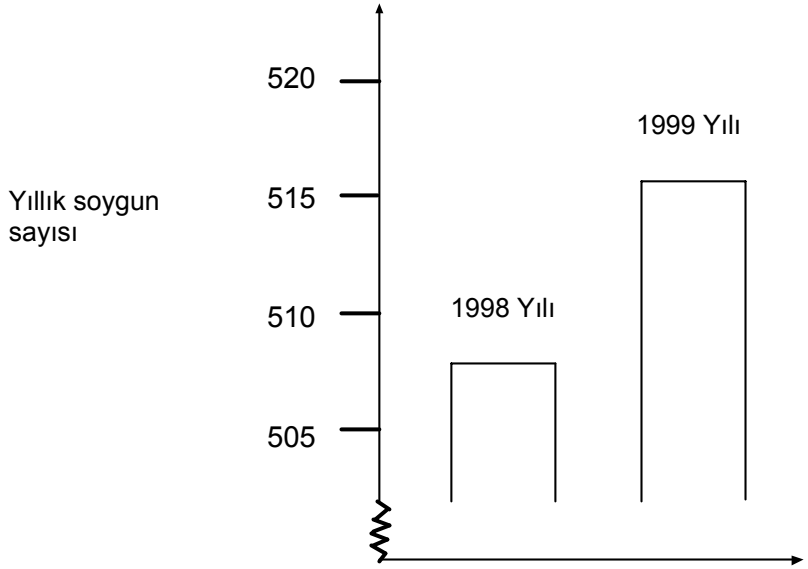
A 1,8 milyon zed.
B 2,3 milyon zed.
C 2,4 milyon zed.
D 3,4 milyon zed.
E 3,8 milyon zed.

M179q01

SOYGUNLAR

Bir televizyon muhabiri, bu grafiği gösterdi ve şöyle dedi:

“Bu grafik 1998 yılından 1999’a kadar soygunların sayısında çok büyük bir artış



olduğunu göstermektedir.“

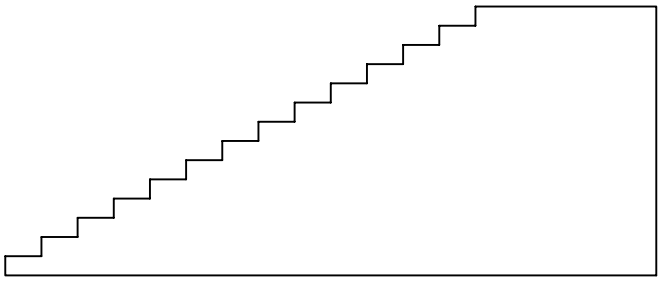
Muhabirin sözlerinin grafiğin kabul edilebilir bir yorumu olduğunu düşünüyor musunuz? Yanıtınızı desteklemek için bir açıklama yapınız.

M547q01

MERDİVEN

Soru 1: MERDİVEN

Aşağıdaki şekil 14 basamaklı ve toplam yüksekliği 252 cm olan bir merdiveni göstermektedir:



Toplam yükseklik 252 cm

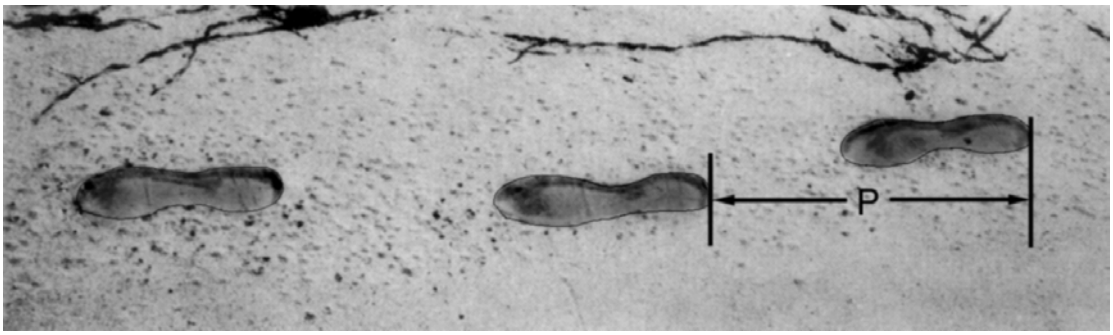
Toplam genişlik 400 cm

14 basamağın her birinin yüksekliği nedir?

Yükseklik: ..... cm.

M124q01

YÜRÜYÜŞ



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

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

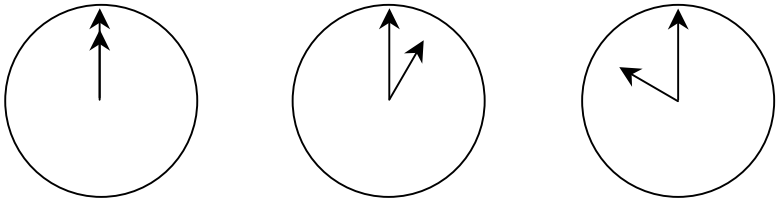
M124Q01- 0 1 2 9

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

INTERNETTE SOHBET

Mark (Avustralya, Sidney'den) ve Hans (Almanya, Berlin'den) internet ortamında "çet" (chat) aracılığıyla haberleşiyorlar. 'Sohbet' edebilmeleri için internete aynı saatte bağlanmaları gerekmektedir.

'Sohbet edebilmek' için uygun bir zaman bulabilmek amacıyla, Mark dünya saat çizelgesine bakarak aşağıdakileri öğrendi:



Greenwich 24:00
(Gece yarısı)

Berlin 1:00
(Sabaha karşı)

Sidney 10:00
(Sabah)

M402q01

Soru 1: İNTERNETTE SOHBET

Sidney'de saat akşam 7:00 iken, Berlin'de saat kaçtır?

Yanıt : .....

M402q02

Soru 2: İNTERNETTE SOHBET

Mark ve Hans okula gitmek zorunda oldukları için yerel saatleriyle 9:00 ve 16:30 arasında sohbet edemiyorlar. Ayrıca, yerel saatleriyle 23:00'ten 07:00'ye kadar uyuyor olacakları için sohbet edemiyorlar.

Mark ve Hans'ın sohbet edebilmeleri için hangi saatler uygun olacaktır? Tabloya yerel saatleri yazınız.

Yer	Saatler
Sidney	
Berlin	

APPENDIX G1

An Example of The LISREL Syntax for Multi-Gorup Analysis of Mathematics Literacy Model

```
Group Turkey
Observed Variables:
m124q01 m124q03t m144q03 m155q01 m155q02t m155q04t m420q01t m421q03 m438q02 m442q02
m447q01 m462q01t m468q01t m474q01 m484q01t m496q02 m505q01 m509q01 m510q01t
m547q01t m559q01 m571q01 m704q01t
Means from File tur..ME
Covariance Matrix from File tur.CM
Asymptotic Covariance Matrix from File tur.ACC
Sample Size: 379
Latent Variables: MathLit
Relationships:
m124q01= CONST 1*MathLit
m124q03t= CONST MathLit
m144q03= CONST MathLit
m155q01= CONST MathLit
m155q02t= CONST MathLit
m155q04t= CONST MathLit
m420q01t= CONST MathLit
m421q03= CONST MathLit
m438q02= CONST MathLit
m442q02= CONST MathLit
m447q01= CONST MathLit
m462q01t= CONST MathLit
m468q01t= CONST MathLit
m474q01= CONST MathLit
m484q01t= CONST MathLit
m496q02= CONST MathLit
m505q01= CONST MathLit
m509q01= CONST MathLit
m510q01t= CONST MathLit
m547q01t= CONST MathLit
m559q01= CONST MathLit
m571q01= CONST MathLit
m704q01t= CONST MathLit

Group USA
Observed Variables:
m124q01 m124q03t m144q03 m155q01 m155q02t m155q04t m420q01t m421q03 m438q02 m442q02
m447q01 m462q01t m468q01t m474q01 m484q01t m496q02 m505q01 m509q01 m510q01t
m547q01t m559q01 m571q01 m704q01t
Means from File usa.ME
Covariance Matrix from File usa.CM
Asymptotic Covariance Matrix from File usa.ACC
Sample Size: 420
Latent Variables: MathLit
Relationships:
MathLit =CONST
set the variance of MathLit free
Method of Estimation: Weighted Least Squares
Path Diagram
End of Problem
```

APPENDIX G2

Logistic Regression Identification of Differential Item Functioning SPSS 10.0 SYNTAX

```
* Michael Jodoin
                                                    *
* Room 160, Hills South
                                                *
* University of Massachusetts
                                                *
* Amherst, MA 01004
                                                *
* April 22, 2001
                                                *
* email: mjodoin@psych.umass.edu
                                                *
* phone: 413-545-1539
                                                *
*This syntax does not work for version prior to SPSS 10.0.  Please
contact me if you need earlier versions.              *
* Look for boxes surrounded by ***** to see where you need to
modify the syntax for your particular application.                *
*                                                                 *
* You will need a datasheet with the following variables:
                                                      *
* Group     (This is the grouping variable e.g., males versus
females)                                              *
* Total       (This is the matching or ability variable e.g., Total
test score).                                                     *
* i1 to i...   (These are the test items e.g.,  i1 through to i50
for items 1 through item 50)                          *
*
* To run this SYNTAX code you need the following files in the
following locations:                                             *
* Place the Script LR_DIF10.sbs in the following path
      C:\Program Files\SPSS\Scripts\LR_DIF10.sbs                 *
* Place the Script DELETE.sbs in the following path
      C:\Program Files\SPSS\Scripts\DELETE.sbs                   *
* Ensure you have the needed "Look file" in the following path
      C:\Program Files\SPSS\Looks\LRDIF.tlo                      *
*You also require a temp folder in the C drive              C:\Temp
                                                    *
*
*The following procedure must also be done before running this
syntax.                                               *

*Open a new Output sheet (.SPO)
                                                    *
*FILE -> EXPORT -> EXPORT OPTIONS (all visible output), and Export
Format  Text (.TXT)     *                               -> OPTIONS
BUTTON  -> Select Produce space separated output,
                *
```

```
*                              -> Cell formatting autofit
*                              -> DO NOT select Export footnotes and
captions, Export Layers or Insert page breaks   *
*A NOTE:  Unfortunately SPSS does not yet support changing these
settings in syntax....when they do I will update this syntax.     *
*Output variable definitions:
                                                *
*  LLM1 Loglikelihood model 1 (e.g., z=bo+b1 theta)
                                                *
*  LLM2 Loglikelihood model 2 (e.g., z=bo+b1 theta+b2 group)
                                                *
*  LLM3 Loglikelihood model 3 (e.g., z=bo+b1 theta+b2 group*theta)
                                                *
*  CSR2M1Cox & Snell R^2 model 1
                                        *
*  CSR2M2Cox & Snell R^2 model 2
                                        *
*  CSR2M3Cox & Snell R^2 model 3
                                        *
*  NR2M1 Nagerlke R^2 model 1
                                                *
*  NR2M2 Nagerlke R^2 model 2
                                                *
*  NR2M3 Nagerlke R^2 model 3
                                                *
*  P_CHI_32 Probability for Chi square value for 1 df test for
uniform DIF                                     *
*  RsqM2 Jodoin & Gierl (2001) R^2 model 2
                                                *
*  RsqM3 Jodoin & Gierl (2001) R^2 model 3
                                                *
*  B0 Beta Weight for constant
                                                *
*  B1 Beta Weight for theta (ability)
                                                *
*  B2 Beta Weight for group         (uniform DIF)
                                                *
*  B3 Beta Weight for ability by group    (Nonuniform DIF)
                                                *
*  CHI_21 Chi square value for 1 df test for uniform DIF
                                                *
*  P_CHI_21 Probability for Chi square value for 1 df test for
uniform DIF                                     *
*  CHI_32 Chi square value for 1 df test for Nonuniform DIF
                                                *
*  P_CHI_32 Probability for Chi square value for 1 df test for
uniform DIF                                     *
*  R2Un  Uniform DIF effect size measure Jodoin & Gierl (2001)
                                                *
*  R2NonUn  NonUniform DIF effect size measure Jodoin & Gierl
(2001)
*  ABC  DIF Classification procedure for LR DIF effect size measure
Jodoin & Gierl (2001)                           *
*****************************************************************
```

```
*****************************************************************
***********
*Saves present settings and then changes setting for datalist
command to come -  Settings are restored later.
PRESERVE.
SET  TLOOK = 'C:\Program Files\SPSS\Looks\LRDIF.tlo'  /BLANKS =
SYSMIS / COMPRESSION = ON
  /DECIMAL = DOT / EPOCH = AUTOMATIC / ERRORS = LISTING /
EXTENSIONS = OFF /FORMAT = F8.2
  /HEADER = BLANK / JOURNAL = ON / LENGTH = NONE / MESSAGES = NONE
/ MEXPAND = ON / MITERATE = 1000
  /MNEST = 50 / MPRINT = OFF / MXCELLS = AUTOMATIC / MXLOOPS = 3000
/ MXMEMORY = 32000
  /MXWARNS = 1000 / TNUMBERS = LABELS / TVARS = LABELS / PRINTBACK
= NONE / RESULTS = LISTING
  /COMPRESSION = ON / SEED = 2000000 /TFIT = LABELS / TNUMBERS =
LABELS / TVARS = LABELS / UNDEFINED = WARN
  / WIDTH = 132 /WORKSPACE = 32000  /CTEMPLATE = NONE .

*****************************************************************
**************
*Place your input file path below.                                *
*Note the single quotation marks and the final period .
     *
* EG  FILE = 'C:\input.sav'.
     *
*****************************************************************
**************

*Loads the inputfile.
GET
  FILE = 'C:\input.sav'.
EXECUTE .

*Center all variables in the analysis via zscores of group and
total.
DESCRIPTIVES
  VARIABLES=group total  /SAVE

Save outfile=temp1.

*The following is a Macro which repeatedly calculates the Chi
Square and R Squared Values for the models.
DEFINE MACNAME (start =  !TOKENS(1) /stop= !TOKENS(1)).
!DO !i = !start !TO !stop   .
Get file=temp1.

*Compute the LR Models to evaluate model fit.
LOGISTIC REGRESSION VAR= !CONCAT(i,!i)
  /METHOD=ENTER ztotal
  /METHOD=ENTER ztotal zgroup
  /METHOD=ENTER ztotal zgroup ztotal*zgroup
  /CRITERIA PIN(.05) POUT(.10) ITERATE(20) CUT(.5)
  /SAVE PRED(!CONCAT(pre,!i)) LRESID (!CONCAT(lre,!i)).
```

```
*Compute variables for R2 test.
AGGREGATE
  /OUTFILE='c:\temp\temp.SAV'
  /BREAK=zgroup ztotal
  /item = SUM(!CONCAT(i,!i)) /!CONCAT(pre,!i) =
MEAN(!CONCAT(pre,!i))
  /Ni=N.

GET
  FILE='c:\temp\temp.SAV'.
EXECUTE.
COMPUTE interact=zgroup*ztotal.
EXECUTE.
COMPUTE !CONCAT(V,!i) = Ni*!CONCAT(pre,!i) *(1 - !CONCAT(pre,!i)) .
EXECUTE.
COMPUTE !CONCAT(Z,!i) = LN(!CONCAT(pre,!i)/(1-!CONCAT(pre,!i)))+
(item-Ni*!CONCAT(pre,!i))/Ni/!CONCAT(pre,!i)/(1-!CONCAT(pre,!i)) .
EXECUTE.
FORMATS !CONCAT(V,!i), !CONCAT(Z,!i), interact (F8.4).
EXECUTE.

Regression
  /Missing listwise
  /Regwgt= !CONCAT(V,!i)
  /Descriptives=corr
  /statistics coeff outs r anova collin tol cha
  /noorigin
  /dependent !CONCAT(Z,!i)
  /method=enter ztotal
  /method=enter zgroup
  /method=enter interact.
execute.

Get file=temp1.

!DOEND .
!ENDDEFINE .

/*****************************************************************
*****************************/
/*Below is the call to the Macro which runs the Logistic Regression
on each item.     */
/*To run a series of items change the start and stop values below.
                  */
/*Again notice the period at the end of the line.
      */
/* EG  MACNAME start = 1 stop = 10 .   *Runs items 1 to 10
inclusive.              */
/*If you need to skip an item for some reason include two lines as
follows.          */
/* EG  MACNAME start = 1 stop = 3 .
           */
/* EG  MACNAME start = 5 stop = 10 .   *Runs items 1 to 3 and 5-10
inclusive.  */
```

```
/*****************************************************************
*****************************/

MACNAME start =  1  stop = 40 .

SCRIPT 'C:\Program Files\SPSS\Scripts\LR_DIF10.sbs'.

DATA LIST
  FILE='C:\temp\temp2.TXT'  RECORDS=158
      /6  itemno 33-39(A)
      /32 LLM1    10-20(3)   CSR2M1    30-36(3)  NR2M1    53-58(3)
     /69  LLM2    10-20(3)   CSR2M2    30-36(3)   NR2M2    53-58(3)
     /108 LLM3   10-20(3)   CSR2M3    30-36(3)   NR2M3    53-58(3)
     /134 B1       30-34(3)   /136 B2      30-34(3)   /138 B3
30-34(3)  /140 B0       30-34(3)
     /151 rsqm1     19-23(3)  /153  rsqm2    19-23(3)  /155  rsqm3
19-23(3) .
EXECUTE.

*Calculates Chisquares and rsquared for uniform and nonuniform DIF.
COMPUTE CHI21 = LLM1 - LLM2 .
COMPUTE CHI32 = LLM2 - LLM3 .
EXECUTE.
COMPUTE P_CHI21 = 1 - CDF.CHISQ(CHI21,1) .
COMPUTE P_CHI32 = 1 - CDF.CHISQ(CHI32,1) .
EXECUTE.
COMPUTE R2UN = RSQm2 - RSQm1 .
COMPUTE R2NONUN = RSQm3 - RSQm2 .
EXECUTE .

IF ((r2un < ABS( 0.035)) OR (P_chi21 > 0.05) OR (r2nonun <ABS(
0.035)) OR (P_chi32 > 0.05) ) dif = 1 .
EXECUTE .
IF ((r2un < ABS(0.07)  &  r2un >=  ABS(0.035) & P_CHI21 < 0.05) OR
(r2nonun < ABS(0.07)  &  r2nonun >=  ABS(0.035) & P_CHI32 < 0.05))
dif = 2 .
EXECUTE.
IF ((r2un >= ABS( 0.07) & P_CHI21 < 0.05) OR (r2nonun >= ABS( 0.07)
& P_CHI32 < 0.05)) dif = 3 .
EXECUTE.

STRING ABC (A1) .
RECODE
  dif
  (1='A')  (2='B')  (3='C')  INTO  ABC .
EXECUTE .

/*****************************************************************
***************/
/*Place your output file path  below.
*/
/*Note the single quotation marks and the final period .
       */
/* EG  SAVE OUTFILE = 'C:\output.sav'.
            */
```

```
/*****************************************************************
***************/

SAVE OUTFILE='C:\output1.sav'
   /KEEP = LLM1 to LLM3, CSR2M1 to CSR2M3, NR2M1 to NR2M3,
                 rsqm1 to rsqm3, CHI21, P_CHI21, CHI32, P_CHI32,
R2UN, R2NONUN, dif, ABC, b0, b1, b2, b3 .
EXECUTE.
*Restores SPSS default settings.
RESTORE.
```

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Çet, Selda
Nationality: Turkish (TC)
Date and Place of Birth: 7 March 1975 , Ankara
Marital Status: Single
Phone: +90 312 287 73 81
email: selda_cet@hotmail.com

EDUCATION

Degree	Institution	Year of Graduation
MS	Marmara University Graduate School of Applied and Natural Sciences Mathematics Education	2000
BS	Marmara University Faculty of Education Mathenatics Education	1997

WORK EXPERIENCE

Year	Place	Enrollment
2002- Present	METU Foundation Schools Ankara	Mathematics Teacher
2000-2002	FMV Private Ayazağa Işık Primary School Istanbul	Mathematics Teacher
1999- 2000	ISTEK Private Belde High School Istanbul	Mathematics Teacher
1998- 1999	Yakacık Vocational School Istanbul	Mathematics Teacher

FOREIGN LANGUAGES

English