THE ANALYSES OF SECONDARY EDUCATION INSTITUTIONS STUDENT SELECTION AND PLACEMENT TEST'S VERBAL SECTION WITH RESPECT TO ITEM RESPONSE THEORY MODELS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF SOCIAL SCIENCES
OF
THE MIDDLE EAST TECHNICAL UNIVERSITY

BY

SEDA CAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
DEGREE OF MASTER OF SCIENCE
IN
THE DEPARTMENT OF EDUCATIONAL SCIENCES

JANUARY 2003

Approval of the Graduate School Science of Social Sciences

Prof. Dr. Bahattin Akşit
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science

Prof. Dr. Hasan Şimşek
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 Master of Science.

Prof. Dr. Giray Berberoğlu
Supervisor

Examining Committee Members

Prof. Dr. Giray Berberoğlu

Prof. Dr. Füsun Akkök

Prof. Dr. Ömer Geban

ABSTRACT

THE ANALYSES OF SECONDARY EDUCATION INSTITUTIONS STUDENT SELECTION AND PLACEMENT TEST'S VERBAL SECTION WITH RESPECT TO ITEM RESPONSE THEORY MODELS

Can, Seda

M.S., Department of Educational Sciences

Supervisor: Prof. Dr. Giray Berberoğlu

January 2003, 93 pages

The purpose of this study is to analyze the Verbal Section (Turkish and Social Sciences subtests) of MONE- Secondary Education Institutions Student Selection and Placement Exam by using Item Response Theory. The results of the MONE- Secondary Education Institutions Student Selection and Placement Exam, which was administered in 2001, were used as the data in the study. The sample included 5585 examinees composed of 2693 (48.2%) male and 2892 (51.8%) female examinees. First of all, whether the assumptions of IRT models were met, was investigated. Then, the item parameter estimates across different groups of examinees and ability parameter estimates across different sets of items were compared in each subtest in order to examine the invariance of item and ability parameter estimates. Lastly, by using chi square statistics the observed and theoretical distribution on each subtest were reviewed. The results of the study indicated that, equality of item discrimination assumption of the one-parameter model was not met by the subtests. The Turkish subtest was a non-speeded test. The two-parameter model item discrimination estimates for high ability vs. low ability group were invariant in the Turkish subtest. The discrimination parameter estimates in the high ability vs. low ability groups and pseudo-chance level parameter estimates for the three-parameter model across two different groups were not invariant in the Turkish subtest. In the Social Sciences subtest, the item difficulty parameter estimates of both one-, two- and three-parameter models were highly invariant. Moreover, the three-parameter pseudo-chance level parameter estimates of Social Sciences subtest were invariant for the two different groups in contrast to Turkish subtest. The three-parameter ability parameter estimates were slightly more invariant than the one- and two-parameter models both in Turkish and Social Sciences Subtests. Chi square statistics results indicated that the fit of the one-parameter model to the Secondary Education Institutions Student Selection and Placement Exam was better than the two- and three-parameter models.

Keywords: Item Response Theory, one-, two- and three-parameter models

ÖZ

ORTA ÖĞRETİM KURUMLARI ÖĞRENCİ SEÇME VE YERLEŞTİRME SINAVI SÖZEL BÖLÜMÜNÜN MADDE TEPKİ KURAMI MODELLERİNE GÖRE ANALİZİ

Can, Seda

Yüksek Lisans, Eğitim Bilimleri Bölümü

Tez Yöneticisi: Prof. Dr. Giray Berberoğlu

Ocak 2003, 93 sayfa

Bu çalışmanın amacı, Madde Tepki Kuramı kullanılarak Milli Eğitim Bakanlığı Orta Öğretim Kurumları Öğrenci Seçme ve Yerleştirme Sınavı Sözel Bölümünün (Türkçe ve Sosyal Bilgiler) analiz edilmesidir. 2001 yılı Milli Eğitim Bakanlığı Orta Öğretim Kurumları Öğrenci Seçme ve Yerleştirme Sınav sonuçları veri olarak kullanılmıştır. Örneklem 2693 (% 48.2) erkek ve 2892 (%51.8) kız öğrenciden oluşmaktadır. İlk olarak, MTK modellerinin sayıltılarının karşılanıp karşılanmadığı araştırılmıştır. Sonra, testi alan farklı gruplardan elde edilen madde parametreleri ile farklı madde gruplarından elde edilen yetenek parametreleri her bir alt test için karşılaştırılmıştır. Son olarak, khi-kare istatistiği kullanılarak, gözlenen ve kuramsal dağılım her bir alt testte incelenmiştir. Çalışmanın sonuçları, bir parametreli modelin eşit madde ayırıcılık indeksleri sayıltısının alt testler tarafından karşılanmadığını göstermiştir. Türkçe alt testi hız testi olarak görünmemektedir. Türkçe alt testinde iki parametreli modelin madde ayırıcılık parametreleri yüksek-düşük yetenek grubunda değişmez özelliktedir. Türkçe alt testinde üç parametreli modelde yüksek-düşük yetenek grubunda madde ayırıcılık parametreleri ve pseudo şans parametreleri iki ayrı grupta da değişmez özellikte değildir. Sosyal Bilimler alt testinde, madde zorluk parametreleri bir, iki ve üç parametreli modellerin hepsinde yüksek değişmezlik göstermiştir. Ayrıca, üç parametreli modelin pseudo şans parametre ölçüleri de Türkçe alt testinden farklı olarak değişmezlik göstermiştir. Türkçe ve Sosyal Bilgiler alt testlerinde üç parametreli modelin yetenek ölçüleri, bir ve iki parametreli modellerden daha değişmezdir. Khi-kare istatistiği sonuçları, bir parametreli modelin Orta Öğretim Kurumları Öğrenci Seçme ve Yerleştirme Sınavına uyumunun iki ve üç parametreli modellere gore daha iyi olduğunu göstermektedir.

Anahtar kelimeler: Madde Tepki Kuramı, bir, iki, üç parametreli modeller

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Prof. Dr. Giray Berberoğlu for his trust, support and valuable suggestions throughout this study. I am also grateful to Prof. Dr. Füsun Akkök and Prof. Dr. Ömer Geban for their suggestions and comments.

I offer thanks to my family especially my sister Sevgi for their faith and I thank them for being there when I need them.

To my friends, I offer thanks to all of you because of your encouragement and support.

To My Father and Mother

TABLE OF CONTENTS

LIST OF TABLES

TABLE

LIST OF FIGURES

FIGURE

CHAPTER 1

INTRODUCTION

Tests have an important role in the lives of people throughout the world. Most of everyone has taken one or more tests in his or her life such as, for attending high school or university or applying for a job especially in recent years.

Various kinds of tests are used in educational, psychological, business or military situations in the field of testing, which has grown rapidly during the twentieth century (Aiken, 1997).

According to Aiken (1997), the main purpose of the testing is to evaluate behavior, cognitive abilities, personality traits and other individual and group characteristics in order to assist in making judgments, predictions and decisions about people. One of these decisions is to classify and place people in an educational context.

In Turkey, there are some institutions, which are responsible of selecting and placing the students in schools. Ministry of National Education (MONE), which has department named Evaluation and Assessment Center studies on public and private selection and placement test in transition to secondary education, state boarding and scholarship tests at all levels, open primary education school, open education high school term exams, personnel allocation and promotion tests of all public institutions and agencies, selection and assessment tests of managers of MONE at all levels and appointment and transfer operations of teachers are realized. Shortly, Evaluation and Assessment Center of MONE, is responsible of placing the students in high schools such as Anatolian, Natural Sciences or Private high schools by preparing the examinations of the schools in Turkey.

This department basically used classical test theory techniques in constructing the tests and in the data analysis. However, there is another theory called Item Response Theory, which is defined as an alternative theory of Classical Test Theory (CTT).

Classical Test Theory (CTT) and Item Response Theory (IRT) are basic two measurement models for constructing tests and interpreting the test scores. Classical test theory and its concepts have been used for a long time in testing, however; classical measurement models and procedures have provided less ideal solutions to many testing problems when compared to Item Response Theory. (Hambleton, et. al., 1991)

There are some shortcomings of classical measurement model, which make researchers to study on Item Response Theory. According to Hambleton, Swaminathan and Rogers (1991) one of these shortcomings and the most important one is that examinee characteristics and test characteristics can not be separated, each can be interpreted only in the context of the other. The examinee characteristic mentioned is, "ability" measured by the test. In CTT, the true score expresses ability; which is explained as "the expected value of observed performance on the test" (Hambleton, et. al., 1991). An examinee's ability is obtained only by the means of a particular test. When the test is "hard", the examinee will appear to have low ability; when the test is "easy" the examinee will appear to have higher ability. What is meant by "hard" and "easy" tests can be found in the definition of "difficulty of a test item" which is defined as the "the proportion of examinees in a group of interest who answer the item correctly", in the CTT. An items' being hard or easy depends on the ability of examinees and the ability of examinees being measured depends on the items' being hard or easy.

The other concepts of CTT such as item discrimination, test score reliability and validity are also defined in terms of a particular group of examinees. Test and item characteristics change as the examinee context changes and examinee characteristics change as the test context changes. For this reason it is very difficult to compare examinees that take different tests and very difficult to compare items whose characteristics obtained using different groups of examinees.

The two shortcomings of classical test theory come from the definition of reliability and its concept known as standard error of measurement. In CTT, test reliability is defined in terms of parallel forms. Reliability is known as the correlation between test scores on parallel forms of a test however to achieve the concept of parallel measures is difficult because individuals may not be exactly the same on a second administration of a test. They may forget things, they may develop new skills or their level of a motivation or anxiety may change (Hambleton and van der Linden, 1982). Hence, the comparison of examinees on the nonparallel test score scales is a shortcoming in CTT. The problem with the standard error of measurement is that it is assumed to be equal for all examinees. However, the errors of measurement on difficult test are not equal for low and high ability examinees.

The last limitation of CTT is its being test oriented rather than item oriented, which does not enable us to make prediction about how an examinee or a group of examinees will perform on a given item. As its name indicates, IRT primarily focuses on the item-level information in contrast to the CTT's focus on test-level information (Fan, 1998). The estimation of the probability that an examinee will answer a particular question correctly is of considerable value when adapting a test to match the examinee's ability level. Such information is necessary if a test designer desires to predict test score characteristics in one or more populations of examinees or to design tests with particular characteristics for certain populations of examinees.

All these limitations, made researchers to investigate alternative theories of measurement, which will include;

- Item characteristics that are not group- dependent
- Scores describing examinee proficiency that are not test- dependent
- A model which is expressed at the item level rather than at the test level
- A model that does not require strictly parallel tests for assessing reliability
- A model that provides a measure of precision for each ability score

Hambleton, Swaminathan and Rogers (1991) not only mentioned about these desirable features that should be in the alternative test theory, they also stated that these features can be obtained within the framework of an alternative test theory known as Item Response Theory.

All these advantages of IRT over CTT in testing made IRT being studied more in recent years. For this reason, this study aimed, by the use of IRT, to analyze the data obtained from the administration of Secondary Education Institutions Student Selection and Placement Exam, which takes an important role in determining the high school that the students continue their education life.

1.1. Statement of the Purpose

The main purpose of this study is to analyze the Verbal Section (Turkish and Social Sciences subtests) of MONE- Secondary Education Institutions Student Selection and Placement Exam by using Item Response Theory. The other purposes which underlies in the main purpose are:

- To investigate whether assumptions of the IRT models are met with the MONE-Secondary Education Institutions Student Selection and Placement Exam's Verbal Section (Turkish and Social Sciences subtests) data
- To compare item parameter estimates obtained in two or more subgroups of the population for whom the test is intended (for example; high and low test performers, males and females)

- To compare ability parameter estimates for different samples of test items (for example; hard and easy test items)
- To compare observed and theoretical distribution on Turkish and Social Sciences subtests

1.2. Statement of the Main and Subproblems

1.2.1. Does the Turkish subtest of the MONE- Secondary Education Institutions Student Selection and Placement Exam meet the assumptions of IRT?

1.2.1.1. Do the Turkish subtest of the MONE- Secondary Education Institutions Student Selection and Placement Exam data meet the unidimensionality assumption?

1.2.1.2. Is the Turkish subtest of the MONE- Secondary Education Institutions Student Selection and Placement Exam a non-speeded test?

1.2.1.3. Are the items locally independent in the Turkish subtest of the MONE- Secondary Education Institutions Student Selection and Placement Exam?

1.2.1.4. Does the MONE- Secondary Education Institutions Student Selection and Placement Exam Turkish subtest meet the equality of item discrimination indices assumption of one-parameter model?

1.2.1.5. Does the MONE- Secondary Education Institutions Student Selection and Placement Exam Turkish subtest meet minimal guessing of one- and two-parameter models?

1.2.2. Are the item parameters as estimated by one of the IRT models invariant across different samples of examinees?

1.2.2.1. Are the item parameters as estimated by one of the IRT models invariant across male vs. female groups in the Turkish subtest of the MONE-Secondary Education Institutions Student Selection and Placement Exam?

1.2.2.2. Are the item parameters as estimated by one of the IRT models invariant across high vs. low-test performers in the Turkish subtest of the MONE-Secondary Education Institutions Student Selection and Placement Exam?

1.2.3. Are the ability parameter estimates invariant across different sets of the items in the Turkish subtest of the MONE- Secondary Education Institutions Student Selection and Placement Exam?

1.2.3.1.Are the ability parameter estimates invariant across hard vs. easy items in the Turkish subtest of the MONE- Secondary Education Institutions Student Selection and Placement Exam?

1.2.3.2. Are the ability parameter estimates invariant across even vs. odd items in the Turkish subtest of the MONE- Secondary Education Institutions Student Selection and Placement Exam?

1.2.4. Is the observed distribution on the Turkish subtest of the MONE-Secondary Education Institutions Student Selection and Placement Exam fits theoretical distribution in one, two and three parameter models?

1.2.5. Do the Social Sciences subtest of the MONE- Secondary Education Institutions Student Selection and Placement Exam meet the assumptions of IRT?

1.2.5.1. Do the Social Sciences subtest of the MONE- Secondary Education Institutions Student Selection and Placement Exam data meet the unidimensionality assumption?

1.2.5.2. Is the Social Sciences subtest of the MONE- Secondary Education Institutions Student Selection and Placement Exam a non-speeded test?

1.2.5.3. Are the items locally independent in the Social Sciences subtest of the MONE- Secondary Education Institutions Student Selection and Placement Exam?

1.2.5.4. Does the MONE- Secondary Education Institutions Student Selection and Placement Exam Social Sciences subtest meet the equality of item discrimination indices assumption of one-parameter model?

1.2.5.5. Does the MONE- Secondary Education Institutions Student Selection and Placement Exam Social Sciences subtest meet minimal guessing of one- and two-parameter models?

1.2.6. Are the item parameters as estimated by one of the IRT models invariant across different samples of examinees?

1.2.6.1. Are the item parameters as estimated by one of the IRT models invariant across male vs. female groups in the Social Sciences subtest of the MONE- Secondary Education Institutions Student Selection and Placement Exam?

1.2.6.2. Are the item parameters as estimated by one of the IRT models invariant across high vs. low-test performers in the Social Sciences subtest of the MONE- Secondary Education Institutions Student Selection and Placement Exam?

1.2.7. Are the ability parameter estimates invariant across different sets of the items in the Social Sciences subtest of the MONE- Secondary Education Institutions Student Selection and Placement Exam?

1.2.7.1. Are the ability parameter estimates invariant across hard vs. easy items in the Social Sciences subtest of the MONE- Secondary Education Institutions Student Selection and Placement Exam?

1.2.7.2. Are the ability parameter estimates invariant across even vs. odd items in the Social Sciences subtest of the MONE- Secondary Education Institutions Student Selection and Placement Exam?

1.2.8. Is the observed distribution on the Social Sciences subtest of the MONE- Secondary Education Institutions Student Selection and Placement Exam fits theoretical distribution in one, two and three parameter models?

1.3. Significance of the study

In the model-data fit studies, the advantages of IRT models can be obtained if there is a satisfactory fit between the model and the test data. By the analysis of the Secondary Education Institutions Student Selection and Placement Exam Turkish and Social Sciences subtests data, more detailed information on the test items can be obtained. Thus, having such information for each item will help to produce more good items in the testing procedure.

Additionally, in the near future, most of the testing programs will be organized in the computer environment. To use computers will tailor test difficulty with the ability of the examinees. This is only possible by the use of Item Response Theory Models. This study may be effective in adaptive testing program.

1.4. Definition of terms

Item Response Theory: A sophisticated approach to item analysis using mathematical models to predict the probability of a correct response to an item based on the learner's ability (as indicated by performance on the test) and certain characteristics of the item (Mc Daniel, 1994, p. 346)

Item Characteristic Curve (ICC): An ICC plots the probability of responding correctly to an item as a function of the latent trait (denoted by θ) underlying performance on the items on the test (Crocker and Algina, 1986, p. 340)

CHAPTER 2

REVIEW OF LITERATURE

This chapter contains both the theoretical background of the IRT and the research studies of IRT in social sciences.

2.1. Theoretical Background

The Item Response Theory has attracted considerable interest from measurement specialists in spite of its mathematical complexity, the lack of convenient and efficient computer programs to analyze the data. In 1970's and 1980's measurement specialists had made more study through applications of IRT, such as test score equating adaptive testing, differential item functioning analysis.

Today, Item response Theories being used to construct both norm-referenced and criterion-referenced tests, to investigate item bias, to equate the test and to report test score information (Hambleton and Swaninathan, 1985).

2.1.1. Basic IRT Concepts

Item Response Theory (IRT) consists of a series of models for describing and explaining examinees' response behavior to educational and psychological test items (Mellenbergh, 1994). An item response model specifies a relationship between the observable examinee test performance and the unobservable traits or abilities assumed to underlie performance on the test (Hambleton and Swaninathan, 1985). The relationship between the “observable” and the “unobservable” quantities is described by a mathematical function. For this reason, item response models are mathematical models, which are based on specific assumptions about the test data.

In the framework of IRT, many possible IRT models exist, differing in the mathematical form of the item characteristics function and/or the number of parameters specified in the model. IRT bases on two basic postulates:

a) The performance of an examinee on a test item can be predicted (or explained) by a set of factors, which are called trades, latent traits or abilities;

b) The relationship between examinees’ item performance and the set of traits underlying item performance can be described by a monotically increasing function which is called item characteristic function or Item Characteristic Curve (ICC) (Hambleton et. al., 1991).

Latent trait is symbolized by θ and it refers to a statistical construct. The latent trait is generally called the ability measured by the test in the cognitive tests. The total score is generally taken as initial estimate of that ability (Anastasi and Urbina, 1997). The Item Characteristic Curve (ICC) is S-shaped curve and shows the probability of giving correct answer to an item as a function of latent trait. This function specifies that as the level of trait increases, the probability of a correct response to an item increases (Hambleton et. al., 1991). In other words, examinee with more ability has higher probabilities for giving correct answers to items than examinee with lower ability.

As mentioned before, many possible IRT models exist and the applicability of each model in a particular situation depends on the nature of the test items and the viability of different theoretical assumptions about the test items (Fan, 1998). There are three IRT models for test items that are dichotomously scored.

The parameters of the IRT that usually describe the ICC are referred to in many applications as the "b, a and c" parameters of the items, which refer to difficulty, discrimination and lower asymptote parameters respectively.

When a test item is described in terms of all three parameters, a three-parameter item response model is used. The IRT three-parameter model takes the following form;

$$P_i(\theta) = c_i + (1 - c_i)\left\{e^{Da_i(\theta - b_i)} / (1 + e^{Da_i(\theta - b_i)})\right\},$$

where c is called pseudochance level parameter which represents the "probability of examinees with low ability answering the item correctly" (Hambleton, et. al., 1991,p.17). a is item discrimination parameter commonly known as item slope, b is the item difficulty parameter, D is a scaling factor which makes the logistic function as close as possible to the normal ogive function (normally D=1.7), e is a transcendental number (like Π) whose value is 2.718, θ is the ability level of a particular examinee and P(θ) is an S-shaped curve with values between 0 and 1 over the ability scale.

If it is assumed that the c parameter is assumed to be 0, and only a and b parameters are used to describe the ICC, the three-parameter model is reduced to two-parameter model and the formula takes the following form;

$$P_i(\theta) = e^{Da_i(\theta - b_i)} / (1 + e^{Da_i(\theta - b_i)}),$$

When the items are described only in terms of their item difficulties, assuming that all the discriminations are the same (equal) and the c parameters are 0, a one-parameter (Rasch) model is being used with the following formula,

$$P_i(\theta) = e^{(\theta - b_i)} / (1 + e^{(\theta - b_i)}).$$

2.1.2. Assumptions of IRT

The most important difference between the popular unidimensional IRT models is in the number of parameters used to describe items. The choice of model depends on the user but this choice involves assumptions about the data that can be verified later by examining how well the model explains the observed results (Hambleton, et. al. 1991). In data-fit studies before the decision of best fitted model, these assumptions should be reviewed.

Two assumptions, which are common for all these models are that the data are unidimensional and the test administration was not speeded. Unidimensionality assumption implies that only one ability is measured by a set of items. This assumption common for all IRT models but it can not be strictly met because several traits such as level of motivation, test anxiety etc. always affect the test performance of the examinee. The other assumption, which relates to unidimensionality is local independence. Local independence means that the probability of a correct response of an examinee to an item is not affected by responses to other items in the test. In other words, items are uncorrelated for individuals having the same ability level.

There are also specific assumptions of IRT models. An additional assumption of the two-parameter model is minimal guessing. According to this assumption; probability of giving correct responses to items for low ability examinees has to be minimal. Another assumption for the one-parameter model is that all item discrimination indices are equal.

Furthermore, ability and item invariance are important in Item Response Theory, which implies that the item parameters do not depend on the ability of the examinees and the ability parameters do not depend on the set of items administered to the examinees. By the ability and item invariance concepts of IRT, person or sample free estimates of item parameters and item free ability estimates can be obtained in IRT models which make the examinees comparable who took different items and item parameters would not change for different ability groups.

2.2. Research Studies Regarding IRT

This part of the chapter presents the review of research about fit studies of IRT models and different applications of IRT in testing in social sciences.

The development of IRT and its adoption in major testing programs represent significant advances in testing area (Thissen, 1987), which made researchers to study IRT in many different topics. Assessing model-data fit is an important aspect of these IRT applications in which the fit of the model to the data is used to ensure that the appropriate IRT model is selected.

The researchers (Hambleton, 1983) studied IRT for solving a number of testing problems in the Maryland Functional Reading Program (MFRP). The purpose of the study was to investigate the fit of one-, two-, and three-parameter models to the test results obtained from the administration of the 1982 Maryland Functional Reading Test (MFRT). The results showed that two-parameter model was able to account for examinee performance on the MFRT adequately. The one-parameter model could not get the substantial variation among the discriminating power of test items. The data fit the three-parameter model slightly because of the minimum amount of guessing on the test.

In the study of Raju and Goldman (1986), the researchers aimed to determine the appropriateness of one- and two-parameter models to a well-known attitude survey and to assess the effect of sample size on the estimation of item and person parameters. A group of 3000 individuals was the sample of the study. The results showed that for item discrimination indices in two-parameter model, 1000 subjects were required for accurate estimation as a sample, and for item difficulty parameters, a sample of 250 subjects in the one-parameter model, 500 subjects in the two-parameter model found to be adequate for estimation. In addition to these findings, two-parameter model fitted Attitude Survey data better than one-parameter model.

The model-data fit studies of IRT could be found in Turkey, too. A study was conducted by Kılıç (1999) to investigate the fit of the one-, two- and three-parameter models of IRT to the Student Selection Test (SST). The results of 1993 SST were used as data. The sample included 2121 examinees selected from seven different state high schools in Ankara. The results showed the fit of the three-parameter model to the SST was better than the other models.

The fit of one-, two- and three- parameter models of IRT to Education Research and Development Directorate's (ERDD) achievement test data was studied by Yalçın in 1999. The study aimed to carry out the statistics related to fit of one-, two- and three- parameter models of IRT to ERDD's Turkish language tests. In the study, 1997 ERDD's examination results, which were obtained from the administration of achievement test to a total number of 7287 (2816 grade level five, 4431 grade level eight) students from 192 primary schools in two provinces of Turkey. The results showed that four items in grade level five, and five items in grade level eight did not fit the three-parameter model. In spite of these misfits, the results of the study showed that three-parameter model worked better for the data.

A model-data fit research was conducted by Choi (1989) to explore the appropriateness of IRT in language testing. The study investigated the dimensionality of the reading and vocabulary sections of two widely used language proficiency tests, which are The University of Cambridge First Certificate of English (FCE), The Test of English as a Foreign Language (TOEFL) and the fit of data to the one-, two-, and three-parameter models of IRT. In investigation of dimensionality it was found that the TOEFL reading subtest was multidimensional and the FCE reading and vocabulary subtests were found to be unidimensional. According to the results, Rasch model clearly failed to provide adequate fit for these data and the three-parameter model fit better than the two-parameter model.

There can be found data-fit studies, which focused on usage of different statistical methods. Fanmin's study (1997) was an example for such kind of studies. The purpose of the study was to adapt the z-statistic in a different way that it can be used for studying model-data fit for short tests. The research focused on two-parameter item response model for dichotomously scored items. The strategy was by adapting z-test of proportions to create a new ability-interval forming strategy and to minimize the inaccurate ability estimation to use multiple regression. The a-parameter, b-parameter, test length and sample size were manipulated in a 3x3x3x2 completely crossed design. The results showed that, the new fit statistic, together with the interval-forming strategy and the regression equations, could be used by researchers to assess model fit when a two-parameter model is applied to short tests with dichotomous responses.

IRT is also being used in the development of the tests. Paek and Holland (1999) conducted one of these studies for the development and preliminary analysis of a mathematical test named as Mathematical Aptitude Test (SEMAT) targeted for high mathematical ability elementary school students who attended the Stanford Education Program for Gifted Youth (EPGY). The test was administered to 9-11 years old 248 students. Item Response Theory determined proficiency estimates, which were then used as scores to predict various outcomes in EPGY. The data obtained by the administration of the SEMAT was fitted to the one-parameter (Rasch) model and the results showed that the SEMAT proved to be a strong predictor for gifted youth.

Gumpel et. al. (1998) used IRT model to examine the 28-item Conners Teacher's Rating Scale (CTRS) for the diagnosis of attention-deficit/hyperactivity disorder (ADHD). The sample was 453 pairs of respondents consisting of parents and teachers. This study attempted to use IRT to reexamine the CTRS's psychometric properties and to understand how respondents using the scale view the ADHD. The results showed that using traditional true score measurement models, the CTRS was found to be consistent and reliable scale. Through the use of IRT procedures, it was shown that CTRS did not meet some basic assumptions of the one-parameter model.

The Item Response Theory not only used in achievement tests but also used in personality questionnaires in social sciences. The study conducted by Gernot (1982) dealt with the problem of applying IRT to personality questionnaires. The one-, two- and three-parameter models were studied in two different personality test data. It was concluded that the two-parameter model was the most efficient way of applying the IRT to these two personality test data.

Ferrando (1994) conducted another study in the application of IRT to the area of personality in order to investigate if the item response theory models could provide a good fit to empirical data. The data used in the analysis were gathered from the administration of EPI-A Impulsivity subscale to 2483 subjects (1531 males, 952 females), who ranged in age from 18 to 30 years. The results of the study indicated that two-parameter model fit the data reasonably well.

The study conducted for the fit of model to the data was studied in the area of job performance in the psychology by Armstrong (1990). The purpose of the study was to explore the application of the two-parameter model to a judgmental measure of work performance. The data used in the study were gathered from 2764 sales representatives who work for a large transportation company. Model-data fit and differential item performance were examined in the study. The results revealed that the two-parameter model was successfully applied to a measure of work performance. The data met the assumptions of the model and an acceptable degree of model-data fit was obtained. This study found no evidence of differential item performance due to the sex and race.

Another evaluation study of a scale by using IRT in psychology was conducted by Young et. al. in 1992. The purpose of the study was to apply item response models to two sets of data, which were collected from the administration of the Beck Hopelessness Scale (HS) as a measure of hopelessness. One and two-parameter item response models were studied by the researchers. Results using IRT models indicated that it measures a simple dimension of hopelessness. In the one-parameter model, all scale items showed an acceptably strong relationship to the latent variable of hopelessness when compared the two-parameter model.

Reliability and validity studies, which were conducted by using IRT, could also been found in the literature. The study conducted by Chow and Winzer (1992) aimed to provide information about the reliability and validity of a scale designed to measure attitudes toward exceptional children and mainstreaming. The 25-item scale measuring teachers' attitudes toward mainstreaming was analyzed using Item Response Theory models. A total of 917 teachers participated in the study. The results showed that three-parameter model fit the data well. After the IRT and reliability analysis five items were extracted from the scale.

The study that focused on reliability and validity of the State-Trait Anxiety Inventory for Children (STAIC) was studied by Kirişci and Duncan (1996). The sample was composed of 675, aged from 12 to 18 year old clinical and nonclinical adolescents. The results indicated that three-parameter model fit the data better than the two-parameter model. The item analysis showed that all of the items of the STAIC were highly discriminating. Scores from both confirmatory factor analysis and IRT analysis revealed that the STAIC was applicable to adolescents to measure the state-trait anxiety.

A research was conducted by Ludlow and Guida (1991) in order to investigate the structure of the Test Anxiety Scale for children (TASC) as an instrument to measure academic anxiety. The one-parameter (Rasch) model was applied to the data to provide evidence that TASC may be understood as a continuum, which defines an academic anxiety construct comprised of commonly related items. The sample consisted of 455 seventh and eight-grade students. The fit analysis of Rasch model indicated that the TASC total score provided a valid estimate of student level of anxiety.

Another research was conducted with the aim of constructing the Utrecht Early Mathematical Competence Scales to assess the developmental level of early mathematical competence in children ages 4 to 7 years (Van de Rijt et. al., 1999). An initial pool of 120 items was developed concerning eight mathematically different domains. Three research questions were presented in the study; 1- does the final pool of 120 items represent a unidimensional continuum?, 2- if this question is answered positively, can two scales be extracted of 40 items for future research and practical use? And 3- are individual differences in scale performance a function of age?. The items were administered to 823 boys and girls in the 4 to 7- year age groups. The test statistic regarding the goodness of fit of the model for the whole set of items had a value which together with the results of the factor analysis and correlations allowed the researchers to conclude that the items measure one dimension called early mathematical competence. Secondly the results showed that the set of 80 items can be seen as an item bank for which the responses have been shown to give a good fit with the one-parameter item response model. At last the researchers indicated that for the answer of 3rd question of this study was need of more research on this topic.

When the literature is reviewed, there can be seen studies focused on the differences between IRT and CTT. Humbert's study (1986) was one of these studies. The purpose of the study was to examine the differences between IRT and CTT as applied to occupational licensing examination by comparing the two methods of item analysis with respect to 1- differences in difficulty and discrimination indices, 2- differences in the examinees' raw test score and their latent ability estimates; 3- implications for training and personnel policy. Data were taken from the 900 examinees to whom an occupational licensing test was administered. The results showed that the estimates were not inconsistent with each other, though IRT estimates provided a wider range of information.

An empirical comparison study of IRT and CTT conducted by Fan (1998) yielded different results when compared to Humbert's study. The study focused on two issues; 1- what are the empirical relationships between IRT and CTT based item and person statistics? And 2- to what extent are the item statistics from IRT and those from CTT invariant across different participant samples?. The data used in the study were gathered from the administration of the Texas Assessment of Academic Skills (TAAS) to 11th-grade students. The test item pool was composed of two tests (mathematics and reading) with 60 and 48 dichotomously scored items, and the participant pool had more than 193,000 examinees who took the both tests. The IRT model fit results showed that the data fit the two- and three-parameter IRT models exceptionally well. Additionally; both in terms of the comparability of item and person statistics and in terms of the degree of invariance of item statistics, the findings showed that the two measurement frameworks IRT and CTT produced similar item and person statistics. As the limitations of the research, Fan mentioned about limited item pool and test items' being easy. Fan expressed that the findings of this study could be interpreted as interesting about how to view differences both in theory and in practice testing between IRT and CTT models because of the similarity of the results obtained from CTT and IRT models.

As mentioned before in this part of the chapter to present different applications of IRT was aimed. Another aspect of IRT usage in social sciences is its being studied in the equivalence of measurement provided by the scales among different ethnic, race or sex samples. In the research of Hui et. al. (1983), the equivalence of measurement provided by Overall Modernity (OM) Scale for American Hispanics and Mainstream Americans was examined. 211 Hispanics and 221 Mainstream navy recruits participated in this study. The fit of the two-parameter model was studied by the researchers. After the model-data fit, they tried to examine the properties of the OM scale when applied in a cross-cultural setting and to compare the modernity of the Mainstream recruits with that of the Hispanic recruits. The results indicated that most of the items in the OM Scale were demonstrated to be cross-culturally equivalent. Comparison of the two groups on the modernity, estimated by IRT methods showed that Mainstream population is not different from the Hispanic population.

As it was seen, in the first part of this chapter theoretical background of the IRT was presented concerning basic IRT concepts and assumptions of item response models.

In the second part of the chapter, research studies regarding IRT were presented. These research not only about model data-fit studies but also different applications of IRT in testing area such as; the use of IRT in test development, evaluation of the scales, reliability and validity studies, the comparison between CTT and IRT and usage of IRT in cross-cultural setting.

To conclude, IRT is being studied in a wide range of testing area in recent years. This chapter aimed to present a summary of these different applications in addition to model data-fit studies.

CHAPTER 3

METHOD OF THE STUDY

This chapter presents the methodological procedures of the study, which include the research design, the subjects, the instrument and the data analysis topics. The first section is related to the research design. In the subject section, the subjects who participated in the study explained in detail. In the instrument part the Turkish and social sciences subtests and their properties were presented. Finally, the preliminary analysis, checking model assumptions, checking expected model features, which are invariance of item and ability parameter estimates and Chi square statistics were presented as the Data Analysis.

3.1. Overall research design

This study is an empirical study on the item statistics obtained from the Classical Test Theory (CTT) and Item Response Theory (IRT) measurement framework. The study focused on the fit of different parameter models of IRT to Turkish and Social Sciences achievement tests data.

3.2. Sample

The data were gathered from the examinees of MONE-Secondary Education Institutions Student Selection and Placement Exam, which was administered in 2001 by selecting randomly. The sample included 5585 examinees composed of 2693 (48.2%) male and 2892 (51.8%) female examinees.

3.3. Data Collection Procedures

In this study, model data-fit analysis was carried out in the Turkish and Social Sciences subtests of 2001 Secondary Education Institutions Student Selection and Placement Exam. Turkish subtest items are related to proficiency in Turkish Language and Social Sciences subtest items are related to ability to reason using Social Sciences concepts and generalizations. Each test consists of 25 multiple-choice items with four alternatives.

3.4. The Data Analysis Procedure

3.4.1. Preliminary Analysis

Descriptive statistics of the data was obtained by using SPSS 11.0 for Windows statistical package program and ITEMAN (Assessment System Corporation, 1993) computer program was used for the item analysis.

3.4.2. Checking Model Assumptions

In order to check the assumption of unidimensionality, a principal component analysis was carried out.

Percentages of examinees completing the last 6 items of the MONE-Secondary Education Institutions Student Selection and Placement Exam data were checked for the non-speeded test administration assumption for each subtest of the exam.

For local independence, the inter-item correlation matrices for low ability groups were obtained in Turkish and Social Sciences subtests.

The assumption of the one-parameter model, which is equal discrimination indices, was checked by reviewing the distribution of biserial correlations obtained from item analysis.

The performance of low- ability students on the most difficult items was checked to review whether the minimal guessing assumption met the data.

3.4.3. Checking Expected Model Features

To obtain the item and ability parameter estimates the BILOG computer program was used. BILOG fits the one, two and three-parameter models using marginal maximum likelihood procedures with optional Bayesian procedures. In this estimation procedure the ability parameters are integrated out and the item parameters are estimated. With the item parameter estimates determined, the ability parameters are estimated (Mislevy and Bock, 1984).

One of the model features of invariance of the item parameter estimates was checked by correlating and forming the scatter plots between high vs. low ability groups and male vs. female groups. Ability estimates of different samples or test items such as, easy-hard and even-odd sets of items, were correlated and scatter plots were formed to check the invariance of ability parameter estimates in the Social Sciences and Turkish subtests data of the Secondary Education Institutions Student Selection and Placement Exam.

3.4.4. Checking Model Predictions of Actual and Simulated Test Results

Chi square statistics of the Social Sciences subtest data of the Secondary Education Institutions Student Selection and Placement Exam obtained by the BILOG program were used to determine the fit of models to the data.

CHAPTER 4

RESULTS

This chapter presents the results of the data analysis procedure which includes preliminary analysis, checking model assumptions, checking expected model features, invariance of item parameter estimates, invariance of ability parameter estimates and checking model predictions of actual and simulated test results topics for Turkish and Social Sciences subtests separately.

4.1. Turkish Subtest

4.1.1. Preliminary Analysis

The descriptive statistics of the Turkish subtest are presented in Table 4.1.1. mean score is 13.524 and no one completely gave correct answer to the Turkish subtest. Skewness and kurtosis results reveal that the distribution is negatively skewed and the kurtosis value is -0.681 (Figure 4.1.1.). The item discrimination "r" and the item difficulty parameters "p" of each item are presented in Table 4.1.2. The difficulty indices range from 0.231 to 0.915. Mean difficulty is 0.541 which shows the Turkish subtest is moderately difficult for the examinees. Mean discrimination of the test is 0.521. The items of the Turkish subtest could be considered as moderately discriminating.

Table 4.1.1. *Test Statistics of the Turkish Subtest*

N of items	25
N of examinees	5585
Mean	13.524
Variance	20.941
Std. dev.	4.576
Skewness	-0.011
Kurtosis	-0.681
Minimum	0.000
Maximum	24.000
Median	13.000
Alpha	0.775
Mean difficulty (p)	0.541
Mean discrimination (r)	0.521

Table 4.1.2. *Classical Item Parameters of the Turkish Subtest*

Item No	Difficulty (p)	Discrimination (r)
1	0.443	0.383
2	0.915	0.446
3	0.751	0.625
4	0.250	0.540
5	0.328	0.487
6	0.467	0.599
7	0.690	0.480
8	0.549	0.383
9	0.647	0.446
10	0.612	0.625
11	0.394	0.540
12	0.380	0.487
13	0.745	0.599
14	0.636	0.480
15	0.231	0.383
16	0.449	0.220
17	0.419	0.640
18	0.542	0.621
19	0.787	0.511
20	0.537	0.494
21	0.712	0.672
22	0.413	0.598
23	0.774	0.646
24	0.242	0.390
25	0.613	0.665

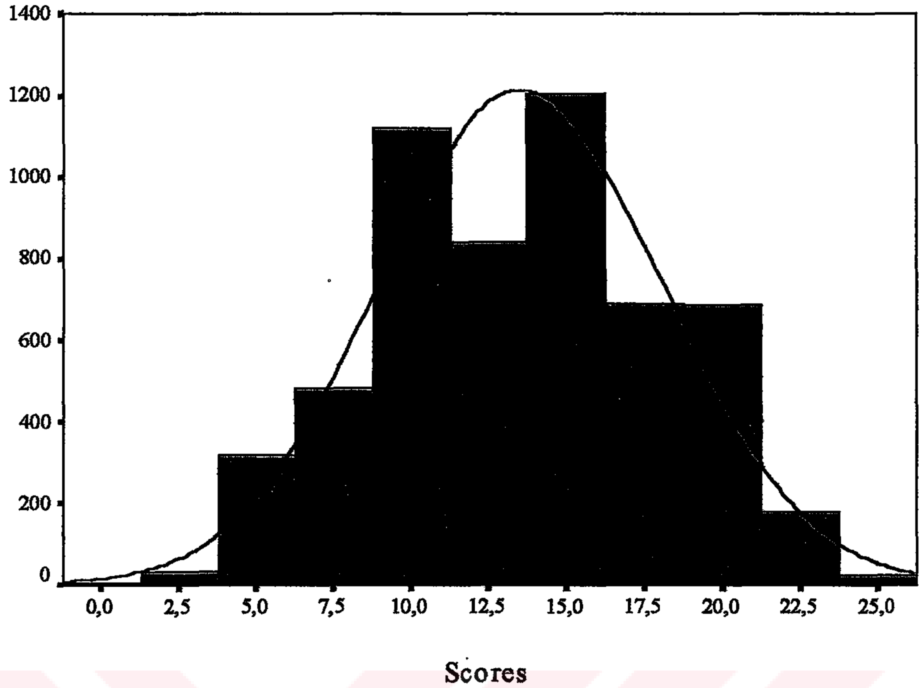

Figure 4.1.1. Frequency Distribution of the Turkish Subtest Scores

4.1.2. Checking Model Assumptions

Principle component analysis was performed in order to see whether the unidimensionality assumption was met in the Turkish Subtest. The results of principal component analysis (Appendix A1) indicated that there are four interpretable factors, which accounted for 30.405% of the total variance. Eigenvalues of these factors are 4.324 (17.296 of the variance), 1.249 (4.995% of the variance), 1.028 (4.114 of the variance) and 1.00 (4.00 of the variance) respectively. As seen in the scree plot of the eigenvalues (Figure 4.1.2.), there is a sharp decrease from first eigenvalue to second, which shows the Turkish subtest could be considered as a unidimensional scale.

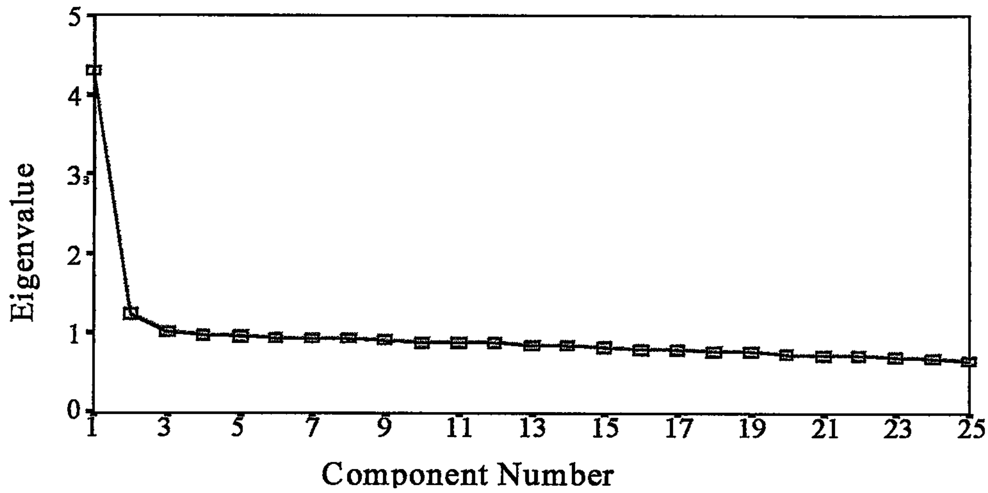

Figure 4.1.2. Plot of Eigenvalues

The non-speeded test administration assumption was investigated by checking the percentages of examinees completing last six items. It was observed that nearly 48% of examinees did not complete these six items. This result can indicate that there is a speeded test administration process for the Turkish Subtest. However, when the percentages reviewed carefully the 24. item was the reason of this result which is one of the most difficult items of the subtest. For this reason it was concluded that there was a non-speeded test administration process for the Turkish Subtest.

To investigate the local independence assumption in the Turkish subtest the inter item correlation matrix of low ability examinees was obtained. The entries in the off-diagonal elements of the matrices approaching to zero (Appendix B1) showed that the items are locally independent in the Turkish subtest.

In order to investigate the equality of item discrimination indices assumption of the one-parameter model, the frequency distribution of the item discrimination indices was used. The item discrimination indices range from 0,220 to 0,672. As seen in Figure 4.1.3., the distribution is not homogenous. Thus, the equality of item discrimination indices assumption of the one-parameter model was not met by the Turkish subtest.

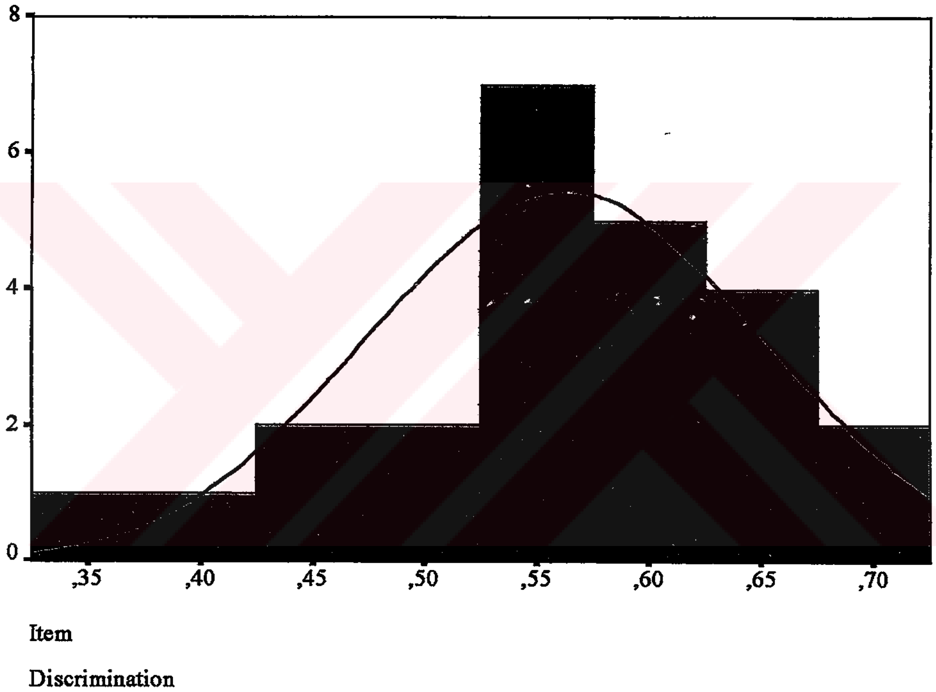

Figure 4.1.3. Frequency Distribution of the Discrimination Indices of the Turkish Items

The minimal guessing assumption of the one- and two-parameter models was investigated by selecting 5 difficult items, taking into account the results of the item analysis results. The performance of the low ability examinees on the five difficult which are 4., 11., 12., 15., and 24. items was reviewed and it was observed that the mean of the frequencies of the examinees who did not give any correct response to these items was approximately 82% which shows the low performance of low ability examinees on these items. Therefore, the minimal guessing assumption was viable for the Turkish subtest data.

Table 4.1.2. *Difficult Item Scores of Low Ability Examinees*

Scores	Frequencies (%)				
	Item 4	Item 11	Item 12	Item 15	Item 24
Missing	7.6	15.6	23.0	6.0	26.7
0 (wrong answer)	75.3	62.5	63.1	80.4	57.5
1 (correct answer)	17.0	21.9	23.0	13.6	15.8
Total	100.0	100.0	100.0	100.0	100.0

4.1.3. Checking Expected Model Features

4.1.3.1. Invariance of Item Parameter Estimates

To investigate the invariance of item parameter estimates of the IRT models, the item parameter estimates across male vs.. female and high ability group vs. low ability group were correlated (Table 4.1.3.1.) in the Turkish subtest and scatterplots were formed. Figures 4.1.3.1. and 4.1.3.2. show scatterplots of item difficulty parameter estimates for the one-parameter model. Figures from 4.1.3.3. to 4.1.3.6. show the plots of the item difficulty and discrimination parameter estimates for the two-parameter model and the figures from 4.1.3.7. to 4.1.3.12 are the scatterplots of the item difficulty, discrimination and guessing parameter estimates for the three-parameter model in the Turkish subtest.

Table 4.1.3.1. *Correlation of Item Parameter Estimates Across Different Samples of Examinees Obtained in Two Groups*

Samples	Bilog		
	One Parameter	Two Parameter	Three Parameter
Male-Female (-b-)	0.282*	0.942**	0.441**
High Ability-Low Ability (-b-)	0.914**	0.874**	0.869**
Male-Female (-a-)		0.816**	0.358*
High Ability-Low Ability (-a-)		0.031	0.005
Male-Female (-c-)			0.212
High Ability-Low Ability (-c-)			0.261

** p<.01
* p<.05

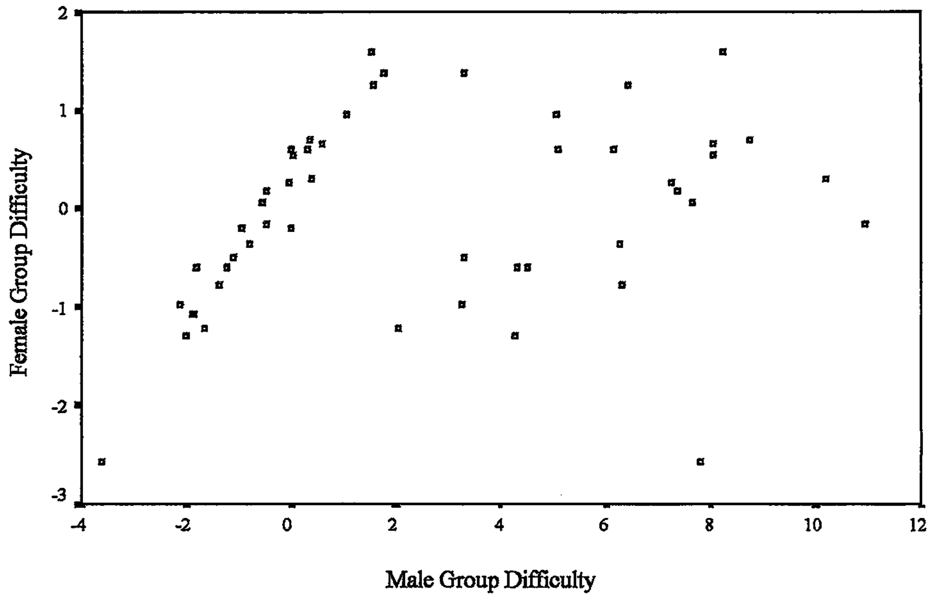

Figure 4.1.3.1. Plot of 1P Item Difficulty Values Based on Female and Male Groups of Examinees

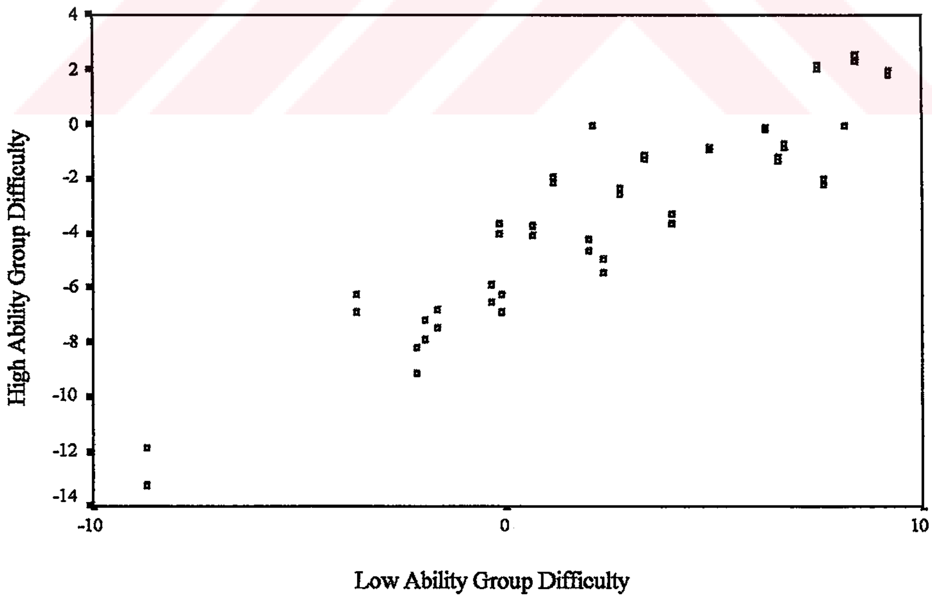

Figure 4.1.3.2. Plot of 1P Item Difficulty Values Based on High and Low Ability Groups of Examinees

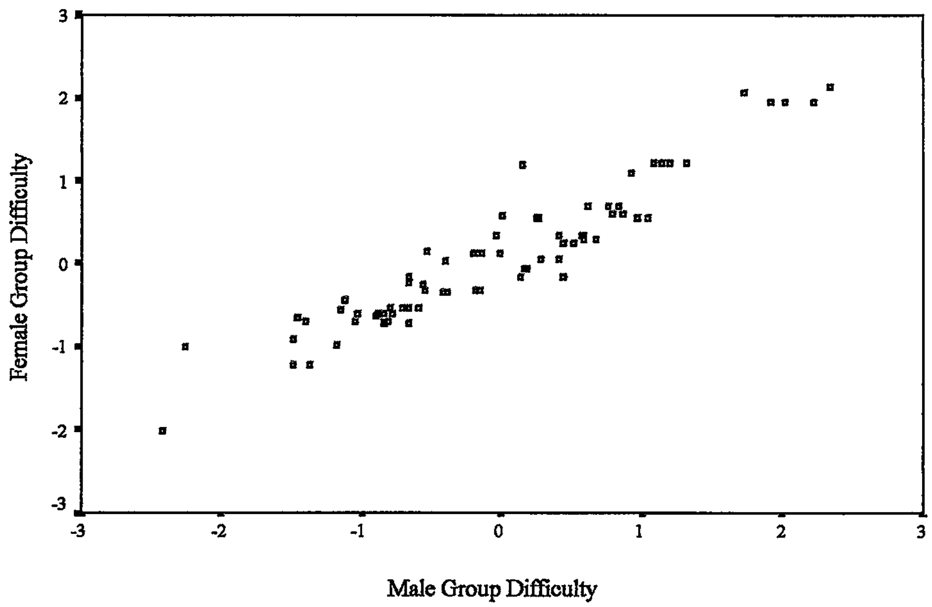

Figure 4.1.3.3. Plot of 2P Item Difficulty Values Based on Female and Male Groups of Examinees

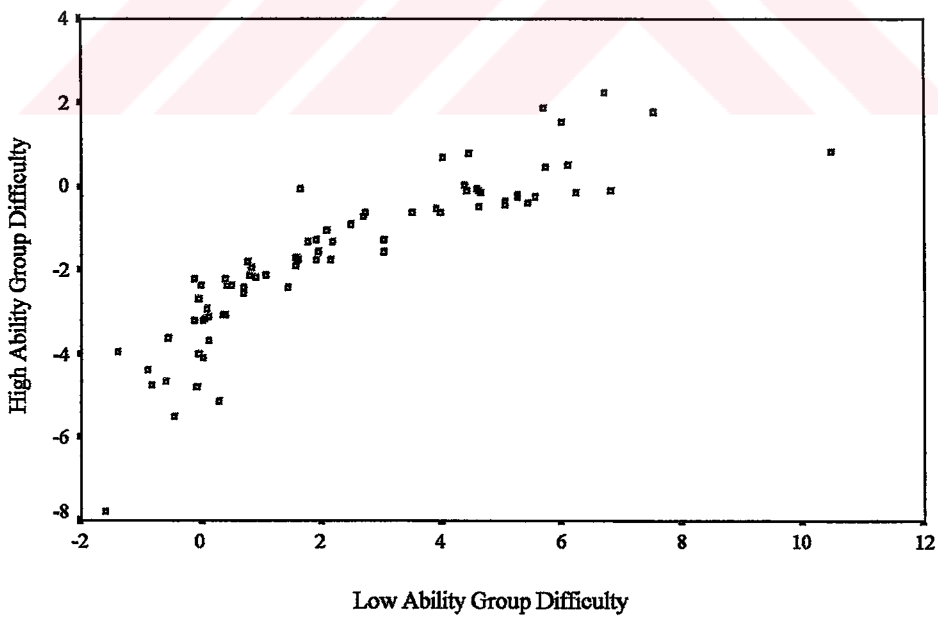

Figure 4.1.3.4. Plot of 2P Item Difficulty Values Based on High and Low Ability Groups of Examinees

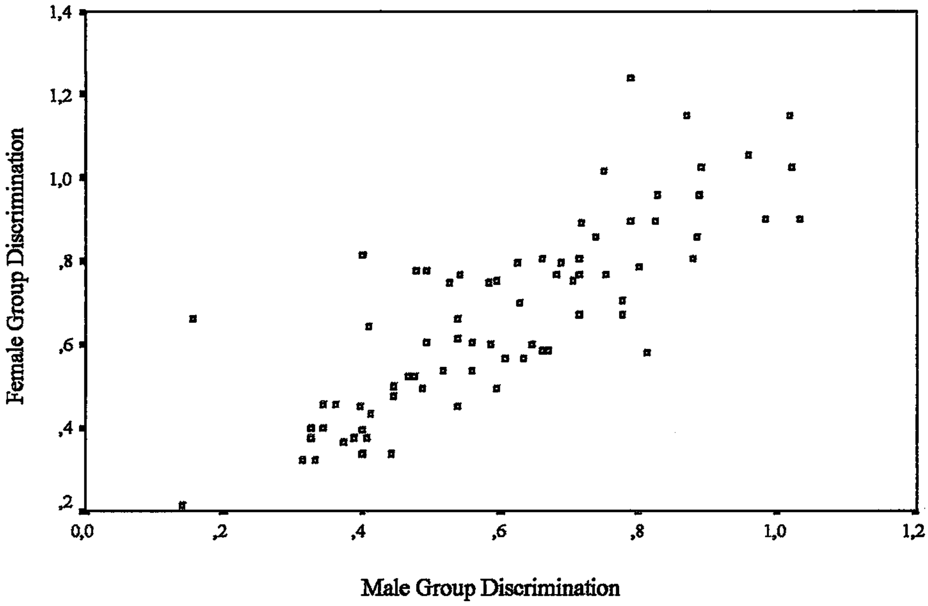

Figure 4.1.3.5. Plot of 2P Item Discrimination Values Based on Male and Female Groups of Examinees

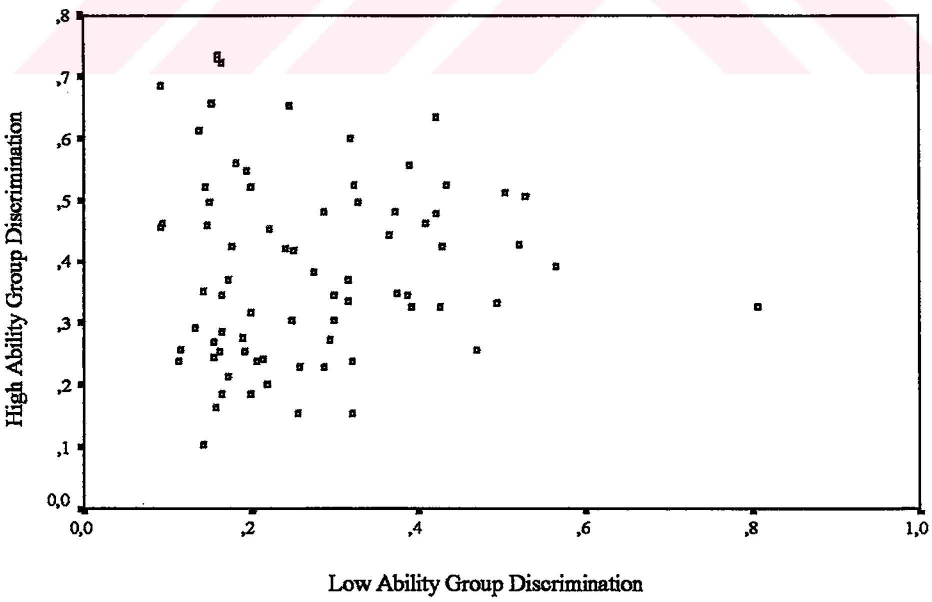

Figure 4.1.3.6. Plot of 2P Item Discrimination Values Based on High and Low Ability Groups of Examinees

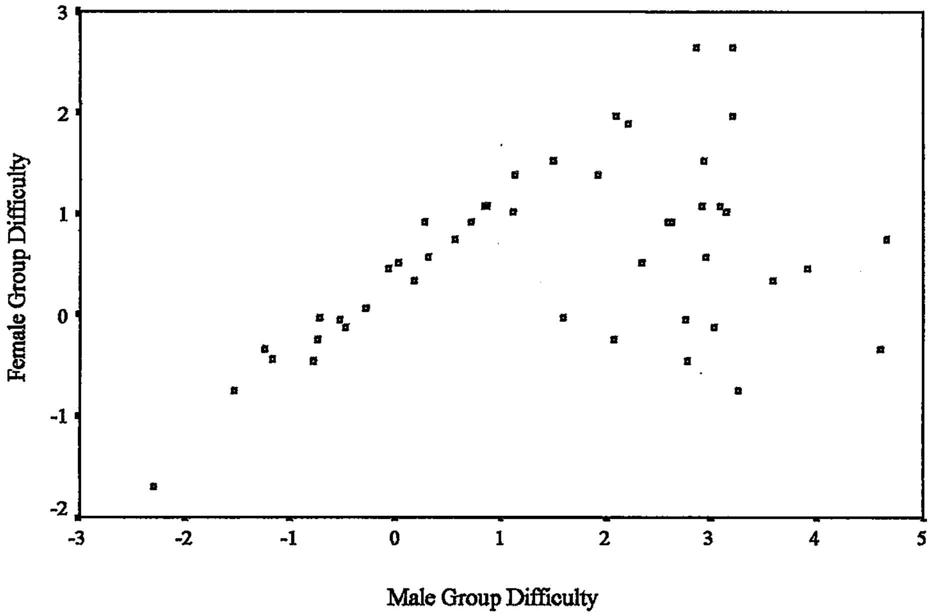

Figure 4.1.3.7. Plot of 3P Item Difficulty Values Based on Female and Male Groups of Examinees

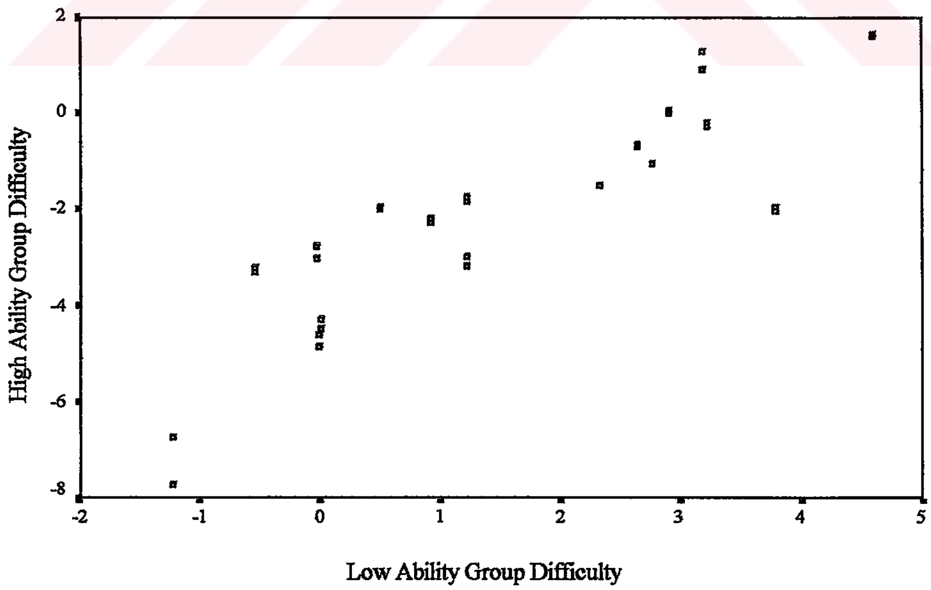

Figure 4.1.3.8. Plot of 3P Item Difficulty Values Based on High and Low Ability Groups of Examinees

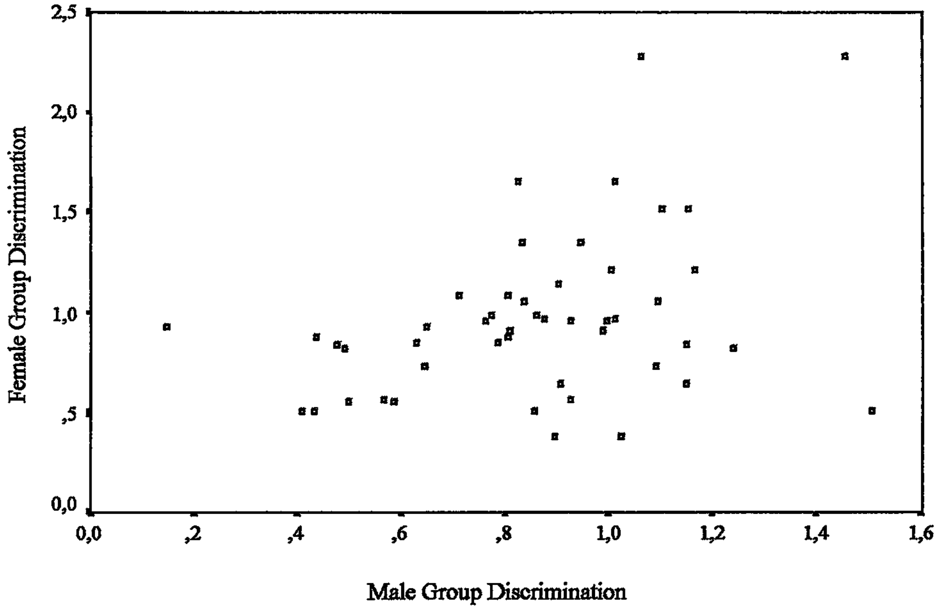

Figure 4.3.1.9. Plot of 3P Item Discrimination Values Based on Male and Female Groups of Examinees

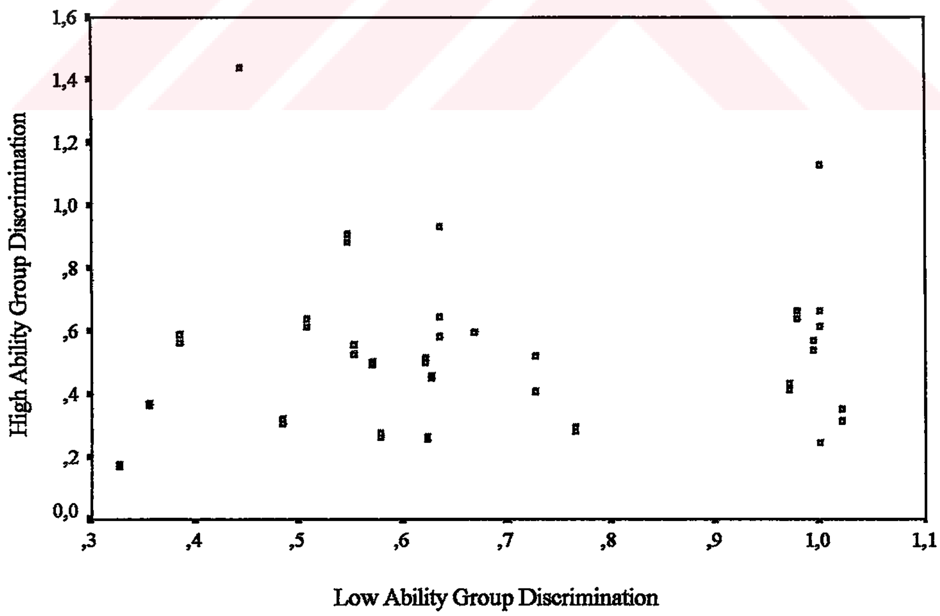

Figure 4.1.3.10. Plot of 3P Item Discrimination Values Based on High and Low Ability Groups of Examinees

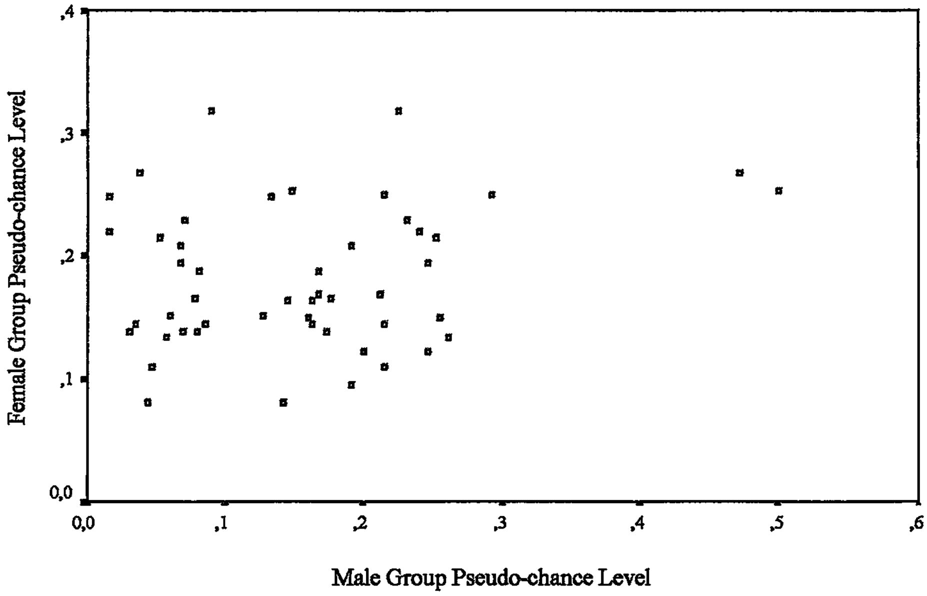

Figure 4.1.3.11. Plot of 3P Item Pseudo-Chance Level Based on Male and Female Groups of Examinees

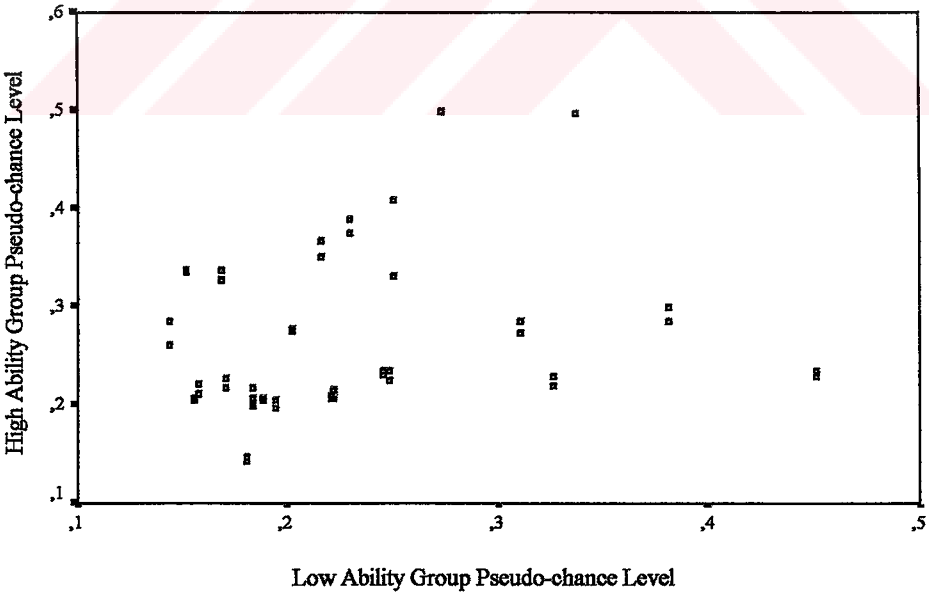

Figure 4.1.3.12. Plot of 3P Item Pseudo-Chance Level Based on High and Low Ability Groups of Examinees

4.1.3.2. Invariance of Ability Parameter Estimates

Table 4.1.3.2. shows the correlations of ability parameter estimates across different sets of items such as; easy vs. difficult and even vs. odd. According to the results, for each IRT model, the ability parameter estimates seems moderately invariant in the Turkish subtest. But, the invariance of ability parameter estimates of the three-parameter model is slightly higher than the one and two parameter models.

Table 4.1.3.2. *Correlations of Ability Parameter Estimates Across Different Sets of Items*

Subtests	Bilog		
	One Parameter	Two Parameter	Three Parameter
Easy- Difficult	0.606**	0.638**	0.654**
Even- Odd	0.647**	0.676**	0.680**

** p<.01

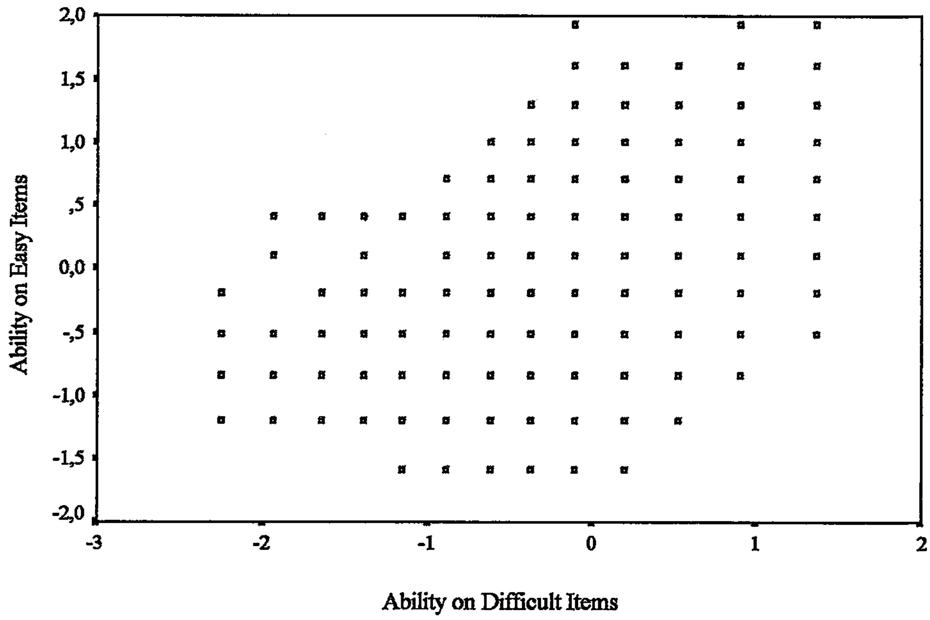

Figure 4.1.3.13. Plot of 1P Ability Estimates Based on Easy and Difficult Items

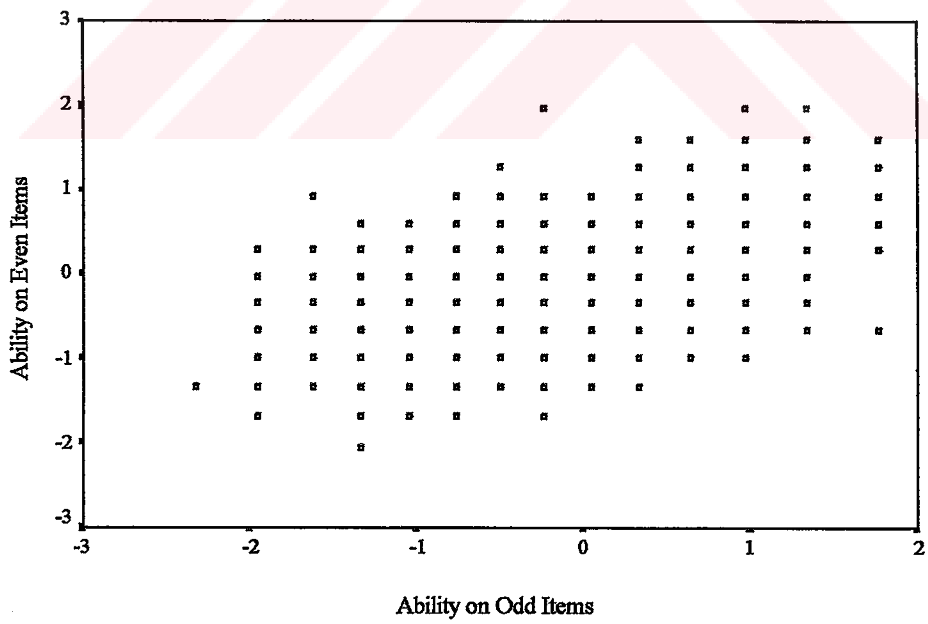

Figure 4.1.3.14. Plot of 1P Ability Estimates Based on Even and Odd Items

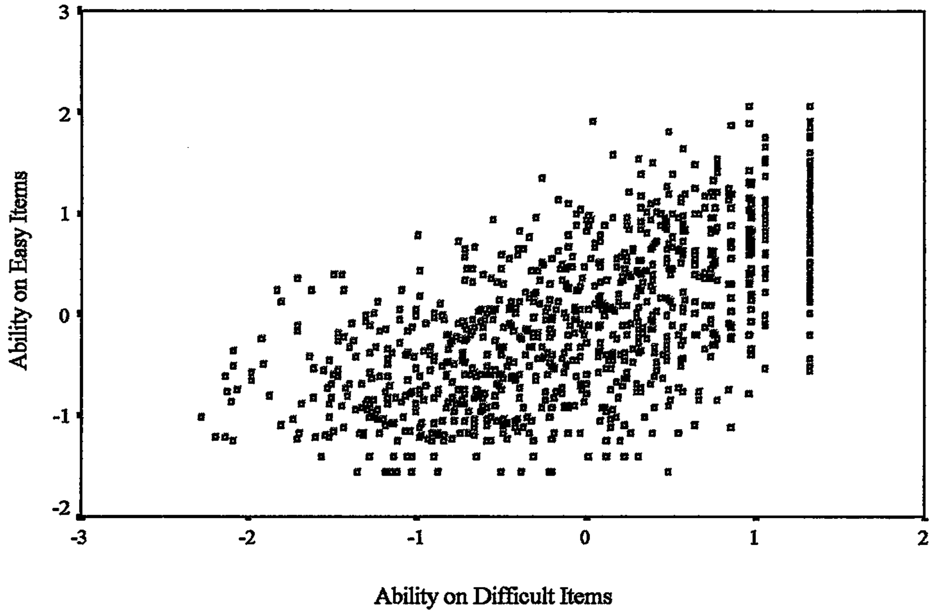

Figure 4.1.3.15. Plot of 2P Ability Estimates Based on Easy and Difficult Items

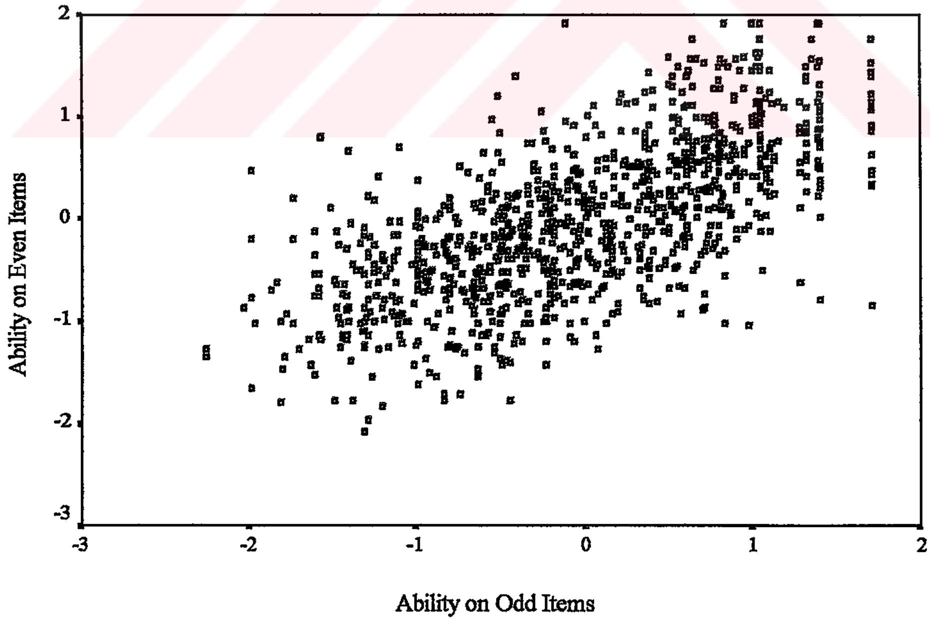

Figure 4.1.3.16. Plot of 2P Ability Estimates Based on Even and Odd Items

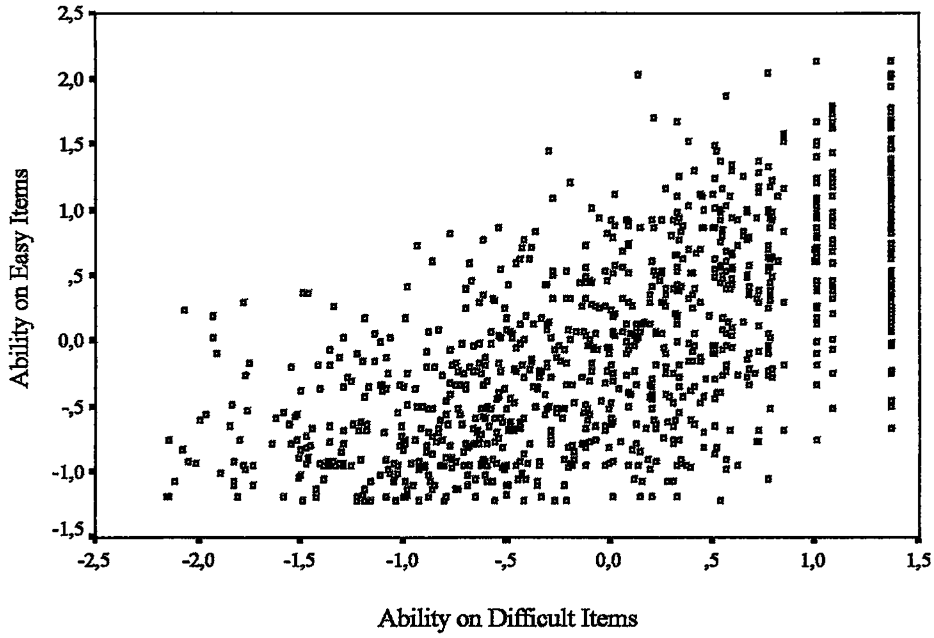

Figure 4.1.3.17. Plot of 3P Ability Estimates Based on Easy and Difficult Items

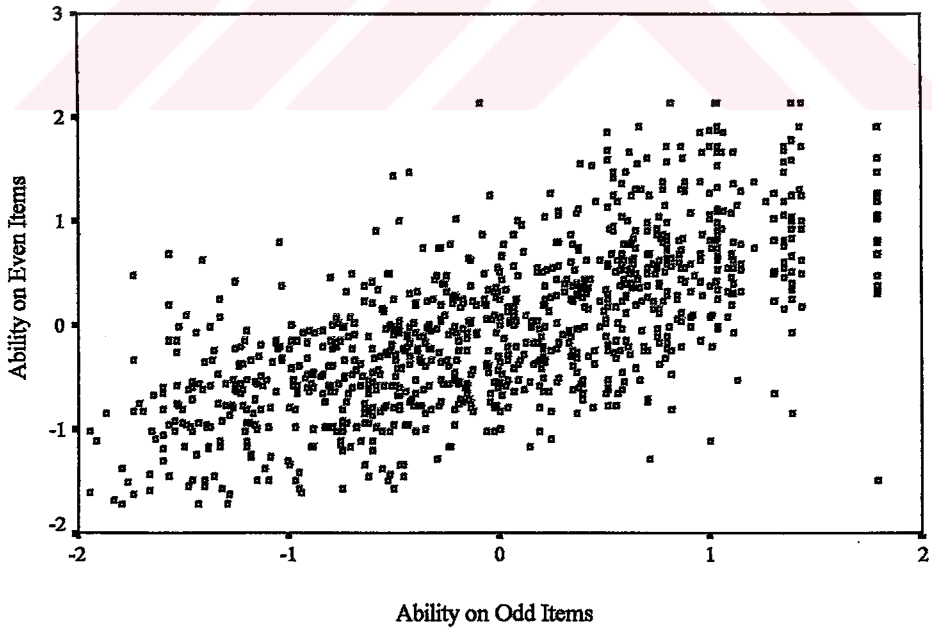

Figure 4.1.3.18. Plot of 3P Ability Estimates Based on Even and Odd Items

4.1.4. Checking Model Predictions of Actual and Simulated Test Results

Chi square statistics was used in order to see how well the observed distribution fits the theoretical distribution. Table 4.1.4. shows the number of misfit items of the IRT models. Tables 4.1.5, 4.1.6. and 4.1.7 present the item parameter estimates for the one, two and three-parameter models. In these tables, bold items indicate misfit items of the IRT models at 0.05 significance level.

Table 4.1.4. *Total Number of Misfit Items For The One, Two and The Three Parameter Models*

Models	N of items	N of misfit items
One parameter	25	12
Two parameter	25	22
Three parameter	25	23

* p< .05

Table 4.1.5. *Item Parameters of the One-Parameter Model*

Item number	One parameter model b- value
1.	0.109
2.	-2.980
3.	-1.436
4.	1.438
5.	0.947
6.	0.304
7.	-1.107
8.	-0.260
9.	-0.800
10.	-0.621
11.	0.465
12.	0.606
13.	-1.303
14.	-0.715
15.	1.425
16.	0.240
17.	0.272
18.	-0.329
19.	-1.755
20.	-0.228
21.	-1.284
22.	0.246
23.	-1.620
24.	1.466
25.	-0.761

Table 4.1.6. *Item Parameters of the Two-Parameter Model*

Item number	Two parameter model	
	b- value	a- value
1.	0.109	0.469
2.	-2.056	0.892
3.	-1.030	0.883
4.	1.733	0.412
5.	0.917	0.540
6.	0.377	0.388
7.	-0.795	0.894
8.	-0.386	0.329
9.	-0.855	0.486
10.	-0.515	0.711
11.	0.410	0.600
12.	0.636	0.483
13.	-1.044	0.729
14.	-0.799	0.460
15.	1.696	0.418
16.	0.654	0.165
17.	0.188	0.824
18.	-0.287	0.676
19.	-1.760	0.523
20.	-0.257	0.469
21.	-0.892	0.946
22.	0.199	0.651
23.	-1.177	0.854
24.	1.977	0.361
25.	-0.556	0.879

Table 4.1.7. *Item Parameters of the Three-Parameter Model*

Item number	Three parameter model		
	b- value	a- value	c- value
1.	0.551	0.561	0.143
2.	-1.930	0.883	0.180
3.	-0.766	0.962	0.161
4.	1.711	0.884	0.136
5.	1.159	0.701	0.100
6.	1.043	0.623	0.214
7.	-0.321	1.199	0.245
8.	0.511	0.424	0.223
9.	-0.412	0.546	0.160
10.	0.172	1.213	0.293
11.	0.755	0.851	0.143
12.	1.013	0.673	0.142
13.	-0.801	0.773	0.134
14.	-0.201	0.545	0.200
15.	1.764	0.721	0.117
16.	3.021	0.697	0.423
17.	0.420	1.048	0.103
18.	0.254	1.055	0.227
19.	-1.176	0.580	0.249
20.	0.223	0.556	0.159
21.	-0.473	1.228	0.229
22.	0.574	0.940	0.158
23.	-0.953	0.883	0.155
24.	1.442	2.159	0.164
25.	-0.178	1.125	0.188

4.2. Social Sciences Subtest

4.2.1. Preliminary Analysis

The descriptive statistics of the Social Sciences subtest are presented in Table 4.4.1. Skewness and kurtosis results indicate that the distribution is positively skewed (Figure 4.4.1.).

Table 4.2.1. *Test Statistics of the Social Sciences Subtest*

N of items	25
N of examinees	5585
Mean	12.744
Variance	27.285
Std. dev.	5.223
Skewness	0.248
Kurtosis	-0.648
Minimum	0.000
Maximum	25.000
Median	12.000
Alpha	0.824
Mean difficulty (p)	0.510
Mean discrimination (r)	0.562

The item difficulty "p" and the discrimination "r" indices were obtained by the use of classical test theory techniques for each items of the Social Sciences subtest (Table 4.2.2.). The item difficulty indices range from 0.234 to 0.809. Mean difficulty is 0.510, which shows the Social Sciences subtest is moderately difficult for the examinees. The Social Sciences subtest could be considered as moderately discriminating because of the value of the mean discrimination, which is 0.562.

Table 4.2.2. *Classical Item Parameters of the Social Sciences Subtest*

Item no	Difficulty (p)	Discrimination (r)
1	0.474	0.432
2	0.633	0.490
3	0.663	0.601
4	0.809	0.594
5	0.636	0.646
6	0.352	0.526
7	0.583	0.617
8	0.327	0.420
9	0.660	0.667
10	0.560	0.701
11	0.367	0.636
12	0.412	0.571
13	0.316	0.644
14	0.401	0.572
15	0.367	0.528
16	0.502	0.566
17	0.234	0.362
18	0.489	0.501
19	0.585	0.608
20	0.332	0.426
21	0.695	0.679
22	0.668	0.558
23	0.490	0.546
24	0.507	0.554
25	0.680	0.606

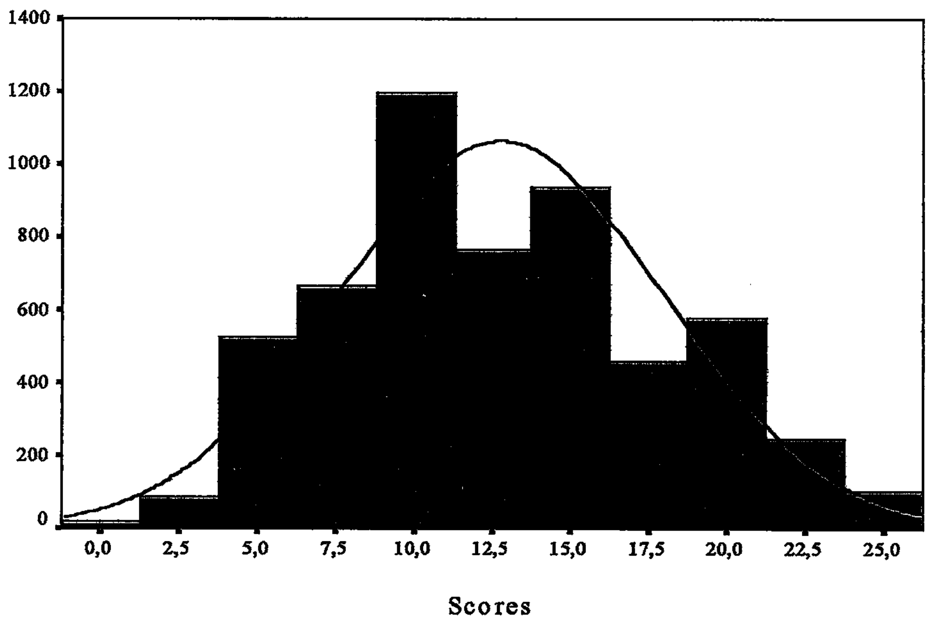

Figure 4.2.1. Frequency Distribution of the Social Sciences Subtest Scores

4.2.2. Checking Model Assumptions

The results of principal component analysis (Appendix A2) indicated that there are three interpretable factors, which accounted for 29.601% of the total variance. Eigenvalues of these factors are 5.112 (20.448 of the variance), 1.274 (5.098% of the variance) and 1.014 (4.055 of the variance) respectively. As seen in the scree plot of the eigenvalues (Figure 4.2.2.), there is a sharp decrease from first eigenvalue to second, which shows the Social Sciences subtest could be considered as a unidimensional scale.

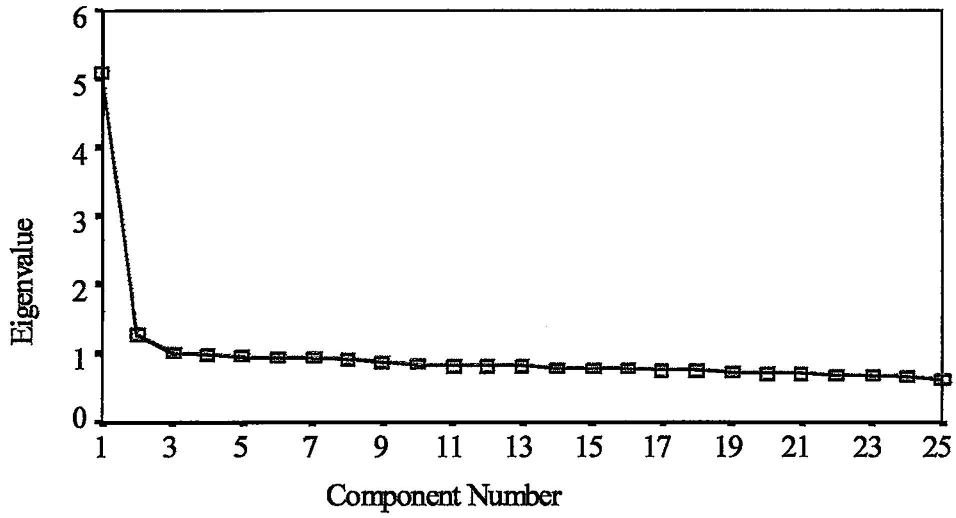

Figure 4.2.2. Plot of Eigenvalues

The non-speeded test administration assumption was investigated by checking the percentages of examinees completing last six items. The results indicated that 38% of examinees did not complete these six items. Therefore, this result shows the speededness of test administration process in the Social Sciences Subtest.

To investigate the local independence assumption in the social sciences subtest the inter item correlation matrix of low ability examinees was checked. The entries in the off-diagonal elements of the matrices approaching to zero (Appendix B2) showed that the items are locally independent in the Social Sciences subtest.

In order to investigate the equality of item discrimination indices assumption of the one-parameter model, the frequency distribution of the item discrimination indices was used. The item discrimination indices range from 0,362 to 0,701. As seen in Figure 4.2.3., the equality of item discrimination indices assumption of the one-parameter model was not met by the Social Sciences subtest.

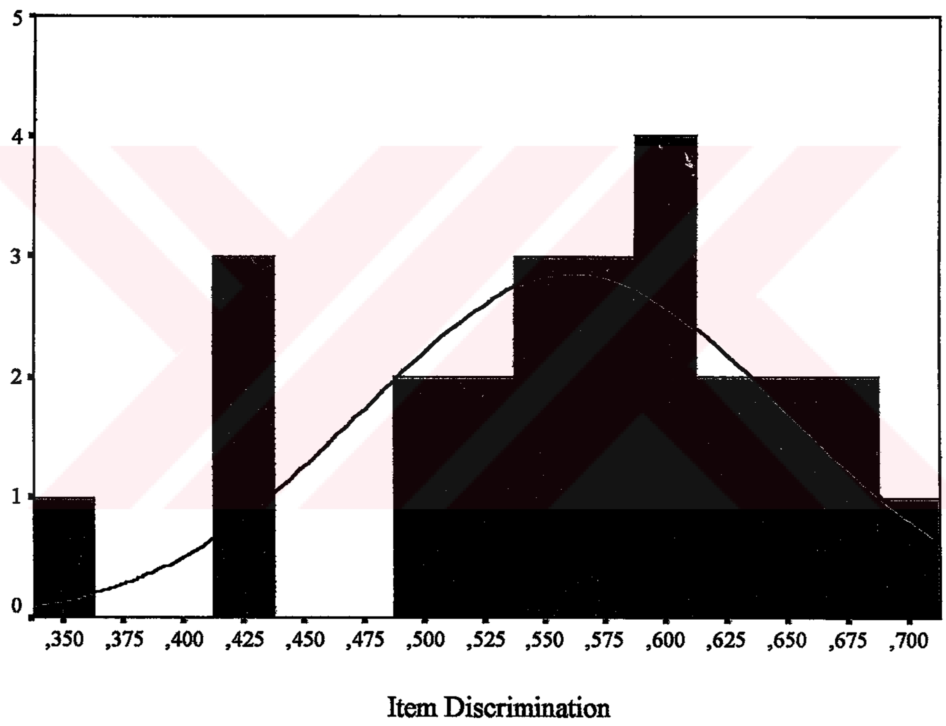

Figure 4.2.3. Frequency Distribution of the Discrimination Indices of the Social Sciences Items

To investigate the minimal guessing assumption of the one- and two-parameter models, the performance of the low ability examinees on the difficult items was reviewed. Five difficult items were selected in order to see the performance of low ability examinees which are the 6., 8., 13., 17., and 20. items of the social sciences subtest. As in the Table 4.2.3., the performance of low ability examinees on the 5 difficult items was checked and the mean of the examinees who did not give any correct response to these items was approximately 82%. The performance of low ability examinees on these items was low. For this reason, it can be concluded that the minimal guessing assumption was met by the Social Sciences subtest data.

Table 4.2.3. *Difficult Item Scores of Low Ability Examinees*

Scores	Frequencies (%)				
	Item 6	Item 8	Item 13	Item 17	Item 20
Missing	14.3	14.3	10.8	27.1	17.6
0 (wrong answer)	65.1	64.7	75.4	56.9	61.2
1 (correct answer)	20.6	21.0	13.8	16.0	21.2
Total	100.0	100.0	100.0	100.0	100.0

4.2.3. Checking Expected Model Features

4.2.3.1. Invariance of Item Parameter Estimates

In order to check the invariance of item parameter estimates in the Social Sciences subtest, the item parameter estimates across male vs. female and high vs. low ability groups were correlated. Figures 4.2.3.1. and 4.2.3.2. are scatter plots of the item difficulty parameters for the one-parameter model. Figures from 4.2.3.3. to 4.2.3.6. show plots of the item difficulty and discrimination parameter estimates for the two-parameter model. Figures from 4.2.3.6. to 4.2.3.12 are the scatter plots of the item difficulty, discrimination and guessing parameter estimates for the three- parameter model.

As seen in Table 4.2.3.1., the correlation of the item difficulty parameter estimates indicate high invariance across different samples of examinees in the Social Sciences subtest for each IRT models.

On the other hand, in contrast to the item discrimination parameter estimates of the three-parameter model, the two-parameter showed no invariance across high vs. low ability samples.

Besides, the pseudo-chance level parameter estimates of the three-parameter model were also invariant across two different samples of examinees in the Social sciences subtest of the Secondary Education Institutions Student Selection and Placement Exam.

Table 4.2.3.1. *Correlation of Item Parameter Estimates Across Different Samples of Examinees Obtained in Two Groups*

Samples	Bilog		
	One Parameter	Two Parameter	Three Parameter
Male-Female (-b-)	0.961**	0.962**	0.901**
High Ability-Low Ability (-b-)	0.914**	0.903**	0.877**
Male-Female (-a-)		0.812**	0.837**
High Ability-Low Ability (-a-)		0.024	0.578**
Male-Female (-c-)			0.502**
High Ability-Low Ability (-c-)			0.319*

** $p<.01$
* $p<.05$

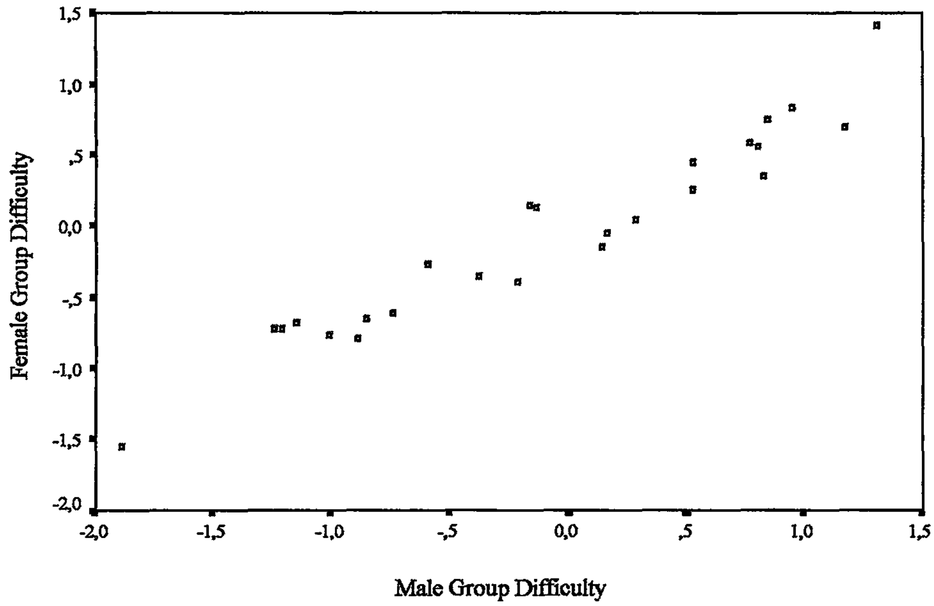

Figure 4.2.3.1. Plot of 1P Item Difficulty Values Based on Female and Male Groups of Examinees

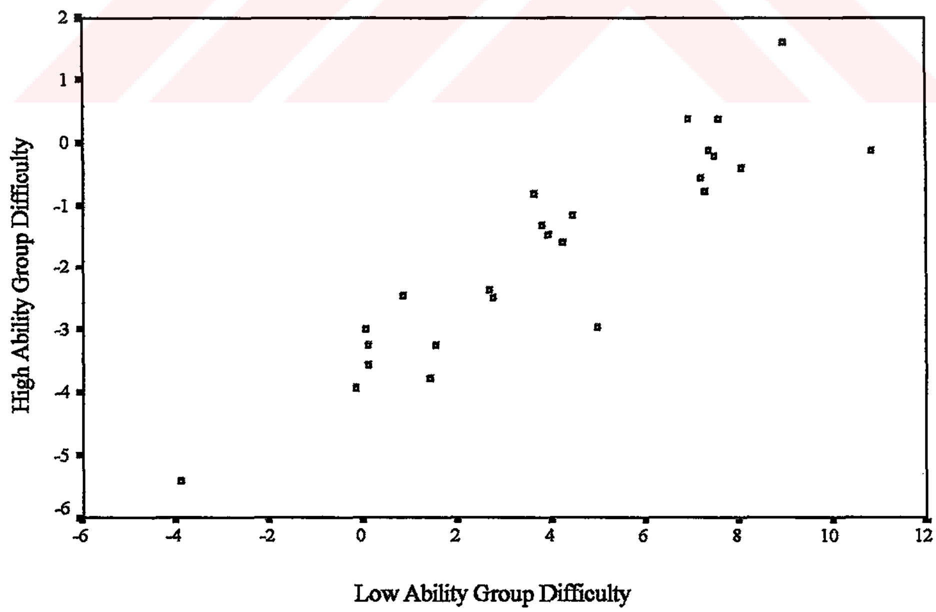

Figure 4.2.3.2. Plot of 1P Item Difficulty Values Based on High and Low Ability Groups of Examinees

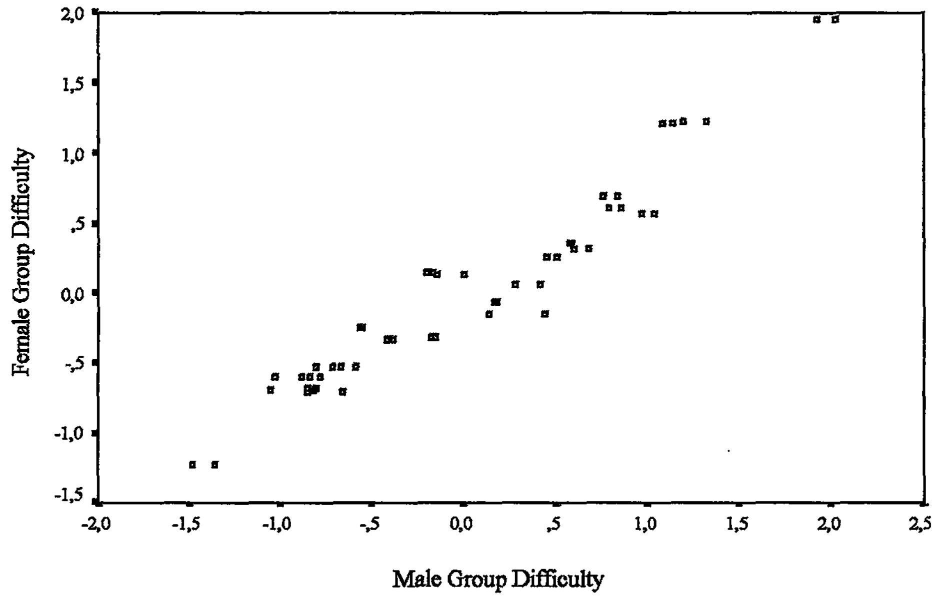

Figure 4.2.3.3. Plot of 2P Item Difficulty Values Based on Male and Female Groups of Examinees

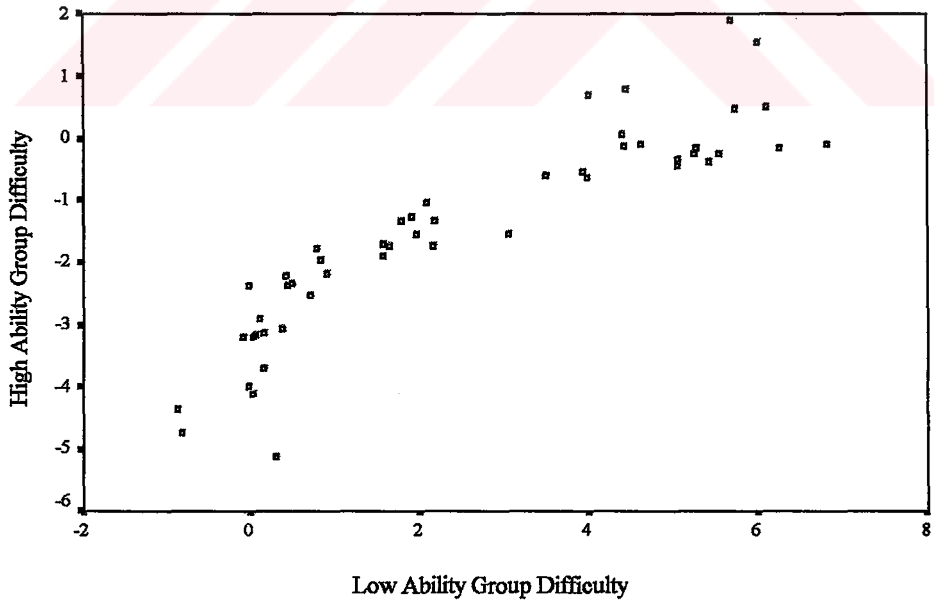

Figure 4.2.3.4. Plot of 2P Item Difficulty Values Based on High and Low Ability Groups of Examinees

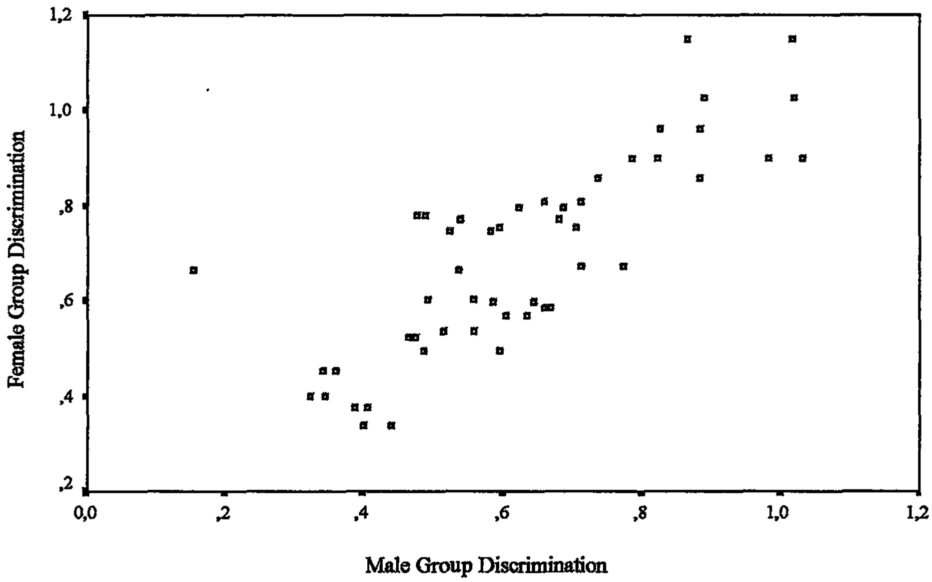

Figure 4.2.3.5. Plot of 2P Item Discrimination Values Based on Male and Female Groups of Examinees

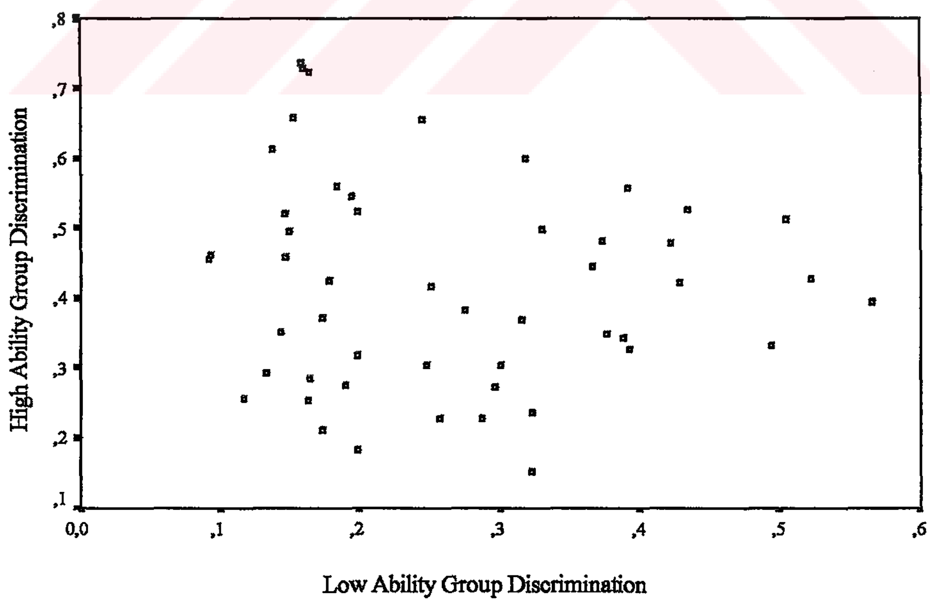

Figure 4.2.3.6. Plot of 2P Item Discrimination Values Based on High and Low Ability Groups of Examinees

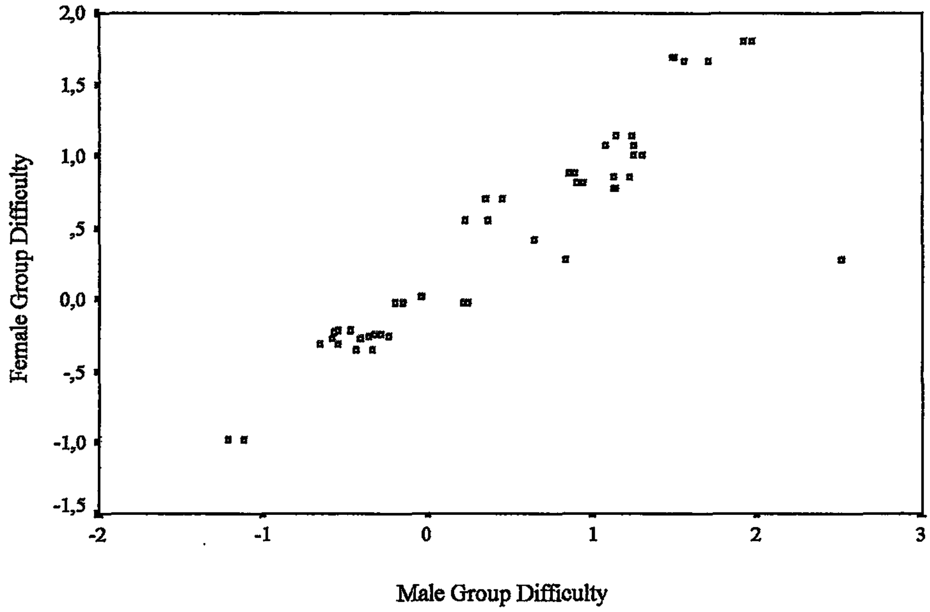

Figure 4.2.3.7. Plot of 3P Item Difficulty Values Based on Female and Male Groups of Examinees

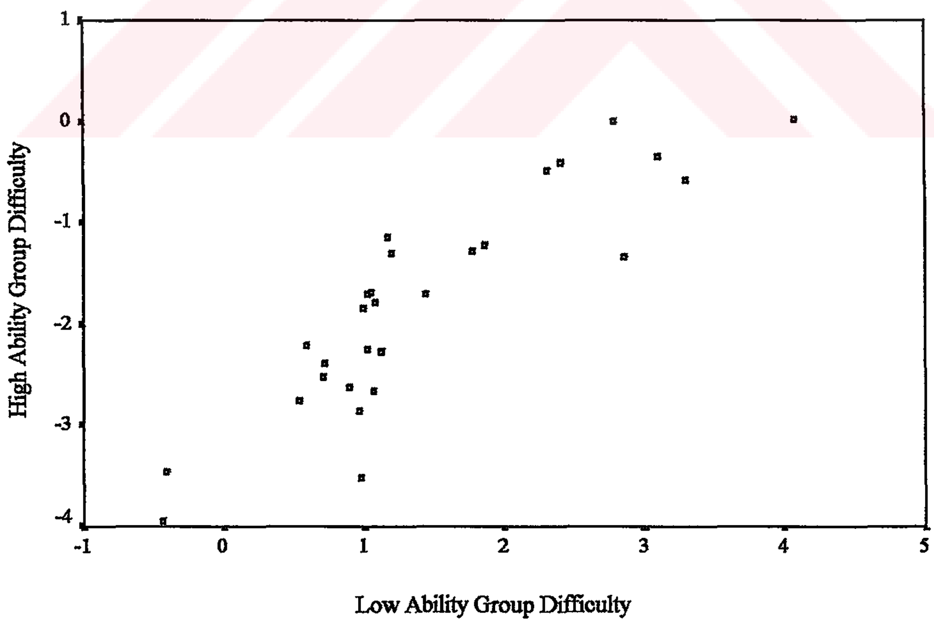

Figure 4.2.3.8. Plot of 3P Item Difficulty Values Based on High and Low Ability Groups of Examinees

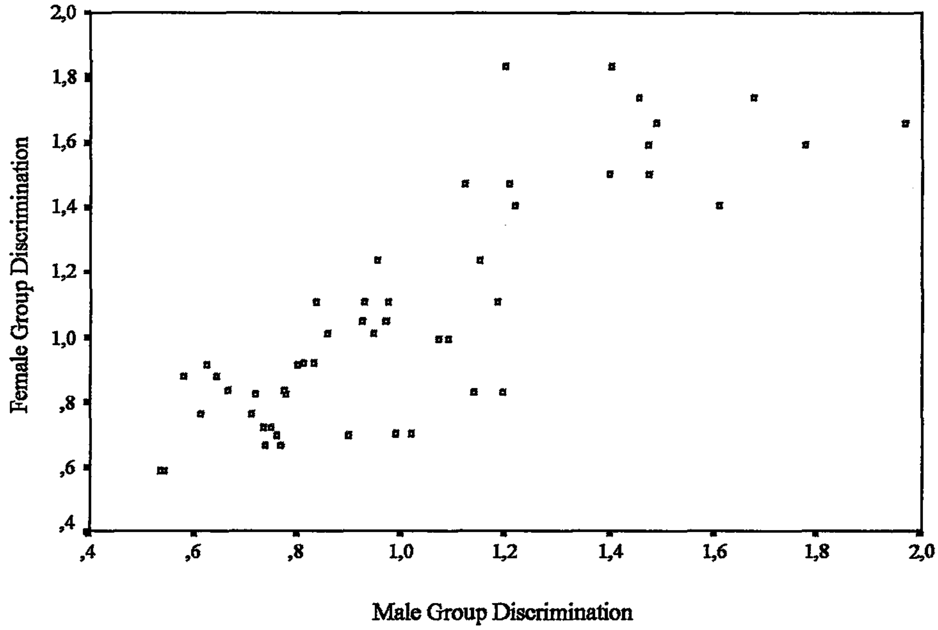

Figure 4.2.3.9. Plot of 3P Item Discrimination Values Based on Male and Female Groups of Examinees

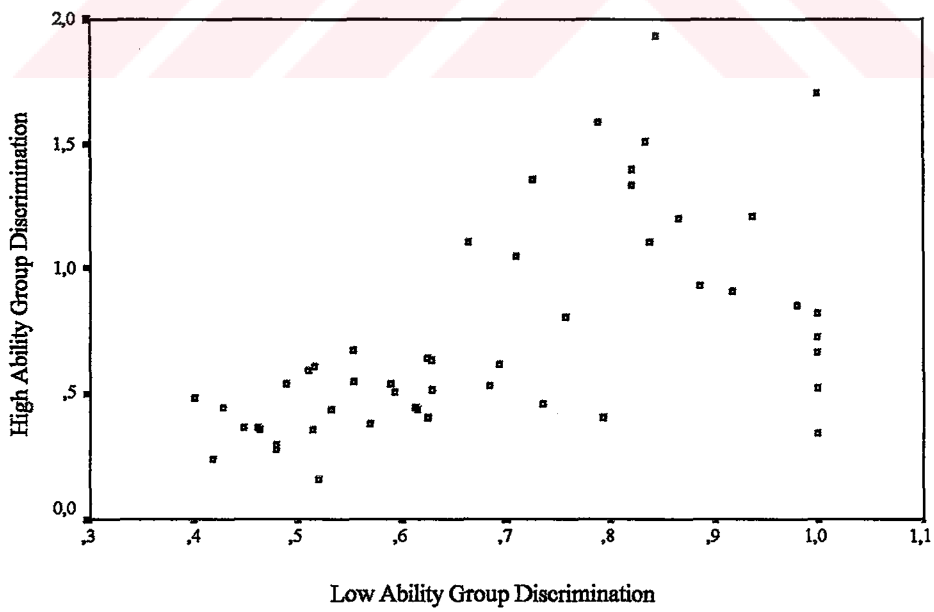

Figure 4.2.3.10. Plot of 3P Item Discrimination Values Based on High and Low Ability Groups of Examinees

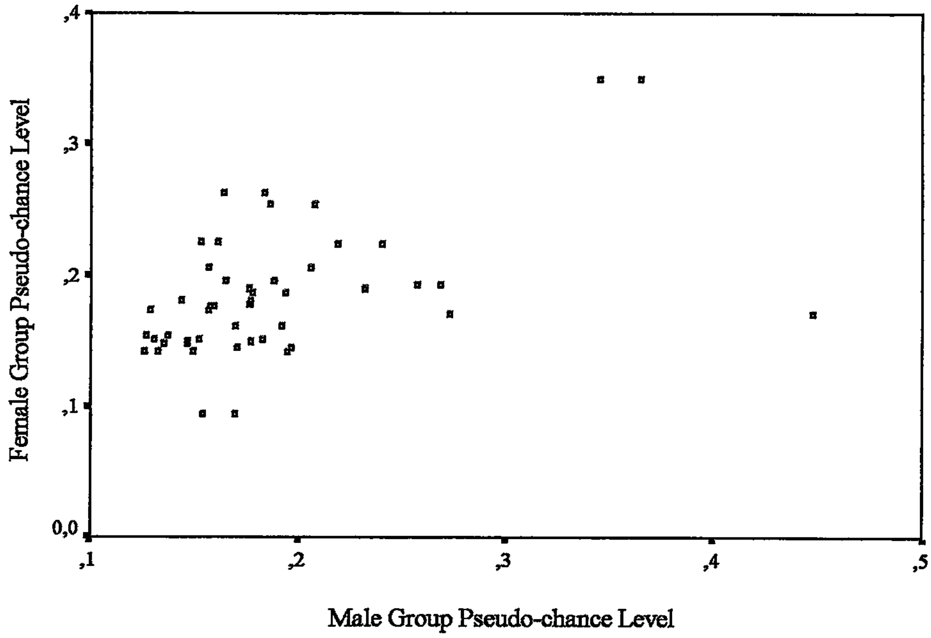

Figure 4.2.3.11. Plot of 3P Item Pseudo-Chance Level Based on Male and Female Groups of Examinees

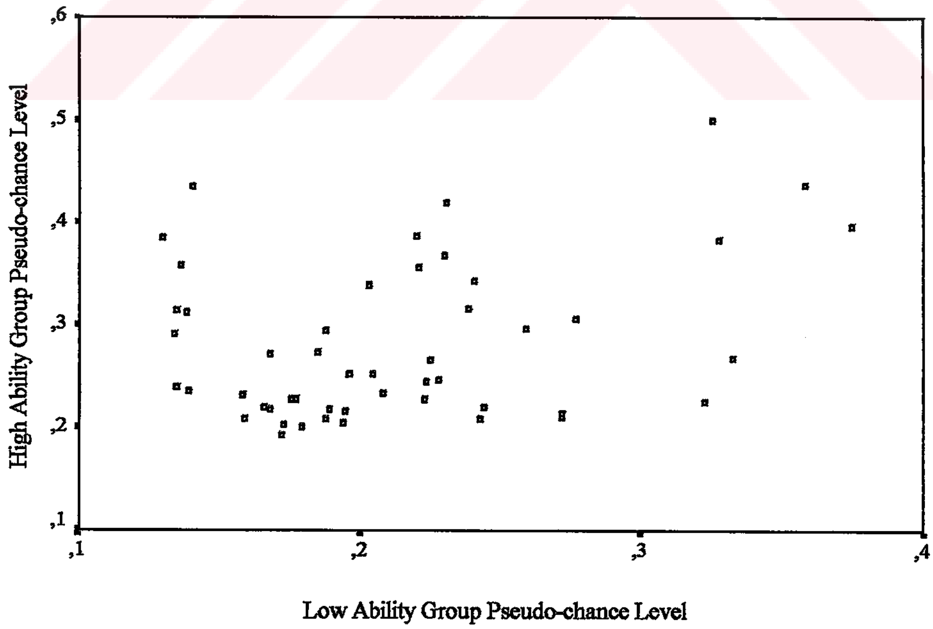

Figure 4.2.3.12. Plot of 3P Item Pseudo-Chance Level Based on High and Low Ability Groups of Examinees

4.2.3.2. Invariance of Ability Parameter Estimates

Correlations of the ability parameter estimates of each IRT model across different sets of items, which are easy vs. difficult and even vs. odd, are presented in Table 4.2.3.2. Each correlation revealed significant results across different sets of items, which indicate the invariance of ability parameter estimates.

For the one-parameter model figures 4.2.3.13 and 4.2.3.14., for the two-parameter model figures 4.2.3.15. and 4.2.3.16. for the three-parameter model figure 4.2.3.17. show the scatterplots of ability parameter estimates across Easy vs. Difficult and Even vs. Odd items of the Social Sciences subtest.

Table 4.2.3.2. *Correlations of Ability Parameter Estimates Across Different Sets of Items.*

Subtests	Bilog		
	One Parameter	Two Parameter	Three Parameter
Easy- Difficult	0.638**	0.642**	0.656**
Even- Odd	0.740**	0.750**	0.761**

** p<.01

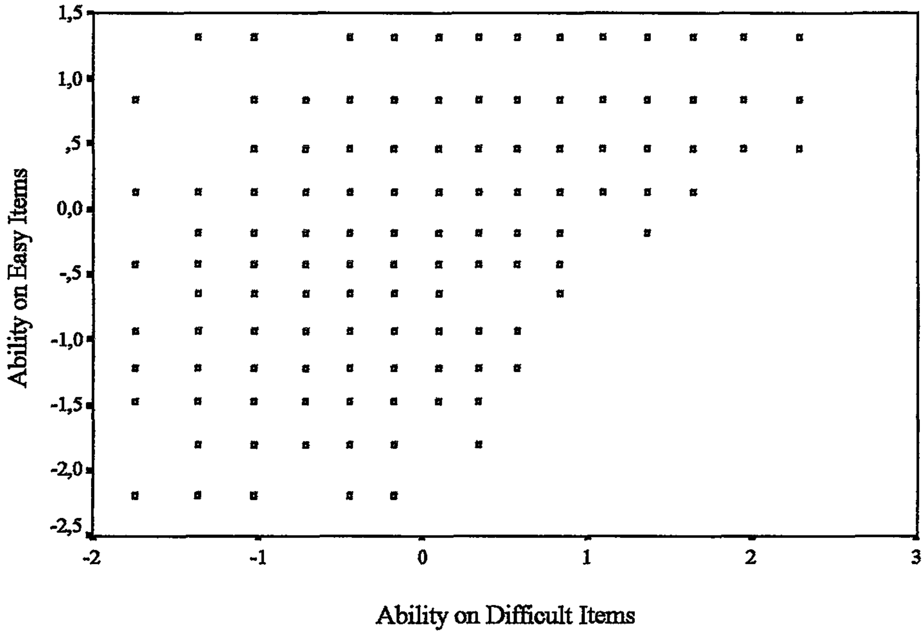

Figure 4.2.3.13. Plot of 1P Ability Estimates Based on Easy and Difficult Items

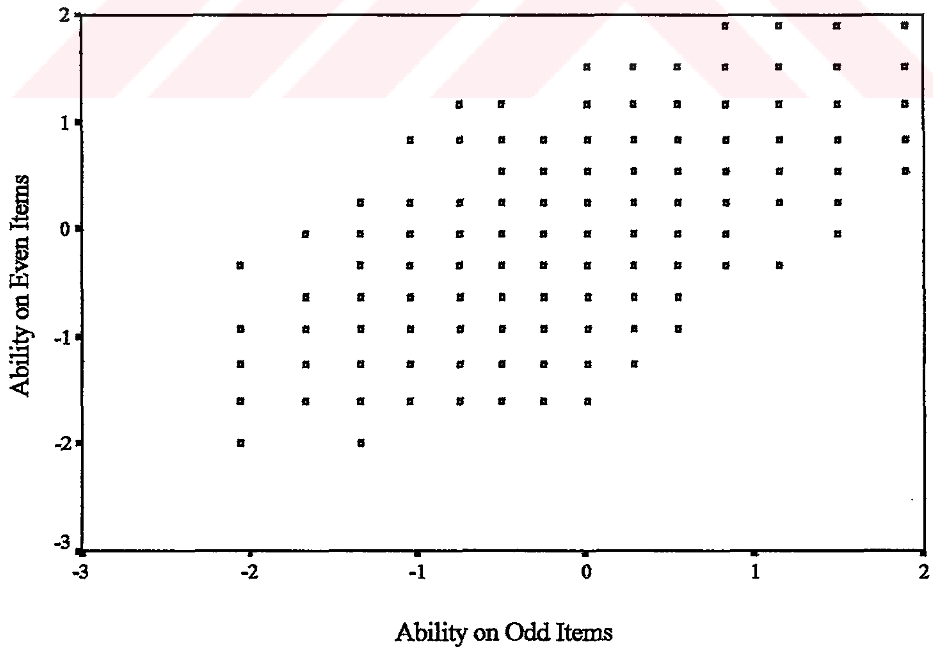

Figure 4.2.3.14. Plot of 1P Ability Estimates Based on Even and Odd Items

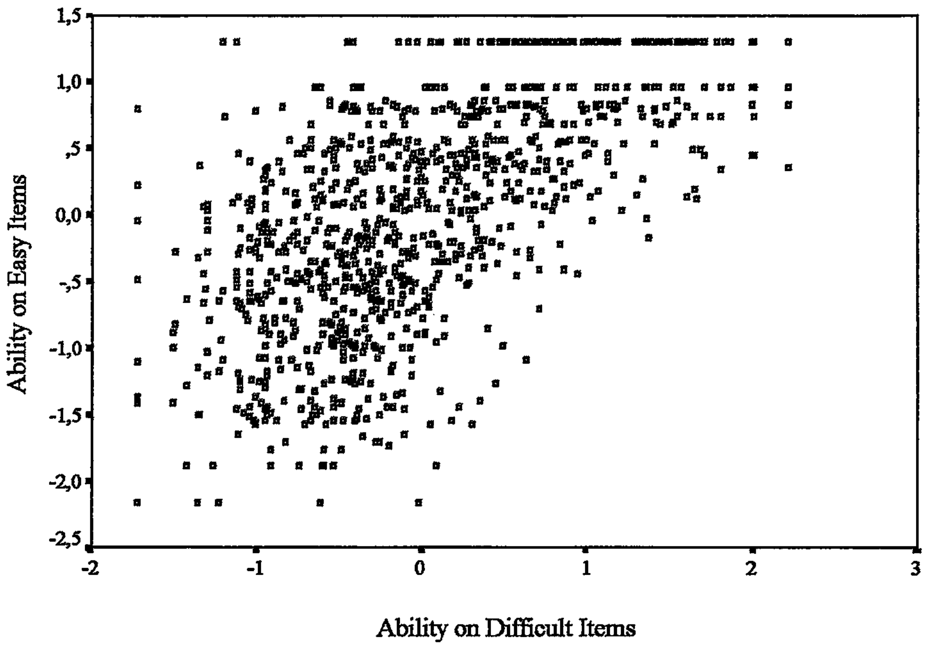

Figure 4.2.3.15. Plot of 2P Ability Estimates Based on Easy and Difficult Items

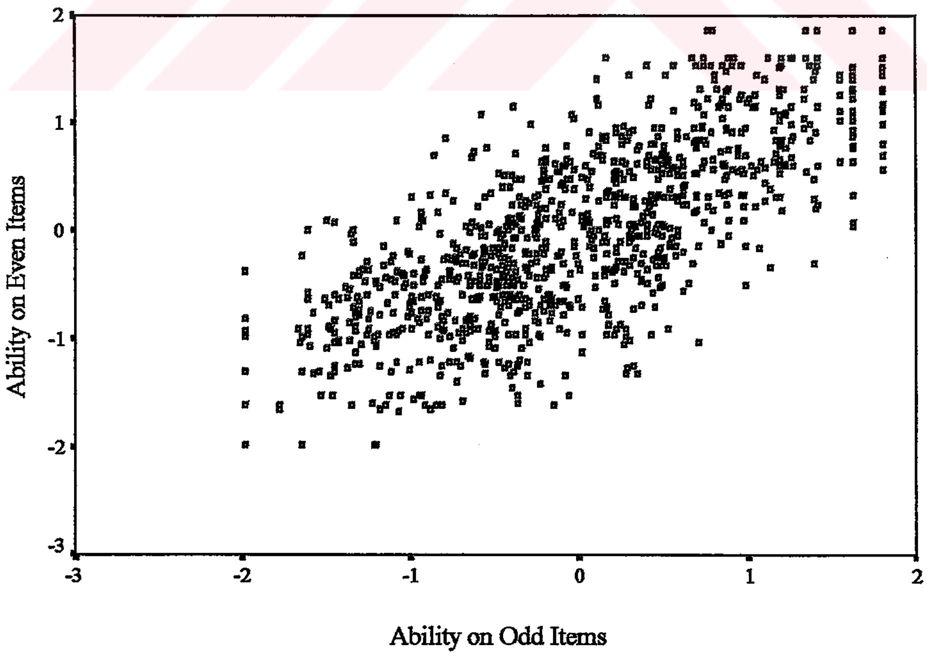

Figure 4.2.3.16. Plot of 2P Ability Estimates Based on Even and Odd Items

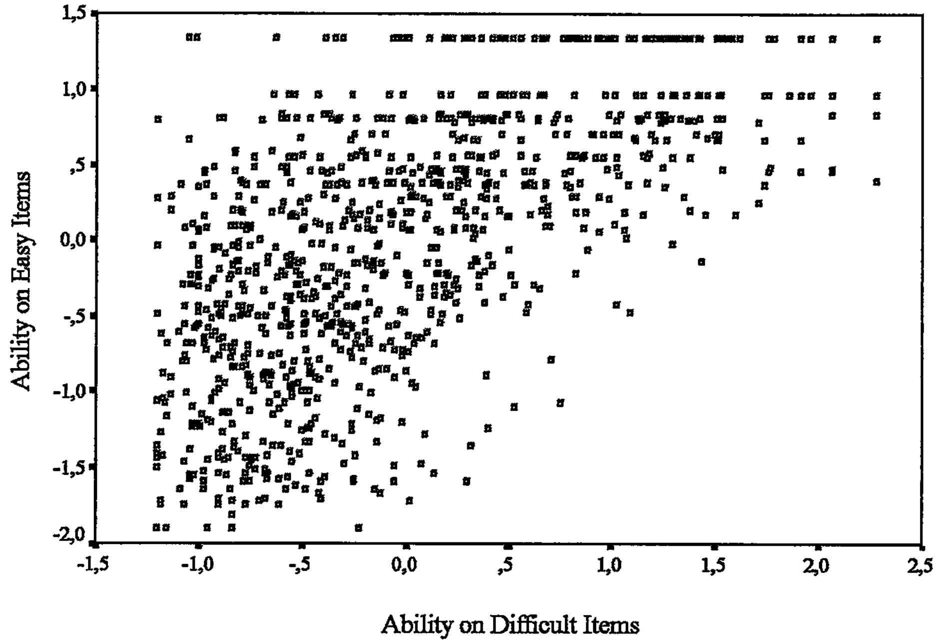

Figure 4.2.3.17. Plot of 3P Ability Estimates Based on Easy and Difficult Items

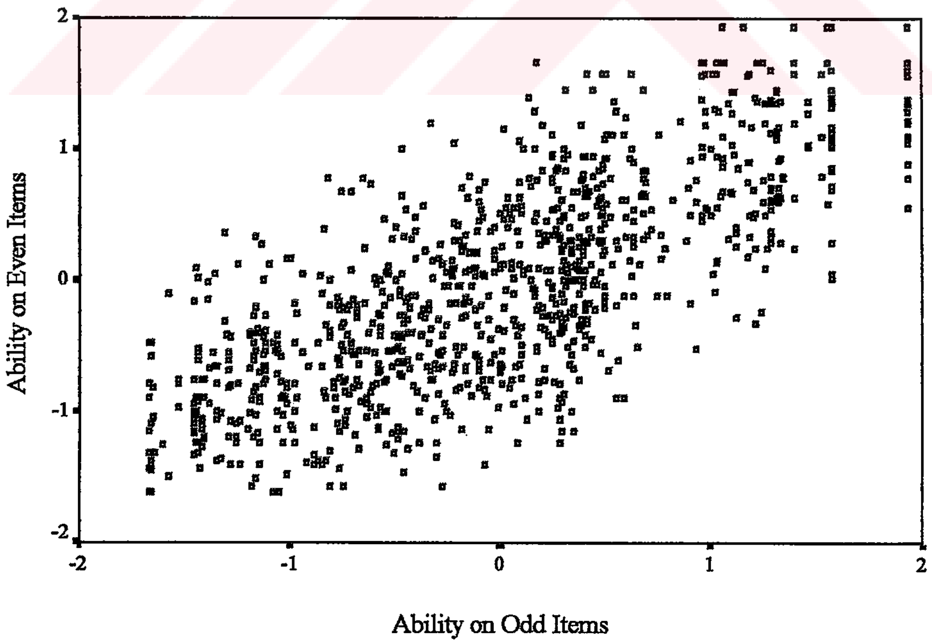

Figure 4.2.3.18. Plot of 3P Ability Estimates Based on Even and Odd Items

4.2.4. Checking Model Predictions of Actual and Simulated Test Results

To see how well the observed distribution fits the theoretical distribution, the chi square statistics was checked. Tables 4.2.5, 4.2.6. and 4.2.7. show the item parameter estimates and misfit items for the one-, two- and three-parameter models respectively. Items, which are written bold, indicate misfit items of the each IRT model at 0.05 significance level. Table 4.2.4. shows the number of misfit items in the models.

Table 4.2.4. *Total Number of Misfit Items For The One, Two and The Three Parameter Models*

Models	N of items	N of misfit items
One parameter	25	14
Two parameter	25	21
Three parameter	25	24

* p<.05

Table 4.2.5. *Item Parameters of the One Parameter Model*

Item number	One parameter model b- value
1.	0.072
2.	-0.709
3.	-0.823
4.	-1.666
5.	-0.664
6.	0.728
7.	-0.336
8.	0.854
9.	-0.912
10.	-0.341
11.	0.565
12.	0.335
13.	0.790
14.	0.296
15.	0.555
16.	-0.056
17.	1.263
18.	0.062
19.	-0.423
20.	0.785
21.	-0.918
22.	-0.891
23.	-0.013
24.	-0.174
25.	-0.934

Table 4.2.6. *Item Parameters of the Two-Parameter Model*

Item number	Two parameter model	
	b- value	a- value
1.	0.112	0.384
2.	-0.811	0.498
3.	-0.709	0.761
4.	-1.280	0.913
5.	-0.525	0.906
6.	0.763	0.568
7.	-0.291	0.771
8.	1.055	0.459
9.	-0.700	0.946
10.	-0.269	0.954
11.	0.531	0.660
12.	0.351	0.568
13.	0.676	0.762
14.	0.311	0.569
15.	0.546	0.619
16.	-0.054	0.632
17.	2.461	0.270
18.	0.069	0.558
19.	-0.368	0.759
20.	1.259	0.336
21.	-0.675	1.038
22.	-0.894	0.598
23.	-0.014	0.676
24.	-0.162	0.688
25.	-0.805	0.758

Table 4.2.7. *Item Parameters of the Three-Parameter Model*

Item number	Three parameter model		
	b- value	a- value	c- value
1.	1.086	1.819	0.367
2.	-0.354	0.545	0.176
3.	-0.421	0.819	0.153
4.	-1.161	0.880	0.144
5.	-0.315	0.942	0.119
6.	1.069	1.000	0.156
7.	-0.048	0.870	0.113
8.	1.368	0.797	0.158
9.	-0.525	0.996	0.111
10.	0.024	1.175	0.145
11.	0.923	2.053	0.216
12.	0.888	1.244	0.235
13.	0.870	1.121	0.101
14.	0.864	1.319	0.244
15.	0.944	1.207	0.188
16.	0.559	1.131	0.251
17.	2.002	1.157	0.199
18.	0.428	0.675	0.134
19.	-0.146	0.823	0.106
20.	1.786	0.550	0.179
21.	-0.531	1.078	0.098
22.	-0.418	0.684	0.207
23.	0.374	0.881	0.161
24.	0.256	0.890	0.176
25.	-0.301	0.961	0.245

CHAPTER 5

CONCLUSIONS AND IMPLICATIONS

This final chapter presents the discussions, conclusions and the implications of the study.

5.1. Discussions

The results of the classical test analysis and descriptive statistics showed that the verbal section of the Secondary Education Institutions student Selection and Placement Test includes moderate difficult items. When the mean scores of Turkish and Social Sciences subtests were examined in detail, it was observed that the Turkish subtest was slightly easier than the Social Sciences subtest for the examinees. The mean difficulties, which were 0.541 and 0.510 for Turkish and Social Sciences subtests respectively, also supported Turkish tests' being easier then the Social Sciences subtest. When the mean discrimination indices were observed, it was seen that the Social Sciences subtest items were more discriminating then the Turkish subtest.

In order to see whether the assumptions of IRT were met by the test data, the principal component analysis and scree plots were checked firstly for the unidimensionality assumption. Each subtests of the verbal section of the

Secondary Education Institutions Student Selection and Placement Test had a dominant underlying factor and the scree plots showed a sharp decrease from first factor to the second one, which showed the unidimensionality of the subtests.

The non-speeded test administration assumption was investigated by checking the percentages of examinees completing the last six items for each subtest. The results indicated that about 48% of examinees did not complete the last six items of the Turkish subtest. However, it was concluded that the value of 48% was because of the 24. items' being one of the most difficult items of the Turkish subtest. For this reason, it was concluded that there was a non-speeded test administration process for the Turkish subtest. The percentage of examinees completing the last six items of the Social Sciences subtest was 38% that pointed out the speededness of the test administration process.

The inter item correlation matrices of the low ability examinees were obtained to check the local independence assumption. The results revealed that the items of the Turkish and Social Sciences subtests were locally independent.

In order to investigate the equality of item discrimination indices assumption of the one-parameter model, the frequency distribution of the item discrimination indices were used. The findings revealed that the distributions for each subtest were not homogenous and the equality of the item discrimination indices assumption of the one-parameter model was not met in the Turkish and Social Sciences subtests of the Secondary Education Institutions student Selection and Placement Test.

The minimal guessing assumption of the one- and two-parameter models was checked by reviewing the performance of the low ability examinees on the most five difficult items of the subtests. The results showed that the performance of the examinees on these items were low for each subtests. Therefore, the minimal guessing assumption of the one- and two-parameter models was viable for the Turkish and Social Sciences subtests.

In order to check the expected model features of IRT, invariance of item and ability parameter estimates were investigated. The item and ability parameter estimates were obtained by the use of BILOG computer program.

The item parameter estimates across male vs. female and high ability vs. low ability group were correlated and the scatter plots were formed in order to investigate the invariance of item parameter estimates. In the Turkish subtest, the results showed that the one-, two- and three parameter models' item difficulty parameter estimates were invariant across male vs. female and high ability vs. low ability group. On the other hand, only the two-parameter model item discrimination estimates for high ability vs. low ability group were invariant in the Turkish subtest. The discrimination parameter estimates in the high ability vs. low ability groups and pseudo-chance level parameter estimates for the three-parameter model across two different groups were not invariant in the Turkish subtest.

In the Social Sciences subtest, the item difficulty parameter estimates of both one-, two- and three-parameter models were highly invariant across male vs.

female and high vs. low ability groups. Moreover, the three-parameter pseudo-chance level parameter estimates of Social Sciences subtest were invariant for the two different groups in contrast to Turkish subtest.

In order to investigate the invariance of ability parameter estimates, the ability estimates across easy vs. difficult and even vs. odd items were correlated and scatter plots were formed. The findings revealed that the three-parameter model ability estimates showed slightly more invariance in the Turkish and Social Sciences subtests when compared to one- and two-parameter models.

Although the equality of item discrimination indices assumption of the one-parameter model was not met in the subtests, the results of the chi square statistics in the study showed that the number of misfit items of the one-parameter item response model at 0.05 significance level were less than the two- and three-parameter models in the Turkish and Social Sciences subtests. But; the number of misfit items was too many for two-, three- and also for one parameter model in the subtests. Not being met the item parameter estimates as expected can be considered as a reason of too many misfit items obtained from the analysis. Another reason can be the sample used in the study. Because Hambleton et.al. (1991) stated that, in the model-data fit studies, there is too much reliance on the statistical tests which are used in the analysis and these tests are sensitive to the sample size. For this reason, the large sample size of the study could affect the number of the fit and misfit items.

The findings of the chi square goodness of fit statistics were not similar with the findings of the study carried out by Kılıç (1999) in which the verbal ability section of the Student Selection Test fit the three-parameter model better and also the percentages of misfit items for both one-, two- and three-parameter models were not high when compared with this study.

5.2. Conclusions

The following conclusions can be drawn from the analysis of the Verbal Section of the Secondary Education Institutions student Selection and Placement Test.

1. The unidimensionality, the local independence and the minimal guessing assumption of the one-parameter models of the IRT were met by the Turkish and Social Sciences subtests.

2. The Turkish and Social Sciences subtests did not meet the equality of item discrimination indices of the one-parameter model.

3. It was observed that the Turkish subtest was non-speeded test but there was a speeded administration process for the Social Sciences subtest.

4. When the invariance of item parameter estimates were considered, except the three-parameter model, similar results were observed for the Turkish and Social Sciences subtests. The item difficulty parameter estimates were

highly invariant across different samples of examinees in the subtests. However, the invariance of the pseudo-chance level parameter estimates could not be achieved in the Turkish subtest. Additionally, the item discrimination parameter estimates across high vs. low ability groups were not invariant both in the Turkish and Social Sciences subtests.

The ability parameter estimates were invariant across different sets of items in each subtests of the verbal section of the Secondary Education Institutions Student Selection and Placement Exam. However, the three-parameter model ability parameter estimates were slightly more invariant than the one- and two-parameter models in both subtests.

5. The results of the Chi square statistics showed that the two- and three-parameter models poorly fit the Turkish and Social Sciences subtests. The one-parameter model fit these subtests better. However, in general the ratio of the misfit items was high in both of the models. The reason can be considered as the sample size. Because, as the Hambleton et.al (1991) states that the statistical tests of model fit are sensitive to the examinee sample size.

5.3. Implications for Research and Practice

In IRT applications, when the fit of the model to the data can be assessed, the model can adequately predict or explain the data (Hambleton, 1991). In other words, the advantages of the IRT models can be obtained if there is a satisfactory fit between the model and the test data. By the use of IRT, more detailed information on the test items can be obtained. For this reason, this study aimed to analyze the Verbal Section (Turkish and Social Sciences subtests) of MONE-Secondary Education Institutions Student Selection and Placement Exam by using Item Response Theory.

One of the results of the study obtained in the investigation of the non-speeded test administration assumption was important. It was seen that there was a speeded test administration process in the Social Sciences subtest. This finding can be accepted as evidence that all of the items could not be completed in the subtest. When the Social Sciences subtests being the last subtest of the exam was considered, the speeded test administration process indicates the inadequacy of the testing time. To be careful in determining the testing time of the exam can be suggested to The Evaluation and Assessment Center of the MONE.

Taking into account the results of the study, some of the item parameter estimates were invariant for different samples of examinees. It would be desirable to select different samples of examinees such as; randomly selected two samples, for the further studies. In addition, the other subtests of the Secondary Education

Institutions Student Selection and Placement Exam could also be studied and the comparison between the verbal section and quantitative section of the exam could be obtained.

In addition, this study will contribute to provide a base for the MONE when they start to use IRT in their applications in the future and also this study will provide a base for the studies of new type of testing named adaptive testing.

REFERENCES

Assessment System Corporation. (1986). *User's manual for ITEMAN: Conventional item analysis program.* St. Paul, MN: Author.

Aiken, L.R. (1997). *Psychological testing and assessment.* Boston: Allyn and Bacon.

Anastasi, A., and Urbina, S. (1997). *Psychological testing.* Upper Suddle River: Prentice Hill Inc.

Armstrong, D.R. (1990). An application of item response theory to the measurement of job performance. *Doctorate Thesis.* Illinois Institute of Technology. (UMI Proquest Digital Dissertations Publication No: AAT 8922162).

Choi, I. (1990). An application of item response theory to language testing: model-data fit studies. *Doctorate Thesis.* University of Illinois. (UMI Proquest Digital Dissertations Publication No: AAT 9010829).

Chow, P., and Winzer M.M. (1992). Reliability and validity of a scale measuring attitudes toward mainstreaming. *Educational and Psychological Measurement, 52,* 223-229.

Crocker, L. and Algina, J. (1986). *Introduction to classical and modern test theory.* Orlando: Holt, Rinehart and Winston, Inc.

Fan, X. (1998). Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58,* 357-381.

Fanmin, G. (1997). A new goodness of fit statistic for assessing model-data fit in item response theory applications. *Doctorate Thesis.* University of Pittsburgh. (UMI Proquest Digital Dissertations Publication No: AAT 9736422).

Ferrando P.J. (1994). Fitting item response theory models to the EPI-A impulsivity subscale. *Educational and Psychological Measurement, 54,* 118-128.

Gernot F. (1996). Applications of item response theory to personality tests. Johannes Kepler Universitaet Linz. *Doctorate Thesis.* (UMI Proquest Digital Dissertations).

Gumpel, T., Wilson, M., and Shalev, R. (1998). An item response theory analysis of Conners teacher's rating scale. *Journal of Learning Disabilities. 31,* 525-533.

Hambleton, R.K. (1983). *Fitting item response models to the Maryland functional reading test results.* Massachusetts University. Annual Meeting of the American Educational Research Association. (ERIC Document Reproduction Service No: ED 230624).

Hambleton, R.K., and Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston. Kluwer Academic Publishers.

Hambleton, R.K., Swaminathan, H., and Rogers, H.J. (1991). *Fundamentals of item response theory.* Newbury Park, CA. Sage Publication.

Hambleton, R.K., and van der Linden, W. J. (1982). Advances in item response theory and applications: An introduction. *Applied Psychological Measurement, 6,* 373-378.

Hui, C.H., Drasgow, F., and Chang, B. (1983). Analysis of the modernity scale: An item response theory approach. *Journal of Cross-cultural Psychology, 14,* 259-278.

Humbert, R.A. (1987). Analyzing results from an occupational licensing examination: A comparison of classical test theory and item response theory findings. *Doctorate Thesis.* West Virginia University. (UMI Proquest Digital Dissertations Publication No: AAT 8627730).

Kılıç, İ. (1999). The fit of one, two and three parameter models of item response theory to the student selection test of the student selection and placement center. *Doctarate Thesis*. Middle East Technical University, Department of Educational Sciences, Ankara.

Kirişci L., and Duncan, B.C. (1996). *Reliability and validity of the state trait anxiety inventory for children in an adolescent sample: Confirmatory factor analysis and item response theory.* Annual Meeting of the American Educational Research Association. (Eric Document Reproduction Service No: ED 400304).

Ludlow, L.H., and Guida, F.V. (1991). The test anxiety scale for children as a generalized measure of academic anxiety. *Educational and Psychological Measurement, 51,*1013-1022.

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin, 115*, 300-308.

Mislevy, R.J., and Bock, D.R. (1986). *PC-BILOG: Item analysis and test scoring with binary logistic models.* Scientific Software Inc.

Paek, P., and Holland, P.W. (1999). Development and analysis of a mathematics aptitude test for gifted elementary school students. *School Science and Mathematics, 99*, 338-348.

Raju, N. S, and Goldman, S.H. (1986). Recovery of one- two-parameter logistic item parameters: An emprical study. *Educational and Psychological Measurement, 46*, 11-21.

Thissen, D. (1985). The new test theory. *Contemporary Psychology, 32*, 336-338.

Young, A.M., Halper, I.S., Clark, D.C., Scheftner, and W., Fawcett, J. (1992). An item response theory evaluation of the Beck hopelessness scale. *Cognitive Therapy and Research, 16,* 579-587.

Van de Rijt, B.A.M., Van Luit, J. E.H., and Pennings, A.H. (1999). The construction of the Utrecht early mathematical competence scales. *Educational and Psychological Measurement, 59*, 289-310.

Yalçın, M. (1999). " The fit of one-, two-, three-, parameter models of item response theory to ERDD's achievement test data." *Doctorate Thesis*. Middle East Technical University, Department of Educational Sciences, Ankara.

APPENDICES

APPENDIX A

Principle Component Analysis of the Turkish and Social Sciences Subtests of The MONE- Secondary Education Institutions Student Selection and Placement Exam

APPENDIX A 1

PRINCIPLE COMPONENT ANALYSIS OF TURKISH SUBTEST

Initial Statistics:

Variable	Communality	Factor	Eigenvalue	Pct of Var	Cum Pct
T1	1,000	1	4,324	17,296	17,296
T2	1,000	2	1,249	4,995	22,291
T3	1,000	3	1,028	4,114	26,405
T4	1,000	4	1,000	4,000	30,405
T5	1,000	5	,977	3,907	34,312
T6	1,000	6	,962	3,849	38,161
T7	1,000	7	,948	3,792	41,954
T8	1,000	8	,933	3,733	45,687
T9	1,000	9	,917	3,668	49,355
T10	1,000	10	,890	3,560	52,915
T11	1,000	11	,884	3,535	56,450
T12	1,000	12	,880	3,521	59,971
T13	1,000	13	,860	3,441	63,412
T14	1,000	14	,852	3,410	66,822
T15	1,000	15	,830	3,319	70,141
T16	1,000	16	,810	3,240	73,381
T17	1,000	17	,797	3,188	76,569
T18	1,000	18	,792	3,167	79,736
T19	1,000	19	,776	3,105	82,841
T20	1,000	20	,757	3,027	85,868
T21	1,000	21	,740	2,959	88,827
T22	1,000	22	,724	2,897	91,724
T23	1,000	23	,718	2,870	94,595
T24	1,000	24	,692	2,766	97,361
T25	1,000	25	,660	2,639	100,000

PC extracted 4 factors.

Factor Matrix:

	Factor 1	Factor 2	Factor 3	Factor 4
T1	,268	3,990E-02	,400	-6,371E-02
T2	,564	-9,723E-02	2,317E-02	,127
T3	,505	5,300E-02	,221	6,179E-02
T4	-9,710E-03	-,204	,708	,102
T5	,203	,260	,300	-3,614E-02
T6	,145	,247	,108	,278
T7	,458	,242	,105	,134
T8	,225	8,763E-02	6,598E-02	,353
T9	,403	5,904E-02	8,436E-02	-3,736E-02
T10	,368	,270	,221	5,537E-02
T11	,162	,407	,280	-,104
T12	,138	,504	1,869E-02	9,756E-02
T13	,547	9,968E-02	7,305E-03	5,794E-02
T14	,223	,357	-3,585E-03	,230
T15	-5,414E-02	,590	-6,431E-02	-3,211E-02
T16	,156	9,334E-02	7,989E-02	-,790
T17	,213	,427	,294	3,334E-02
T18	,319	,334	,203	,195
T19	,281	,261	6,158E-02	-,108
T20	,297	,158	,236	-,106
T21	,456	,200	,196	,190
T22	,234	,270	,410	3,568E-02
T23	,568	,217	-3,846E-02	-5,659E-02
T24	-,219	,388	,505	,159
T25	,379	,333	,231	9,268E-02

Final Statistics:

Variable	Communality	Factor	Eigenvalue	Pct of Var	Cum Pct
T1	,238	1	4,324	17,296	17,296
T2	,344	2	1,249	4,995	22,291
T3	,310	3	1,028	4,114	26,405
T4	,553	4	1,000	4,000	30,405
T5	,200				
T6	,171				
T7	,297				
T8	,187				
T9	,174				
T10	,260				
T11	,281				
T12	,283				
T13	,313				
T14	,230				
T15	,356				
T16	,663				
T17	,316				
T18	,293				
T19	,162				
T20	,180				
T21	,322				
T22	,297				
T23	,375				
T24	,479				
T25	,317				

APPENDIX A 2

PRINCIPLE COMPONENT ANALYSIS OF SOCIAL SCIENCES SUBTEST

Initial Statistics:

Variable	Communality	Factor	Eigenvalue	Pct of Var	Cum Pct
S1	1,000	1	5,112	20,448	20,448
S2	1,000	2	1,274	5,098	25,546
S3	1,000	3	1,014	4,055	29,601
S4	1,000	4	,975	3,901	33,502
S5	1,000	5	,972	3,888	37,390
S6	1,000	6	,948	3,792	41,182
S7	1,000	7	,930	3,720	44,902
S8	1,000	8	,910	3,640	48,542
S9	1,000	9	,879	3,516	52,058
S10	1,000	10	,856	3,424	55,482
S11	1,000	11	,834	3,336	58,818
S12	1,000	12	,820	3,282	62,100
S13	1,000	13	,817	3,267	65,367
S14	1,000	14	,789	3,156	68,524
S15	1,000	15	,786	3,144	71,668
S16	1,000	16	,771	3,085	74,753
S17	1,000	17	,762	3,046	77,800
S18	1,000	18	,748	2,992	80,791
S19	1,000	19	,726	2,904	83,695
S20	1,000	20	,721	2,882	86,577
S21	1,000	21	,706	2,825	89,402
S22	1,000	22	,694	2,774	92,176
S23	1,000	23	,678	2,714	94,890
S24	1,000	24	,660	2,641	97,530
S25	1,000	25	,617	2,470	100,000

PC extracted 3 factors.

Factor Matrix:

	Factor 1	Factor 2	Factor3
S1	5,599E-02	,551	-,159
S2	,553	,200	-,252
S3	,516	,175	4,856E-02
S4	,585	6,861E-02	-7,473E-03
S5	,421	,221	,268
S6	,154	,324	,241
S7	,490	,186	,164
S8	-9,249E-02	,238	,489
S9	,517	,177	,234
S10	,448	,396	,109
S11	,170	,590	,143
S12	,278	,526	,152
S13	,157	,575	,118
S14	,161	,496	,174
S15	,115	,511	5,520E-02
S16	,198	,357	,263
S17	-1,431E-02	,292	,161
S18	,193	,239	,216
S19	,439	,206	,137
S20	,110	,255	,107
S21	,582	3,609E-02	,262
S22	,382	5,531E-02	,336
S23	,203	7,786E-02	,553
S24	,154	,193	,477
S25	,400	5,388E-02	,413

Final Statistics:

Variable	Communality	Factor	Eigenvalue	Pct of Var	Cum Pct
S1	,332	1	5,112	20,448	20,448
S2	,409	2	1,274	5,098	25,546
S3	,300	3	1,014	4,055	29,601
S4	,347				
S5	,298				
S6	,187				
S7	,301				
S8	,304				
S9	,353				
S10	,370				
S11	,397				
S12	,377				
S13	,369				
S14	,302				
S15	,277				
S16	,236				
S17	,112				
S18	,141				
S19	,254				
S20	8,825E-02				
S21	,409				
S22	,262				
S23	,354				
S24	,288				
S25	,333				

APPENDIX B

Inter Item Correlation Matrices of the Turkish and Social Sciences Subtests of The MONE- Secondary Education Institutions Student Selection And Placement Exam

APPENDIX B 1

INTER ITEM CORRELATION MATRICES OF THE TURKISH SUBTEST

Correlation Matrix

	T1	T2	T3	T4	T5
T1	1,0000				
T2	,0510	1,0000			
T3	,0399	,1823	1,0000		
T4	-,0090	-,0162	,0188	1,0000	
T5	,0015	,0344	,0570	,0076	1,0000
T6	,0304	,0334	,0219	,0043	-,0078
T7	-,0039	,0890	,0536	-,0313	,0306
T8	-,0041	,0288	,0148	,0023	-,0776
T9	,0025	,0881	,0335	-,0122	-,0128
T10	,0108	,0440	,0435	,0022	-,0124
T11	-,0354	-,0192	-,0044	-,0150	,0353
T12	-,0138	,0183	-,0252	-,0199	,0023
T13	,0239	,1643	,1323	-,0625	,0312
T14	-,0102	,0508	,0249	-,0211	-,0095
T15	-,0268	-,0166	-,0012	-,0002	,0079
T16	-,0103	-,0036	-,0072	-,0048	-,0105
T17	-,0113	,0129	,0093	-,0485	-,0101
T18	-,0448	,0424	,0210	-,0350	-,0030
T19	-,0232	,0361	,0405	-,0301	-,0164
T20	,0340	,0421	,0403	-,0369	,0061
T21	-,0088	,1194	,0764	-,0215	-,0104
T22	,0240	,0243	,0535	-,0035	-,0022
T23	,0038	,1006	,0904	-,0549	,0017
T24	-,0385	-,0786	-,1007	,0305	-,0027
T25	,0192	,0710	,0589	-,0343	,0010

	T6	T7	T8	T9	T10
T6	1,0000				
T7	,0149	1,0000			
T8	,0184	-,0146	1,0000		
T9	-,0205	,0174	,0083	1,0000	
T10	-,0009	,0804	-,0003	,0695	1,0000
T11	-,0378	-,0142	-,0462	,0147	,0044
T12	-,0288	,0520	-,0460	-,0229	-,0054
T13	,0583	,0980	-,0072	,0573	,0467
T14	,0054	,0291	,0201	,0292	,0291
T15	-,0164	-,0623	,0105	-,0342	-,0303
T16	-,0391	-,0186	-,0076	,0132	-,0062
T17	-,0413	,0292	-,0019	-,0027	,0081
T18	,0163	,0612	,0421	,0084	-,0250
T19	-,0380	,0222	,0296	,0592	,0169
T20	-,0143	,0538	-,0139	-,0116	,0116
T21	,0315	,0569	,0520	,0507	,0412
T22	-,0342	,0377	-,0026	,0112	-,0132
T23	,0111	,0961	-,0014	,0546	,0532
T24	-,0238	-,0595	-,0215	-,0753	-,0275
T25	,0055	,0642	-,0006	-,0048	-,0055

	T11	T12	T13	T14	T15
T11	1,0000				
T12	,0173	1,0000			
T13	-,0176	,0186	1,0000		
T14	-,0297	-,0088	-,0091	1,0000	
T15	-,0058	,0086	-,0011	-,0021	1,0000
T16	,0181	-,0328	,0095	-,0281	-,0121
T17	-,0008	,0334	,0123	,0251	-,0271
T18	-,0230	,0083	,0144	,0131	,0158
T19	-,0112	,0167	,0349	,0196	-,0076
T20	,0210	-,0033	,0498	-,0048	-,0416
T21	-,0241	,0306	,1046	,0187	-,0228
T22	,0150	-,0156	,0158	-,0054	,0043
T23	,0003	,0071	,1512	,0515	-,0176
T24	-,0081	-,0194	-,0950	-,0552	-,0172
T25	-,0042	-,0145	,0877	-,0130	-,0089

	T16	T17	T18	T19	T20
T16	1,0000				
T17	-,0019	1,0000			
T18	-,0130	,0036	1,0000		
T19	,0304	,0550	,0053	1,0000	
T20	,0139	,0607	-,0090	,0586	1,0000
T21	-,0165	-,0029	,1054	,0284	,0590
T22	-,0102	-,0183	-,0341	-,0149	,0049
T23	,0510	,0425	,0590	,1098	,0515
T24	-,0403	-,0050	-,0635	-,0340	-,0585
T25	-,0071	,0151	,0271	-,0064	,0141

	T21	T22	T23	T24	T25
T21	1,0000				
T22	,0469	1,0000			
T23	,1285	,0430	1,0000		
T24	-,0551	-,0010	-,2184	1,0000	
T25	,0521	,0402	,1066	-,0127	1,0000

APPENDIX B 2

INTER ITEM CORRELATION MATRICES OF THE SOCIAL SCIENCES SUBTEST

Correlation Matrix:

	S1	S2	S3	S4	S5
S1	1,0000				
S2	,0563	1,0000			
S3	-,0561	,0917	1,0000		
S4	-,0192	,1105	,1351	1,0000	
S5	-,0046	,0513	,1311	,1154	1,0000
S6	-,0315	-,0008	,0065	,0381	,0841
S7	-,0315	,0539	,1130	,1352	,1074
S8	-,0305	-,0416	-,0120	-,0277	-,0171
S9	-,0504	,0718	,1011	,1085	,1120
S10	,0020	,0896	,0769	,0715	,0415
S11	,0263	-,0341	-,0197	,0195	,0250
S12	,0098	-,0490	,0011	-,0062	-,0157
S13	-,0069	-,0299	,0035	-,0065	-,0199
S14	-,0331	-,0332	-,0273	-,0014	-,0535
S15	,0096	,0263	-,0117	-,0045	-,0112
S16	-,0201	-,0062	-,0054	-,0266	,0058
S17	-,0063	-,0547	-,0304	-,0472	-,0337
S18	-,0082	-,0310	,0011	,0246	,0196
S19	-,0343	,0570	,0474	,0897	,0321
S20	-,0395	-,0276	-,0288	,0179	-,0266
S21	-,0529	,0438	,0835	,1327	,0774
S22	-,0653	,0270	,0041	,0232	,0525
S23	-,0461	-,0153	,0094	-,0030	,0296
S24	-,0622	-,0106	-,0290	-,0100	-,0025
S25	-,0466	,0213	,0206	,0433	,0293

	S6	S7	S8	S9	S10
S6	1,0000				
S7	,0214	1,0000			
S8	-,0115	-,0294	1,0000		
S9	,0224	,1290	-,0112	1,0000	
S10	-,0089	,0836	-,0312	,1144	1,0000
S11	-,0129	-,0413	,0008	-,0247	-,0204
S12	-,0141	-,0411	,0053	-,0336	,0209
S13	,0051	-,0323	-,0132	-,0246	,0194
S14	,0185	-,0620	,0056	-,0508	-,0018
S15	-,0728	-,0226	-,0067	-,0485	-,0171
S16	,0018	,0322	,0217	,0064	-,0144
S17	-,0158	-,0261	-,0008	-,0171	-,0114
S18	-,0100	,0200	-,0033	,0177	-,0237
S19	-,0313	,0505	-,0029	,0365	,0289
S20	,0154	-,0491	-,0059	-,0004	-,0767
S21	-,0050	,0726	-,0544	,1223	,0240
S22	,0164	,0348	-,0573	,0507	,0172
S23	-,0263	,0007	,0030	,0272	-,0011
S24	-,0330	-,0262	-,0060	,0004	-,0273
S25	-,0134	,0378	-,0305	,0459	,0375

	S11	S12	S13	S14	S15
S11	1,0000				
S12	,0643	1,0000			
S13	-,0013	,0465	1,0000		
S14	-,0066	,0517	,0544	1,0000	
S15	,0051	,0143	-,0148	-,0129	1,0000
S16	,0295	,0008	-,0084	-,0078	,0293
S17	-,0342	-,0025	-,0342	,0057	-,0130
S18	-,0139	,0238	,0202	,0058	-,0121
S19	-,0221	-,0114	,0176	-,0332	-,0210
S20	,0290	-,0136	,0144	-,0013	,0133
S21	-,0390	-,0071	-,0062	-,0283	,0014
S22	-,0229	-,0389	-,0218	-,0119	-,0220
S23	-,0327	-,0363	-,0463	-,0241	-,0016
S24	-,0288	-,0304	-,0535	-,0177	-,0415
S25	-,0453	-,0308	-,0619	-,0188	-,0191

	S16	S17	S18	S19	S20
S16	1,0000				
S17	,0295	1,0000			
S18	,0542	-,0062	1,0000		
S19	,0299	,0027	,0676	1,0000	
S20	-,0185	,0226	,0112	,0175	1,0000
S21	,0494	-,0146	,0308	,1539	,0770
S22	,0052	-,0228	,0249	,0633	-,0230
S23	-,0171	-,0141	-,0025	,0170	-,0298
S24	-,0129	-,0186	,0055	,0231	-,0114
S25	-,0299	-,0413	-,0200	,0606	-,0099

	S21	S22	S23	S24	S25
S21	1,0000				
S22	,1445	1,0000			
S23	,1019	,1415	1,0000		
S24	,0471	,0696	,0946	1,0000	
S25	,1264	,1493	,1678	,1432	1,0000