THE PREDICTIVE VALIDITY OF BAŞKENT UNIVERSITY PROFICIENCY EXAM (BUEPE) THROUGH THE USE OF THE THREE-PARAMETER IRT MODEL'S ABILITY ESTIMATES

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

BY

OYA PERİM YEĞİN

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

JUNE 2003

ABSTRACT

THE PREDICTIVE VALIDITY OF BAŞKENT UNIVERSITY PROFICIENCY EXAM (BUEPE) THROUGH THE USE OF THE THREE-PARAMETER IRT MODEL'S ABILITY ESTIMATES

Yeğin, Oya Perim

M. S., Department of Educational Sciences
Supervisor: Prof. Giray Berberoğlu
June 2003, 135 pages

The purpose of the present study is to investigate the predictive validity of the BUEPE through the use of the three-parameter IRT model's ability estimates.

The study made use of the BUEPE September 2000 data which included the responses of 699 students. The predictive validity was established by using the departmental English courses (DEC) passing grades of a total number of 371 students.

As for the prerequisite analysis the best fitted model of IRT was determined by first, checking the assumptions of IRT; second, by analyzing the invariance of ability parameters and item parameters and thirdly, by interpreting the chi-square statistics.

After the prerequisite analyses, the best fitted model's estimates were correlated with DEC passing grades to investigate the predictive power of BUEPE on DEC passing grades.

The findings indicated that the minimal guessing assumption of the one- and two-parameter models was not met. In addition, the chi-square statistics indicated a better fit to the three-parameter model. Therefore, it was concluded that the best fitted model was the three-parameter model.

The findings of the predictive validity analyses revealed that the best predictors for DEC passing grades were the three-parameter model ability estimates. The second best predictor was the ability estimates obtained from sixty high information items. In the third place BUEPE total scores and the total scores obtained from sixty high information items followed with nearly the same correlation coefficients. Among the three sub-tests, the reading sub-test was found to be the best predictor of DEC passing grades.

Keywords: Predictive Validity, Item Response Theory, Item Characteristic Curves, Ability Parameter Estimates, Başkent University English Proficiency Exam.

ÖZ

BAŞKENT ÜNİVERSİTESİ İNGİLİZCE YETERLİK SINAVI'NIN (BÜİYS) MADDE TEPKİ KURAMI'NIN (MTK) ÜÇ PARAMETRELİ MODELİNİN KULLANIMIYLA ELDE EDİLEN YETENEK KESTİRİMLERİNİN YORDAMA GEÇERLİĞİ

Yeğin, Oya Perim

Yüksek Lisans, Eğitim Bilimleri Bölümü
Tez Yöneticisi: Prof. Giray Berberoğlu
Haziran 2003, 135 sayfa

Bu çalışmanın amacı, MTK'nın üç parametreli modelinden elde edilen yetenek kestirimleriyle BÜİYS'nın yordama gerçerliğini araştırmaktır.

Bu çalışmada 699 öğrencinin cevaplarından oluşan BÜİYS 2000 Eylül ayı verileri kullanılmıştır. BÜİYS'nın yordama geçerliği 371 öğrenciden elde edilen bölüm İngilizce dersleri (BİD) geçme notlarının kullanımıyla sağlanmıştır.

Ön hazırlık çalışmalarında MTK'ya en iyi uyan modeli saptamak için MTK'nın sayıtlılarının karşılanıp karşılanmadığına bakılmış, farklı madde gruplarından elde edilen yetenek parametreleri ve farklı öğrenci örneklemlerinden elde edilen madde parametreleri her bir alt testte karşılaştırılmış ve son olarak ki-kare istatistiği yorumlanmıştır.

Ön hazırlık çalışmalarının sonucuda belirlenen en iyi uyan modelin kestirimleriyle BİD geçme notlarının ilişkisine BÜİYS'nin yordama geçerliğini araştırmak amacıyla bakılmıştır.

Bulgular, şans faktöründen arınık olma sayıtlısının karşılanmadığını göstermektedir. Ayrıca, ki-kare istatistikleri de üç parametreli modele daha çok uygunluk göstermiştir. Bu nedenle, üç parametreli modelin BÜİYS verilerine daha uygun olduğuna karar verilmiştir.

BÜİYS'nın yordama geçerliği bulguları ise, BİD geçme notlarının en iyi belirleyicisinin üç parametre yetenek kestirimleri olduğunu göstermiştir. İkinci iyi belirleyicinin en yüksek bilgi veren altmış sorudan elde edilen yetenek kestirimleri olduğu gözlenmiştir. Üçüncü sırayı BÜİYS ham puanları ve en yüksek bilgi veren altmış sorudan elde edilen ham puanların paylaştığı görülmüştür. Üç alt test arasında, okuma alt testinin BİD geçme notlarının en iyi belirleyicisi olduğu saptanmıştır.

Anahtar Kelimeler: Yordama Geçerliği, Madde Tepki Kuramı, Madde Karakteristik Eğrisi, Yetenek Parametresi Kestirimleri, Başkent Üniversitesi İngilizce Yeterlik Sınavı.

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to my thesis supervisor, Prof. Dr. Giray Berberoğlu for his invaluable support and guidance he has devoted to this study.

I am also grateful to Prof. Dr. Hüsnü Enginarlar and Assist. Prof. Oya Yerin Güneri for their valuable suggestions and comments.

My special thanks also go to Murat Güngör for his technical assistance and Taner Yapar for his ongoing encouragement and assistance in this study.

Finally, I would like to express my gratitude to my husband Tolga Yeğin, my mother and father Seçkin and Altan Eray, and my friends Neşe Ekici, Yeliz Canpolat for their continual support and encouragement throughout the study.

TABLE OF CONTENTS

LIST OF TABLES

TABLE

LIST OF FIGURES

FIGURE

CHAPTER 1

INTRODUCTION

1.1. Background of the Study

Tests, which have an important role in language learning, are generally constructed for reinforcing learning, motivating students and above all, for obtaining an assessment of student's performance in language (Heaton, 1988).

Bachman (1990, p.20) defines tests as "instruments designed to elicit specific samples of individuals' language behaviour."

Our purpose and the type of information we want to collect determine the type of test that we administer to students.

Besides Aptitude tests that Heaton (1988) mentions; Alderson, Clapham and Wall (1995), categorize tests under five major headings: Placement tests, Progress tests, Achievement tests, Diagnostic tests and Proficiency tests.

Proficiency tests are mainly constructed for measuring the ability of examinees. They are not designed by using a course content and they try to test the examinees' ability on what they have to perform so that they can be considered as proficient in the language (Hughes, 1989). In addition to proficiency tests for specific purposes, there are also proficiency tests for assessing general language proficiency. The second type of proficiency tests can be constructed in preparatory schools of universities.

Many universities have English Preparatory Schools in Turkey regardless of whether the university is a Turkish medium school. Students study in the preparatory school for a year and then sit a proficiency exam the result of which determines whether they pass to freshmen or not. The systems that are used in these universities are generally similar. One such university is Başkent University (BU).

At Başkent University Preparatory School, most test types mentioned above are used for serving different purposes. One such test is the Başkent University Proficiency Exam (BUEPE), which is administered twice a year.

Each student who registers for studying at Başkent University is required to be at a certain level of English proficiency to start as freshmen. Since the ones who do not succeed to pass the proficiency exam study in the preparatory school for one year, the instrument chosen for this purpose is expected to be not only reliable but also valid.

A test is considered to be valid if it measures what it purports to measure according to Hughes (1989). The concept of validity can be approached from several dimensions, each of which displays a different perspective to the concept of validity.

Hughes (1989) categorizes validity under four main headings which are: Content validity, Construct validity, Face validity and Criterion-related validity (concurrent and predictive validity).

Content validity is related to how adequately a test covers representative behaviour which it is concerned with (Hughes, 1989). Having a table of specifications of the course content can be helpful in judging the content validity of a test; however, everything in the table of specifications may not be expected to appear in the test. A test which reflects the major aspects in a table of specifications is likely to measure what it claims to measure; thus, have content validity (Hughes, 1989).

Gronlund (1985, p.58 in Alderson, Clapham & Wall, 1995) defines construct validity as "how well test performance can be interpreted as a meaningful measure of some characteristic or quality." Alderson, Clapham and Wall (1995) mention different ways for establishing the construct validity of a test. Some of these are assessing whether the test is based on its underlying theory, internal correlations among different components of a test, multitrait-multimethod analysis, convergent-divergent validation and factor analysis.

Face validity that is not considered to be a scientific concept by many authors; however, is regarded as quite important. A test is considered to have face validity if it looks acceptable to examinees, teachers or education authorities (Hughes, 1989). Poor items, unclear instructions or unrealistic time limits may negatively affect the face validity of a test (Alderson, Clapham & Wall, 1995).

Criterion-related validity, according to Hughes (1989), demonstrates whether there is a relationship between the results of a test and some other independent and highly dependable criterion of the examinee's ability. There are two types of criterion-related validity: concurrent and predictive.

In concurrent validity the test scores are compared with another measure for the same examinees and both measures are administered at about the same time (Alderson, Clapham & Wall, 1995).

The second type of criterion-related validity, predictive validity according to Alderson, Clapham and Wall (1995), can be tested for the same examinees by comparing a test score with another measure, which is collected after the test has been given. It is common to look for predictive validity in a proficiency test because predictive validity analyses are important in checking whether the main objective of the proficiency exam, which is to evaluate an examinee's ability to successfully perform in a future course, is achieved or not.

A study named as Predictive validity (2002, p.2), suggests using the following questions as a guideline in a study of predictive validity:

1. What criterion measure(s) have been used in evaluating validity? What is the rationale for choosing this measure? Is this criterion measure appropriate?
2. Is the distribution of scores on the criterion measure adequate?
3. What is the basis for the statistics used to demonstrate predictive validity?

In order to investigate the relationship between BUEPE scores and DEC grades the predictive validity of BUEPE should be examined.

Besides collecting predictive validity evidence by correlation analyses of BUEPE total scores with first and second semester DEC grades, using

Item Response Theory rather then Classical Test Theory indices could provide more information for improving the test's predictive validity.

Statistical analyses which are carried out in order to describe the effectiveness of test items are called item analyses. These analyses enable us to construct more effective test items which also improve the quality of our tests. Two important measurement frameworks by which we can carry out item analyses are Classical Test Theory (CTT), which has satisfied measurement specialists for most of this century and Item Response Theory (IRT), which has been introduced in the past decades in order to fulfill the testers' dissatisfied needs that have arisen due to the weaknesses of CTT.

The most important advantage of CTT that has maximized its usage for long years has been its easy to apply theoretical assumptions (Fan, 1998). Not only CTT but also IRT makes use of item difficulty and item discrimination indices. Item difficulty is the percentage of students to answer an item correctly (Anastasi, 1982). An item's difficulty is directly related to how easy the group of examinees view the item. Thus, when the item difficulty is a large value the item is considered to be easy for that group. The smaller a value gets, the harder the item is. On the other hand, item discrimination index shows how much an item distinguishes among students at different ability levels. If the item is a discriminating item, high achievers are expected to answer the item correctly whereas low achievers fail to answer correctly.

However, the interpretation of the indices that makes the CTT so easy to apply are the very qualities that limit its usage. The major limitation of CTT which has given rise to IRT is its results being group (sample) dependent and test dependent. In addition, the inseparability of the examinee and test characteristics makes it even harder to interpret whether a score that examinees get from a test is a result of their ability or the hard/easy characteristics of an item (Hambleton, Swaminathan &Rogers, 1991). Also, in the CTT context it is not easy to compare two individuals who take different tests because the scores that two examinees get will be on different scales. To compare examinees who take the same or parallel forms of a test may not still be easy at all if examinees are at different ability levels.

In this respect, according to Hambleton, Swaminathan and Rogers (1991), the use of IRT has considerable advantages over CTT in terms of separating an individual's characteristics from that of the test item. In other words, item difficulty and discrimination are not group (sample) dependent in IRT. IRT assumes that an examinee's response to an item is determined by latent traits or abilities. Therefore, the latent trait or the ability of examinees determine their performance on a single item. Rather than assigning a true score to the examinee as in CTT, IRT assigns an ability score θ to the examinee. The score of an examinee is not affected from neither how hard or easy the item is nor whether the group that the examinee is tested in is different from another group which may consist of high and low achievers. This is the basic property of IRT models which means that item parameters are not dependent on the ability level of examinees and similarly, the ability level of an examinee is not dependent on the set of test items administered (Hambleton, Swaminathan & Rogers, 1991). Moreover, instead of giving test level information as in CTT, IRT focuses on item level information. One last difference between CTT and IRT is related to establishing the reliability. Having parallel tests is necessary for establishing reliability in CTT whereas in IRT test forms do not have to be absolutely parallel.

Furthermore, IRT makes use of ICCs which demonstrate the relationship between the probability of correctly answering an item and the ability or trait underlying the performance on the items. This implies that an examinee's probability of correctly answering a question increases if the level of ability increases (Hambleton, Swaminathan & Rogers, 1991).

According to Hambleton, Swaminathan and Rogers (1991), in IRT it is possible to find out an item's individual contribution to the whole test without knowing the characteristics of the other items in the test. However, in CTT it is not possible to interpret the contribution of a single item to test reliability and item discrimination because an individual item in a test is dependent on the other items on the test. The total score is directly related with the items chosen for a test. Therefore, if one of the items in the test is changed the item and test indices change. As a result of this advantage that IRT models share over CTT, terms known as item information and test information can be used for making inferences of the contribution of each item in a test.

The superior characteristics of IRT over CTT are the very characteristics that make it suitable for predictive validity studies. Separating the individual’s characteristics from that of the items, providing Item Information Functions (IIF) and Test Information Functions (TIF) in addition to computing ability scores for each individual make IRT estimates suitable for predictive validity studies. Therefore, if estimates of IRT such as ability estimates for each individual or IIFs are used in predictive validity studies this can enhance the predictive validity coefficient obtained.

There are three main IRT models named as one-, two-, three-parameter models which are suitable for dichotomously scored items. A computer program, BILOG, can be used for running these analyses. These models differ in terms of the number of parameters that they use. The one-parameter model uses only item difficulty (b) parameter whereas the two-parameter model takes item difficulty (b) parameter and item discrimination (a) parameter into consideration. Three-parameter model makes use of item difficulty (b) parameter and item discrimination (a) parameter in addition to a third parameter which is the guessing or chance parameter (c). Deciding on which model to use is a matter of model-data fit, which is an important issue in IRT.

According to Hambleton, Swaminathan and Rogers (1991), besides its different models IRT has important assumptions to be satisfied. These are unidimensionality, local independence, equal discrimination indices, minimal guessing and nonspeeded test administration. Furthermore, IRT has expected model features such as invariance of ability parameters estimates and invariance of item parameter estimates.

In conclusion, the type of evidence we want to collect determines the test type we would use. Language proficiency tests are used for determining whether a student has reached a given level of language ability to perform successfully in future courses. In order to see if a proficiency test accurately predicts future performance of examinees, predictive validity analyses can be performed. To see how the test is functioning or whether it needs improvement or even how much information it gives in terms of the ability levels of students requires other detailed analyses such as IRT analyses.

As opposed to CTT, IRT has some obvious advantages and is more grounded in theory. By separating item characteristics from examinee characteristics, IRT solves the problem of group dependency and item dependency. Having three different models for different circumstances is important because different data sets can be addressed. Finally, models can be selected on basis of basic assumptions.

1.2. Statement of the Purposes

The main purpose of this study is to examine the predictive validity of Başkent University English Proficiency Exam (BUEPE) by using IRT based ability estimates.

The analyses in this study are carried out in two stages. First, the fit of BUEPE data to IRT models is investigated. Secondly, the predictive validity of BUEPE is examined by using the IRT based ability estimates in the first stage.

1.3. Statement of the Main and the Subproblems

As the prerequisite analysis, the best fitted model of IRT was determined. In this process the following steps were carried out for the data obtained on BUEPE.

1.Whether the BUEPE data met the assumptions of IRT was investigated.

1. 1. The unidimensionality of the data was investigated.

1. 2. Whether the data met the the local independence assumption was examined.

1. 3. Whether the data met the equal item discrimination indices assumption for the one parameter model was examined.

1. 4. Whether the data met the minimal guessing assumption of the one parameter and two-parameter model was explored.

1. 5. Whether BUEPE was a non-speeded test was investigated.

2. The invariance of ability parameter estimates of the one-, two-, three-, parameter models obtained across different samples of test items was interpreted.

2.1. Whether the ability parameters of the one-, two-, three-, parameter models were invariant across hard and easy items in BUEPE was examined.

2.2. Whether the ability parameters of the one-, two-, three-, parameter models were invariant across the first fifty and the second fifty items in BUEPE was examined.

3. The invariance of item parameter estimates of the one-, two-, three-, parameter models obtained across different samples of examinees was interpreted.

3.1. Whether the item parameters of the one-, two-, three-, parameter models were invariant across odd and even cases in BUEPE was examined.

4. How well the simulated test results of the one-, two-, three-, parameter model predicted the actual test results was investigated.

4.1. How well the observed distribution of the BUEPE 2000 scores fitted the theoretical distribution of the one-, two-, three-, parameter IRT models was examined.

After deciding on the best fitted model, the following research questions were investigated.

5. Do the ability estimates obtained through the use of the IRT based model on BUEPE 2000 predict success in departmental English courses (DEC) at Başkent University?

5.1. How well do BUEPE 2000 total scores predict freshmen first semester DEC passing grades ?

5. 2. How well do BUEPE 2000 total scores predict freshmen second semester DEC passing grades ?

5. 3. How well do ability estimates of the best fitted model predict the freshmen first semester DEC passing grades ?

5. 4. How well do ability estimates of the best fitted model predict the freshmen second semester DEC passing grades ?

5. 5. How well do the total scores obtained by using the sixty highest-information items in the best fitted model predict first semester DEC passing grades ?

5. 6. How well do the total scores obtained by using the sixty highest-information items in the best fitted model predict second semester DEC passing grades ?

5. 7. How well do the ability estimates obtained by using the sixty highest-information items in the best fitted model predict first semester DEC passing grades ?

5. 8. How well do the ability estimates obtained by using the sixty highest-information items in the best fitted model predict second semester DEC passing grades ?

5. 9. How well do the total scores obtained by using the grammar subtest in the best fitted model predict freshmen first semester DEC passing grades?

5. 10. How well do the total scores obtained by using the grammar subtest in the best fitted model predict freshmen second semester DEC passing grades?

5. 11. How well do the ability estimates obtained by using the grammar subtest in the best fitted model predict freshmen first semester DEC passing grades?

5. 12. How well do the ability estimates obtained by using the grammar subtest in the best fitted model predict freshmen second semester DEC passing grades?

5. 13. How well do the total scores obtained by using the reading subtest in the best fitted model predict freshmen first semester DEC passing grades?

5. 14. How well do the total scores obtained by using the reading subtest in the best fitted model predict freshmen second semester DEC passing grades?

5. 15. How well do the ability estimates obtained by using the reading subtest in the best fitted model predict freshmen first semester DEC passing grades?

5. 16. How well do the ability estimates obtained by using the reading subtest in the best fitted model predict freshmen second semester DEC passing grades ?

5. 17. How well do the total scores obtained by using the vocabulary subtest in the best fitted model predict freshmen first semester DEC passing grades?

5. 18. How well do the total scores obtained by using the vocabulary subtest in the best fitted model predict freshmen second semester DEC passing grades?

5. 19. How well do the ability estimates obtained by using the vocabulary subtest in the best fitted model predict freshmen first semester DEC passing grades?

5. 20. How well do the ability estimates obtained by using the vocabulary subtest in the best fitted model predict freshmen second semester DEC passing grades?

1.4 Significance of the study

Using IRT estimations rather than Classical Test Theory Models in conducting predictive validity studies for English proficiency tests is important in many respects:

1. This study can be a starting point for other studies which investigate language proficiency tests by means of IRT.

2. The construction of high-stake exams such as proficiency exams needs a lot of expertise. The items constructed should serve directly the purpose of the proficiency test. Thus, if such importance is attached to a test or items of a test, then items which give utmost information about an examinee should be selected by using IRT's ICCs and item information functions.

3. Because of the variety of item parameters like item difficulty, item discrimination, ICC and item information function, item level information can be backed up with many different types of statistical support.

4. The use of items which are tested previously and which are definite to produce the desired results can increase the consistency of scores obtained from a test.

5. This study underlines one of the main advantages of IRT which is the invariance property of item parameters. The items used in the test produce positive results in all types of examinee groups rather than in a specific examinee group.

6. This study can initiate creating Item Banks in English preparatory schools because of not only the ease of using IRT models but also the quality of information they provide. Having an item bank full of a variety of items to use may ease the burden of test constructors in terms of time and energy.

CHAPTER 2

REVIEW OF LITERATURE

This chapter presents the basic concepts on the predictive validity of English language tests and background of Item Response Theory. Finally, a review of literature both on research on predictive validity of English language tests and Item Response Theory is provided.

2.1 Predictive Validity

Our purpose and the type of information we want to collect determine the type of test that we administer to students.

Heaton (1988) defines Aptitude Tests as aiming at measuring a student’s probable ability to perform in a foreign language. By measuring the performance of the student in an artificial language, these tests try to predict students’ ability to succeed in a given foreign language, prior to learning that specific language.

In addition, Alderson, Clapham and Wall (1995), categorize tests under five major headings:

1. Placement Tests are constructed to correctly place students to a class or course according to their level of language ability. These tests generally cover the content of the future course.

2. Progress Tests measure students as to newly learnt materials in the course. They are given at regular intervals to see if students have shown progress in what they have learnt at a given interval of time. Similarly, it provides information as a teaching device.

3. Achievement Tests are similar to progress tests because both are based on the course content. However, achievement tests are more formal than progress tests and they cover a longer period of teaching in the course. They are generally given at the end of the course and represent all the material that has been taught.

4. Diagnostic Tests are prepared to find out student difficulties in different areas of language. Achievement, proficiency and even progress tests can be used for diagnostic purposes, to identify students weaknesses and strengths.

5. Proficiency Tests do not test what students have previously learned. Rather they aim to find out the ability of students to see if they will be able to perform successfully in future courses. Since they test students coming from various language backgrounds, they are not based on a syllabus or course content.

Hughes (1989) and Alderson, Clapham and Wall (1995) mention that it is possible to talk about two types of proficiency tests. The first type involves being proficient for specific purpose such as a job or a course. The content of such a proficiency test focuses on what is necessary for that specific purpose. The second type considers the term proficiency in a more general sense. These tests are designed to test if the examinee has reached a certain level of language proficiency.

This second type of proficiency tests are generally constructed by examining bodies or testing boards which are not dependent on schools. Cambridge First Certificate Examination or TOEFL prepared by the Educational Testing Service can be an example. However, besides such testing boards, university preparatory schools also administer their own general proficiency tests.

The main reason for administering BUEPE is to examine if the capacity of students is sufficient to attend and succeed in Departmental English Courses (DEC) which is parallel with the main goal of proficiency exams. However, whether BUEPE is successful in this respect has not yet been studied.

Furthermore, since we are living in a world in which ideas about language proficiency are changing at a great speed, test developers and researchers have to keep up with recent developments and thus, make improvements in their test method or items of the test (Alderson, Clapham & Wall, 1995) by using various analyses.

In order to improve a test, validity analyses can be conducted. Validity analyses that can be carried out include; Content validity, Criterion-related validity (concurrent and predictive validity), Construct validity and Face validity (Hughes, 1989).

One of these analyses, predictive validity analysis, can especially be useful in proficiency testing situations. Crocker and Algina (1986, p.224) define predictive validity as “the degree to which test scores predict critierion measurements that will be made at some point in the future.” In terms of a proficiency test, predictive validity refers to the extent to which a test can be appropriately used to draw inferences regarding proficiency (Predictive validity, 2002). These inferences in our case are about the success in departmental English courses. In order to make sure that the BUEPE is functioning as intended as a proficiency test, BUEPE scores should be correlated with an appropriate criterion measure. According to Bachman (1990, p.251-252), there should be no “mismatch between what the ability the test appears to measure and the performance we are trying to predict”. Therefore, choosing first and second semester DEC passing grades would be an accurate criterion measure to estimate the predictive validity of BUEPE.

In addition to the concern of choosing an appropriate criterion measure, one other common consideration in predictive validity studies is related to the sample size. It is only possible to use a part of the actual test population in predictive validity studies (Alderson, Clapham & Wall, 1995). Since students who are below the cut scores are not available to be included in validation studies, this presents a difficulty for predictive validity research (Smith & Hambleton,1990 in Mikitovics & Crehan, 2002). The spread of students’ scores is reduced and this leads to lower correlations between test scores and other measures.

Correlation and scatterplot analyses can be employed in order to prove predictive validity, because these analyses are considered to be sound indicators of relations between measures according to Tabachnick & Fidel, (1996).

The correlation coefficient used in predictive validity studies is called a validity coefficient. Crocker and Algina (1986) define validity coefficient as a correlation coefficient between two variables: a test score and a criterion score. As an answer to the question of what an acceptable validity coefficient can be is answered by Hughes (1989) who says that a validity coefficient around 0.40 is the highest correlation expected in predictive validity studies. Cronbach (1990), states that a correlation as low as 0.30 may definitely have a practical value and correlations below that value may help improve decisions. The reason for having such low validity coefficients is because students who are below the cut scores in the test are not included in predictive validity studies. Still, Cronbach (1990) argues that test validities ranging from 0.30 to 0.50 contribute considerably despite the fact that they may wrongly predict many students. Whereas, validity coefficients around .40 can be accepted as sufficient with any suitable external measure, higher validity coefficients may be expected among the measure and the criterion if both are measuring similar traits. For example, an English proficiency exam is expected to yield a higher predictive validity coefficient across the grades obtained from a freshmen English course.

Predictive validity studies can be carried out by using statistics obtained from various analyses. For almost the first half of the previous century CTT served the needs of psychometricians; however, new measurement systems arose due to the lack of satisfaction from the characteristics of indices. IRT claims to fulfill the needs of test developers mainly because the item characteristics and examinee characteristics can be individually evaluated.

Since IRT analyses permits a wider range of comparisons implementing a predictive validity study by means of the indices obtained from IRT seems quite reasonable.

2.2. Background of Item Response Theory

Since the beginning of the last century test theory has been developing at a great speed. The first decades of the century witnessed the emergence of CTT and its many well-known concepts such as true score, item difficulty or item discrimination indices. The second half of the century; however, witnessed a move from CTT to modern test theory which brought a new perspective to the field of testing and provided solutions to the dissatisfactions encountered in the CTT. The inseparability of item and examinee characteristics in CTT gave way to a theory, later to be known as IRT, dwelling on item and examinee characteristics individually.

Binet and Simon set the first example of an item-based test theory (1916) by presenting the relationship between proportion of correct response to an item and chronological age in a tabular fashion in their intelligence test. Terman (1916) and Terman and Merrill (1937) used the same approach in order to plot the curves of two variables. Today's terminology defines these plots as ICCs. Test theory based on the items of a test begins with Lawly (1943) who redefines the true score as for the items of a test and demonstrates how to get the parameters of ICCs by using maximum likelihood estimates. Lord (1952) developed Lawly's work and showed that the parameters of ICCs obtained from test items could be used for explaining concepts of CTT. The ideas of Lawly and Lord formed the basis of the modern test theory which we now know as IRT.

This new test theory was named as Latent Trait Theory from 1950s to 1970s. Then, for a short time, it was referred to as Item Characteristic Curve Theory, because of the important role of the curves in the theory. Recently, IRT has been accepted as a sufficient term because the theory is based on an examinee's response to test items.

In CTT, the term true score is used to define the observed performance of an examinee on a test. However, for defining the performance of an examinee IRT exploits the term ability or latent trait. According to IRT, an examinee should have the achievement variable to answer an item correctly. This achievement variable, also called ability or latent trait, in IRT reflects an underlying hypothetical variable which exists but is not observed (Baker, 1992). If this variable, the latent trait, is present in the examinee, the examinee will demonstrate it by correctly responding to an item. Thus, the examinee's correct response would directly show the individual's ability or proficiency level independent of the characteristics of the items in the test.

In IRT a unique plot called Item Characteristic Curve (ICC) can be created for each item. This curve shows the probability of correct response in relation to an individuals' ability (Item Analysis, 2002). The shape of the ICC reflects the influence of the three factors according to Item Analysis, 2002, p.2 :

> Increasing the *difficulty* of an item causes the curve to shift right - as candidates need to be more able to have the same chance of passing.

> Increasing the *discrimination* of an item causes the gradient of the curve to increase. Candidates below a given ability are less likely to answer correctly, whilst candidates above a given ability are more likely to answer correctly.
>
> Increasing the *chance* raises the baseline of the curve - for instance for a MCQ with four possible answers, candidates of even the lowest ability will have a one in four chance of getting the right answer and so the baseline is 0.25

If the level of latent trait or ability increases, this causes the probability of correct response to increase. The advantage of ICC when compared to item difficulty and discrimination indices of the CTT is that it shows the direct relationship between the latent trait and probability of correct response (Crocker & Algina, 1986).

Another strength of IRT lies in its making use of item and test information functions. Item and test information functions demonstrate how well the single items and the test as a whole estimate ability along the ability scale (Baker, 1992). The test information function is the total of all the item information functions for all the items on a test (Crocker & Algina, 1986). By using the item information and test information functions, a more comprehensive analysis of the items can be made and different tests can be compared. Depending on the purpose of the test, a more precise estimate can be achieved by means of item information function.

Hambleton, Swaminathan and Rogers (1991, p.91) describe the general trend in item information as follows:

a) information is higher when the b value is close to θ than when the b value is far from θ,
b) information is generally higher when the a parameter is high, and

c) information increases as the c parameter goes to zero.

Test developers can select test items that fulfill their needs for a test that they construct since item information function is used for evaluating the contribution of each item in estimating examinees ability. Crocker and Algina (1986) indicate that for constructing a useful test, determining the regions of the ability scale which is needed for discrimination among scale points is important. Then, a test which locates examinees in the desired regions can be developed. The points at which the test provides more information is helpful in discriminating among individuals who have ability scores falling in these regions (Crocker & Algina, 1986). Therefore, if an item gives more information at one end of the ability continuum it may not be useful in providing information at the other end of the continuum (Hambleton, Swaminathan & Rogers, 1991). Furthermore, the more a test gives information at a particular ability level, the closer the ability estimates focus around the true ability level and this results in precise estimates (Baker, 1992).

Hambleton, Swaminathan and Rogers (1991) underlines that the information functions in a test is related to ICCs because the fit of the ICCs to the test data determines the usefulness of the item information functions. Having a poor fit of the data and the ICCs give rise to misleading item statistics and item information functions. Also, because higher c parameters result in lower item information researcher may be inclined to use one or two parameter models which take the c parameter as zero in order to have higher item information functions. However, this procedure can only create correct results if the ICCs of the one and two parameter models fit the data as mentioned above.

Test information function is the total of all the item information function of the items in the test. It measures how precisely the total number of items in the test estimate ability at any point on the ability scale (Baker, 1992). This role of the test information function is quite similar to that of reliability in CTT. Test information function is advantageous in that, it demonstrates how well ability is estimated at every ability level. However, reliability coefficient in CTT is a global measure of consistency (Baker, 1992).

Since the IRT analysis procedures are mathematically complicated and since data processing is involved, it is necessary to carry out the analyses by means of a computer program. LOGIST, BICAL, BILOG and MULTILOG are available computer programs for IRT analyses. In a comparative study with LOGIST (Tang, Way & Carey, 1993) found out that, the item parameters and item characteristic curves estimated by BILOG were closer in magnitude to the "true" parameter values when compared with the LOGIST estimates. Hambleton, Swaminathan and Rogers (1991) also mention some disadvantages of LOGIST estimates. The main advantage of BILOG is that it has consistent item parameter estimates as the number of examinees increases. For the purpose of this study, analyses have been carried out by running BILOG which was developed in 1981 and implemented in the computer program by Mislevy and Bock in 1984 (Hambleton, Swaminathan & Rogers, 1991).

There are three main IRT models named as one-, two-, three-parameter models which are suitable for dichotomously scored items. These models differ in terms of the number of parameters that they use. The one-parameter model, also called the Rasch model and is the simplest of the three models. It uses only item difficulty (b) as a parameter which affects examinees' performance and is limited for this reason. The one-parameter model assumes that all items are equally discriminating and suggests that there is zero probability for answering an item correctly for low ability examinees because it does not take guessing as a factor.

Two-parameter model, takes item difficulty (b) parameter and item discrimination (a) parameter into consideration while carrying out the analyses. Still, this model does not take guessing into consideration but it is more complex than the one-parameter model.

Lastly, three-parameter model makes use of item difficulty (b) parameter and item discrimination (a) parameter in addition to pseudo guessing parameter (c). Among the three models, it is the most complex one and perhaps the one which is the most suitable to real-life because it takes guessing as a factor influencing examinees' performance. According to Crocker and Algina (1986), on multiple-choice items guessing must be taken into consideration.

Each of the models is suitable for different data sets and deciding on which model to use is a major consideration in IRT.

Hambleton and Swaminathan (1985) (in Hambleton, Swaminathan & Rogers, 1991) has recommended three types of evidence for assessing model data fit: 1) the assumptions for the model should be valid for the test data 2) to which degree are item and ability parameters invariant? 3) The model predictions which use real or simulated data should be accurate.

In order to collect evidence on the fit of the model to test data, IRT models require some assumptions to be fulfilled. Hambleton, Swaminathan and Rogers (1991) discuss these assumptions. Unidimensionality is a common assumption of IRT models. To be defined as unidimensional a test should have a prominent factor that explains the performance of an examinee and all other factors should be functionally insignificant (Stark et.al, 2001). Thus, only one ability is measured by the items of the test.

A second assumption is called local independence. According to the local independence assumption, when the ability level is held constant, an examinee's probability of correctly answering any two items in a test is equal (Lord & Novick, 1968) or in other words it is independent of the other items (Hambleton, Swaminathan & Rogers, 1991). If a data set is unidimensional it is said to be locally independent. However, if there is local independence it does not necessitate unidimensionality.

A third important assumption is non-speededness. For a test to be considered as non-speeded, the time limit given for the test should be sufficient for examinees to complete all the items. Thus, speed does not affect the test performance of examinees.

Besides the common three assumptions of all IRT models, there are other assumptions for different models.

Another assumption, for the one-parameter model is equal discrimination indices which requires a homogeneous distribution of item-test score correlations.

Minimal guessing is an assumption of the one- and two- parameter models as the three-parameter model inherently takes guessing into consideration. The probability of a low-ability examinee to give a correct response to the most difficult items should be close to zero.

Invariance of item and ability parameters is one of the basic premises of IRT. The invariance of item parameters implies that the a, b or c parameters are not influenced from subgroups of the test population such as males and females or high and low ability groups. Ability invariance similarly, indicates that the ability of an examinee is not influenced by the set of items (such as hard and easy items) that the examinees are administered. In other words, test characteristics are not group dependent and the examinee characteristics are not test dependent. It is important to note that it is only possible to achieve invariance if the IRT model that is selected fits the data.

In addition, since it is hard to observe invariance in the strict sense, it is possible to talk about the “degree” of invariance by evaluating either the correlation between the two sets of estimates or by examining the scatterplot (Hambleton, Swaminathan & Rogers, 1991). Poor model-data fit or poor item parameter estimation may cause a large amount of scatter which shows lack of invariance (Hambleton, Swaminathan & Rogers, 1991). Invariance is the most important characteristic of IRT which enables many important applications such as equating, item banking, item bias investigation and adaptive testing.

Lastly, predictions of actual or simulated test results could be checked by using residuals, standardized residuals or as in our case chi-square statistics.

2.3. Studies on Predictive Validity

A study was conducted by Prapphal (1990) to find out the predictive validity of three sub-tests of the National English Entrance Examination in Thailand on academic achievement in Freshmen General English and English for Academic Purposes courses at two different universities. The study involved 264 randomly selected science students who had taken the National English Entrance Examination in Thailand in 1982. The results indicated that all three tests correlated significantly and substantially with university English achievement. However, the Matching Cloze Test correlated better with university achievement (.691 with General English and .602 with EAP) at both Chulalongkorn and Mahidol than the other two sub-tests. Since the content of all three tests involved general English, the three tests account for more variance with the General English Course than with the English for Academic Purposes Course. Prapphal (1990) also suggested that test format can play an important role in predicting future academic achievement in English.

Another study conducted by Prapphal (1990) examined the relationship between the test of General English (GE) which aimed at assessing the students' ability in understanding general English and the English for the Academic Purposes Test (EAP) which had a more discipline specific content. The formats of both exams were the same. The study was conducted with 320 Chulalongkorn University students. Significant indirect relationships between the subskills of General English and English for Academic Purposes were found. This study suggests that all language subskills are related to one another, no matter what the format is. A transfer of subskills from one content (General English content) to another (English for Academic Purposes content) is possible.

A third study that Prapphal (1990) conducted involved one hundred first year students. The study was conducted to find out to what extent did the EAP subtests, the EAP Department Test and the University English AB Entrance Examination which assessed the general proficiency determined the academic achievement which is represented by GPA. The results showed that even if all the tests were able to predict the academic achievement, the EAP tests were more successful when compared with the General English Test. It is suggested that EAP tests may predict achievement in EAP programs more effectively than General English tests.

A study carried out by Stofflet, Fenton,and Strough (2001) examined the predictive validity of the Alaska State High School Graduation Qualifying Examination (HSGQE) and Benchmark Examinations on the performance on California Achievement Tests (CAT). The results showed a strong and direct relationship between performances on the Benchmark Test or HSGQE Reading scores and Writing scores and performances on the CAT Total Reading scores and Total Language and Arts scores, respectively.

Doey (1999) carried out a study which aimed at answering the question of whether IELTS is an accurate predictor of performance and success in Business, Science and Engineering. Business was particularly chosen as 'linguistically demanding' as opposed to Science and Engineering which was considered to be 'less linguistically demanding'. This provided the opportunity to compare students in different disciplines. The study was conducted on a total of 89 students in their first years. The results indicated that the only consistently positive correlation between IELTS scores and academic results was in the reading subtest which was a moderately significant correlation in the second semester of Business which was considered to be 'more linguistically demanding' discipline. Among the four modules of IELTS, Reading had the highest correlation. However, the study does not show evidence about the validity of IELTS as a predictor of academic success in freshmen.

A study conducted by Educational Testing Unit researchers, Ramist Lewis and McCauley-Jenkins (2002) investigated the correlations between SAT II Subject Tests and freshmen GPA. The results pointed out that English composition had a correlation of .51 with freshmen GPA. This was the highest correlation among SAT II Subject Tests. French, German, Hebrew, Latin and Spanish sub-tests showed lower correlations with freshmen GPA.

Kuncel et al. (2002) conducted a study to find out the relationship between Graduate Record Examination (GRE) scores and graduate performance. GRE Verbal section showed moderate correlations with performance measures identified especially with Comprehensive Exam Scores. GRE Subject Tests showed higher correlations with most of the six performance measures, namely, Comprehensive Exam Scores, Faculty Ratings and first year GPA.

Some studies on the relationship of TOEFL with other English Proficiency Tests have been cited in Marvin and Simner (1999) in order to justify the use of the TOEFL for decision making. In addition, according to Marvin and Simner (1999) a relationship between TOEFL scores and first year performance in university English courses can be possible; however, the relationship may not continue beyond first year. Pack (1972) (in Marvin & Simner, 1999) carried out a study on 402 students and found out that, TOEFL scores were "significantly related to the grade obtained in the first English course taken, however, they are not related to grades obtained in subsequent English courses nor are they related to the probability that an examinee will graduate" (Hale et al. p. 161) (in Marvin &Simner, 1999).

A study conducted by Huong (2001) investigated the predictive validity of IELTS scores. The relationship between IELTS scores and subsequent academic performance was examined. 202 Vietnamese students who studied in different Australian universities were participants. Huang found a significant and positive correlation between IELTS scores and first and second semester GPA's which was considered to be satisfactory by Alderson, Clapham and Wall (1995). Moreover, among the four sub-tests of IELTS (Listening, Reading, Writing and Speaking) the highest correlation was observed between Reading and first semester GPA; Reading and Listening subtests compared to Writing and Speaking had higher correlations with first and second semester GPA's. The findings suggested a correspondence with the first and second semester GPA in terms of both IELTS total scores and sub-test scores.

Breland, Kubota and Bonner (1999) carried out a study in order to examine the relationship between scores on the SAT II: Writing Subject Test and performance in writing in the first year of university. 222 students participated with all the required writing samples; however, more cases were available for some variables when compared to others. The results of the study revealed high correlations between SAT I Verbal score and university course grades. Also, a high correlation was achieved for SAT II : Writing Test. However, the SAT Writing Test Essay score had a lower correlation for predicting course grades when compared to SAT I Verbal score and SAT II : Writing Test.

In a study, Heard and Ayers (1988) examined the validity of the American College Test (ACT) in predicting success on the Pre-Professional

Skills Test (PPST). PPST is designed to measure proficiency in reading, writing and mathematics. ACT, which is used for admission, consists of English, mathematics, natural science, social science subtests in addition to a composite test score. 202 students took part in the study. These students had taken the PPST as a requirement for admission to the Professional component of the teacher education program at Tennessee Technological University. The students had also completed the ACT. It was concluded that the ACT composite score was the best predictor of success on three tests of PPST. ACT composite scores, subtest scores together with GPA in college English courses improved the prediction of achievement. These results indicate that scores from ACT are a reasonable predictor of success on the PPST.

Stricker, Rock and Burton (1996) carried out a study with the aim of appraising the utility of SAT scores in combination with collateral variables: grades in high school courses, and the number and quality of theses courses in predicting college grades in various fields of study in order to provide students with predictions of their academic performance for guidance purposes. 981 students participated in the study. The SAT and the collateral variables were found to be predicting college grades in different areas by taking account of marked variations in grade distributions among fields.

2.4. Studies on Item Response Theory

Fan (1998) examined how comparable the item and person statistics derived from two measurement frameworks: CTT and IRT were and how invariant the items statistics of CTT and IRT were across examinee samples. He used the data of 193,000 participants in his study. Random samples consisting of 1000 examinees were drawn from the participant pool for invariance studies. Fan found out that the item and person statistics obtained from CTT and IRT were quite comparable. Similarly, invariance of item statistics were comparable. These findings contrasted the widely accepted view that IRT was superior over CTT.

ETS staff Chyn, Tang and Way (1995) carried out a study to investigate the feasibility of the Automated Item Selection Procedure (AIS) for the Test of English as a Foreign Language (TOEFL). Statistical specifications, which were IRT based, were developed. By using the AIS procedure two final forms of TOEFL were assembled and the statistical and content related properties were checked. The results of the study showed the superiority of the AIS technique over traditional test assembly procedures in terms of statistical parallelism. In addition, it was also found out that the TOEFL tests assembled by using the AIS procedure successfully met the IRT specifications. Efficiency in several sections of the test was achieved.

Another study by ETS researchers Way and Rease (1991) compared the uses of the one- and two- parameter logistic IRT estimation models with the three-parameter IRT model for scaling and equating the TOEFL test. The design of the study involved the simulation of typical TOEFL equating by using artificial data. The results of the study supported the use of the three-parameter model and emphasized that the differences between the score conversions of the compared models had a tendency to appear at the lower and upper ends of the score scales. The simulated equatings of the three-parameter model did not seem to be sensitive to the sample sizes used in the study.

Tang and Eignor (2001) carried out a study which aimed at investigating whether classical item statistics could be used for collecting collateral information in the IRT calibration of pretest items for the computer-based TOEFL and reduce examinee sample sizes. The study was conducted by using BILOG computer program for analysing data. The data was taken from three TOEFL pre-test item pools used in implementing Computer Based Testing (CBT) were used to simulate the conditions required for the purpose of the study. At least 600 examinee responses per item were used in all three pre-testing item pools. However, the results of the study showed that the classical item statistics did not provide a sufficient level of collateral information to support a reduction in pre-test sample sizes.

A study that Kılıç (1999) carried out investigated the fit of IRT models to the Mathematics, Natural Sciences, Turkish and Social Sciences sub-tests of the 1993 Student Selection Test (SST). The data of 2121 examinees were used in the analyses. After determining the fit of the data to the IRT models, invariance was checked by using ability estimates and item parameter estimates obtained from different samples. Finally, the observed and theoretical distribution of each sub-test was examined. The results indicated that the homogeneous item discrimination indices assumption of the one-parameter model and the all the sub-tests except for the Turkish sub-test were speeded. The Turkish sub-test was more invariant in the Turkish sub-test than the Mathematics, Natural Sciences and Social Sciences sub-tests. The results showed that the three-parameter model indicated a better fit according to chi-square statistics.

Çalışkan (2000) investigated the fit of the one-, two- and three-parameter IRT models to the MNE-ERDD's Science Achievement Tests data. The data was obtained from 2912 students from grade level 5, 4477 students from grade level 8 and 2187 students from grade level 11. First, whether the assumptions of IRT was met was examined. Secondly, the invariance of item parameter estimates and the ability parameter estimates collected from different groups was investigated. Thirdly, the chi-square statistics were interpreted to check the observed and theoretical distributions of the test data. The results indicated that equal item discrimination of the one-parameter model was not met in all three grade levels of the test data. Minimal guessing assumption was only met in grade level 11 Science Achievement Test. All the tests seemed to be non-speeded. The ability parameter estimates all three models were more invariant across different sets of items in both grade level 5 and 11 Science Achievement Tests. For the one parameter model the item difficulty parameter estimates across different samples of students were quite invariant in all Science Achievement Tests while in the item discrimination parameter estimates across low and high ability groups invariance did not hold. Chi-square statistics indicated a better fit of the three-parameter model to the MNE-ERDD's Science Achievement Tests data.

Karataş (2001) examined the use of IRT models in the evaluation of the items of the English Proficiency Test of Erciyes University; moreover, the fit of the ELT test data to one- two three-parameter models was investigated. The data collected from Erciyes University Preparatory School English Proficiency Exam made use of 468 examinee responses to the Form A of the test. First, the fit of the data to IRT assumptions was examined. Secondly, the invariance of item parameter estimates and the ability parameter estimates collected from different groups was investigated. Thirdly, the chi-square statistics were interpreted to check the observed and theoretical distributions of the test data. The results of the study showed that the test data met the assumptions of the IRT models. Moreover, the item parameter estimates and ability parameter estimates obtained from different samples were found to be invariant. Still, the one- and two- parameter models were slightly more invariant across different groups when compared with the three parameter model. It was concluded that the two- and three parameter models provided better fit to the data than the one- parameter model according to chi-square statistics.

Özkurt (2002) carried out a similar study on the fit of one-, two-, and three- parameter models to English Proficiency Test of a state university. 361 students who studied for at least one year in the Preparatory School were used in the data analyses. First, the assumptions of the IRT models were investigated to determine the fit of data to the assumptions. Secondly, ability parameter estimates and item parameter estimates obtained from different samples were compared to determine if they were invariant. The results of the study showed that the data met the unidimensionality, non-speededness and local independence assumptions. However, the ability parameters and the item parameters were not invariant across different groups. It was concluded that the data met the two-parameter model according to the results of the chi square statistics.

2.5. Summary

BUEPE tests whether the capacity of students is sufficient to attend and succeed in the DEC. However, if it successfully performs this job was not investigated up to now. In order to examine whether the BUEPE predicts DEC scores the predictive validity of BUEPE over the DEC scores can be investigated. Since IRT estimates provide a wider range of comparisons other than total scores, examining the predictive validity of IRT based ability estimates across DEC scores may enhance the predictive validity coefficient obtained.

IRT emerged in the first decades of the previous century with superior characteristics over CTT such as separability of item and examinee characteristics and the use of ICCs, IIFs and TIFs. There are three main IRT models namely; the one-, two-, three- parameter models. Deciding on which model to use is an important decision. Moreover, IRT has some assumptions to be met by the data set. These are the unidimensionality assumption, local independence assumption, equal discrimination indices, minimal guessing and non-speeded test administration assumptions. Furthermore, the invariance of ability parameters, item parameters and the results of the chi-square statistics play an important role in determining whether the IRT model fits the data.

Studies on predictive validity of English tests and fit of IRT models to test data exist separately; however, there are virtually no studies on the IRT estimated predictive validity of English proficiency exams. Thus, this study may initiate predictive validity studies with IRT estimates.

CHAPTER III

METHOD

This chapter reviews the methodological procedures in the study. The main titles in this chapter consist of overall research design, research questions, the data collection instrument, population and sample selection, data collection procedure and data analysis procedure.

3.1. Overall Research Design

The main purpose of this study was to examine the predictive validity of Başkent University English Proficiency Exam (BUEPE) by using IRT based ability estimates. First, the fit of BUEPE data to IRT models was investigated as prerequisite analysis in order to examine the predictive validity of BUEPE by using the IRT based ability estimates obtained in the prerequisite analyses. The BUEPE was administered to 699 examinees in September 2000 and the data collected by means of the BUEPE was analyzed by statistical procedures of IRT. The freshmen DEC passing grades of the 371 students who had passed the BUEPE that year were also collected for predictive validity analyses. Correlation coefficients were computed for examining the relationship between the BUEPE IRT based ability scores and DEC passing grades.

3.2. Research Questions

1. Does the BUEPE data meet the assumptions of IRT?

2. Are the obtained ability parameter estimates of the one-, two-, and three- parameter models invariant across different samples of test items ?

3. Are the obtained item parameter estimates of the one-, two-, and three-parameter models invariant across different samples of examinees ?

4. How well do the simulated test results of the one-, two-, and three-parameter models predict the actual test results ?

5. Do the ability estimates obtained through the use of the IRT based model on BUEPE 2000 predict success in departmental English courses (DEC) at Başkent University ?

3.3. Data Collection Instrument

This study used the results of the data collected from Başkent University English Proficiency Exam (BUEPE) in September 2000.

BUEPE 2000. The BUEPE 2000 consists of 100 items all of which are multiple choice with four alternatives like all other proficiency exams of Başkent University. The exam has 3 sections: Grammar, reading and vocabulary, respectively. Different question types are exploited in each section for different purposes. Examples of item types mentioned below can be seen in Appendix A.

In the grammar section items 1-15 make up the modified cloze test. Items 16-36 are discrete point items. Items 37-40 are spot the mistake type of grammar items.

In the reading section items 41-45 are sentence completion items. Items 46-50 are paragraph completion items. After these, items related to three different reading texts follow. Reading Text 1: items 51-60 are sentence completion, guessing vocabulary and reference type items. Reading Text 2: items 61-70 are sentence completion, guessing vocabulary and reference type items. Reading Text 3: items 71-80 are sentence completion, guessing vocabulary and reference type items.

The last section is the vocabulary section. Items 81-100 are sentential level fill in the blank type multiple-choice items.

So as to study the predictive validity of the BUEPE 2000, the passing grades in the first and second semester of Departmental English Courses (DEC) were used.

First and Second Semester Departmental English Grades in Freshmen.

These grades collected in the 2000-2001 academic year were obtained by adding the following weightings of the exams. 30% Midterm Exam (Achievement Exam testing grammar, reading comprehension, vocabulary and writing). 10% Project Exam (Alternative Assessment testing reading comprehension or speaking skills depending on the department).10% Teacher Evaluation (Evaluation of the class teacher according to four criteria: Participation, Attendance, Homework, Preparation). 50% Final Exam (Achievement Exam testing grammar, reading comprehension, vocabulary and writing).

3.4. Population and Sample Selection

All 699 students who took the BUEPE administered in September 2000 were selected for the purpose of the study. This group included students who had failed in the summer school proficiency exam, students who did not attend the summer school programme after failing from June proficiency Exam, those who failed due to unattendance to preparatory school and new enrollments to the university.

Since, only 371 students had managed to pass the BUEPE in September 2000, predictive validity studies were carried out on a sample of 371 students who had 2000-2001 freshmen year passing grades.

3.5. Data Collection Procedure

In Başkent University each student registering is required to be at a certain level of English proficiency to start as freshmen. Therefore, students are administered two exams before they are admitted to freshmen. Students take the placement exam first; those who pass can take the proficiency exam, the others start the preparatory school of BU at C-level. Students who pass the proficiency exam start as freshmen. The students who fail to pass the proficiency exam start the preparatory school at B-level. For both exams the passing score is 60. During the one-year period in the preparatory school, students are tested on eight progress tests. They have to reach an overall of 60 to be able to take the proficiency exam. Those students who get a minimum score of 60 from BUEPE pass to the freshmen year. The ones who fail to get the minimum score can register to the BU summer school program. All students who attend the summer school are given an extra chance of taking a proficiency exam. If they get a score of 60 and above, they pass to freshmen. The others have a final chance and take another proficiency exam.

Therefore, students who attend the summer school take the final proficiency exam administered in September together with the students who did not attend the summer school course and with those who have just enrolled at the university and those who failed in the preparatory school due to unattendance.

3.6. Data Analysis Procedure

The data analyses were carried out in two dimensions. In order to carry out the predictive validity analyses at which this study mainly aims first, the fit of the BUEPE 2000 was established; secondly, the predictive validity of BUEPE 2000 was assessed by using the results obtained from the first phase of the study.

This chapter describes the data analyses procedures under five major headings; preliminary analyses, checking model assumptions, checking expected model features, checking model features of actual and simulated test results, and predictive validity analyses.

3.6.1. Preliminary Analyses

The data obtained from the optic forms of the September 2000 BUEPE were coded dichotomously on the SPSS processor as 0 for incorrect and 1 for correct responses. Then the descriptive statistics including measures of central tendency (mean, mode, median) and measures of variation (standard deviation, variance, skewness, kurtosis), minimum-maximum scores and frequency distribution with a normal curve were obtained to demonstrate an overall picture of the proficiency exam results.

Secondly, the reliability of the scores obtained from BUEPE was calculated by using the alpha coefficient, which Green, Salkino and Akey (1997) view as the most appropriate index for estimating the reliability of dichotomously scored items.

Item analysis was conducted on the 100 items proficiency exam by using SPSS. To demonstrate how each item functioned, the item difficulty (Item means) and discrimination (Corrected-item total correlation) indices for each item as well as for the whole test were obtained by using CTT techniques provided by SPSS.

3.6.2. Checking Model Assumptions

In order to check the unidimensionality of the 100 items in BUEPE, principle component analysis was run. The eigenvalues and the scree test results were interpreted to decide whether the exam was unidimensional. Varimax rotation procedure was used to rotate the factors.

To see if the items of the BUEPE were locally independent, the total inter-item correlations were compared with the inter-item correlations of examinees in restricted range ability groups; in this case the examinees in the high performers group and the low performers group were selected. The high and low performers were selected according to the scores they obtained. Thus, examinees in the first quartile with total scores of 45 and below and examinees in the fourth quartile with total scores 70 and above were selected. Local independence holds if the means of the inter-item correlations in the high and low performers groups are close to zero.

The item discrimination indices obtained by classical item analysis of SPSS were reviewed and plotted in a histogram to examine whether the distribution is reasonably homogeneous. A reasonable homogeneous distribution would imply that the one parameter model could be appropriate for this data set.

To see whether there was a guessing factor affecting the results obtained from the BUEPE the most difficult 5 items were selected and tested on the students in the first quartile. The means of the items were compared to see if they were close to zero. If guessing appears to be a factor affecting the scores, it would be wise to take the three parameter model into consideration.

In order to check whether the proficiency exam functioned as a speed test or not, the ratio of the variance of omitted items to the variance of the items answered incorrectly was calculated. The ratio is expected to be close to zero. However, if the results indicate that BUEPE 2000 is speeded, none of the IRT models (one-, two-, three-, parameter models) would fit our data since speededness is an assumption common to all models.

3.6.3. Checking Expected Model Features

The procedures for this part were carried out by using BILOG (Mislevy & Bock, 1984). BILOG computes item parameter estimates and ability parameter estimates according to the selected model.

First, the invariance of ability parameters was analysed. Ability estimates for different samples of test items in one-, two-, and three- parameter models were computed separately, compared and the scatterplots were displayed. The hard and easy items were selected according to the item difficulty indices and the data obtained from the hard items and the easy items were run separately. The resulting ability estimates obtained these hard versus easy items were compared. In addition, the first fifty versus second fifty items in one-, two-, and three- parameter models were compared with respect to the ability estimates they produce. Invariance is said to be established when the plot is linear with little scatter.

Secondly, the invariance of the item parameters (b-values, a-values and c-values) was examined. Item parameter estimates of one-, two-, and three-parameter models obtained from odd versus even cases were correlated and scatterplots were obtained. The estimates are considered to be invariant if the plots are linear and the correlations are reasonable.

3.6.4. Checking Model Predictions of Actual and Simulated Test Results

To find out the best fitting IRT model to BUEPE data chi square statistics which is one of the goodness of fit analyses were obtained from one-, two-, and three- parameter models. The model with the least number of misfitted or insignificant items is said to fit the data.

3.6.5. Predictive Validity Analyses

After BUEPE data is analysed by using IRT and a fit is obtained, the indices obtained from these analyses which provide a variety of information are used in establishing the predictive validity of the BUEPE. Pearson product-moment correlation and scatterplot analyses were employed in order to study whether BUEPE predicts the first and second semester DEC passing grades in freshmen.

First of all BUEPE total scores were correlated with the first and second semester DEC passing grades in freshmen and scatterplots were obtained to display the relationship.

Secondly, ability estimates of the best fitted model which were obtained in prior analyses were correlated with first and second semester DEC passing grades in freshmen and scatterplots were obtained to demonstrate the relationship.

Third, by using the item information functions of the best fitted model, sixty items which give the highest information indices were selected previously. Another total score, which summed up only the sixty highest information items, was computed afterwards. Then, the total scores obtained by using the sixty highest information yielding items in the best fitted model were correlated with the first and second semester DEC passing grades in freshmen and scatterplots were obtained to display the relationship.

Next, BILOG was run in the best fitted model with the data of the sixty highest information yielding items and ability estimates were obtained. The ability estimates obtained by using sixty highest information yielding items in the best fitted model were correlated with the first and second semester DEC passing grades in freshmen and scatterplots were obtained to display the relationship.

The same procedure followed for selecting the sixty high information items was used for choosing the thirty-five highest information items then a total score that summed only those thirty-five items was calculated. The total scores obtained by using the thirty-five high information items were correlated with the first and second semester DEC passing grades in freshmen and scatterplots were obtained to display the relationship.

Similarly, the ability estimates obtained by using the highest thirty-five information yielding items in the best fitted model were correlated with the first and second semester DEC passing grades in freshmen and scatterplots were obtained to display the relationship.

The content sampling of both the sixty highest information items and the thirty-five highest information items were checked in order to see if the distribution of the number of items in the sub-tests was kept the same when the number of items was reduced. The alpha coefficients were computed for the highest information yielding sixty and thirty-five items.

Furthermore, the function of grammar, reading and vocabulary sub-tests in predicting DEC passing grades was analyzed. In order to carry out these analyses firstly three different total scores were calculated for each student by using each sub-test (grammar, reading, vocabulary) at a time. Then, three different ability estimates were obtained for each student by running BILOG in the best fitted model for the grammar, reading and vocabulary sub-tests, respectively. The Test Information Functions (TIF) of the three sub-tests were interpreted.

The total scores obtained by the using the grammar sub-test was correlated with the first and second semester DEC passing grades in freshmen.

The ability estimates obtained by the using the grammar sub-test in the best fitted model were correlated with the first and second semester DEC passing grades in freshmen.

In addition, the total scores obtained by the using the reading sub-test were correlated with the first and second semester DEC passing grades in freshmen.

Next, the ability estimates obtained by the using the grammar sub-test in the best fitted model were correlated with the first and second semester DEC passing grades in freshmen.

Then, the total scores obtained by the using the vocabulary sub-test were correlated with the first and second semester DEC passing grades in freshmen.

Finally, the ability estimates obtained by the using the vocabulary sub-test in the best fitted model were correlated with the first and second semester DEC passing grades in freshmen.

CHAPTER IV

RESULTS

This chapter presents the results of the study which was conducted on the predictive validity of the BUEPE 2000 data by means of IRT estimates.

4.1. Preliminary Analyses

The descriptive statistics and the frequency distribution are presented in Table 4.1.1 and Figure 4.1.1.

Table 4.1.1. Descriptive Statistics for The Whole Data

Number of items	100
Number of examinees	699
Mean	57.42
Median	57
Mode	50
Variance	273.78
Standard Deviation	16.56
Skewness	.090
Kurtosis	-.72
Minimum	15
Maximum	95
Cronbach Alpha	.9328
Mean Difficulty (p)	.574
Mean Item-Total (r)	.334

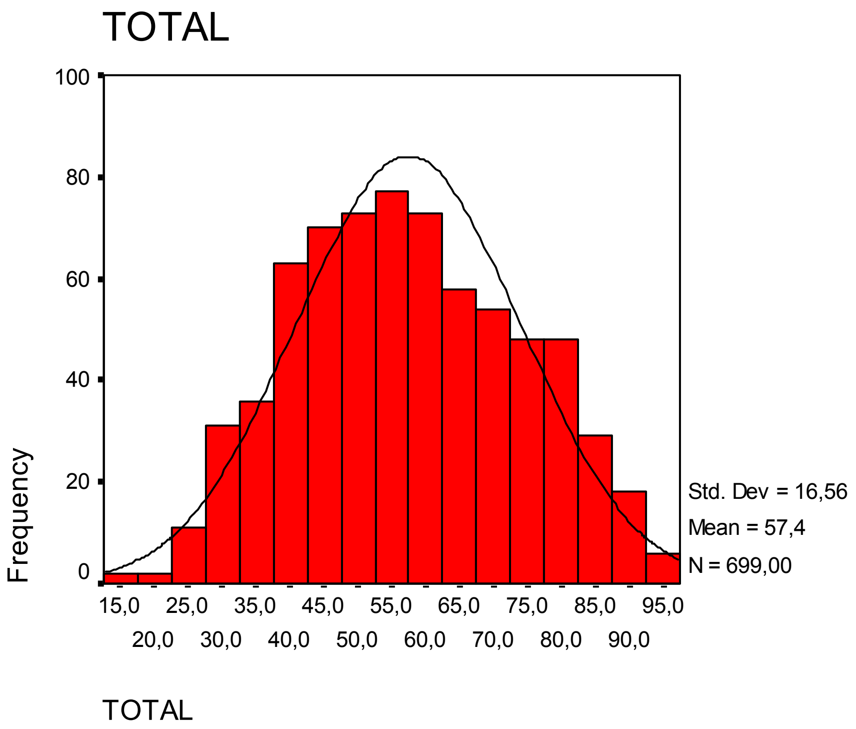


Figure 4.1.1. Frequency Distribution of Whole Test Scores

The mean score of the whole group was 57.4 which is not a very high score considering the cut off score of 60. The standard deviation was 16.56 and the variance was 273.78 which indicate a large and desirable distribution. The total score distribution was positively skewed with a value of .089 and flat with a kurtosis of -.726. The minimum score was 15 and the maximum score was 95, indicating a range of 80, which is quite high. The reliability of the scores obtained from the exam was quite high with a cronbach alpha of .93. This provides evidence to support the fact that BUEPE produces reliable scores.

The mean item difficulty was .57, which indicated that the exam was not very hard for the examinees. Figure 4.1.2. displays the frequency distribution of the item difficulty indices.

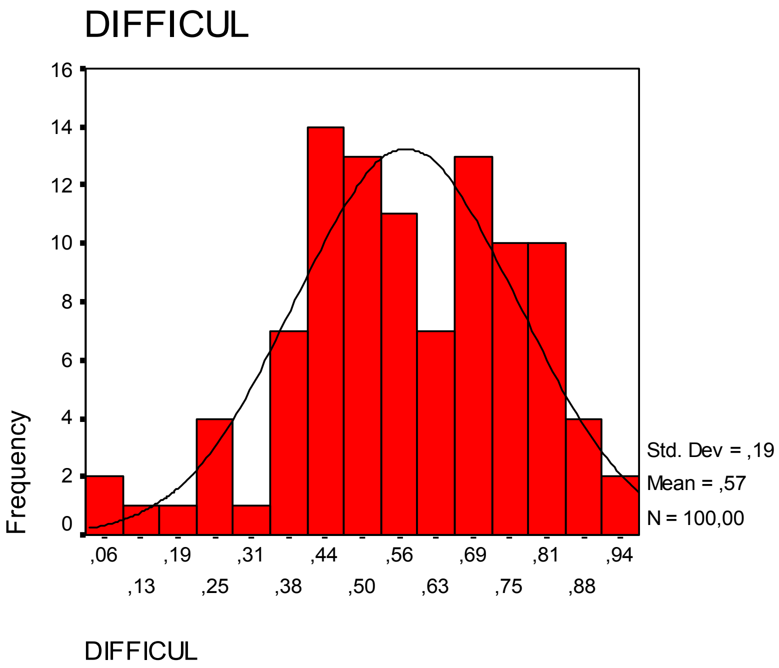


Figure 4.1.2. Frequency Distribution of Item Difficulty Indices

The mean item discrimination was .33, which showed that the items in the test were moderately discriminating among high achievers and low achievers, considering the minimum acceptable level as .20 in item discrimination. Appendix B presents the item means (difficulty) and the corrected item-total correlations (discrimination).

4.2. Checking Model Assumptions

Principle component analysis results were interpreted in order to find out whether the test data met the unidimensionality assumption. Firstly, the eigenvalues were examined. The first eigenvalue explained a total variance of 14.26, the second eigenvalue explained 2.90 per cent and the third explained 1.79 per cent of the total variance as seen in Appendix C. The sharp drop from the first eigenvalue to the second one shows that the data is unidimensional. Figure 4.2.1. below supports the findings.

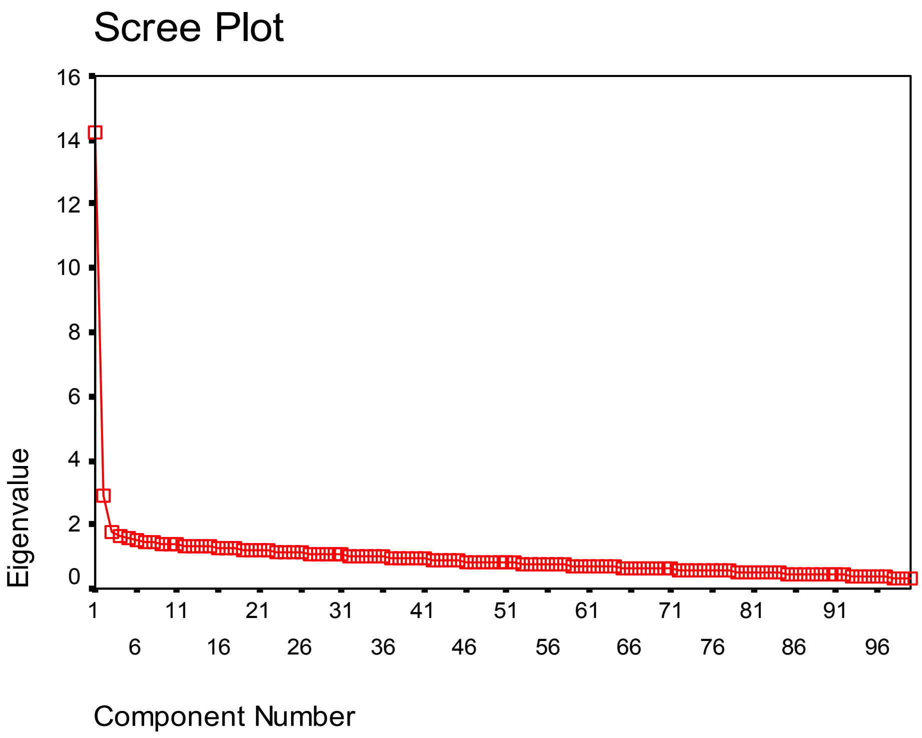


Figure 4.2.1. Plot of Eigenvalues

To check whether the items of BUEPE are locally independent, the total inter-item correlation was compared with the inter-item correlations of the examinees in the high performers group (>=70) and the low performers group (<=45). The inter-item correlation means in the high and low ability groups, .0191 and .0094, respectively were lower than that of the total group's mean which was .1196, and were close to zero. Table 4.2.1. shows total inter-item correlations with respect to restricted range ability groups. The results indicate that there is evidence for local independence since the inter-item correlation means in restricted range groups were lower than that of the total group's mean and were close to zero. Therefore, unidimensionality entails local independence.

Table 4.2.1. Inter-Item Correlations of Total and Restricted Range Groups
Inter-item

Correlations	Mean	Min	Max	Range	Min/Max	Variance
TOTAL	,1196	-,1062	,3814	,4876	-3, 5917	,0053
>=70	,0191	-,2060	,3645	,5705	-1, 7700	,0063
<=45	,0094	-,2796	,3337	,6133	-1, 1933	,0062

The item discrimination indices can be seen in Figure 4.2.2. below. The distribution is negatively skewed and is not homogeneous, implying that the equal discrimination indices assumption of the one-parameter model is not met with BUEPE data.

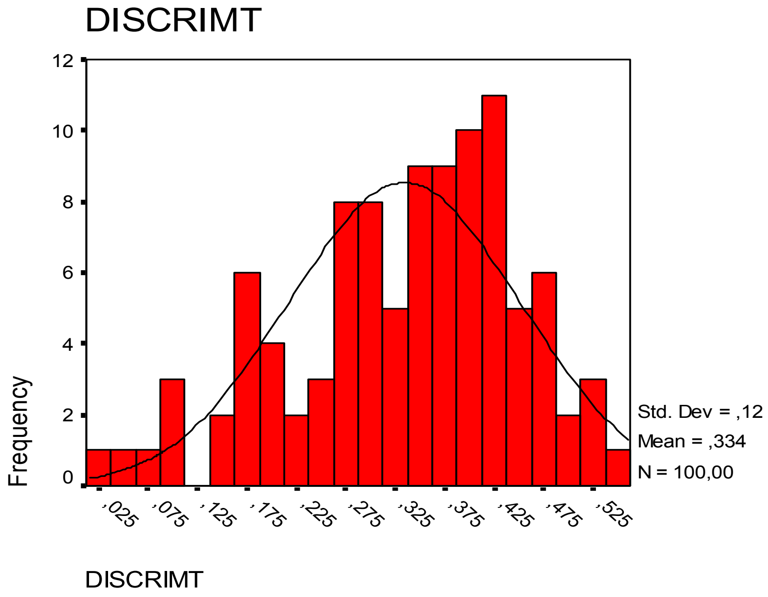


Figure 4.2.2. Frequency Distribution of Item Discrimination Indices

In order to examine whether there was a guessing factor in the exam, the 183 low ability examinees' performance on the most difficult 5 items was checked. The results are displayed in the Table 4.2.3. below.

Table 4.2.3. Low Ability Student Scores on Most Difficult 5 Items

Item no	Item difficulty	Percent incorrect	Mean
Item 9	.0916	90.7	9.29 E-02
Item 20	.2160	93.4	6.56 E-02
Item 25	.0787	94.5	5.46 E-02
Item 71	.1559	96.7	3.28 E-02
Item 100	.2632	76.0	.24

Since this is a multiple-choice exam with four alternatives and since the low ability student has a one in four chance factor for correctly answering the item, the means of the items must be below .25 to conclude that there is no guessing factor involved. In other words, the items means must be close to zero. Items 9, 20, 25, and 71 present no problem as to this rule because they are all lower than .25. However, the mean of item 100 is close to .25 and may be problematic. Also, it has a lower per cent of incorrect items. Therefore, the three-parameter model can be taken into consideration.

Non-speededness which is an assumption common to all three models can be examined by interpreting the ratio of the variance of omitted items to the variance of the items answered incorrectly. The variance of omitted items was 39.418 and the variance of incorrect answers was 245.103 which yielded a ratio of .16 when calculated. The value was close to zero which emphasizes that BUEPE was a non-speeded exam.

4.3. Checking Expected Model Features

4.3.1. Invariance of Ability Parameter Estimates

In order to establish the invariance of ability parameters 50 easy and 50 hard items were compared. Moreover, the first 50 items were compared with the second fifty items. Table 4.3.1.1. presents the correlation coefficients across one-, two-, and three- parameter models.

Table 4.3.1.1. Correlation Coefficients Across Different Sets of Items in One, Two and Three Parameter Models

	1 Parameter Model	2 Parameter Model	3 Parameter Model
Easy vs. Hard Items	.788*	.798*	.797*
First fifty vs. Second fifty	.813*	.822*	.833*

* Correlation is significant at 0.01 level (two-tailed)

According to the results, the above correlations between easy and hard items seem moderately high. The one-parameter model has a slightly lower correlation coefficient of r = .788 when compared to the two- and three-parameter models which produce similar correlation coefficients for easy and hard items r =.798 and r = .797, respectively. The correlation coefficients obtained from comparing the first fifty items and the second fifty items of the exam seem to yield higher values in general when compared with that of easy versus hard items. The one-parameter model has the lowest correlation (r =.813) among the correlations of the first versus second fifty items, whereas the three-parameter model has the highest correlation coefficient with a value of r =.833. The scatterplots presented in Figures 4.3.1.1., 4.3.1.2., 4.3.1.3. support the relationship between ability parameters obtained from easy and hard items. The one-parameter model displays a slightly greater amount of scatter whereas the scatterplots of the two- and three- parameter models seem to be quite similar. These findings indicate that two- and three- parameter model's ability estimates obtained from easy and hard items yield higher invariance when compared with the one- parameter model's ability estimates.

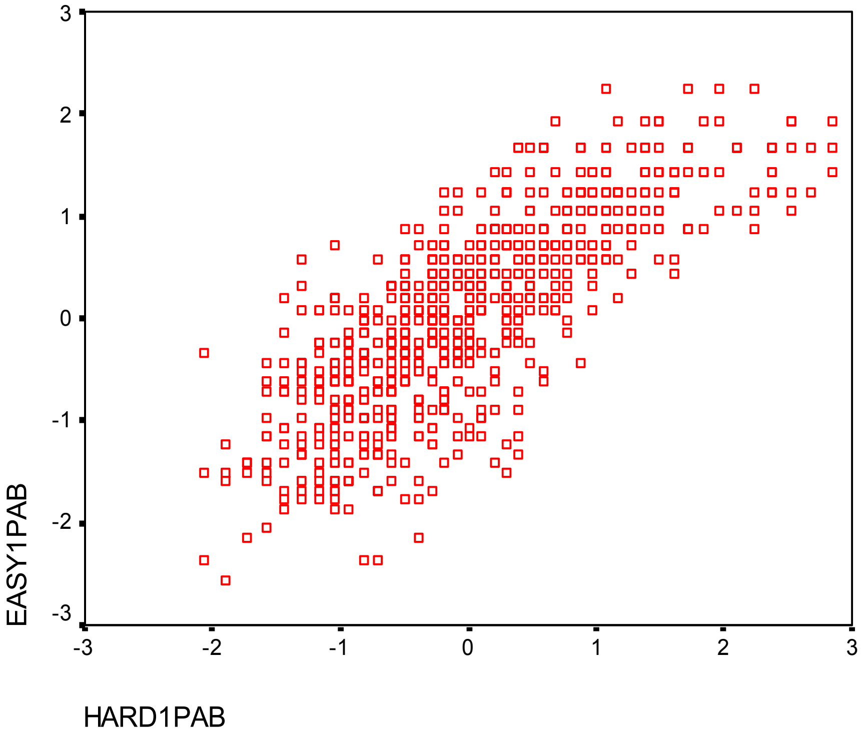


Figure 4.3.1.1. Scatterplot of 1P Ability Estimates (Easy vs. Hard)

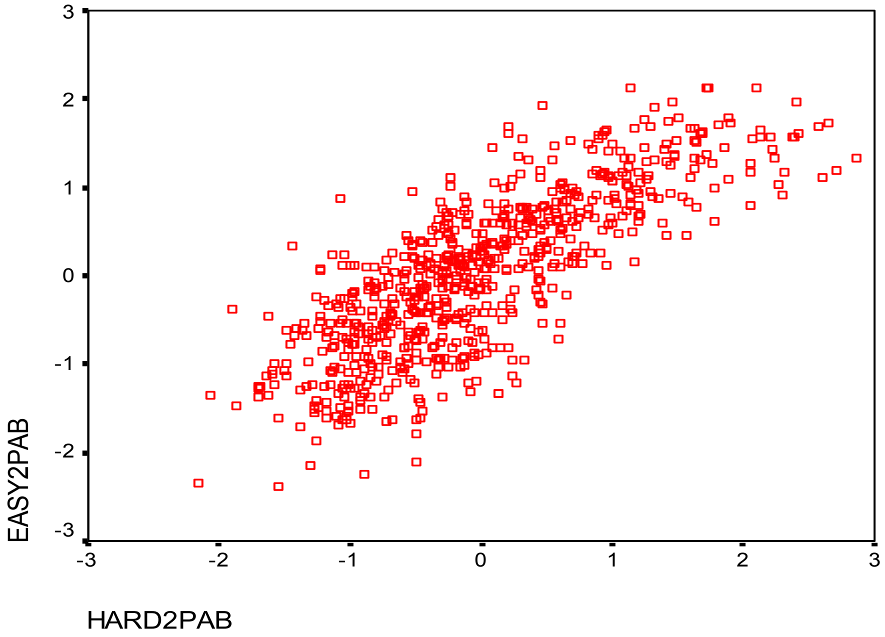


Figure 4.3.1.2. Scatterplot of 2P Ability Estimates (Easy vs. Hard)

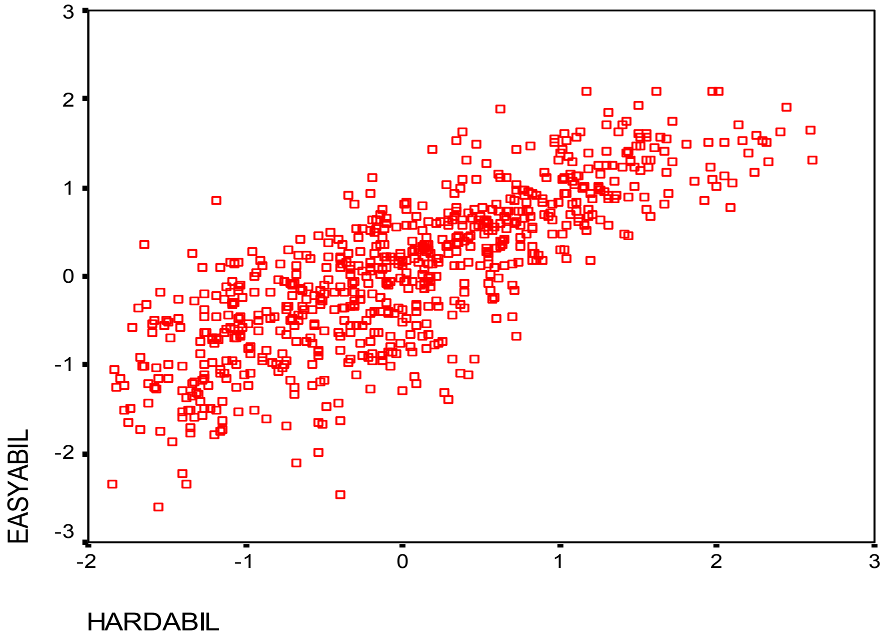


Figure 4.3.1.3. Scatterplot of 3P Ability Estimates (Easy vs. Hard)

Similarly, the scatterplots 4.3.1.4., 4.3.1.5. and 4.3.1.6. obtained from the comparisons between the first fifty and second fifty items support the findings obtained from the comparisons of easy and hard items. The scatterplot of the one-parameter model, Figure 4.3.1.4. displays a slight scatter when compared with the plots of the two- and three- parameter models. However, the ability estimates obtained from the comparisons of first and second fifty items in the three-parameter model, as seen in Figure 4.3.1.6, yield the highest degree of invariance.

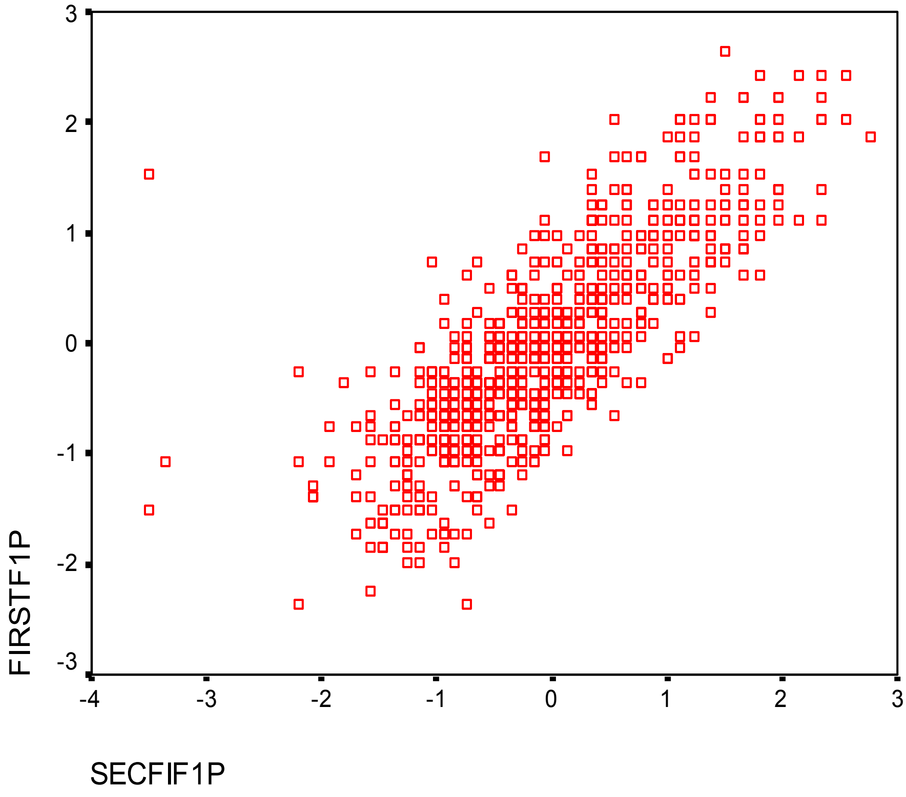


Figure 4.3.1.4. Scatterplot of 1P Ability Estimates (First 50 vs. Second 50)

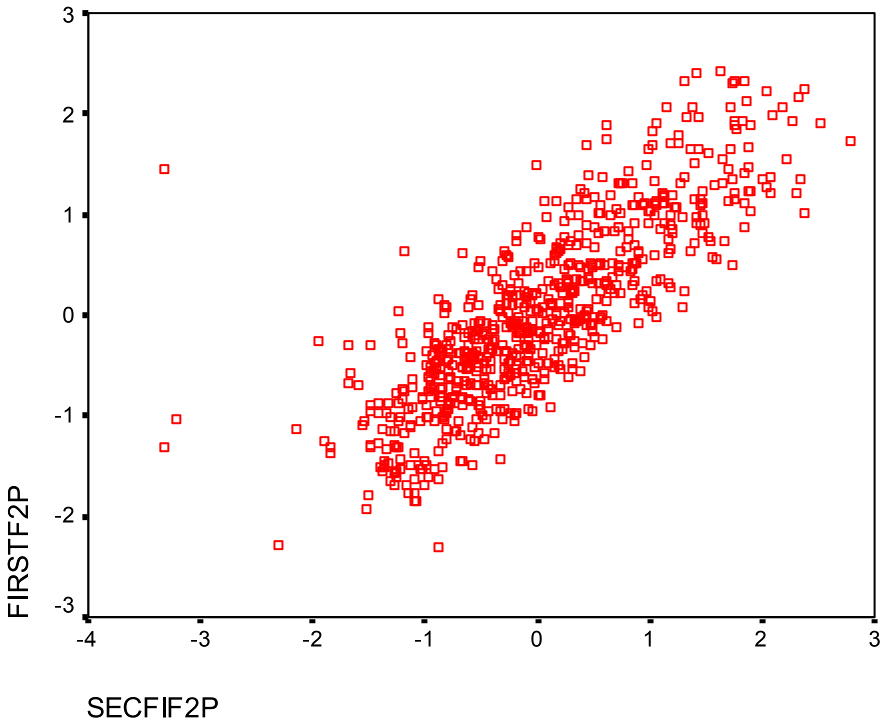


Figure 4.3.1.5. Scatterplot of 2P Ability Estimates (First 50 vs. Second 50)

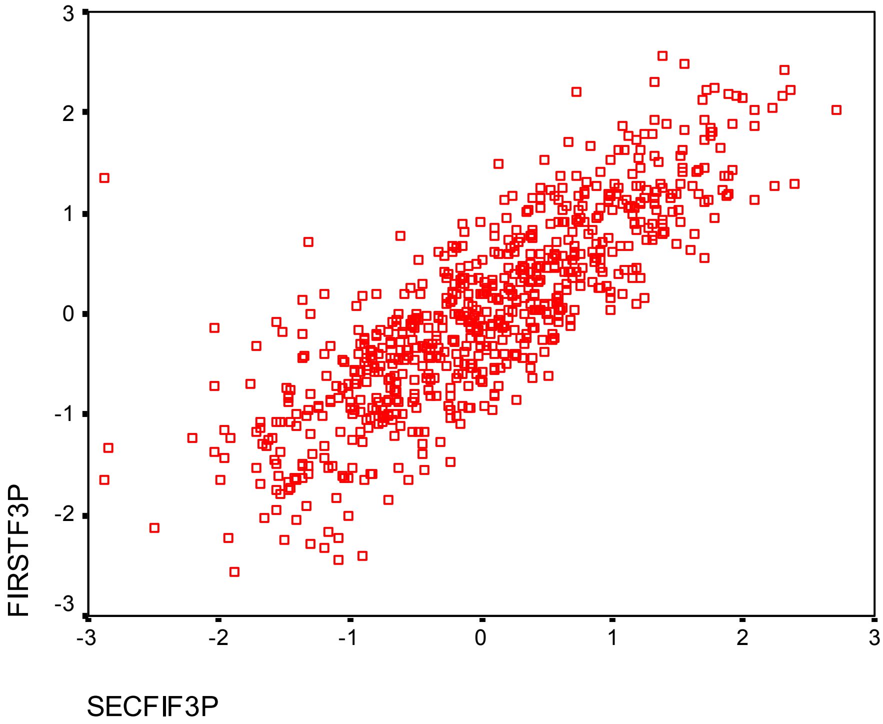


Figure 4.3.1.6. Scatterplot of 3P Ability Estimates (First 50 vs. Second 50)

4.4. Invariance of Item Parameter Estimates

Item parameters in all three models were examined across odd and even cases ability groups to determine if they were invariant.

Table 4.4.1. Correlation Coefficients of Item Parameters in Three Models

Subgroups	Item Parameters	1 Parameter Model	2 Parameter Model	3 Parameter Model
Odd vs. Even Cases	b	.988*	.972*	.971*
	a		.781*	.708*
	c			.695*

* Correlation is significant at 0.01 level (two-tailed)

The results can be seen in Table 4.4.1 which presents the correlation coefficients of item parameters in all three models. In the one-parameter model the correlation between the b parameters of the odd and even cases was r = .988, p=.000 significant at the .01 alpha level. Figure 4.4.1. displays the relationship between odd versus even items in a scatterplot.

In the two-parameter model the correlation between the b parameter estimates obtained from the odd and even cases was r = .972, p=.000 significant at the .01 alpha level. The correlation between the a parameters estimates in the odd and even cases group was r = .781, p=.000 significant at the .01 alpha level. Figure 4.4.2. below displays the scatterplots for the discrimination estimates of the two-parameter model.

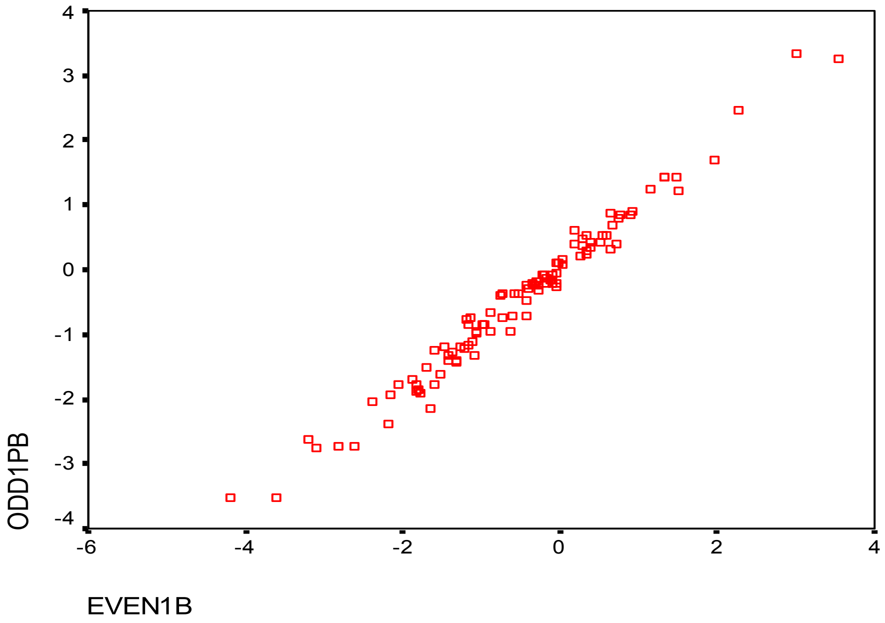


Figure 4.4.1. Scatterplot of 1P Difficulty Estimates (Odd vs. Even)

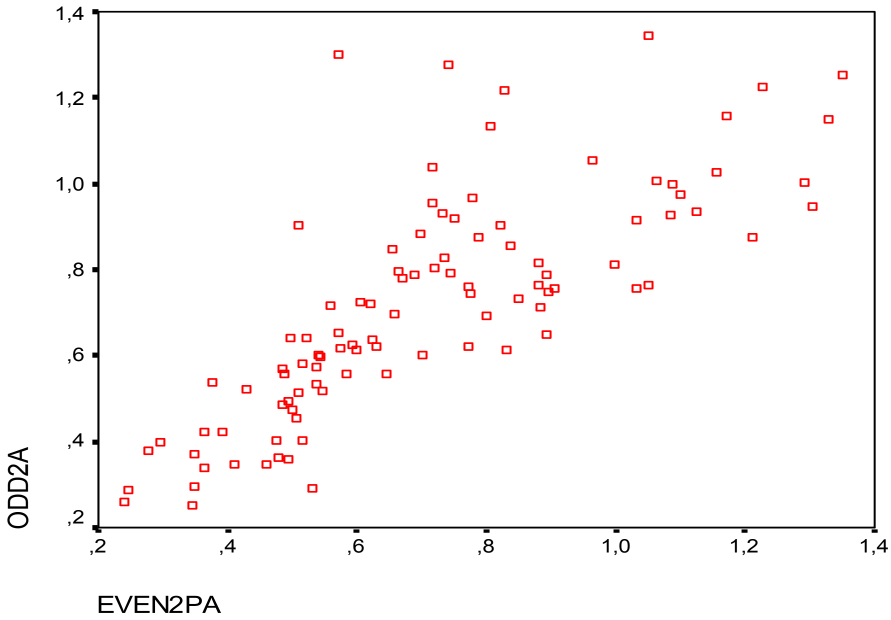


Figure 4.4.2. Scatterplot of 2P Discrimination Estimates (Odd vs. Even)

In the three parameter model the correlation between the b parameter estimates of the odd and even cases was r = .971, p=.000 significant at the .01 alpha level, whereas the correlation between the a parameters of the odd and even cases was r = .708, p=.000 significant at the .01 alpha level. The correlation between the c parameters of the same ability groups was r = .695, p=.000 significant at the .01 alpha level. Figure 4.4.3. displays the discrimination indices of odd versus even ability groups. Figure 4.4.4. displays the pseudo-chance factor indices of odd versus even ability groups in the three-parameter model. It seems that the odd versus even cases groups yield high degree of invariance in the three-parameter model.

These results imply that whereas relatively lower correlation coefficients are observable in the a parameters of the two- and three- parameter models, the b parameters of all three models yield high correlations in odd and even cases ability groups. The high correlation coefficients across odd and even cases ability groups imply that item parameters are invariant across different groups. The scatterplots in Appendix D display this relationship graphically in Figures D1 and D2.

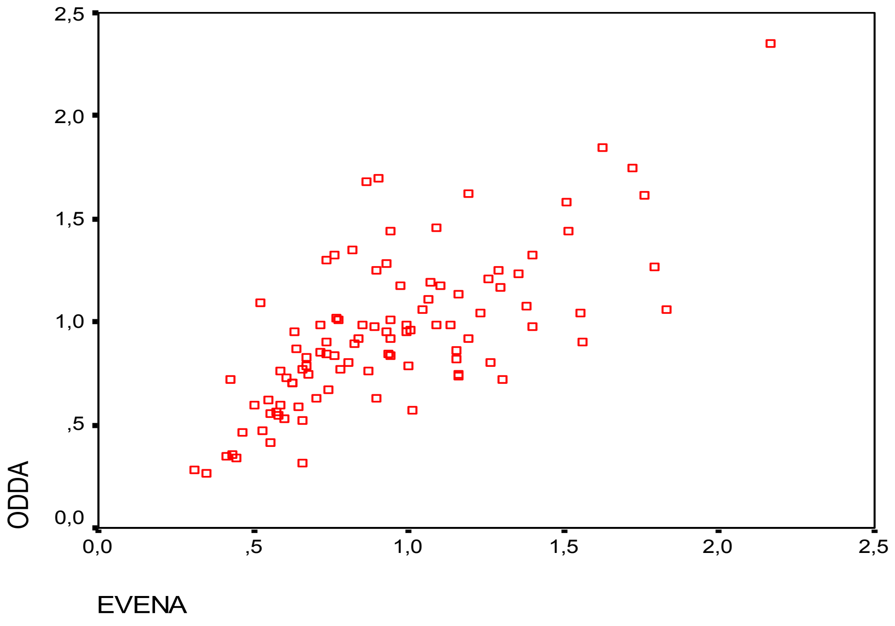


Figure 4.4.3. Scatterplot of 3P Discrimination Estimates (Odd vs. Even)

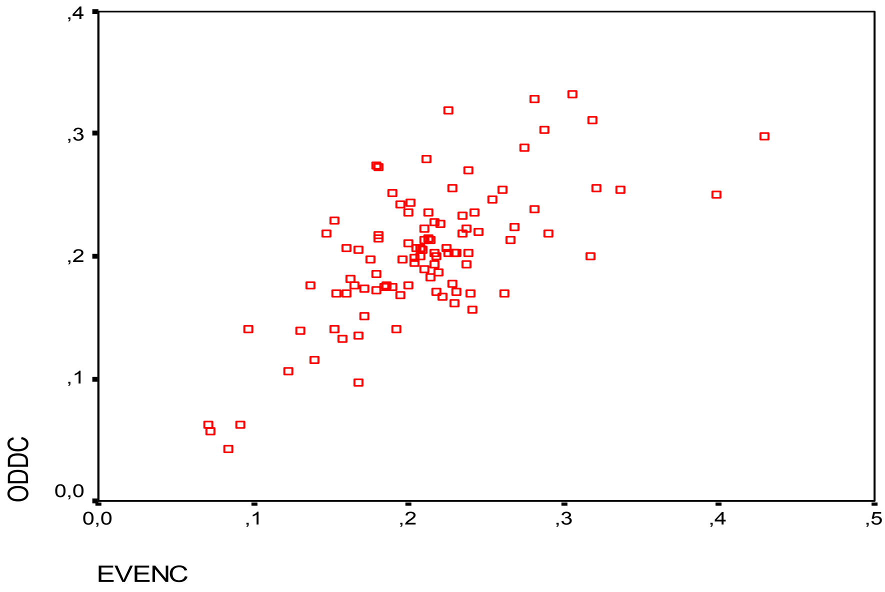


Figure 4.4.4. Scatterplot of 3P Pseudo-Chance Factor Estimates (Odd vs.Even

4.5. Checking Model Predictions of Actual and Simulated Test Results

Chi Square statistics of one-, two-, and three- parameter models were computed to determine which model fits the data best. Table 4.5.1. shows the number and percent of misfitted items in all three IRT models.

Table 4.5.1. The Number and Percent of Misfitted Items in Three IRT Models

Models	Number of misfitted Items	% of Fitted Items
One parameter	38*	62%
Two parameter	11*	89%
Three parameter	6*	94%

*Significant at 0.05 level

The results in the table above indicate that the three-parameter model has the smallest number of misfitted items, consequently, the highest percent of fitting items. It is concluded that the three-parameter model fits the data well according to final, chi square statistics.

Tables 4.5.2., 4.5.3. and 4.5.4. display the item parameters for the one-, two- and three- parameter models, respectively. Misfitted items are bolded in all three tables.

Figures 4.5.1., 4.5.2. and 4.5.3. represent the test information curves for the one-, two- and three- parameter models, respectively.

Table 4.5.2. IRT Item Parameters for The One-Parameter Model

ONE PARAMETER MODEL					
Item Number	**b-values**	**a-values**	**Item Number**	**b-values**	**a-values**
1	-1,889	,487	51	-,780	,487
2	-,290	,487	52	-1,913	,487
3	-,552	,487	**53**	-,347	,487
4	-1,796	,487	**54**	-,972	,487
5	-,160	,487	55	-1,200	,487
6	-1,413	,487	56	-,453	,487
7	-1,819	,487	**57**	-,461	,487
8	,763	,487	58	-2,289	,487
9	3,164	,487	59	-3,831	,487
10	,524	,487	60	-2,662	,487
11	,359	,487	61	,903	,487
12	-1,116	,487	62	,392	,487
13	-1,785	,487	63	-1,008	,487
14	-,152	,487	64	-,478	,487
15	-1,172	,487	**65**	,433	,487
16	-,999	,487	66	,408	,487
17	-,290	,487	**67**	-,330	,487
18	-1,842	,487	**68**	-1,373	,487
19	,541	,487	**69**	-2,206	,487
20	1,814	,487	70	-2,768	,487
21	,236	,487	71	2,359	,487
22	,685	,487	72	1,383	,487
23	,762	,487	**73**	1,197	,487
24	-,569	,487	74	-,569	,487
25	3,380	,487	75	,090	,487
26	-,192	,487	**76**	-1,210	,487
27	-1,016	,487	77	,375	,487
28	-3,570	,487	78	-,928	,487
29	-,265	,487	79	-1,598	,487
30	-1,556	,487	80	-1,373	,487
31	-1,228	,487	81	1,383	,487
32	-2,920	,487	**82**	-,910	,487
33	,318	,487	83	-,737	,487
34	-2,049	,487	84	1,353	,487
35	,025	,487	85	,334	,487
36	,466	,487	86	-,120	,487
37	-,095	,487	**87**	-,047	,487
38	,876	,487	88	,285	,487
39	,466	,487	89	-,927	,487
40	,558	,487	**90**	,050	,487
41	,285	,487	**91**	,815	,487
42	-,160	,487	**92**	-,152	,487
43	-1,363	,487	93	-,241	,487
44	-,661	,487	**94**	-1,413	,487
45	-,780	,487	95	-,273	,487
46	-1,325	,487	96	,050	,487
47	-1,315	,487	**97**	-,144	,487
48	-2,881	,487	98	-,919	,487
49	-1,685	,487	99	-,144	,487
50	-1,842	,487	**100**	1,454	,487

The mean of the item difficulty estimates (threshold) of the one-parameter model in Table 4.5.2. is -, 470 (Standard deviation = 1.274). The item difficulty estimates of the one-parameter model range from -3.831 to 3.380. The single fixed item discrimination estimate of the one-parameter model is .487.

```
TEST:   prof00

  STANDARD                                                              INFOR-
  ERROR                                                                 MATION
       ------------------------------------------------------------------
    .43|   *                      ++++++                     *          |14.1258
       |    *                   ++      ++                              |
    .41|                      ++          +                 *           |13.4195
       |     *               +             +               *            |
    .39|      *             +               +             *             |12.7132
       |       *           +                 +           *              |
    .36|        *         +                   +         *               |12.0069
       |         *       +                     +       *                |
    .34|          *                             +     *                 |11.3006
       |           **   +                        +   *                  |
    .32|             * +                           **                   |10.5943
       |              +*                          +                     |
    .30|             +  *                       ** +                    | 9.8880
       |                 ***                  **    +                   |
    .28|            +       ***           ****                          | 9.1817
       |           +           ***********           +                  |
    .26|          +                                   +                 | 8.4755
       |         +                                     +                |
    .24|                                                +               | 7.7692
       |        +                                                       |
    .21|       +                                         +              | 7.0629
       |      +                                           +             |
    .19|     +                                             +            | 6.3566
       |    +                                               +           |
    .17|                                                     +          | 5.6503
       |   +                                                  +         |
    .15|  +                                                    +        | 4.9440
       | +                                                      +       |
    .13|+                                                        +      | 4.2377
       |                                                          +     |
    .11|                                                           +    | 3.5314
       |                                                            ++  |
    .09|                                                              + | 2.8252
       |                                                               +|
    .06|                                                                | 2.1189
       |                                                                |
    .04|                                                                | 1.4126
       |                                                                |
    .02|                                                                |  .7063
       |                                                                |
    .00|                                                                |  .0000
       -+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
      -4.00   -3.00   -2.00   -1.00     .00    1.00    2.00    3.00    4.00

 MAXIMUM INFORMATION APPROXIMATELY   .1413D+02  AT     -2.0000
```

Figure 4.5.1. One- Parameter Model Test Information Curve

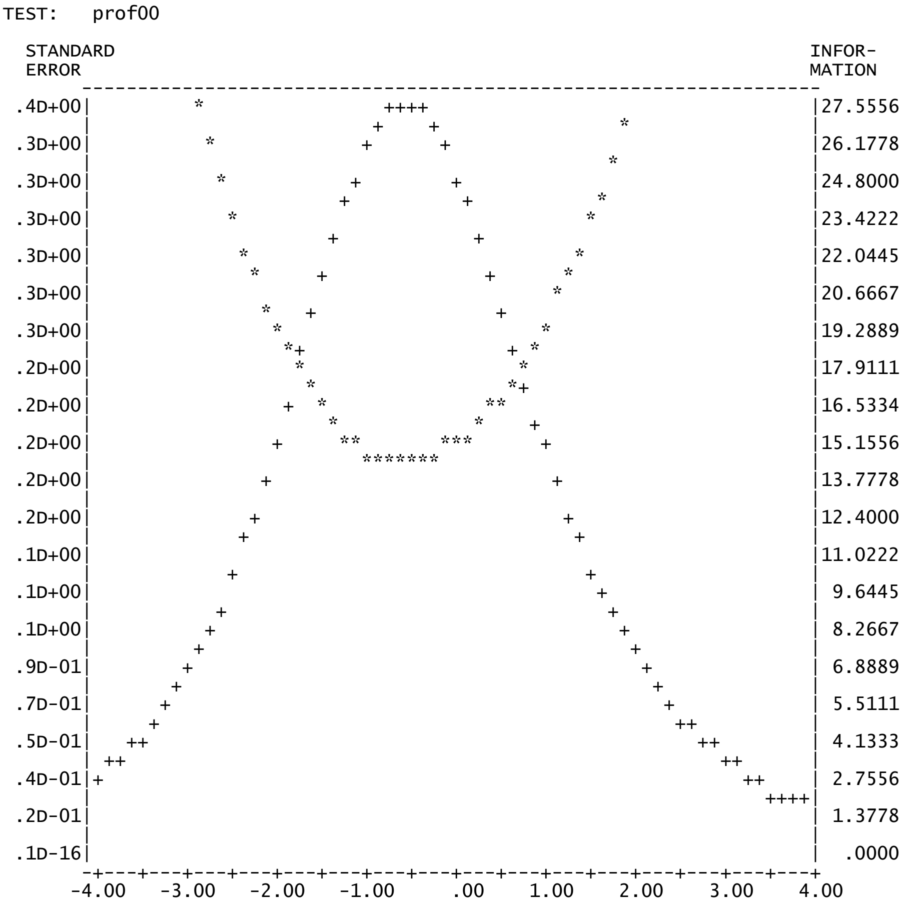
The Test Information Curve (TIF) of the one-parameter model is represented in Figure 4.5.1. above. The TIF of the 100 items displays the amount of information available at each θ level. Maximum information is provided at the ability score of -0.4, which the peak of the curve corresponds to. According to the U-shaped Test Standard Error of Measurement Function (TSEMF) of the 100 items, it seems that the lowest values of TSEM is provided between the ability levels of -1,00 and 0.5.

The mean of item difficulty indices (threshold) of the two-parameter model in Table 4.5.3. is -,334 (Standard deviation = 1.309). The item difficulty indices of the two-parameter model range from -5.502 to 4.312. The mean of item discrimination indices (slope) in the two-parameter model is.635 (Standard deviation = .236). The item discrimination indices of the two-parameter model range from a minimum of .185 to a maximum of 1.200.

Table 4.5.3. IRT Item Parameters for The Two-Parameter Model

TWO PARAMETER MODEL							
Item number	b-value	a-value	c-value	Item number	b-value	a-value	c-value
1	-2,084	,419	.0000	51	-,556	,787	.0000
2	-,219	,821	.0000	52	-1,167	,960	.0000
3	-,467	,604	.0000	53	-,249	,892	.0000
4	-1,337	,697	.0000	54	-,623	,933	.0000
5	-,130	,951	.0000	55	-,872	,739	.0000
6	-2,009	,313	.0000	56	-,342	,748	.0000
7	-1,352	,698	.0000	57	-,310	,994	.0000
8	1,117	,299	.0000	58	-1,615	,746	.0000
9	4,312	,329	.0000	59	-2,964	,645	.0000
10	,520	,465	.0000	60	-1,845	,759	.0000
11	,243	,715	.0000	61	,884	,475	.0000
12	-1,125	,467	.0000	62	,448	,392	.0000
13	-1,522	,577	.0000	**63**	-,750	,717	.0000
14	-,125	1,012	.0000	64	-,380	,673	.0000
15	-1,578	,332	.0000	65	,378	,537	.0000
16	-,649	,917	.0000	66	,313	,624	.0000
17	-,221	,800	.0000	**67**	-,301	,552	.0000
18	-1,294	,760	.0000	68	-,983	,749	.0000
19	,506	,498	.0000	69	-1,275	1,042	.0000
20	1,251	,782	.0000	70	-2,293	,595	.0000
21	,179	,608	.0000	71	1,681	,747	.0000
22	,631	,509	.0000	72	1,074	,646	.0000
23	,811	,429	.0000	**73**	,975	,603	.0000
24	-,383	,928	.0000	74	-,427	,736	.0000
25	4,126	,373	.0000	75	,036	,812	.0000
26	-,187	,517	.0000	76	-,689	1,200	.0000
27	-1,933	,228	.0000	77	,201	,950	.0000
28	-2,662	,678	.0000	78	-,672	,752	.0000
29	-,292	,425	.0000	79	-1,126	,762	.0000
30	-1,156	,704	.0000	80	-1,231	,542	.0000
31	-1,192	,490	.0000	81	1,343	,483	.0000
32	-5,502	,236	.0000	82	-1,253	,323	.0000
33	,378	,375	.0000	83	-,923	,360	.0000
34	-1,137	1,147	.0000	84	1,162	,565	.0000
35	-,022	1,068	.0000	85	,341	,446	.0000
36	,429	,506	.0000	86	-,109	,668	.0000
37	-,095	,562	.0000	87	-,072	,290	.0000
38	1,816	,206	.0000	88	,180	,753	.0000
39	,515	,409	.0000	89	-1,025	,417	.0000
40	,553	,464	.0000	**90**	,082	,262	.0000
41	,203	,660	.0000	91	1,092	,330	.0000
42	-,139	,686	.0000	92	-,236	,283	.0000
43	-1,284	,508	.0000	93	-,199	,690	.0000
44	-,442	,918	.0000	94	-,827	1,087	.0000
45	-,529	,868	.0000	95	-,248	,566	.0000
46	-,860	,881	.0000	96	,033	,545	.0000
47	-1,197	,531	.0000	97	-,125	,721	.0000
48	-1,547	1,150	.0000	98	-,645	,798	.0000
49	-1,063	,915	.0000	99	-,127	,677	.0000
50	-2,413	,345	.0000	**100**	3,348	,185	.0000

The TIF of the 100 items in the two-parameter model can be seen in Figure 4.5.2. The test provides maximum information at the ability level of -0.5, since the peak of the curve corresponds to this ability level. For the two-parameter model, the lowest parts of the U-shaped TSEMF falls between the ability scores of -1.00 and 0.00, which indicates the minimum TSEM.



MAXIMUM INFORMATION APPROXIMATELY .2756D+02 AT -2.0714

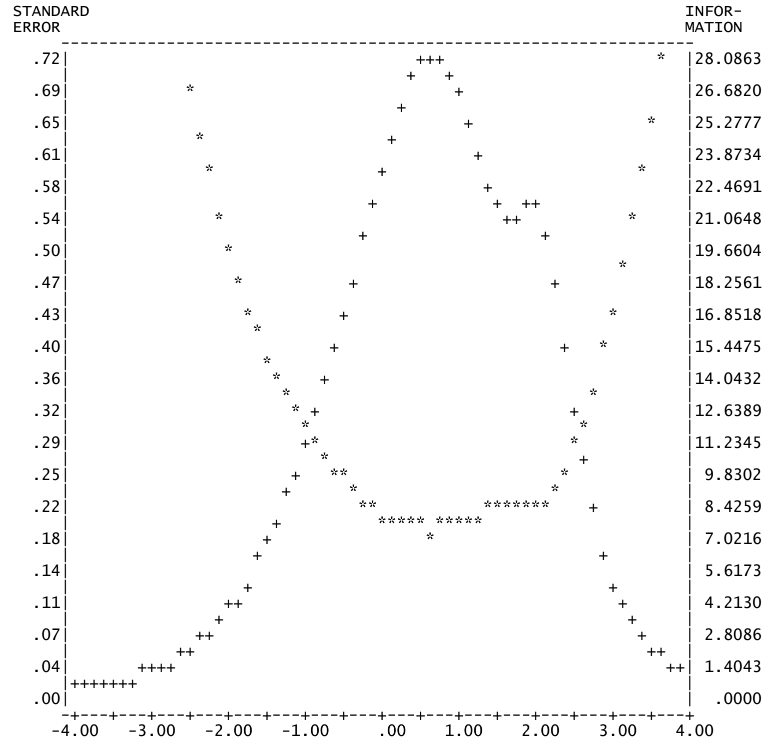
Figure 4.5.2. Two- Parameter Model Test Information Curve

Table 4.5.4. IRT Item Parameters for The Three-Parameter Model

THREE PARAMETER MODEL							
Item number	b-value	a-value	c-value	Item number	b-value	a-value	c-value
1	-1,592	,413	,233	51	,146	,919	,200
2	,170	1,040	,176	52	,932	,960	,216
3	-,029	,693	,178	53	,101	1,097	,164
4	-,832	,787	,276	54	,446	,934	,118
5	,185	1,081	,143	55	,249	,932	,297
6	-1,251	,336	,215	56	,078	,881	,187
7	-,936	,754	,246	57	,056	1,244	,181
8	1,811	,779	,269	58	1,467	,691	,206
9	2,034	2,750	,073	59	3,276	,534	,199
10	,993	,660	,167	60	1,800	,672	,208
11	,728	1,210	,206	61	1,293	,979	,198
12	-,522	,521	,223	62	1,253	1,127	,306
13	-1,097	,588	,237	**63**	,369	,779	,193
14	,284	1,556	,196	64	,145	,865	,220
15	-,808	,349	,225	65	,789	,682	,145
16	-,348	,966	,178	66	,769	,944	,182
17	,300	1,215	,232	**67**	,066	,623	,139
18	-,967	,759	,238	68	,679	,745	,196
19	,813	,613	,110	69	1,136	1,006	,198
20	1,274	1,051	,049	70	2,317	,506	,212
21	,672	,891	,192	71	1,556	1,324	,058
22	1,104	1,448	,240	72	1,235	1,070	,110
23	1,287	,869	,210	73	1,151	1,255	,142
24	,094	1,342	,233	74	,099	,949	,232
25	2,229	1,973	,063	75	,538	1,508	,226
26	,358	,620	,189	76	,464	1,258	,163
27	-,574	,252	,258	77	,440	1,171	,100
28	-2,722	,617	,194	78	,306	,819	,187
29	,781	,770	,331	79	,785	,781	,223
30	-,821	,687	,220	80	,805	,557	,206
31	-,510	,522	,267	81	1,478	1,133	,160
32	-4,893	,234	,208	82	,135	,384	,279
33	1,299	,951	,313	83	,144	,460	,285
34	-1,067	1,023	,167	84	1,330	1,078	,134
35	,295	1,456	,146	85	,975	,676	,211
36	1,040	1,002	,241	86	,657	1,853	,324
37	,690	,946	,286	87	1,369	,561	,340
38	2,449	,970	,315	88	,624	1,203	,191
39	1,088	,576	,181	89	-,409	,455	,207
40	1,163	1,095	,254	90	1,519	,446	,305
41	,747	1,294	,233	91	1,677	,912	,262
42	,287	,849	,175	92	1,441	,876	,422
43	-,811	,527	,213	93	,258	,871	,189
44	-,064	1,114	,191	94	,663	1,040	,151
45	-,071	1,123	,227	95	,487	,882	,272
46	-,683	,848	,146	96	,639	,798	,221
47	-,765	,544	,204	97	,113	,776	,100
48	-1,601	1,020	,169	98	,348	,848	,161
49	-,814	,932	,199	99	,517	1,119	,262
50	-1,928	,339	,202	**100**	2,064	2,248	,242

According to the indices observed above in the three-parameter model the mean of item difficulty indices is 0.06 (Standard deviation = 1.1934). Item difficulty indices range from -4.893 to 2.449. The mean of item discrimination indices is .913 (Standard deviation = .3961), with a minimum of .234 to a maximum of 2.750. The mean of the pseudo-chance factor indices (asymptote) is .206 (Standard deviation 6.28E-02). This index ranges from .049 to .422 in the three-parameter model.

Figure 4.5.3. shows the TIF of the three-parameter model. It can clearly be seen that the test provides the maximum information at the .05 ability level, which the peak of the curve corresponds to. The TSEMF of the 100 items is lowest between the ability scores of 0.00 and 1.50. It seems that higher information is obtained in the three-parameter model.



```
MAXIMUM INFORMATION APPROXIMATELY    .2809D+02  AT     -1.3571
```

Figure 4.5.3. Three- Parameter Model Test Information Curve

Examples of good and poor IIFs are presented in Appendix E. Items 71, 9, 25 and 77 provide good examples of IIFs with their high difficulty indices, steep slopes and low involvement of guessing.

The increasing difficulty of these items causes the S-shaped curve to shift to the higher end of the ability scale. Item 77, when compared to the other three items seems less difficult in this respect.

The steepness of the curves indicates the discrimination parameter of the item. The steeper the curve, the more discriminating it is. Among the four items, item 9 seems to be the most discriminating and item 77 seems to be the least discriminating among the four items.

The guessing factor causes the baseline of the curve to raise. Guessing is lowest in item 71, whereas it is slightly more in item 77.

Items 27, 32 and 50 are examples of poor IIFs because no information is provided. All three items seem to have negative difficulty, no slope at all and a high guessing factor.

4.6. Predictive Validity Analyses

After establishing the fit of BUEPE data to the three-parameter model, the predictive validity analyses were carried out by using the estimates obtained from the three-parameter model.

To determine whether BUEPE predicts success in DEC, various relationships were examined by computing correlation coefficients and scatterplots. Appendix F contains the scatterplots obtained. Table 4.6.1. presents the correlation coefficients obtained between different variables.

Table 4.6.1. Correlation Between DEC Grades versus Total Scores and Ability Estimates

	DEC1	DEC2
BUEPE Total Scores	.754	.687
3P Model Ability Estimates	.772	.701
60 High Information Items Total Scores	.749	.680
60 High Information Items 3P Ability Estimates	.768	.692
35 High Information Items 3P Total Scores	.715	.659
35 High Information Items 3P Ability Estimates	.729	.661
Grammar Sub-test Total Scores	.624	.579
Grammar Sub-test 3P Ability Estimates	.629	.586
Reading Sub-test Total Scores	.724	.644
Reading Sub-test 3P Ability Estimates	.742	.658
Vocabulary Sub-test Total Scores	.528	.461
Vocabulary Sub-test 3P Ability Estimates	.572	.502

Correlation significant at 0.01 level

4.6.1. Total Scores and Three-parameter Ability Estimates vs. DEC Grades

First, BUEPE total scores were correlated with the first and second semester DEC passing grades in freshmen. The BUEPE total scores had a correlation of r = .754, p= .000 significant at 0.01 alpha level with the first semester DEC passing grades. On the other hand, the correlation between the BUEPE total scores and DEC second semester passing grades was r = .687 significant at the 0.01 alpha level.

Secondly, the correlation between the ability estimates of the three-parameter model and the first semester DEC passing grades was r = .772, p= .000 significant at 0.01 alpha level. The ability estimates of the three-parameter model and the second semester DEC passing grades had a correlation of r = .701, p= .000 significant at 0.01 alpha level.

The Item Information Functions (IIF) obtained from the three-parameter model were examined. The IIFs of the one hundred items of the BUEPE ranged from a high of 4.7365 to a low of .0266. The IIFs of the three-parameter model can be seen in Appendix G1 and G2.

4.6.2. Total Scores and Ability Estimates of Sixty High Information Items vs. DEC Grades

Thirdly, the highest information yielding sixty items, which had information functions of ≥ .3026 were selected, see Appendix G3. The correlation between the total scores obtained from the sixty highest-information items in the three-parameter model and the first semester DEC passing grades was r = .749, p= .000 significant at 0.01 alpha level. However, the second semester DEC passing grades had a correlation of r = .680, p= .000 significant at 0.01 alpha level with the same total scores.

Next, the ability estimates obtained by using the sixty highest-information items in the three-parameter model was correlated with the first semester DEC passing grades. The correlation that was found out was r = . 768, p= .000 significant at 0.01 alpha level. The same ability estimates had a correlation of r = .692, p= .000 significant at alpha level 0.01 with the second semester DEC passing grades.

The total scores and the ability estimates obtained by using the sixty high information items seem to predict the DEC first and second semester passing grades similarly well when compared with the correlations of the general total scores and three-parameter ability scores with the DEC first and second semester passing grades.

The TIF obtained from running the 60 high information items in the best fitting three-parameter model is presented in Figure 4.6.1. According to the TIF the sixty high information items provide information between the -0.5 and 2.5 ability levels. The maximum information is provided for the 0.5 ability level with minimum TSEM.

```
   STANDARD                                                               INFOR-
 ERROR                                                                  MATION
      -----------------------------------------------------------------
  1.06|            *                      +++                         |21.5868
      |                                  +   +                       *|
  1.01|                                                               |20.5075
      |                                 +     +                       |
   .96|                                        +                      |19.4281
      |             *                  +                            * |
   .91|                                         +                     |18.3488
      |                               +                               |
   .85|                                          +                    |17.2695
      |              *                                             *  |
   .80|                              +            +                   |16.1901
      |                                            +                  |
   .75|               *                             ++++          *   |15.1108
      |                             +                   +             |
   .69|                                                               |14.0314
      |                *           +                             *    |
   .64|                                                  +            |12.9521
      |                 *                                             |
   .59|                           +                             *     |11.8728
      |                  *                                +           |
   .53|                          +                                    |10.7934
      |                   *                                    *      |
   .48|                    *                               +          | 9.7141
      |                         +                             *       |
   .43|                     *                                         | 8.6347
      |                      * +                             *        |
   .37|                       *                             +         | 7.5554
      |                       +*                                      |
   .32|                      +  **                         *          | 6.4760
      |                           *                       *  +        |
   .27|                     +      **               ******            | 5.3967
      |                              ***       *****          +       |
   .21|                    +            *******                       | 4.3174
      |                   +                                    +      |
   .16|                  +                                            | 3.2380
      |                 +                                       +     |
   .11|                +                                         +    | 2.1587
      |              ++                                           ++  |
   .05|            ++                                               ++| 1.0793
      |         ++++                                                  |
   .00|+++++++++                                                      |  .0000
      -+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
     -4.00   -3.00   -2.00   -1.00     .00    1.00    2.00    3.00    4.00

MAXIMUM INFORMATION APPROXIMATELY   .2159D+02  AT     -1.3571
```

Figure 4.6.1. Test Information Curve of 3P 60 High Information Items

4.6.3. Total Scores and Ability Estimates of Thirty-five High Information Items vs. DEC Grades

Furthermore, the correlation between the total scores obtained from the thirty-five highest-information items with information functions ≥ .5003, in the three-parameter model and the first semester DEC passing grades was found to be r = .715, p= .000 significant at 0.01 alpha level. Appendix G4 presents the thirty-five highest information items. On the other hand, the second semester DEC passing grades had a correlation of r = . 659, p= .000 significant at 0.01 alpha level with the same total scores.

The ability estimates obtained by using the thirty-five highest-information items in the three-parameter model were correlated with the first and second semester DEC passing grades. While the correlation with the first semester DEC passing grades was r = .729, the correlation with second semester DEC passing grades was r = .661 p= .000 significant at 0.01 alpha level.

The correlations between the total scores and ability estimates obtained by using the thirty-five high information items with the DEC first and second semester passing grades seem to yield slightly lower correlations when compared with the correlations of the general total scores and three-parameter ability scores with the DEC first and second semester passing grades.

```
STANDARD                                                              INFOR-
ERROR                                                                 MATION
     ------------------------------------------------------------------
 1.45|                                     ++                         |13.3783
     |                                    +  ++                       |
 1.37|             *                     +                            |12.7094
     |                                  +      +                      |
 1.30|                                          +                    *|12.0405
     |                                                 +              |
 1.23|              *                  +         +    + +             |11.3716
     |                                            +                   |
 1.16|                                +            + +   +            |10.7027
     |               *                              +               * |
 1.08|                                                                |10.0338
     |                               +                                |
 1.01|                *                                   +           | 9.3648
     |                                                             *  |
  .94|                              +                                 | 8.6959
     |                 *                                              |
  .87|                                                     +      *   | 8.0270
     |                  *          +                                  |
  .80|                                                                | 7.3581
     |                   *        +                              *    |
  .72|                                                                | 6.6892
     |                    *                                 +         |
  .65|                     *     +                              *     | 6.0203
     |                                                                |
  .58|                      *   +                              *      | 5.3513
     |                       *                               +        |
  .51|                        *+                              *       | 4.6824
     |                         *                                      |
  .43|                        + *                            *+       | 4.0135
     |                           **                         *         |
  .36|                       +     *                       *          | 3.3446
     |                              **             **     *    +      |
  .29|                      +         *************  *****            | 2.6757
     |                     +                                    +     |
  .22|                    +                                           | 2.0068
     |                   +                                       +    |
  .14|                 ++                                         +   | 1.3378
     |                +                                            +  |
  .07|             +++                                              ++|  .6689
     |         ++++                                                   |
  .00|+++++++++                                                       |  .0000
     -+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
    -4.00   -3.00   -2.00   -1.00     .00    1.00    2.00    3.00    4.00

MAXIMUM INFORMATION APPROXIMATELY   .1338D+02  AT     -1.2857
```

Figure 4.6.3. Test Information Curve of 3P 35 High Information Items

The TIF obtained from running the thirty-five high information items in the best fitting three-parameter model is presented in Figure 4.6.3.

According to the TIF the sixty high information items provide information between the -0.5 and 2.8 ability levels. The maximum information is provided for the 1.00 ability level with minimum TSEM.

4.6.4. Content Sampling of Sixty and Thirty-five High Information Items

The content sampling of not only the sixty highest information items but also the thirty-five highest information items was examined. The original one hundred items exam had 40 grammar, 40 reading and 20 vocabulary items. The highest sixty items was composed of 20 grammar, 28 reading and 12 vocabulary items. The proportion of items seem to be approximately the same, still the reading section has borrowed 4 items from the grammar section. As for the highest thirty-five items, the number of items in the grammar section is 13, reading 15 and vocabulary 7. Similarly, the distribution of the number of items in the sub-tests seem to be approximately the same. There is an extra item in the reading section that is only one item more than expected according to the proportion. Therefore, the content sampling remains the same when the number of items in the exam is reduced by making use of the highest items in the exam.

Moreover, the reliability of the scores obtained from the sixty high information items was α = .92, which is considered to be quite high. For the thirty-five high information items this alpha coefficient dropped to α = .88, which is considered to be a moderate correlation for good tests of vocabulary, structure and reading (Lado, 1961 in Hughes, 1989).

These results imply that reduction in the number of items affects the proportion of the items in the sub-tests at a minimum level. Moreover, without sacrificing the reliability of the scores obtained from the exam, the number of items in the BUEPE can be reduced by using the highest information items in the exam but only by careful consideration of what the number of sufficient items can be.

4.6.5. Total Scores and Ability Estimates of Sub-tests vs. DEC Grades

The following results display the relationship of BUEPE sub-tests with the first and second semester DEC passing grades presented in Table 4.6.1.

The total scores obtained by using the grammar sub-test was correlated first, with the first semester and then with the second semester DEC passing grades. The results were, r = .624 and r =. 579 respectively, p= .000 significant at 0.01 alpha level.

The ability estimates obtained by using the grammar sub-test in the three-parameter model was correlated with the first semester DEC passing grades and a correlation of r = . 629, p= .000 significant at alpha level 0.01 was found. Also, a correlation of r = . 586, p= .000 significant at alpha level 0.01 was found between the total score obtained by the grammar sub-test and the second semester DEC passing grades.

The TIF obtained from running the grammar items in the best fitting three-parameter model is presented in Figure 4.6.5. The grammar items provide maximum information at both the -0.5 and 2.5 ability levels.

```
TEST:   prof00

  STANDARD                                                              INFOR-
  ERROR                                                                 MATION
       ------------------------------------------------------------------
   1.21|          *                                      ++             |10.1783
       |                                                                |
   1.15|           *                      ++++                          | 9.6694
       |                                 +              +  +           *|
   1.09|            *                         +                         | 9.1605
       |                                +      +                        |
   1.03|             *                                                * | 8.6515
       |                                        +      +                |
    .97|              *                +         +          +           | 8.1426
       |                                                             *  |
    .91|               *              +           +   +                 | 7.6337
       |                                           + +                  |
    .85|                *                           +               *   | 7.1248
       |                             +                       +          |
    .79|                 *                                              | 6.6159
       |                  *                                        *    |
    .73|                            +                                   | 6.1070
       |                   *                                      *     |
    .67|                    *      +                                    | 5.5981
       |                                                      +         |
    .61|                     *                                   *      | 5.0891
       |                      *   +                                     |
    .55|                       *                                *       | 4.5802
       |                        *+                             +        |
    .49|                         *                             *        | 4.0713
       |                        + *                                     |
    .42|                           *                          *         | 3.5624
       |                       +    **                       *  +       |
    .36|                      +       *          ******                 | 3.0535
       |                               **********      **  **           |
    .30|                     +                           **      +      | 2.5446
       |                    +                                           |
    .24|                   +                                      +     | 2.0357
       |                  +                                        +    |
    .18|                 +                                              | 1.5267
       |               ++                                           +   |
    .12|             ++                                              ++ | 1.0178
       |          +++                                                  +|
    .06|     +++++                                                      |  .5089
       |+++++                                                           |
    .00|                                                                |  .0000
       -+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
      -4.00   -3.00   -2.00   -1.00     .00    1.00    2.00    3.00    4.00

 MAXIMUM INFORMATION APPROXIMATELY   .1018D+02  AT     -.4286
```

Figure 4.6.5. Test Information Curve 3P Grammar Items

The total scores obtained by using the reading sub-test were correlated with the first semester DEC passing grades. The correlation obtained was r = .724, p= .000 significant at alpha level 0.01. The correlation of the total scores obtained by the reading sub-test with the second semester DEC passing grades was r = .644, p= .000 significant at alpha level 0.01.

The ability estimates obtained by the items in the reading sub-test was correlated with the first semester DEC passing grades and a correlation of r = .742, was found p= .000 significant at 0.01 alpha level. The correlation between the ability estimates obtained by the items in the reading sub-test and the second semester DEC passing grades fell to r = .658, p= .000 significant at 0.01 alpha level.

The TIF obtained from running the items in the reading sub-test in the best fitting three-parameter model is presented in Figure 4.6.6. below. According to TIF, the items in the reading sub-test provide information between the -1.0 and 1.6 ability levels. The maximum information is provided for the 0.5 ability level with minimum TSEM between -0.4 and -1.0.

```
TEST:   prof00

  STANDARD                                                              INFOR-
  ERROR                                                                 MATION
       ------------------------------------------------------------------
    .92|                                  +++                           |11.9801
       |            *                    +   +                          |
    .88|                                +     +                         |11.3811
       |                               +       +               *        |
    .83|             *                +                                 |10.7821
       |                                        +                       |
    .78|                             +                                  |10.1831
       |              *                          +            *         |
    .74|                            +                                   | 9.5841
       |                                          +                     |
    .69|               *           +                         *          | 8.9851
       |                                           +                    |
    .65|                *         +                         *           | 8.3861
       |                                            +                   |
    .60|                 *                                              | 7.7871
       |                         +                         *            |
    .55|                  *                          +                  | 7.1880
       |                        +                         *             |
    .51|                   *                          +                 | 6.5890
       |                    *  +                         *              |
    .46|                     *                                          | 5.9900
       |                      *                        +*               |
    .42|                      +                        *                | 5.3910
       |                       **                     * +               |
    .37|                     +   *                   *                  | 4.7920
       |                          *                **                   |
    .32|                    +      ***           **      +              | 4.1930
       |                   +          ***********                       |
    .28|                                                  +             | 3.5940
       |                  +                                             |
    .23|                                                   +            | 2.9950
       |                 +                                              |
    .18|                +                                   +           | 2.3960
       |               +                                     +          |
    .14|              +                                       +         | 1.7970
       |             +                                         +        |
    .09|           ++                                           +       | 1.1980
       |         ++                                              ++     |
    .05|     ++++                                                  ++   |  .5990
       |+++++                                                        +++|
    .00|                                                                |  .0000
       -+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
      -4.00   -3.00   -2.00   -1.00     .00    1.00    2.00    3.00    4.00

 MAXIMUM INFORMATION APPROXIMATELY   .1198D+02  AT     -1.5000
```

Figure 4.6.6. Test Information Curve of 3P Reading Items

The correlations between the total scores obtained by using the vocabulary sub-test and the first semester and the second semester DEC passing grades were the lowest among all three sub-tests. The results were, r = .528 and r =.461 respectively, p= .000 significant at 0.01 alpha level.

The correlation between the ability estimates obtained by the items in the vocabulary sub-test with the first semester DEC passing grades yielded a correlation of r = .572, p= .000 significant at 0.01 alpha level, whereas, the same ability estimates' correlation with second semester DEC passing grades was r = .502, p= .000 significant at 0.01 alpha level.

The TIF obtained from running the items in the vocabulary sub-test in the best fitting three-parameter model is presented in Figure 4.6.7. below. According to the TIF the vocabulary items provide information between the -0.5 and 2.5 ability levels. The maximum information is provided for the 0.8 ability level with minimum TSEM between 0.4 and 1.5.

```
TEST:   prof00

  STANDARD                                                              INFOR-
  ERROR                                                                 MATION
       ------------------------------------------------------------------
   2.45|                                     +++                        | 5.3800
       |             *                          +                       |
   2.33|                                    +                           | 5.1110
       |                                   +     +                      |
   2.20|                                                                | 4.8420
       |              *                           +                     |
   2.08|                                  +        +                    | 4.5730
       |                                                                |
   1.96|                                 +          +                   | 4.3040
       |               *                                                |
   1.84|                                             +                  | 4.0350
       |                                +                               |
   1.71|                *                             +                 | 3.7660
       |                                               +                |
   1.59|                               +                +               | 3.4970
       |                 *                               +              |
   1.47|                              +                   +             | 3.2280
       |                  *                                            *|
   1.35|                                                   +            | 2.9590
       |                             +                      +         * |
   1.22|                   *                                            | 2.6900
       |                    *       +                        +       *  |
   1.10|                                                            *   | 2.4210
       |                     *     +                          +         |
    .98|                                                           *    | 2.1520
       |                      *   +                            +  *     |
    .86|                       *                                 *      | 1.8830
       |                        *+                              +       |
    .73|                         *                             *        | 1.6140
       |                        + **                          *  +      |
    .61|                       +    **                      **          | 1.3450
       |                              **               *****      +     |
    .49|                      +         ***        ****            +    | 1.0760
       |                     +             ********                     |
    .37|                    +                                       +   |  .8070
       |                   +                                         +  |
    .24|                  +                                           ++|  .5380
       |                ++                                              |
    .12|              ++                                                |  .2690
       |         +++++                                                  |
    .00|+++++++++                                                       |  .0000
       -+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
      -4.00   -3.00   -2.00   -1.00     .00    1.00    2.00    3.00    4.00

 MAXIMUM INFORMATION APPROXIMATELY   .5380D+01  AT     -1.2857
```

Figure 4.6.7. Test Information Curve of 3P Vocabulary Items

All the correlation coefficients computed for establishing the predictive validity of the BUEPE on the DEC first and second semester passing grades have two general implications. All the correlations with the ability scores seem to be higher as opposed to total score correlations. In addition, all correlation coefficients computed with DEC first semester grades are higher than correlation coefficients computed with DEC second semester passing grades.

CHAPTER V

CONCLUSIONS AND IMPLICATIONS

This final chapter presents the discussion of the findings of the study, draws conclusions and suggests implications for further research.

5.1. Discussion of the Findings

In this section first the prerequisite analyses are reviewed. The assumptions of IRT models are discussed. The discussion of the results of the expected model features is followed by the discussion of model predictions of actual test results. Finally, the results of the predictive validity analyses are discussed.

The analyses in this study were carried out on the data obtained from 699 subjects' results from the BUEPE. The mean score in the exam was 57 and the exam in general was viewed to be at a moderate difficulty level.

To check whether the unidimensionality assumption was met by the BUEPE data, principal component analysis was run. As Hambleton, Swaminathan and Rogers (1991) indicate a dominant factor underlying the test performance of the examinees is necessary to claim that the data meets the unidimentionality assumption. With this purpose, the eigenvalues and the scree-test were interpreted. The first eigenvalue was 14.26, the second was 2.90 and the third was 1.79. The sharp drop from the first eigenvalue to the second was evidence to claim that the BUEPE data was unidimensional. The scree-test in Figure 4.2.1. supported the findings.

As for the local independence assumption, the total inter-item correlations were compared with the inter-item correlations of high performers and low performers. The means of the inter-item correlations of the high and low ability groups were found to be close to zero and lower than the total inter-item correlations. Therefore, the items of the BUEPE were considered to be locally independent.

To check whether the equal discrimination indices assumption of the one-parameter model is viable, the discrimination indices obtained from classical item analysis were plotted. The histogram was slightly skewed to the left, implying that the equal discrimination indices assumption of the one-parameter model was not met. This finding was consistent with the findings of Kılıç (1999).

Hughes (1989) mentions the "unknowable effect" of guessing on the scores that the students get when multiple-choice item format is used. Since BUEPE is a multiple-choice exam it was relevant to question the role of guessing in the exam. Whether there was a guessing factor affecting the results obtained from the BUEPE was examined by selecting the most difficult five items and testing these items on the students in the low ability group. The means of four out of the five items were close to zero and below .25. However, the mean of item 100 was not close to zero but close to .25, with a lower percentage of incorrect answers when compared with the other four items. The results indicate that there might be a guessing factor in the exam so the three-parameter model was taken into consideration.

To decide if the BUEPE was a non-speeded exam the ratio of the variance of omitted items to the variance of the items answered incorrectly was interpreted. The ratio was calculated as .16. Since this value was close to zero it was concluded that the data met the assumption of non-speededness.

As for checking the expected model features the invariance of ability and item parameters different samples of items and different subgroups, respectively were examined.

The invariance of ability parameter estimates was established on the fifty easy and fifty hard items, in addition to the first and second fifty items in the exam and scatterplots were plotted. The results, as shown in Table 4.3.1.1. indicated that the two- and three- parameter models yielded slightly higher correlations when compared with the one-parameter model. However, all the correlations across different samples of items in the one-, two-, and three- parameter models were quite high, emphasizing that invariance of ability parameter estimates holds across easy versus hard and first fifty versus second fifty items.

The invariance of item parameter estimates was examined across odd cases versus even cases ability groups. Table 4.4.1. displays the correlation coefficients computed across these groups. Fan (1998) indicated that as the dissimilarity between samples increased, the invariance of item discrimination indices decreased. Fan's findings and Çalışkan's (2000) similar findings were supported in this study.

As for the odd cases versus even cases ability groups, which are considered to be less dissimilar according to Fan (1998), it can be said that invariance is established with quite high correlation coefficients. Since Hambleton, Swaminathan and Rogers (1991) suggest that it is possible to talk about the "degree" of invariance it is possible to conclude that invariance holds at a high degree. The b parameters across all three models yielded very high correlations. The a parameters of the two- and three-parameter models and the c parameter of the three-parameter model yielded relatively lower correlations but still were invariant across odd versus even cases ability groups.

For checking the model predictions of actual test results chi-square statistics were examined. The number of misfitted items and percentage of fitting items are presented in Table 4.5.1. Since the three-parameter model has the smallest number of misfitted items, it was concluded that the three-parameter model best fitted the data.

In addition, the TIFs of all three models were examined to view the range of ability levels that the BUEPE yields highest information in, because knowing the points which the exam provides more information at is helpful in discriminating among individuals who have ability scores falling in these regions according to Crocker and Algina, 1986. In the one parameter model the BUEPE yielded information between the -1.5 and 1.5 ability levels with a maximum information at the -0.4 ability level, whereas in the two-parameter model the exam provided information between the -1.5 and 1.00 ability levels with a maximum information of -0.5. The best fitting three-parameter model provided information between the -0.5 and 2.5 ability levels and the maximum information was provided at the 1.00 ability level.

As a result, the fit of the data to the three-parameter model of IRT was established on basis of the fact that there could be a guessing factor needed to be taken into consideration also because higher chi-square statistics and higher information were obtained in the three-parameter model.

Once the fit of the data to the three-parameter model of IRT was established, the estimates obtained from the analyses were used in determining the predictive validity of BUEPE on DEC freshmen first and second semester passing grades. The relationship was examined at various dimensions by computing correlation coefficients and scatterplots. Table 4.6.1. presents the correlation coefficients computed with different estimates. Appendix F contains the scatterplots of these correlations.

Cronbach (1990) and Alderson, Clapham and Wall (1995) emphasized that .30 or .40 are satisfactory correlation coefficients for predictive validity studies. However, the results of this study displayed higher correlation coefficients across all the different variables since the measure and the criterion were measuring similar traits.

The results of the correlations obtained from BUEPE total scores versus DEC first semester and second semester passing grades were considered to be moderately high, with a slightly lower correlation for DEC second semester passing grades.

In comparison with the BUEPE total scores the ability estimates of the three-parameter model seemed to yield slightly higher correlations for both DEC freshmen first and second semester passing grades. Also, this correlation of the ability estimates with DEC first and second semester passing grades with r = .772 and r = .701, respectively, seemed to yield the highest correlation among others in Table 4.6.1. This also indicated that the three-parameter model ability estimates were the best predictors of DEC passing grades.

The item information functions (IIF) obtained from the three-parameter model were ordered from the highest to the lowest and the sixty highest information items were selected, total scores and ability estimates computed only from these items were obtained. The total scores of the sixty items and the ability estimates obtained from those sixty items were correlated with DEC first and second semester passing grades. The correlation coefficients obtained were closely comparable with that of BUEPE total scores and three-parameter ability estimates versus DEC first and second semester passing grades. Therefore, it was concluded that the estimates of the sixty items were similarly good estimates of the DEC first and second semester passing grades.

As for the correlations obtained from the thirty-five highest information items versus DEC first and second semester passing grades, slightly lower correlations were observed. The correlation between total scores of thirty-five items was r = .715 and r = .659, respectively for DEC first and second semester passing grades whereas ability estimates of thirty-five items was r =.729 and r =.661, respectively.

In order to interpret the correlations obtained from the sixty and thirty-five highest information items accurately the content sampling and the reliability of the scores obtained from the total scores and the ability scores of sixty and thirty-five items should be taken into consideration.

As for the content sampling, it was concluded that the content sampling remained approximately the same when the number of items in the exam was reduced by making use of the highest items in the exam.

Moreover, the reliability of the scores obtained from the sixty high information items of the exam was α = .92, which was slightly lower than the reliability coefficient observed for the whole test (α = .93). The reliability of the scores obtained from the thirty-five high information items of the exam was α = .88, which was considered to be a moderately high correlation coefficient. While consistency of the scores obtained from sixty highest items was nearly the same, the consistency of the scores obtained from thirty-five highest items seemed to drop to some extent. This explained the slight decrease in the correlation coefficients between the thirty-five highest information items total scores and ability estimates versus DEC first and second semester passing grades.

Hughes (1989) underlines the fact that we must demand greater reliability when taking decisions that are more important. Since whether a student passes or fails a proficiency test is an important decision to be given, a decrease to an alpha level of α = .88 when the number of items in the test is reduced to thirty-five could be troublesome. Hughes (1989) emphasizes that it is important to construct a test long enough so that satisfactory reliability is achieved; however, it should not be too long to make the examinees bored or tired because then the results may be unrepresentative of their ability. It can be concluded that the highest information sixty item form of the exam may yield more reliable results when compared with the results obtained from the highest information thirty-five items of the exam.

The grammar, reading and vocabulary sub-tests of the BUEPE were also run and examined separately. The results presented in Table 4.6.1. indicated that the grammar and vocabulary sub-tests had moderate correlations with DEC first and second semester passing grades. Among the sub-tests, the reading sub-test ability estimates were the best predictor of DEC first and second semester passing grades. This finding was consistent with what Huang (2001) found for sub-tests of IELTS since he observed the highest correlation between Reading and first semester GPA.

The results of the predictive validity analyses show that the best predictor for DEC first and second semester passing grades is the three-parameter model ability estimates. The second best predictor is the ability estimates obtained from sixty high information items. In the third place BUEPE total scores and the total scores obtained from sixty high information items follow with nearly the same correlation coefficients. In terms of the sub-tests, the fourth best predictor is found to be the reading subtest.

As a final remark, all the correlation coefficients computed for the ability estimates seemed to yield higher results when compared to total scores estimates. Similarly, all the correlation coefficients with DEC first semester passing grades were higher when compared to DEC second semester passing grades, a finding which was also backed up by Pack (1972 in Marvin & Simner, 1999) that emphasized TOEFL scores were related to the grade obtained in the first English course taken but not related to grades obtained in subsequent English courses.

5. 2. Limitations of the study

1. The equal discrimination indices assumption of the one-parameter model was not met by the BUEPE data.

2. Since ESP is taught in Departmental English Courses (DEC), every department is given a different ESP course in freshmen. The study has not taken the differences between the DEC into consideration while conducting the predictive validity analyses.

3. Every student did not have both DEC first semester and second semester passing grades because some departments do not offer English two consecutive terms in the freshmen year. Therefore, while some students have two consecutive DEC passing grades for the first semester and second semester; the others only have either the first semester or the second semester DEC passing grades.

4. An important problem that is encountered in predictive validity studies is related to only being able to use a part of the whole test population because

students below the cut scores are not available to be included in the study. This results in lower validity coefficients. In the present study while the whole test population was 699, the sample size in the predictive validity analyses decreased to 371, which may have lowered the validity coefficients obtained.

5.3. Conclusions

1. The unidimensionality, local independence and non-speededness assumptions of Item Response Theory were met in the BUEPE data.

2. The equal discrimination indices assumption of the one-parameter model and the minimal guessing assumptions of the one- and two- parameter models were not met by the BUEPE data.

3. Invariance of ability parameter estimates was established across easy versus hard and first fifty versus second fifty items with quite high correlation coefficients in the one-, two-, and three- parameter models.

4. Invariance of item parameter estimates is established at a high degree across the odd cases versus even cases ability groups.

5. The three-parameter model is the best fitted model to the BUEPE data according to chi-square statistics.

6. The best fitted three-parameter model provided information between the -0.5 and 2.5 ability levels and the maximum information was provided at the 1.00 ability level.

7. All predictive validity coefficients obtained across different ability estimates versus DEC grades were not lower than .50.

8. The best predictor for DEC first and second semester passing grades was the three-parameter model ability estimates. The second best predictor was the ability estimates obtained from sixty high information items. In the third place BUEPE total scores and the total scores obtained from sixty high information items follow with nearly the same correlation coefficients. Among the three sub-tests, the reading sub-test had the highest correlations; thus, ranked as the fourth best predictor of DEC first and second semester passing grades.

9. All the ability scores yielded higher correlations when compared with total score correlations.

10. All correlation coefficients computed with DEC first semester grades were higher than correlation coefficients computed with DEC second semester passing grades.

5.4. Implications of the study

Since there aren't many studies which investigate the predictive validity of IRT estimates, the results of this study may set an example for test constructors at preparatory schools.

The findings of this study indicated that the three-parameter model displays a better fit to the BUEPE data.

Invariance of item and ability parameters was established at a high degree across different samples. The high degree of invariance implies that the items in the exam match with the ability level of the examinees. This is parallel with the purpose of contemporary testing techniques such as adaptive testing, which mainly focuses on confronting the examinees with the items that meet their ability level.

The best fitted three-parameter model provided information between the -0.5 and 2.5 ability levels and the maximum information was provided at the 1.00 ability level. Since the Test Information Functions (TIF) give valuable feedback about the ability levels that the exam gives more information in, the range of the ability levels in which the test gives more information can be widened by including items which give information at other ability levels as well.

Furthermore, Item Characteristic Curves (ICC) can be interpreted to view how each item functions and at which ability level it gives the highest information. Thus, items with similar formats can be constructed when there is insufficient number of items at that specific ability level.

Consequently, by interpreting the ICCs and the TIFs Item Banks can be initiated in English preparatory schools. Item banks full of items varying in terms of their previously determined characteristics such as difficulty, discrimination, ability level that they give utmost information at and subject area can ease the burden of test constructors in terms of time and energy.

The main purpose of BUEPE is to determine if the capacity of students is sufficient to attend and succeed in DEC. Since deciding on which students will be exempted is an important decision the exam must be proven to be functioning as intended as a proficiency test. This can be achieved with predictive validity studies. Advanced statistical analyses such as IRT have enhanced prediction analyses. Instead of using simple total score correlations, this study has made extensive use of IRT estimates which are proven to be more precisely predicting success in DEC.

It is possible to shorten an exam by using the highest information items that IRT models compute. However, deciding on what constitutes the appropriate number of items without jeopardizing reliability of scores (obtained from the shortened form of the test) is an important decision to be taken. The findings of this study have indicated that shortening the BUEPE by using the highest information yielding sixty items is appropriate because the scores obtained from those sixty items yield reliable results.

Since the reading sub-test had the highest predictive validity coefficient, increasing the number of reading items in the exam may further strengthen the predictive validity of BUEPE.

This study can be a starting point for other studies, which may focus on examining TIFs and IIFs for test development and improvement purposes, in addition to using IRT estimates for increasing predictive validity across future English courses.

REFERENCES

Anastasi, A. (1982). Psychological Testing. New York: Macmillan Publishing Co., Inc.

Alderson, J. C., Clapham, C, Wall, D. (1995). Language Test Construction and Evaluation. Cambridge: Cambridge University Press.

Bachman, L. F. (1990). Fundamental Considerations in Language Testing. Oxford: Oxford University Pres.

Baker,B. F. (1992). Item Response Theory : Parameter Estimation Techniques . New York: Marcel Dekker, Inc.

Breland, H., M., Kubota, M., Y., Bonner, M., W. (1999). "The Performance Assessment Study in Writing: Analysis of the SAT II: Writing Test." Retrieved in Jan 5, 2003 from http://www.collegeboard.com/prod_downloads/about/news_ info/cbsenior/ yr2002/pdf/seventeen.pdf

Crocker, L., Algina, J. (1986). Introduction to Classical and Modern Test Theory. Orlando: Holt, Rinehart and Winston, Inc.

Cronbach, L. J. (1990). Essentials of Psychological Testing. (5th ed.). New York: Harper Collins Publishers, Inc.

Çalışkan, M. (2000)." The Fit of one-, two-, three- parameter Models of IRT to the Ministery of National Education-Educational Research and Development Directorate's Science Achievement Test Data." Master Thesis. The Middle East Technical University, The Department of Educational Sciences, Ankara.

Dooey, P. (1999). An investigation into the predictive validity of the IELTS Test as an indicator of future academic success. In K. Martin, N. Stanley and N. Davison (Eds), Teaching in the Disciplines/ Learning in Context, 114-118. Proceedings of the 8th Annual Teaching Learning Forum, The University of Western Australia, February 1999. Perth: UWA. Retreived Dec 5 from http://cea.curtin.edu.au/tlf/tlf1999/dooey.html

Fan, X. (1998). "Item Response Theory and Classical Test Theory: An Empirical Comparison of their Item/ Person Statistics." Educational and Psychological Measurement. 58, pp.357-381.

Green,S.B., Salkino, N. J., Akey, T. M. (1997). Using SPSS for Windows. (2nd ed.). New Jersey: Prentice Hall, Inc.

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

Heard, S. A., Ayers, J. B. (1988). "Validity of the American College Test in Predicting Success on the the Pre-Professional Skills Test." Educational and Psychological Measurement. 48, pp.197-200.

Heaton, J.B. (1988). Writing English Language Tests. Essex: Longman Group UK Limited.

Hughes, A. (1989). Testing for Language Teachers. Cambridge: Cambridge University Press.

Huong, T., T., T. (2001). "The Predictive Validity of the International English Language Testing System (IELTS) Test." Postscript 2, 1, 66-96. Retrieved from www.idp.edu.au/conference/conf2001/pres/28_1400_GlobCapEng LngTst_ IELTSRschViet_pres_Huong.pdf

"Item Analysis" Retrieved Jan 5, 2003, from http://www.scrolla.hw.ac.uk/focus/ia.html

Karataş, A.G. (2001). "The Use of Item Response Theory Models to Scale an English Proficiency Test." Master Thesis. The Middle East Technical University, The Department of Educational Sciences, Ankara.

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." Master Thesis. The Middle East Technical University, The Department of Educational Sciences, Ankara

Lord, F. M., Novick, M. R. (1968). Statistical Theories of Mental Test Scores. Massachusetts: Addison Publishing Company.

Lord, F. L. (1980). Applications of Item Response Theory To Practical Testing Problems. New Jersey: Lawrence Erlbaum Associates, Inc., Publishers.

Marvin, L., Simner, C. (1999). "Postscript to the Canadian Psychological Association's Position Statement on the TOEFL." Retrieved Dec 5, 2002 from www.cpa.ca/documents/TOEFL.html

Mikitovics, A., Crehan, K.D. (2002). "Pre-professional Skills Test Scores as College of Education Admission Criteria." The Journal of Educational Research. 95, Number 4 pp.215-223.

Özkurt, S. (2002). "The fit of one-, Two-, Three-, Parameter Models of Item Response Theory to an English Proficiency Achievement Test Data." Master Thesis. The Middle East Technical University, The Department of Educational Sciences, Ankara.

Prapphal, K. (1990). "The Relevance of Language Testing Research in the Planning of Language Programmes." Retrieved Jan 5, 2003 from http://pioneer.netserv.chula.ac.th/~pkanchan/html/testres.htm

"Predictive validity" Retrieved Oct 25, 2002 from http://ericae.net/seltips.txt

Ramist, L., Lewis, C., & McCauley-Jenkins, C. (2002). "Validity of the SAT II Science Tests." Science Insights, 6, 5. Retrieved Dec 5, 2002 from http://www.nas.org

Stofflet, F., Fenton, R., Strough, T. (2001, April). "Construct and Predictive Validity of the Alaska State High School Graduation Qualifying Examination: First Administration." A paper presented at the 2001 American Educational Research Association Convention. Retrieved Dec 5, 2002, from http://www.asd.k12.ak.us/Depts/assess_eval/hsgqe/

Tabachnick, B., Fidell, L.s. (1996). Using Multivariate Statistics. (3rd ed.). New York: Harper collins Publishers Inc.

Tang, K., L., Eignor, D., R. (2001). "A Study of the Use of the Collateral Statistical Information in Attempting to Reduce TOEFL IRT Item Parameter Estimation Sample Sizes." Retrieved March 14, 2003 from ftp://ftp.ets.org/pub/toefl/989084.pdf

Tang, K., L., Way, D. (1995). "Investigation of IRT-Based Assembly of the TOEFL Test." Retrieved March 15, 2003 from http://www.toefl.org/research/rtecrpts.html

Way, D., Rease, R. (1991). "An Investigation of the Use of Simplified IRT Models for Scaling and Equating the TOEFL Test." Retrieved March 15, 2003 from http://www.toefl.org/research/rtecrpts.html

APPENDICES

APPENDIX A

EXAMPLES OF ITEM TYPES

A. GRAMMAR SECTION

Items 1-15 Modified Cloze Test

One evening, several years ago I was walking through a forest in Switzerland. A few meters away there was an old man and a woman who _________ (1) to him in a loud voice. It was not until I got closer that I realised _________ (2).

1)	a) had talked	b) was talking	c)she talked	d)she was talking
2)	a)who the man was	b)that the man was	c)that was the man	d)who was the man

Items 16-36 Discrete Point Grammar Items

16. I'm not sure about the exact time of the concert; you __________ look it up in the newspaper.

a) had beter b) would like to c) would d) had to

Items 37-40 Spot the Mistake

37. He is always leaving his books in the living room and his bedroom is in
a
a mess all the time. I wish he isn't so untidy ! I think I will have to talk to
b c d
him once more.

B. READING SECTION

Items 41-45 Sentence Completion

41. Some people eat large quantities of food, __________________ .

a) because they need a well balanced diet

b) as they haven't had a bite for hours

c) although they don't need to diet

d) yet never gain any weight

Items 46-50 Paragraph Completion

46. Have you ever had to decide whether to go shopping or stay at home and watch TV on a weekend?________________. Home shopping TV networks have become a way for many people to shop without ever having to leave their homes.

a) You can communicate with the shops by means of yor computer

b) They can buy anything by phoning and giving their credit card number

c) Some shoppers are tired of department stores and shopping malls

d) Now you can do both at the same time and save more time

Items 51-80 Sentence Completion, Guessing Vocabulary from Context and Reference type items related to three different reading texts

51. At the beginning of the nineteenth century _________________ .

a) alcohol killed a lot of patients

b) operations were done very slowly

c) surgeons had to work quite fast

d) patients died of blood poisoning

56. In line 27 “depise” means to consider someone as ___________ .

a) helpful b) courageous c) worthless d) selfish

59. In line 39 “it” refers to ____________ .

a) chloroform b) heart disease c) ether d) heart beating

C. VOCABULARY

Items 81-100 Sentential Level Fill-in-the –Blanks

83. The agent was _________ for carrying a false passport at the airport.

a) charged b) arrested c) banned d) prohibited

100. When her report was not approved, she didn’t say a word but I know she felt really sorry. The look on her face gave her feelings __________ .

a) out b) through c) away d)in

APPENDIX B

RELIABILITY ANALYSIS

SCALE (ALPHA)

	Mean	Std Dev	Cases
ITEM1	,7983	,4016	699,0
ITEM2	,5522	,4976	699,0
ITEM3	,5980	,4907	699,0
ITEM4	,7868	,4098	699,0
ITEM5	,5293	,4995	699,0
ITEM6	,7353	,4415	699,0
ITEM7	,7897	,4078	699,0
ITEM8	,3691	,4829	699,0
ITEM9	,0916	,2886	699,0
ITEM10	,4092	,4920	699,0
ITEM11	,4378	,4965	699,0
ITEM12	,6910	,4624	699,0
ITEM13	,7854	,4108	699,0
ITEM14	,5279	,4996	699,0
ITEM15	,6996	,4588	699,0
ITEM16	,6724	,4697	699,0
ITEM17	,5522	,4976	699,0
ITEM18	,7926	,4058	699,0
ITEM19	,4063	,4915	699,0
ITEM20	,2160	,4118	699,0
ITEM21	,4592	,4987	699,0
ITEM22	,3820	,4862	699,0
ITEM23	,3691	,4829	699,0
ITEM24	,6009	,4901	699,0
ITEM25	,0787	,2694	699,0
ITEM26	,5351	,4991	699,0
ITEM27	,6753	,4686	699,0
ITEM28	,9356	,2456	699,0
ITEM29	,5479	,4981	699,0
ITEM30	,7554	,4302	699,0
ITEM31	,7082	,4549	699,0
ITEM32	,8970	,3042	699,0
ITEM33	,4449	,4973	699,0
ITEM34	,8169	,3870	699,0
ITEM35	,4964	,5003	699,0
ITEM36	,4192	,4938	699,0
ITEM37	,5179	,5000	699,0
ITEM38	,3505	,4775	699,0

(Table continued)

ITEM39	,4192	,4938	699,0
ITEM40	,4034	,4909	699,0
ITEM41	,4506	,4979	699,0
ITEM42	,5293	,4995	699,0
ITEM43	,7282	,4452	699,0
ITEM44	,6166	,4866	699,0
ITEM45	,6366	,4813	699,0
ITEM46	,7225	,4481	699,0
ITEM47	,7210	,4488	699,0
ITEM48	,8941	,3079	699,0
ITEM49	,7725	,4195	699,0
ITEM50	,7926	,4058	699,0
ITEM51	,6366	,4813	699,0
ITEM52	,8011	,3994	699,0
ITEM53	,5622	,4965	699,0
ITEM54	,6681	,4712	699,0
ITEM55	,7039	,4569	699,0
ITEM56	,5808	,4938	699,0
ITEM57	,5823	,4935	699,0
ITEM58	,8426	,3644	699,0
ITEM59	,9471	,2241	699,0
ITEM60	,8770	,3287	699,0
ITEM61	,3462	,4761	699,0
ITEM62	,4320	,4957	699,0
ITEM63	,6738	,4691	699,0
ITEM64	,5851	,4931	699,0
ITEM65	,4249	,4947	699,0
ITEM66	,4292	,4953	699,0
ITEM67	,5594	,4968	699,0
ITEM68	,7296	,4445	699,0
ITEM69	,8340	,3723	699,0
ITEM70	,8856	,3186	699,0
ITEM71	,1559	,3631	699,0
ITEM72	,2732	,4459	699,0
ITEM73	,3004	,4588	699,0
ITEM74	,6009	,4901	699,0
ITEM75	,4850	,5001	699,0
ITEM76	,7053	,4562	699,0
ITEM77	,4349	,4961	699,0
ITEM78	,6609	,4737	699,0
ITEM79	,7611	,4267	699,0
ITEM80	,7296	,4445	699,0
ITEM81	,2732	,4459	699,0
ITEM82	,6581	,4747	699,0
ITEM83	,6295	,4833	699,0
ITEM84	,2775	,4481	699,0
ITEM85	,4421	,4970	699,0

(Table continued)

ITEM86	,5222	,4999	699,0
ITEM87	,5093	,5003	699,0
ITEM88	,4506	,4979	699,0
ITEM89	,6609	,4737	699,0
ITEM90	,4921	,5003	699,0
ITEM91	,3605	,4805	699,0
ITEM92	,5279	,4996	699,0
ITEM93	,5436	,4984	699,0
ITEM94	,7353	,4415	699,0
ITEM95	,5494	,4979	699,0
ITEM96	,4921	,5003	699,0
ITEM97	,5265	,4997	699,0
ITEM98	,6595	,4742	699,0
ITEM99	,5265	,4997	699,0
ITEM100	,2632	,4407	699,0

RELIABILITY ANALYSIS - SCALE (ALPHA)

N of Cases = 699,0

Statistics for	Mean	Variance	Std Dev	N of Variables
Scale	57,4263	274,1676	16,5580	100

Item Means	Mean	Minimum	Maximum	Range	Max/Min	Variance
	,5743	,0787	,9471	,8684	12,0364	,0353

Inter-item Correlations	Mean	Minimum	Maximum	Range	Max/Min	Variance
	,1196	-,1062	,3814	,4876	-3,5917	,0053

RELIABILITY ANALYSIS - SCALE (ALPHA)

Item-total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Alpha if Item Deleted
ITEM1	56,6280	271,4116	,1961	.	,9327
ITEM2	56,8741	266,5457	,4538	.	,9316
ITEM3	56,8283	268,0679	,3647	.	,9320
ITEM4	56,6395	269,6263	,3249	.	,9322
ITEM5	56,8970	265,4306	,5215	.	,9313
ITEM6	56,6910	271,5892	,1638	.	,9329

(Table continued)

ITEM7	56,6366	269,7589	,3167	.	,9322
ITEM8	57,0572	271,0712	,1801	.	,9329
ITEM9	57,3348	273,3262	,0794	.	,9330
ITEM10	57,0172	268,9252	,3099	.	,9323
ITEM11	56,9886	266,8451	,4362	.	,9317
ITEM12	56,7353	269,9112	,2661	.	,9325
ITEM13	56,6409	270,1961	,2815	.	,9324
ITEM14	56,8984	265,3206	,5283	.	,9313
ITEM15	56,7268	270,9467	,1993	.	,9328
ITEM16	56,7539	266,7531	,4689	.	,9316
ITEM17	56,8741	266,5601	,4529	.	,9316
ITEM18	56,6338	269,3184	,3517	.	,9321
ITEM19	57,0200	268,6730	,3260	.	,9322
ITEM20	57,2103	268,1634	,4326	.	,9318
ITEM21	56,9671	267,6221	,3859	.	,9319
ITEM22	57,0443	268,8189	,3206	.	,9322
ITEM23	57,0572	269,6386	,2709	.	,9325
ITEM24	56,8255	266,1643	,4855	.	,9315
ITEM25	57,3476	273,1784	,1029	.	,9329
ITEM26	56,8913	268,3750	,3390	.	,9322
ITEM27	56,7511	272,2503	,1098	.	,9332
ITEM28	56,4907	272,5483	,1922	.	,9326
ITEM29	56,8784	269,4680	,2722	.	,9325
ITEM30	56,6710	268,8285	,3654	.	,9320
ITEM31	56,7182	269,7127	,2843	.	,9324
ITEM32	56,5293	273,7395	,0333	.	,9331
ITEM33	56,9814	269,8378	,2499	.	,9326
ITEM34	56,6094	268,6395	,4239	.	,9319
ITEM35	56,9299	264,8446	,5571	.	,9311
ITEM36	57,0072	268,5630	,3312	.	,9322
ITEM37	56,9084	267,9572	,3641	.	,9320
ITEM38	57,0758	272,4771	,0928	.	,9333
ITEM39	57,0072	269,4025	,2789	.	,9324
ITEM40	57,0229	269,2860	,2880	.	,9324
ITEM41	56,9757	267,3304	,4047	.	,9318
ITEM42	56,8970	267,0410	,4213	.	,9318
ITEM43	56,6981	269,9474	,2749	.	,9324
ITEM44	56,8097	266,4408	,4715	.	,9315
ITEM45	56,7897	266,7422	,4576	.	,9316
ITEM46	56,7039	267,6701	,4294	.	,9318
ITEM47	56,7053	269,6809	,2907	.	,9324
ITEM48	56,5322	270,5616	,3467	.	,9322
ITEM49	56,6538	268,4215	,4052	.	,9319
ITEM50	56,6338	271,9144	,1561	.	,9329
ITEM51	56,7897	267,1978	,4282	.	,9317
ITEM52	56,6252	268,8393	,3946	.	,9320
ITEM53	56,8641	266,1348	,4807	.	,9315

(Table continued)

ITEM54	56,7582	266,7309	,4687	.	,9316
ITEM55	56,7225	268,1721	,3867	.	,9319
ITEM56	56,8455	267,0621	,4252	.	,9318
ITEM57	56,8441	265,7479	,5081	.	,9314
ITEM58	56,5837	270,3551	,3071	.	,9323
ITEM59	56,4793	272,8975	,1648	.	,9327
ITEM60	56,5494	271,0187	,2810	.	,9324
ITEM61	57,0801	269,3919	,2911	.	,9324
ITEM62	56,9943	269,8968	,2471	.	,9326
ITEM63	56,7525	267,9000	,3938	.	,9319
ITEM64	56,8412	267,5865	,3929	.	,9319
ITEM65	57,0014	268,3625	,3431	.	,9321
ITEM66	56,9971	267,5587	,3927	.	,9319
ITEM67	56,8670	268,0668	,3598	.	,9321
ITEM68	56,6967	268,5067	,3751	.	,9320
ITEM69	56,5923	269,2218	,3935	.	,9320
ITEM70	56,5408	271,6928	,2260	.	,9326
ITEM71	57,2704	269,5901	,3729	.	,9321
ITEM72	57,1531	268,5166	,3730	.	,9320
ITEM73	57,1259	268,5515	,3595	.	,9321
ITEM74	56,8255	267,1414	,4236	.	,9318
ITEM75	56,9413	266,4421	,4578	.	,9316
ITEM76	56,7210	266,4479	,5043	.	,9314
ITEM77	56,9914	265,6131	,5138	.	,9313
ITEM78	56,7654	267,7558	,3991	.	,9319
ITEM79	56,6652	268,8362	,3680	.	,9320
ITEM80	56,6967	269,6471	,2961	.	,9323
ITEM81	57,1531	269,7287	,2895	.	,9324
ITEM82	56,7682	271,0637	,1842	.	,9328
ITEM83	56,7969	270,4572	,2187	.	,9327
ITEM84	57,1488	268,8374	,3491	.	,9321
ITEM85	56,9843	269,1989	,2895	.	,9324
ITEM86	56,9041	267,3676	,4007	.	,9319
ITEM87	56,9170	271,0275	,1754	.	,9329
ITEM88	56,9757	266,4736	,4580	.	,9316
ITEM89	56,7654	269,9850	,2543	.	,9325
ITEM90	56,9342	271,2220	,1636	.	,9330
ITEM91	57,0658	270,7464	,2018	.	,9328
ITEM92	56,8984	271,3063	,1587	.	,9330
ITEM93	56,8827	267,1696	,4142	.	,9318
ITEM94	56,6910	267,1451	,4731	.	,9316
ITEM95	56,8770	268,2284	,3490	.	,9321
ITEM96	56,9342	268,2220	,3475	.	,9321
ITEM97	56,8999	266,9527	,4266	.	,9317
ITEM98	56,7668	267,2392	,4324	.	,9317
ITEM99	56,8999	267,1991	,4113	.	,9318
ITEM100	57,1631	273,4089	,0387	.	,9334

APPENDIX C
FACTOR ANALYSIS

Total Variance Explained

	Initial Eigen-values			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	14,268	14,268	14,268	14,268	14,268	14,268	9,803	9,803	9,803
2	2,906	2,906	17,174	2,906	2,906	17,174	7,370	7,370	17,174
3	1,792	1,792	18,965						
4	1,655	1,655	20,620						
5	1,605	1,605	22,225						
6	1,520	1,520	23,745						
7	1,464	1,464	25,209						
8	1,422	1,422	26,631						
9	1,415	1,415	28,046						
10	1,392	1,392	29,438						
11	1,371	1,371	30,809						
12	1,344	1,344	32,153						
13	1,342	1,342	33,494						
14	1,320	1,320	34,814						
15	1,302	1,302	36,116						
16	1,274	1,274	37,390						
17	1,260	1,260	38,649						
18	1,237	1,237	39,886						
19	1,222	1,222	41,108						
20	1,190	1,190	42,299						
21	1,184	1,184	43,482						
22	1,172	1,172	44,654						
23	1,158	1,158	45,812						
24	1,135	1,135	46,947						
25	1,121	1,121	48,068						
26	1,107	1,107	49,175						
27	1,100	1,100	50,275						
28	1,079	1,079	51,354						
29	1,063	1,063	52,418						
30	1,047	1,047	53,465						
31	1,045	1,045	54,510						
32	1,033	1,033	55,543						
33	1,022	1,022	56,565						
34	,996	,996	57,561						
35	,991	,991	58,552						
36	,977	,977	59,529						
37	,958	,958	60,487						
38	,945	,945	61,432						
39	,941	,941	62,374						
40	,933	,933	63,306						
41	,919	,919	64,225						
42	,908	,908	65,133						
43	,895	,895	66,028						
44	,872	,872	66,901						
45	,864	,864	67,764						
46	,845	,845	68,610						

(Table continued)

47	,841	,841	69,451
48	,838	,838	70,288
49	,829	,829	71,118
50	,814	,814	71,931
51	,799	,799	72,730
52	,788	,788	73,518
53	,774	,774	74,292
54	,767	,767	75,059
55	,752	,752	75,811
56	,742	,742	76,553
57	,735	,735	77,288
58	,725	,725	78,013
59	,718	,718	78,731
60	,705	,705	79,436
61	,690	,690	80,125
62	,685	,685	80,810
63	,681	,681	81,492
64	,662	,662	82,154
65	,657	,657	82,811
66	,650	,650	83,461
67	,641	,641	84,102
68	,627	,627	84,729
69	,613	,613	85,343
70	,604	,604	85,947
71	,602	,602	86,549
72	,591	,591	87,139
73	,581	,581	87,720
74	,577	,577	88,298
75	,565	,565	88,862
76	,558	,558	89,420
77	,544	,544	89,965
78	,540	,540	90,505
79	,531	,531	91,036
80	,522	,522	91,558
81	,506	,506	92,064
82	,503	,503	92,567
83	,489	,489	93,057
84	,475	,475	93,531
85	,470	,470	94,001
86	,466	,466	94,467
87	,457	,457	94,924
88	,451	,451	95,375
89	,443	,443	95,818
90	,428	,428	96,245
91	,425	,425	96,670
92	,422	,422	97,092
93	,399	,399	97,490
94	,398	,398	97,888
95	,388	,388	98,276
96	,381	,381	98,658
97	,363	,363	99,021
98	,345	,345	99,365
99	,323	,323	99,688
100	,312	,312	100,000

Extraction Method: Principal Component Analysis.

Rotated Component Matrix

	Component	
	1	2
ITEM76	,626	9,130E-02
ITEM54	,544	,129
ITEM52	,524	3,376E-02
ITEM49	,519	5,553E-02
ITEM69	,517	3,532E-02
ITEM78	,513	6,138E-02
ITEM46	,503	,107
ITEM94	,490	,202
ITEM48	,488	-5,103E-03
ITEM7	,487	-4,012E-02
ITEM44	,476	,220
ITEM4	,469	-1,085E-02
ITEM77	,466	,299
ITEM34	,458	,165
ITEM57	,454	,298
ITEM24	,453	,262
ITEM53	,451	,270
ITEM97	,443	,195
ITEM12	,422	-5,562E-02
ITEM2	,418	,267
ITEM67	,409	,112
ITEM45	,409	,278
ITEM51	,403	,239
ITEM63	,392	,193
ITEM74	,389	,230
ITEM98	,388	,261
ITEM56	,387	,242
ITEM68	,380	,174
ITEM79	,369	,172
ITEM64	,364	,225
ITEM75	,364	,329
ITEM19	,359	,130
ITEM55	,357	,223
ITEM93	,353	,270
ITEM58	,349	,108
ITEM3	,346	,194
ITEM13	,326	7,992E-02
ITEM43	,326	7,310E-02
ITEM18	,299	,230
ITEM60	,297	,116
ITEM65	,295	,221
ITEM80	,292	,145
ITEM47	,289	,144
ITEM72	,286	,277
ITEM96	,272	,256
ITEM28	,271	-6,749E-03
ITEM10	,259	,201
ITEM6	,255	-3,471E-02
ITEM61	,249	,189
ITEM89	,248	,122
ITEM23	,224	,184
ITEM1	,193	,103
ITEM83	,191	,132

ITEM50	,170	5,508E-02
ITEM70	,169	,169
ITEM59	,141	9,847E-02
ITEM32	2,935E-02	1,529E-02
ITEM5	,245	,562
ITEM11	,137	,554
ITEM37	7,262E-02	,513
ITEM35	,340	,508
ITEM36	4,954E-02	,494
ITEM88	,210	,492
ITEM71	,114	,470
ITEM33	-3,185E-02	,449
ITEM99	,205	,435
ITEM14	,373	,424
ITEM20	,250	,415
ITEM41	,215	,413
ITEM16	,309	,410
ITEM84	,140	,405
ITEM66	,219	,390
ITEM81	8,665E-02	,378
ITEM22	,131	,366
ITEM30	,203	,365
ITEM21	,233	,364
ITEM73	,194	,364
ITEM17	,327	,363
ITEM31	9,082E-02	,359
ITEM42	,284	,357
ITEM26	,174	,354
ITEM86	,261	,344
ITEM95	,228	,312
ITEM91	4,654E-02	,282
ITEM39	,148	,282
ITEM85	,174	,282
ITEM40	,181	,266
ITEM15	5,723E-02	,252
ITEM9	-,112	,250
ITEM29	,171	,247
ITEM90	1,876E-02	,246
ITEM62	,158	,218
ITEM25	-3,818E-02	,206
ITEM8	8,190E-02	,197
ITEM38	-4,023E-02	,193
ITEM92	7,453E-02	,180
ITEM82	,116	,169
ITEM100	-8,797E-02	,154
ITEM87	,121	,148
ITEM27	5,208E-02	,114

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

A Rotation converged in 3 iterations.

APPENDIX D
SCATTERPLOTS OF INVARIANCE OF ITEM PARAMETERS

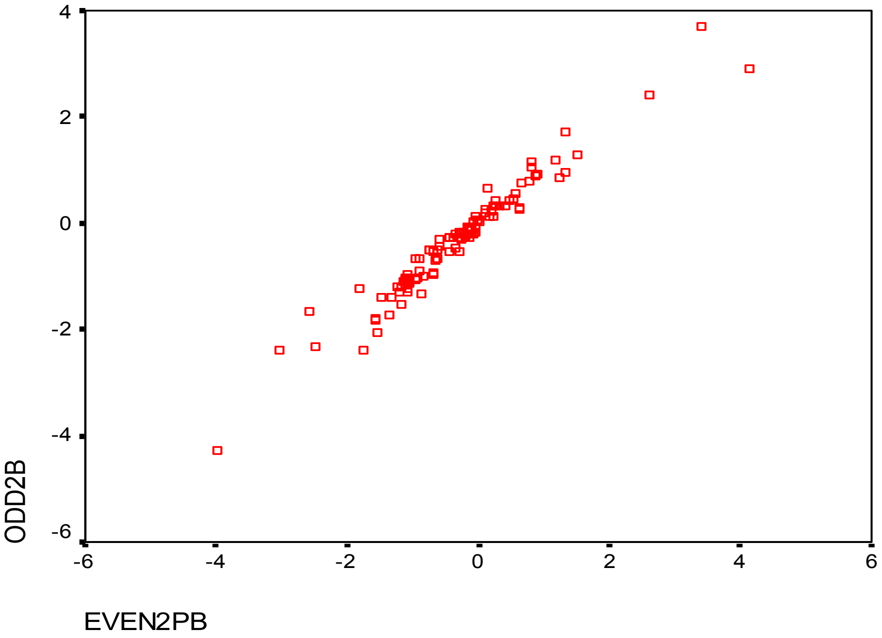


Figure D1 Scatterplot of 2P Difficulty Estimates (Odd vs. Even)

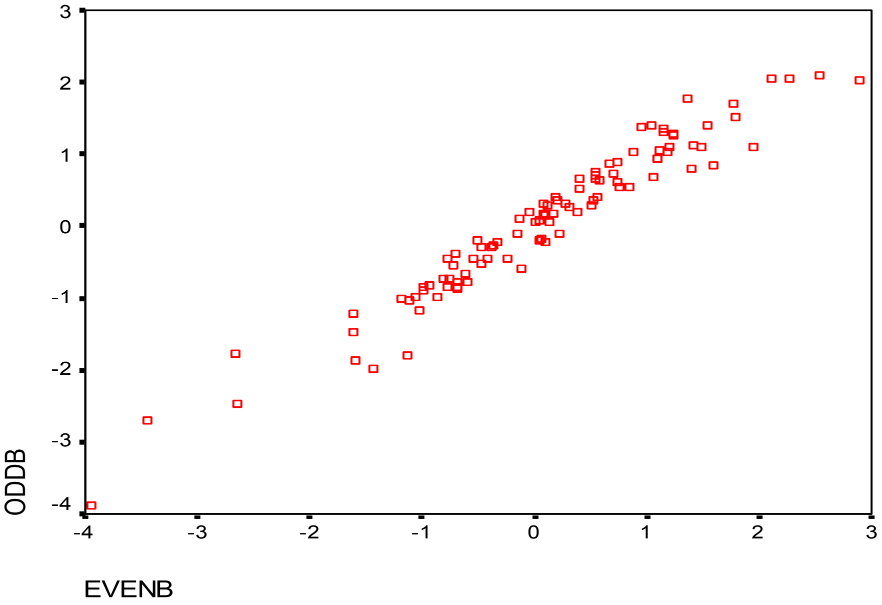


Figure D2 Scatterplot of 3P Difficulty Estimates (Odd vs. Even)

APPENDIX E
ITEMS WITH GOOD AND POOR ITEM INFORMATION FUNCTIONS IN THE 3P MODEL

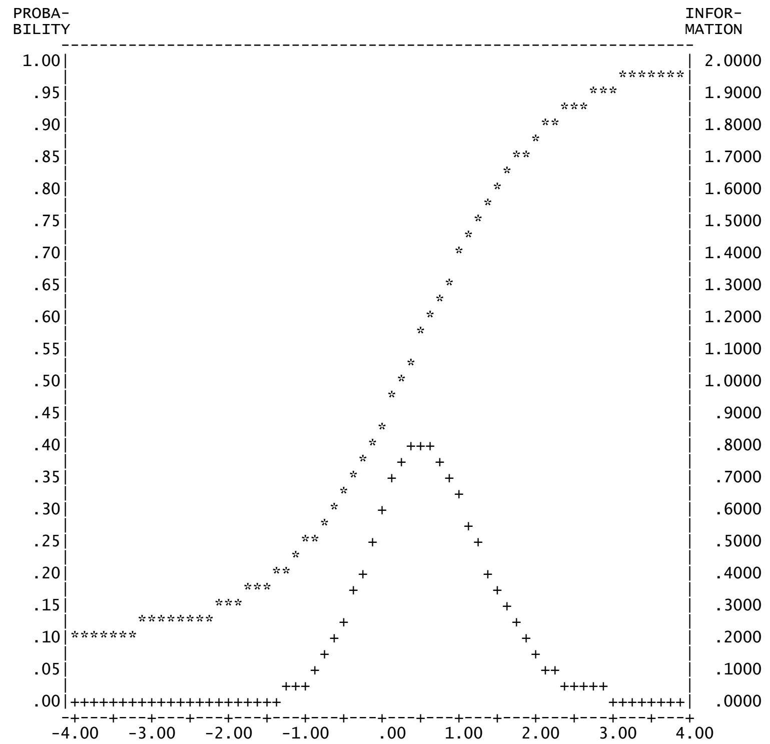
Items with Good IIFs in the 3 P Model

```
                                      ITEM:   0071
 PROBA-                                                                INFOR-
 BILITY                                                                MATION
      ------------------------------------------------------------------
  1.00|                                                                | 2.0000
      |                                                                |
   .95|                                                             ***| 1.9000
      |                                                           **   |
   .90|                                                         **     | 1.8000
      |                                                        *       |
   .85|                                                      **        | 1.7000
      |                                                     *          |
   .80|                                                    *           | 1.6000
      |                                                                |
   .75|                                                   *            | 1.5000
      |                                                  *             |
   .70|                                                 *              | 1.4000
      |                                                *               |
   .65|                                                                | 1.3000
      |                                               *                |
   .60|                                              *                 | 1.2000
      |                                             +                  |
   .55|                                            +*+                 | 1.1000
      |                                           +                    |
   .50|                                            *  +                | 1.0000
      |                                          +*                    |
   .45|                                                +               |  .9000
      |                                         +*                     |
   .40|                                         *       +              |  .8000
      |                                        *                       |
   .35|                                        +         +             |  .7000
      |                                       *                        |
   .30|                                      *                         |  .6000
      |                                     * +           +            |
   .25|                                    *                           |  .5000
      |                                   *                +           |
   .20|                                  *   +                         |  .4000
      |                                **                   +          |
   .15|                               *     +                +         |  .3000
      |                            ***     +                           |
   .10|                        ****                           +        |  .2000
      |             ***********           +                    ++      |
   .05|*************                     +                       ++    |  .1000
      |                               +++                          ++++|
   .00|+++++++++++++++++++++++++++++++                                 |  .0000
      -+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
     -4.00   -3.00   -2.00   -1.00     .00    1.00    2.00    3.00    4.00

POINT OF MAXIMUM INFORMATION & STANDARD ERROR:      1.6004    (       .0390)
VALUE OF MAXIMUM INFORMATION & STANDARD ERROR:      1.1311    (       .0462)
```

```
                                      ITEM:   0009
 PROBA-                                                                 INFOR-
 BILITY                                                                 MATION
      ------------------------------------------------------------------
  1.00|                                                             ***| 2.0000
      |                                                          ***   |
   .95|                                                        **      | 1.9000
      |                                                    +  *        |
   .90|                                                                | 1.8000
      |                                                      *         |
   .85|                                                     *          | 1.7000
      |                                                                |
   .80|                                                    *           | 1.6000
      |                                                                |
   .75|                                             +     *            | 1.5000
      |                                                                |
   .70|                                                                | 1.4000
      |                                                  *             |
   .65|                                                                | 1.3000
      |                                                                |
   .60|                                                 *              | 1.2000
      |                                                                |
   .55|                                                     +          | 1.1000
      |                                                *               |
   .50|                                                                | 1.0000
      |                                                                |
   .45|                                                                |  .9000
      |                                               *                |
   .40|                                                                |  .8000
      |                                            + *                 |
   .35|                                                                |  .7000
      |                                                      +         |
   .30|                                             *                  |  .6000
      |                                                                |
   .25|                                            *                   |  .5000
      |                                                                |
   .20|                                           *           +        |  .4000
      |                                          *+                    |
   .15|                                         *                      |  .3000
      |                                        *                       |
   .10|                                     ***                +       |  .2000
      |*************************************     +                     |
   .05|                                                         +      |  .1000
      |                                         +                ++    |
   .00|+++++++++++++++++++++++++++++++++++++++++                   ++++|  .0000
      -+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
     -4.00   -3.00   -2.00   -1.00    .00     1.00    2.00    3.00    4.00

POINT OF MAXIMUM INFORMATION & STANDARD ERROR:      2.0599    (       .0497)
VALUE OF MAXIMUM INFORMATION & STANDARD ERROR:      4.7365    (       .1819)
```

```
                                      ITEM:   0025

 PROBA-                                                                 INFOR-
 BILITY                                                                 MATION
      ------------------------------------------------------------------
  1.00|                                                +               | 2.0000
      |                                                               *|
   .95|                                                             ** | 1.9000
      |                                                            *   |
   .90|                                                          **    | 1.8000
      |                                                         *      |
   .85|                                                     +          | 1.7000
      |                                                        *       |
   .80|                                                       *        | 1.6000
      |                                               +                |
   .75|                                                      *         | 1.5000
      |                                                                |
   .70|                                                     *          | 1.4000
      |                                                                |
   .65|                                                    * +         | 1.3000
      |                                                                |
   .60|                                                   *            | 1.2000
      |                                                                |
   .55|                                              +   *             | 1.1000
      |                                                                |
   .50|                                                                | 1.0000
      |                                                 *     +        |
   .45|                                                                |  .9000
      |                                                *               |
   .40|                                                                |  .8000
      |                                             + *                |
   .35|                                                        +       |  .7000
      |                                              *                 |
   .30|                                                                |  .6000
      |                                             *                  |
   .25|                                            *                   |  .5000
      |                                            +            +      |
   .20|                                           *                    |  .4000
      |                                          *                     |
   .15|                                        **                +     |  .3000
      |                                       *   +                    |
   .10|                                    ***                    +    |  .2000
      |************************************      +                 +   |
   .05|                                                             +  |  .1000
      |                                        ++                    ++|
   .00|++++++++++++++++++++++++++++++++++++++++                        |  .0000
      -+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
     -4.00   -3.00   -2.00   -1.00    .00     1.00    2.00    3.00    4.00

POINT OF MAXIMUM INFORMATION & STANDARD ERROR:      2.2614    (       .0620)
VALUE OF MAXIMUM INFORMATION & STANDARD ERROR:      2.4846    (       .1112)
```

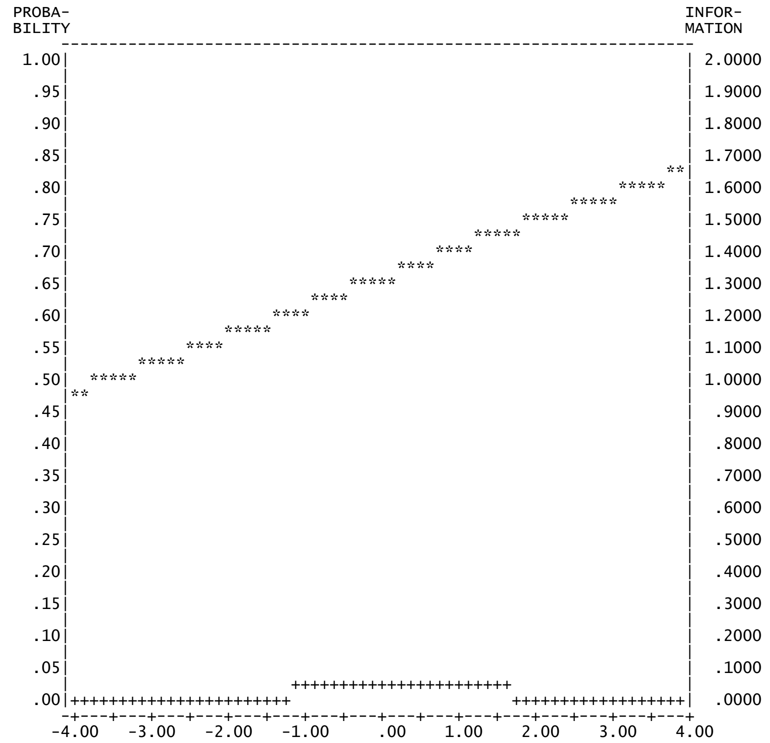
```
ITEM:   0077
```



```
POINT OF MAXIMUM INFORMATION & STANDARD ERROR:       .5189    (       .0239)
VALUE OF MAXIMUM INFORMATION & STANDARD ERROR:       .8167    (       .0487
```

Items with Poor IIFs in the 3 P Model

```
ITEM:   0027
```



```
POINT OF MAXIMUM INFORMATION & STANDARD ERROR:       .1682    (       .0783)
VALUE OF MAXIMUM INFORMATION & STANDARD ERROR:       .0280    (       .0041)
```

```
                                      ITEM:   0032

 PROBA-                                                                 INFOR-
 BILITY                                                                 MATION
      ------------------------------------------------------------------
  1.00|                                                                | 2.0000
      |                                                                |
   .95|                                                                | 1.9000
      |                                                                |
   .90|                                                       *********| 1.8000
      |                                               ********         |
   .85|                                        *******                 | 1.7000
      |                                 *******                        |
   .80|                            *****                               | 1.6000
      |                      ******                                    |
   .75|                 *****                                          | 1.5000
      |             ****                                               |
   .70|        *****                                                   | 1.4000
      |    ****                                                        |
   .65|****                                                            | 1.3000
      |                                                                |
   .60|                                                                | 1.2000
      |                                                                |
   .55|                                                                | 1.1000
      |                                                                |
   .50|                                                                | 1.0000
      |                                                                |
   .45|                                                                |  .9000
      |                                                                |
   .40|                                                                |  .8000
      |                                                                |
   .35|                                                                |  .7000
      |                                                                |
   .30|                                                                |  .6000
      |                                                                |
   .25|                                                                |  .5000
      |                                                                |
   .20|                                                                |  .4000
      |                                                                |
   .15|                                                                |  .3000
      |                                                                |
   .10|                                                                |  .2000
      |                                                                |
   .05|                                                                |  .1000
      |++++++++                                                        |
   .00|        ++++++++++++++++++++++++++++++++++++++++++++++++++++++++|  .0000
      -+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
     -4.00   -3.00   -2.00   -1.00     .00    1.00    2.00    3.00    4.00

POINT OF MAXIMUM INFORMATION & STANDARD ERROR:     -4.2025    (       .4355)
VALUE OF MAXIMUM INFORMATION & STANDARD ERROR:       .0266    (       .0052)
```

```
                                      ITEM:   0050

 PROBA-                                                                 INFOR-
 BILITY                                                                 MATION
      ------------------------------------------------------------------
  1.00|                                                                | 2.0000
      |                                                                |
   .95|                                                                | 1.9000
      |                                                                |
   .90|                                                            ****| 1.8000
      |                                                      ******    |
   .85|                                                 *****          | 1.7000
      |                                             ****               |
   .80|                                         ****                   | 1.6000
      |                                     ****                       |
   .75|                                  ***                           | 1.5000
      |                               ***                              |
   .70|                           ****                                 | 1.4000
      |                        ***                                     |
   .65|                     ***                                        | 1.3000
      |                  ***                                           |
   .60|                **                                              | 1.2000
      |             ***                                                |
   .55|          ***                                                   | 1.1000
      |      ****                                                      |
   .50|   ***                                                          | 1.0000
      |***                                                             |
   .45|                                                                |  .9000
      |                                                                |
   .40|                                                                |  .8000
      |                                                                |
   .35|                                                                |  .7000
      |                                                                |
   .30|                                                                |  .6000
      |                                                                |
   .25|                                                                |  .5000
      |                                                                |
   .20|                                                                |  .4000
      |                                                                |
   .15|                                                                |  .3000
      |                                                                |
   .10|                                                                |  .2000
      |                                                                |
   .05|                                                                |  .1000
      |++++++++++++++++++++++++++++++++++++++++++++++                  |
   .00|                                              ++++++++++++++++++|  .0000
      -+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
     -4.00   -3.00   -2.00   -1.00     .00    1.00    2.00    3.00    4.00

POINT OF MAXIMUM INFORMATION & STANDARD ERROR:     -1.4623    (       .1113)
VALUE OF MAXIMUM INFORMATION & STANDARD ERROR:       .0565    (       .0071
```

APPENDIX F
PREDICTIVE VALIDITY SCATTERPLOTS

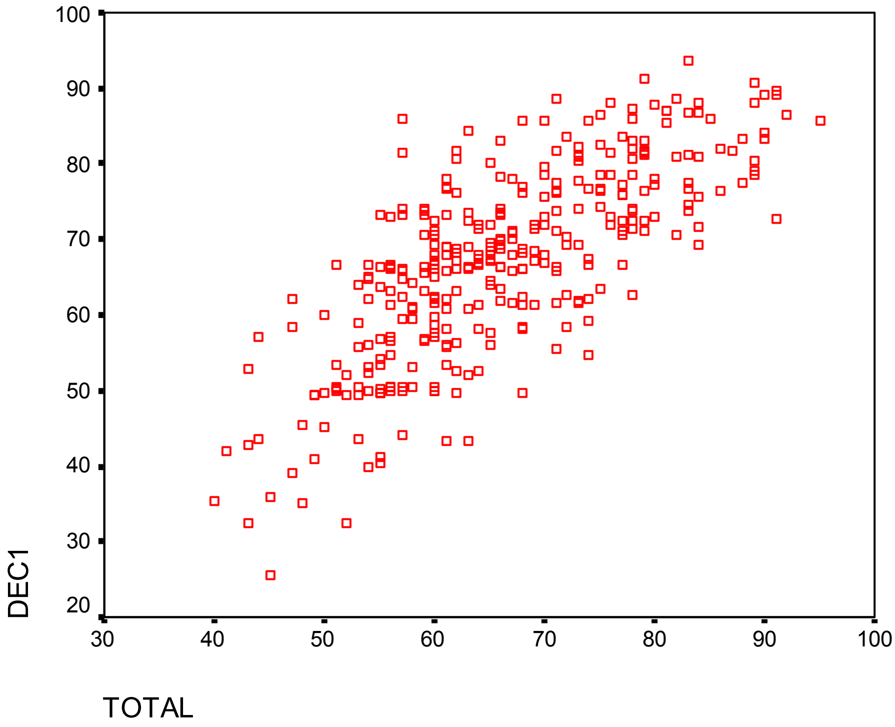


Figure F1 BUEPE Total Scores and First Semester DEC Passing Grades

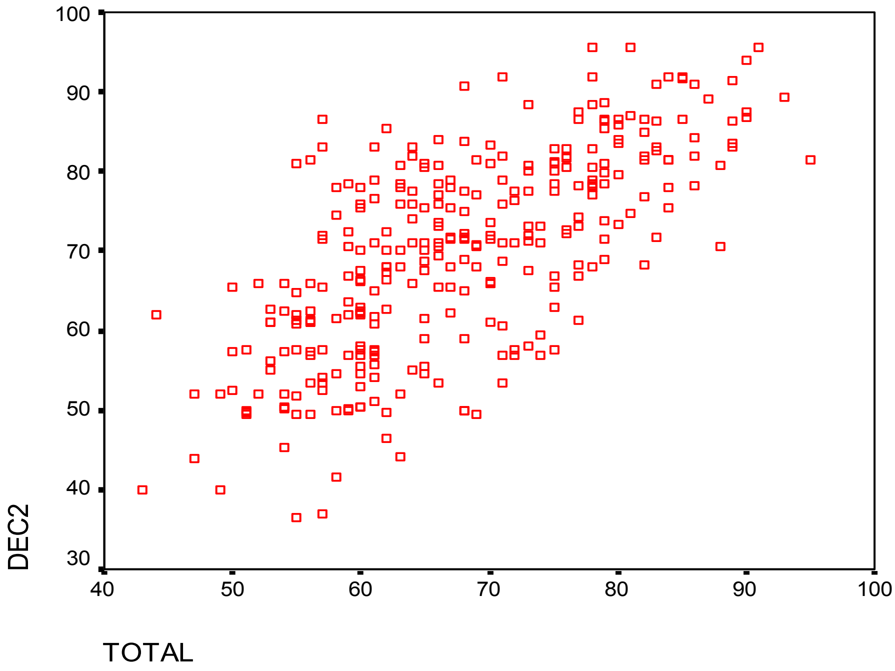


Figure F2 BUEPE Total Scores and Second Semester DEC Passing Grades

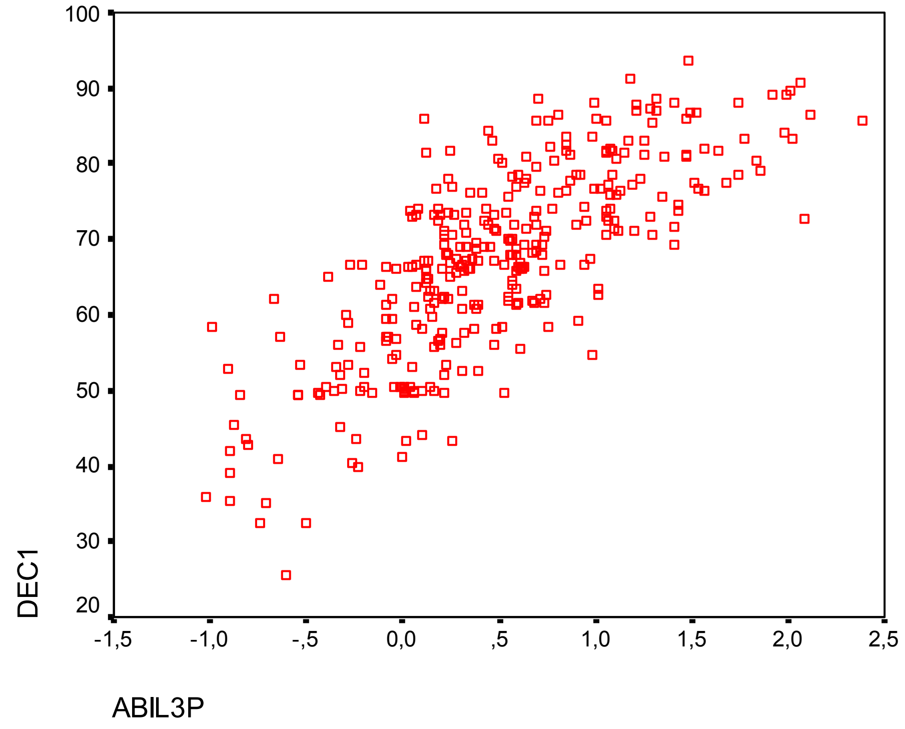


Figure F3 3 P Ability Scores and First Semester DEC Passing Grades

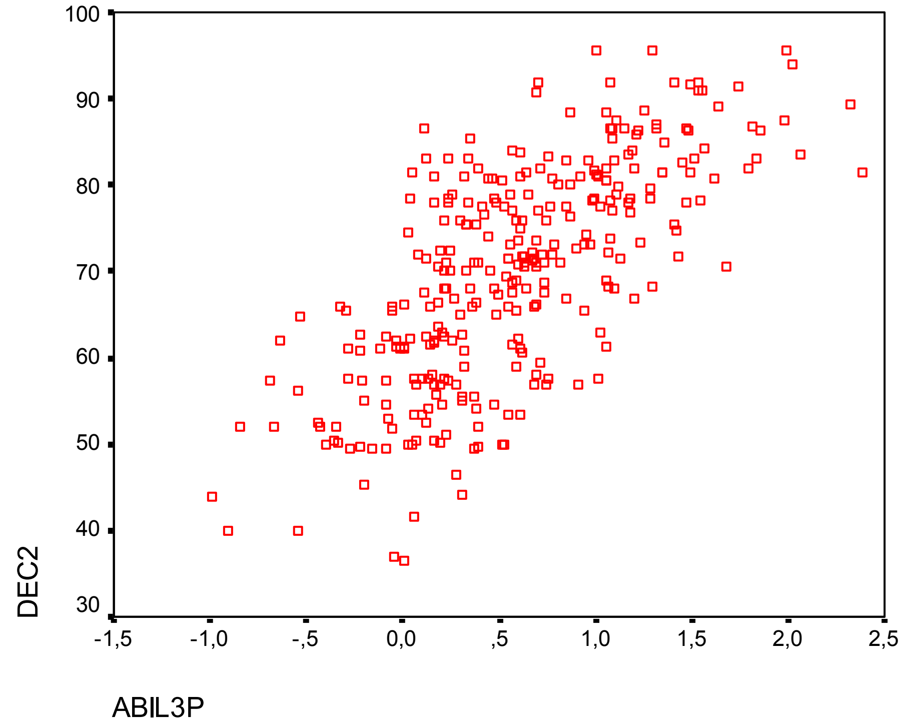


Figure F4 3 P Ability Scores and Second Semester DEC Passing Grades

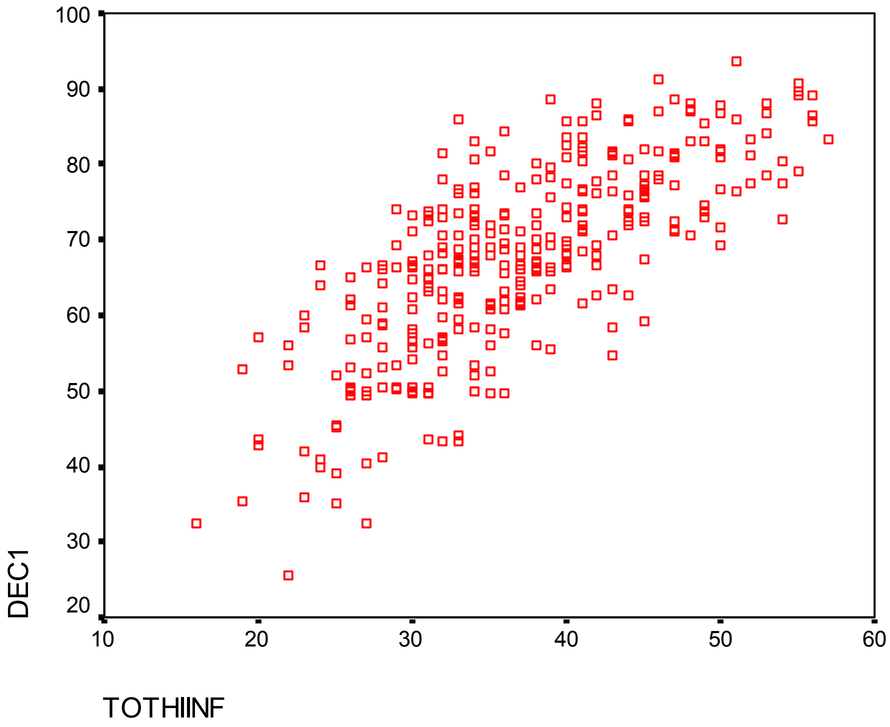


Figure F5 3 P BUEPE Total Scores with 60 High Information Items vs. First Semester DEC Passing Grades

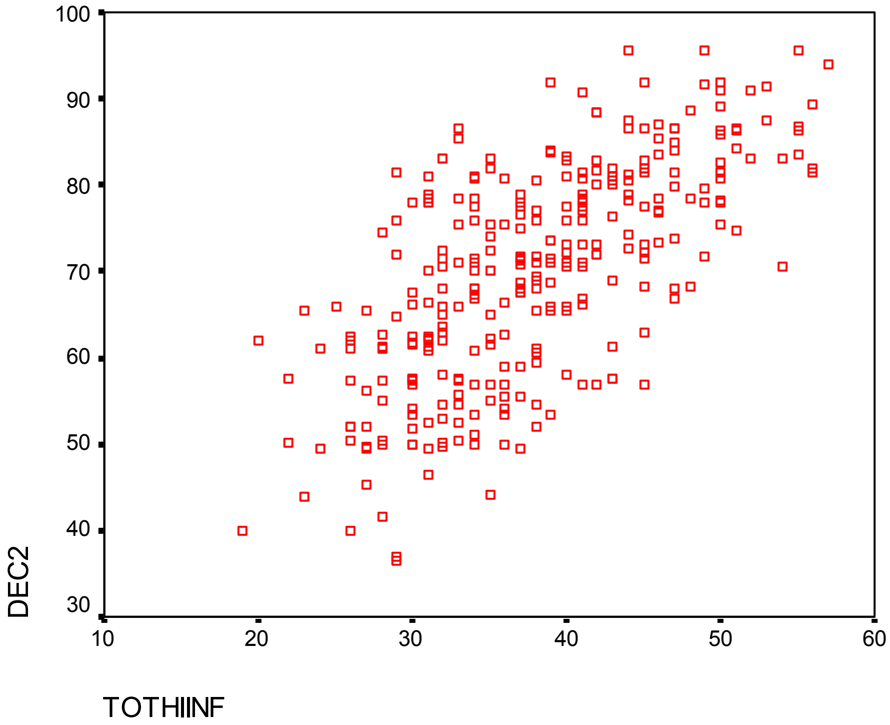


Figure F6 3 P BUEPE Total Scores with 60 High Information Items vs. Second Semester DEC Passing Grades

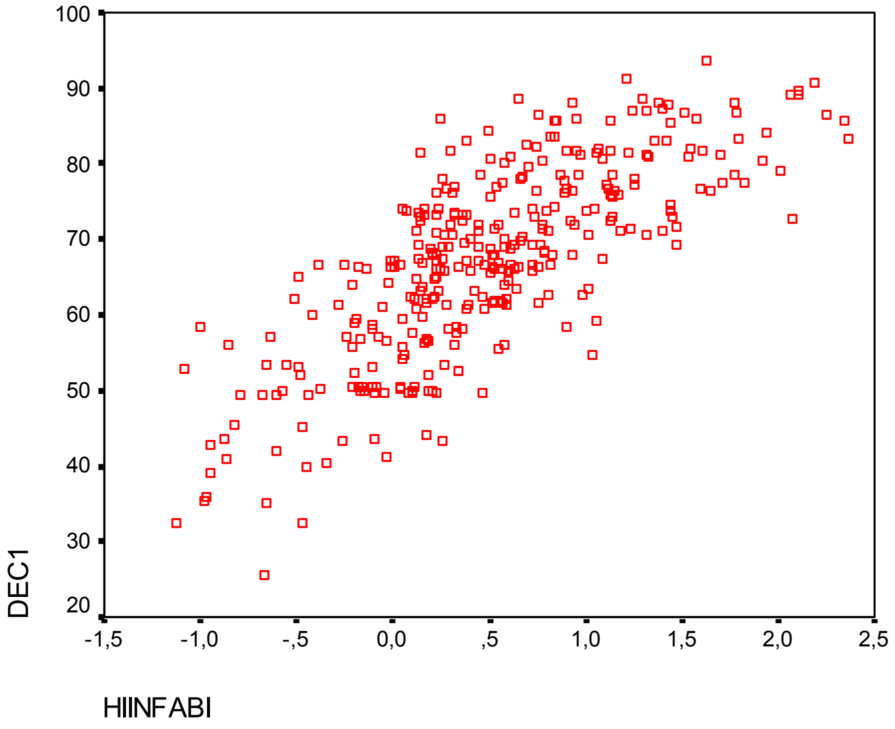


Figure F7 Ability Scores with 60 High Information Items vs. First Semester DEC Passing Grades

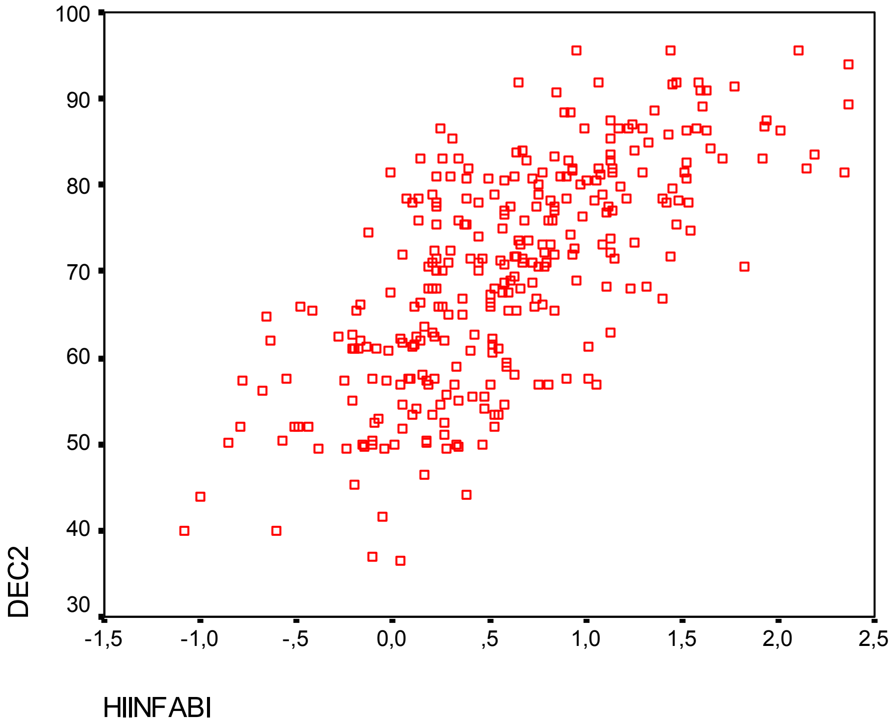


Figure F8 Ability Scores with 60 High Information Items vs. Second Semester DEC Passing Grades

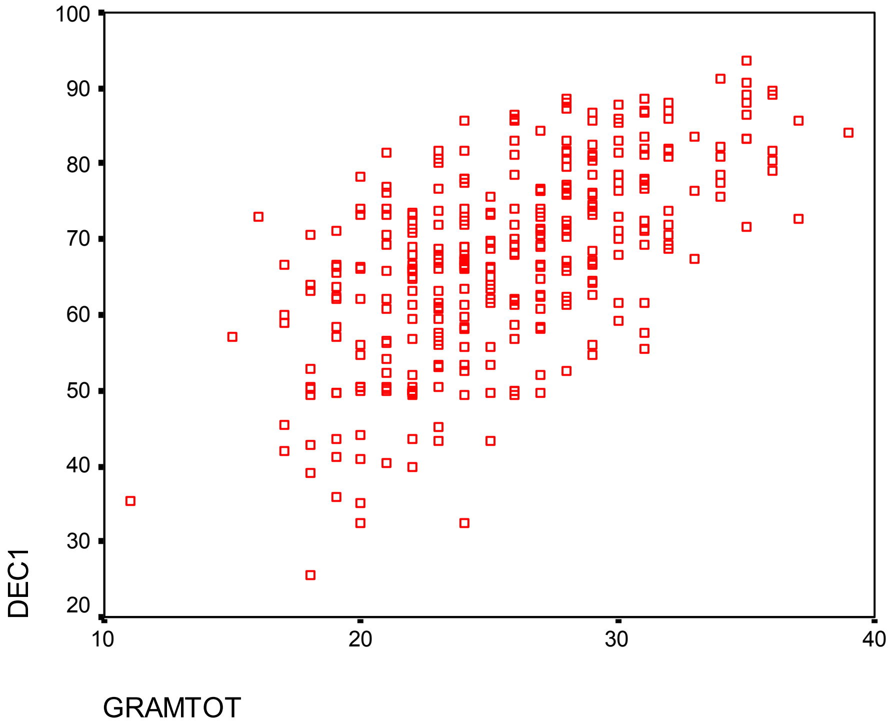


Figure F9 Total Score with Grammar Sub-test vs. First Semester DEC Passing Grades

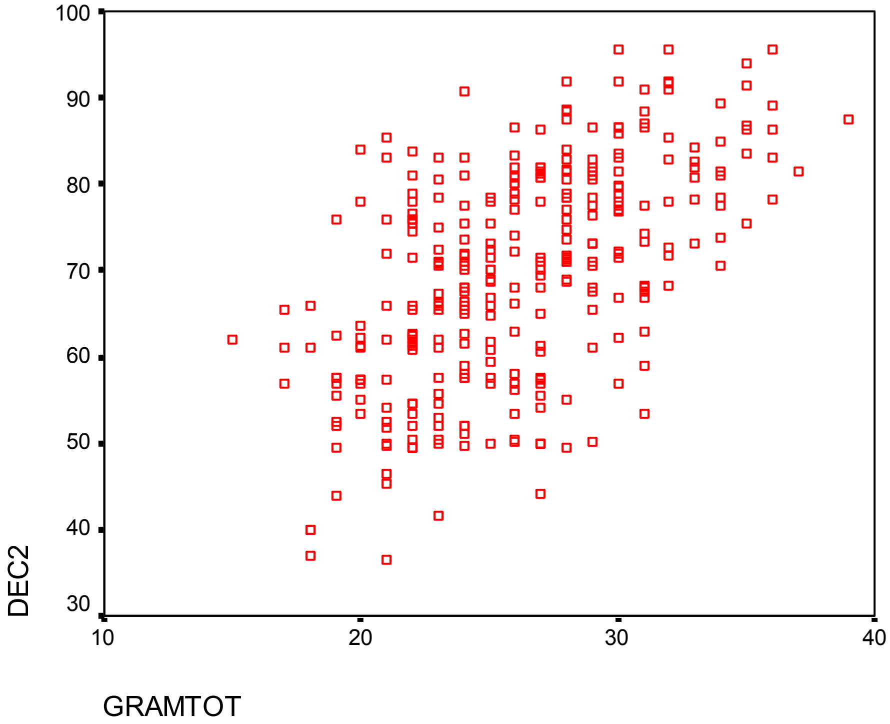


Figure F10 Total Score with Grammar Sub-Test vs. Second Semester DEC Passing Grades

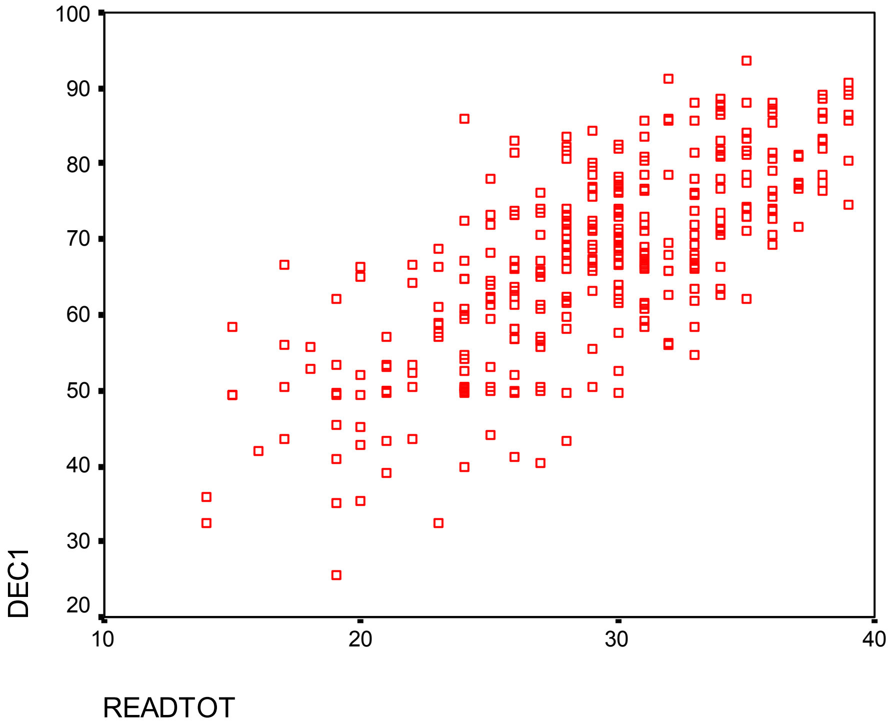


Figure F11 Total Score with Reading Sub-Test vs. First Semester DEC Passing Grades

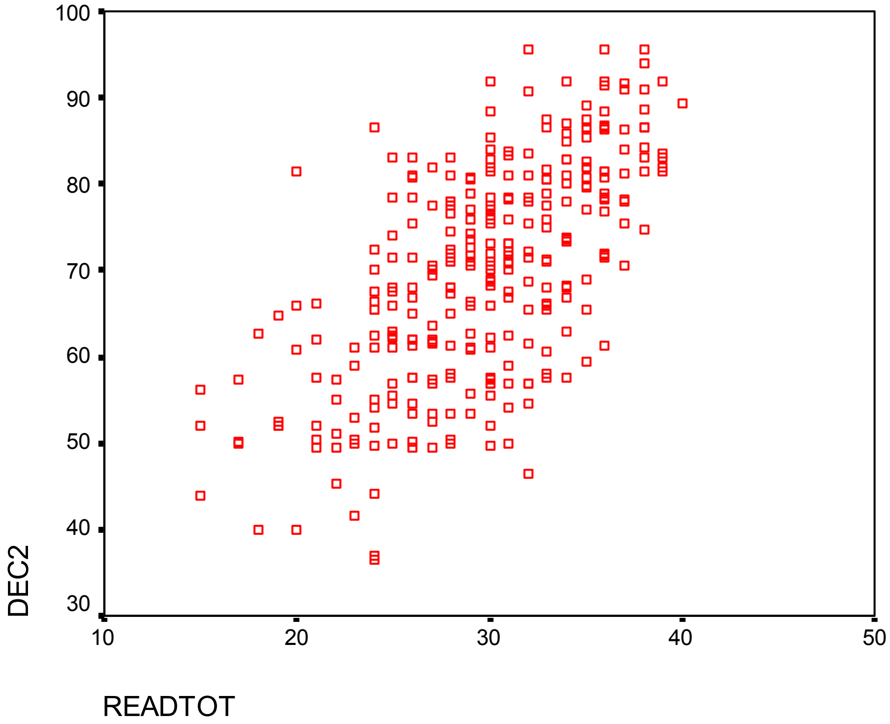


Figure F12 Total Score with Reading Sub-Test vs. Second Semester DEC Passing Grades

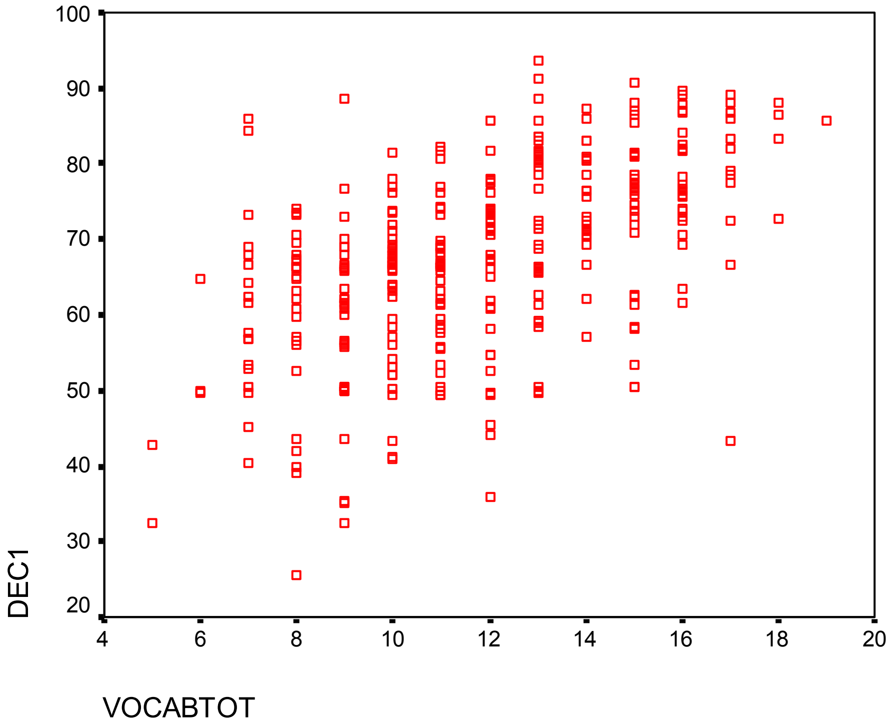


Figure F13 Total Score with Vocabulary Sub-Test vs. First Semester DEC Passing Grades

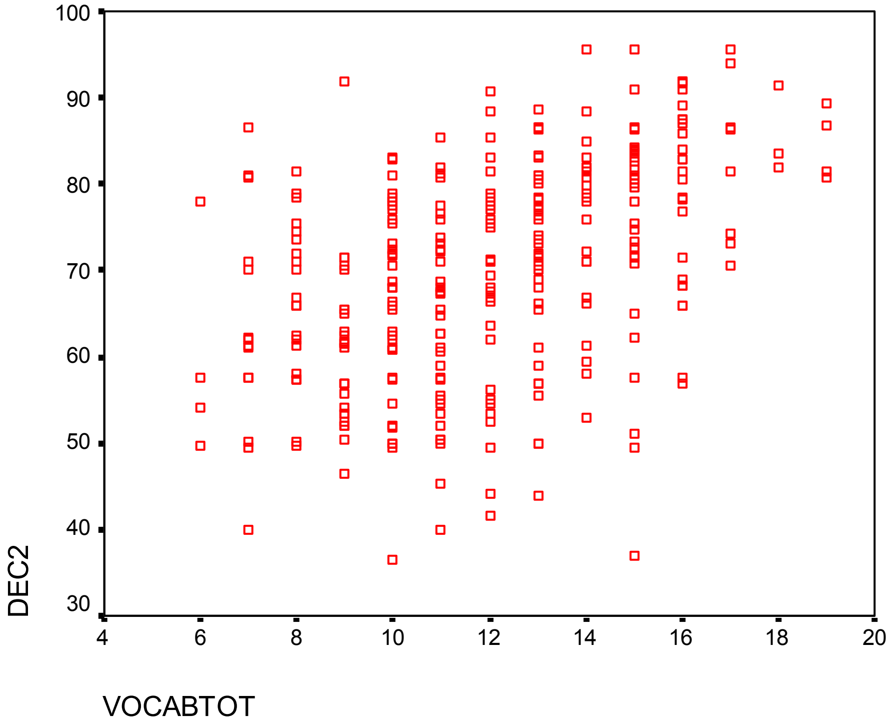


Figure F14 Total Score with Vocabulary Sub-Test vs. Second Semester DEC Passing Grades

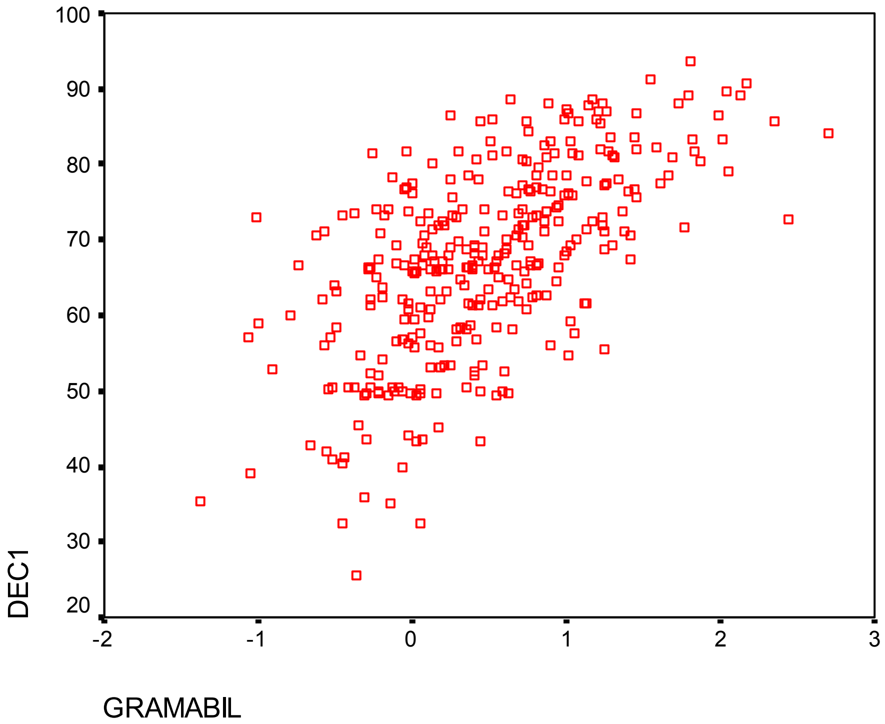


Figure F15 Ability Scores with Grammar Sub-Test vs. First Semester DEC Passing Grades

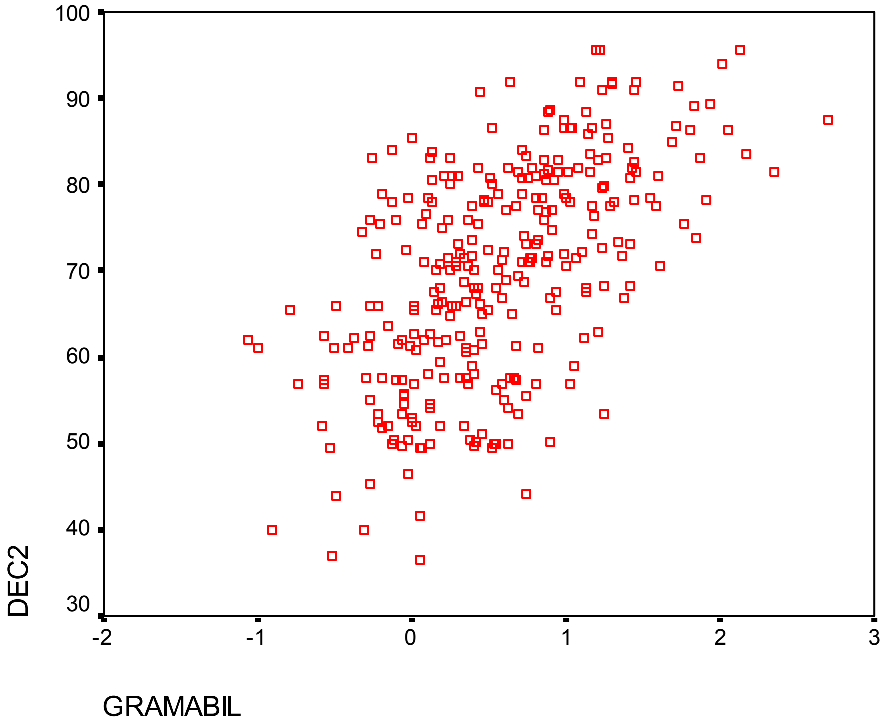


Figure F16 Ability Scores with Grammar Sub-Test vs. Second Semester DEC Passing Grades

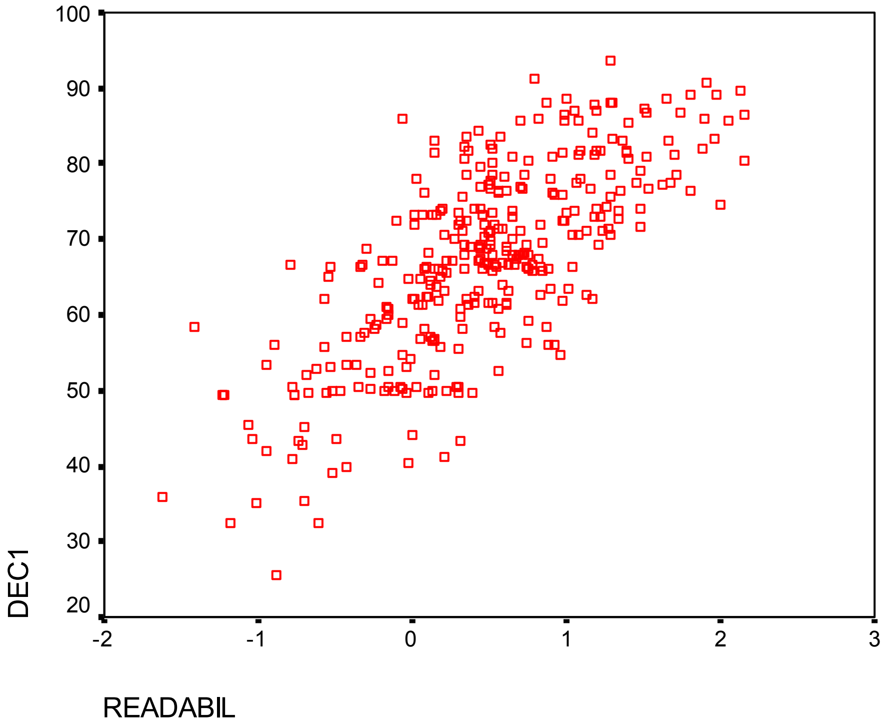


Figure F17 Ability Scores with Reading Sub-Test vs. First Semester DEC Passing Grades

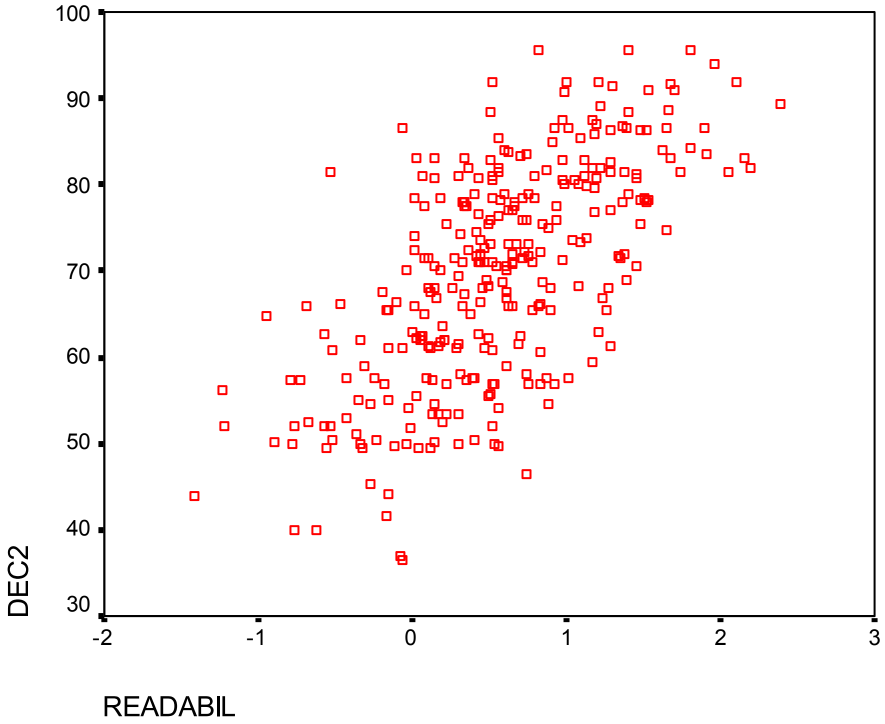


Figure F18 Ability Scores with Reading Sub-Test vs. Second Semester DEC Passing Grades

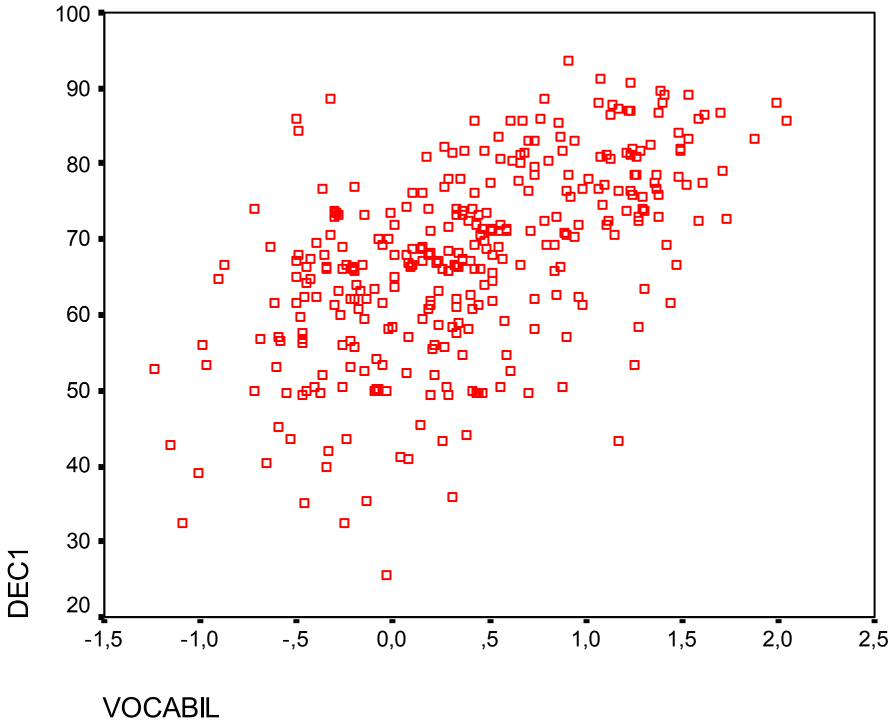


Figure F19 Ability Scores with Vocabulary Sub-Test vs. First Semester DEC Passing Grades

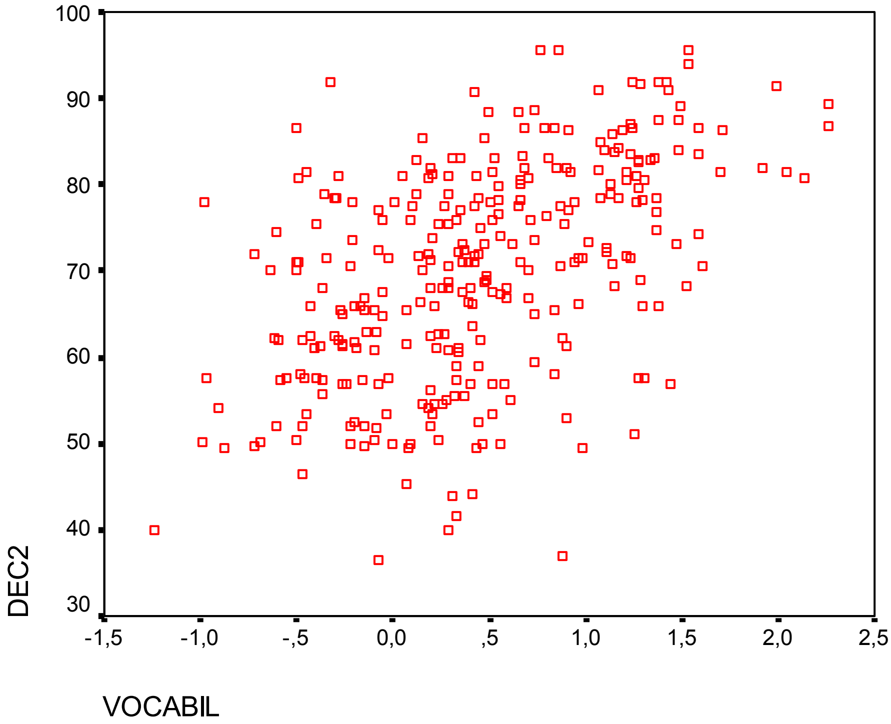


Figure F20 Ability Scores with Vocabulary Sub-Test vs. Second Semester DEC Passing Grades

APPENDIX G
APPENDIX G1
ITEM INFORMATION FUNCTIONS (IIF) OF THE 3 PARAMETER MODEL IN ITEM NUMBER ORDER

Item Number	IIF	Item Number	IIF
1,00	,0786	51,00	,4158
2,00	,5579	52,00	,4404
3,00	,2465	53,00	,6346
4,00	,2630	54,00	,5016
5,00	,6414	55,00	,3537
6,00	,0539	56,00	,3917
7,00	,2558	57,00	,7898
8,00	,2612	58,00	,2323
9,00	4,7365	59,00	,1406
10,00	,2284	60,00	,2185
11,00	,7129	61,00	,4735
12,00	,1276	62,00	,5082
13,00	,1582	63,00	,3026
14,00	1,2025	64,00	,3544
15,00	,0573	65,00	,2544
16,00	,4799	66,00	,4538
17,00	,6839	67,00	,2142
18,00	,2633	68,00	,2751
19,00	,2195	69,00	,5003
20,00	,7257	70,00	,1233
21,00	,3968	71,00	1,1311
22,00	,9545	72,00	,6695
23,00	,3647	73,00	,8667
24,00	,8317	74,00	,4172
25,00	2,4846	75,00	1,0649
26,00	,1930	76,00	,8360
27,00	,0280	77,00	,8167
28,00	,1898	78,00	,3386
29,00	,2256	79,00	,2875
30,00	,2236	80,00	,1509
31,00	,1181	81,00	,6813
32,00	,0266	82,00	,0622
33,00	,3568	83,00	,0882
34,00	,5480	84,00	,6477
35,00	1,1553	85,00	,2205
36,00	,4571	86,00	1,3248
37,00	,3728	87,00	,1172
38,00	,3694	88,00	,7251
39,00	,1692	89,00	,1009
40,00	,5319	90,00	,0796
41,00	,7736	91,00	,3630
42,00	,3725	92,00	,2401
43,00	,1332	93,00	,3813
44,00	,6216	94,00	,5847
45,00	,5900	95,00	,3327
46,00	,3925	96,00	,3011
47,00	,1446	97,00	,3586
48,00	,5439	98,00	,3813
49,00	,4289	99,00	,5465
50,00	,0565	100,0	2,2947

APPENDIX G2
ITEM INFORMATION FUNCTIONS (IIF) OF THE 3 PARAMETER MODEL FROM HIGHEST TO LOWEST

Item Number	IIF	Item Number	IIF
9,00	4,7365	38,00	,3694
25,00	2,4846	23,00	,3647
100,0	2,2947	91,00	,3630
86,00	1,3248	97,00	,3586
14,00	1,2025	33,00	,3568
35,00	1,1553	64,00	,3544
71,00	1,1311	55,00	,3537
75,00	1,0649	78,00	,3386
22,00	,9545	95,00	,3327
73,00	,8667	63,00	,3026
76,00	,8360	96,00	,3011
24,00	,8317	79,00	,2875
77,00	,8167	68,00	,2751
57,00	,7898	18,00	,2633
41,00	,7736	4,00	,2630
20,00	,7257	8,00	,2612
88,00	,7251	7,00	,2558
11,00	,7129	65,00	,2544
17,00	,6839	3,00	,2465
81,00	,6813	92,00	,2401
72,00	,6695	58,00	,2323
84,00	,6477	10,00	,2284
5,00	,6414	29,00	,2256
53,00	,6346	30,00	,2236
44,00	,6216	85,00	,2205
45,00	,5900	19,00	,2195
94,00	,5847	60,00	,2185
2,00	,5579	67,00	,2142
34,00	,5480	26,00	,1930
99,00	,5465	28,00	,1898
48,00	,5439	39,00	,1692
40,00	,5319	13,00	,1582
62,00	,5082	80,00	,1509
54,00	,5016	47,00	,1446
69,00	,5003	59,00	,1406
16,00	,4799	43,00	,1332
61,00	,4735	12,00	,1276
36,00	,4571	70,00	,1233
66,00	,4538	31,00	,1181
52,00	,4404	87,00	,1172
49,00	,4289	89,00	,1009
74,00	,4172	83,00	,0882
51,00	,4158	90,00	,0796
21,00	,3968	1,00	,0786
46,00	,3925	82,00	,0622
56,00	,3917	15,00	,0573
93,00	,3813	50,00	,0565
98,00	,3813	6,00	,0539
37,00	,3728	27,00	,0280
42,00	,3725	32,00	,0266

APPENDIX G3
ITEM INFORMATION FUNCTIONS (IIF) OF THE HIGHEST 60 ITEMS IN THE 3 PARAMETER MODEL

Item Number	IIF	Item Number	IIF
2,0	,5579	54,0	,5016
5,0	,6414	55,0	,3537
9,0	4,7365	56,0	,3917
11,0	,7129	57,0	,7898
14,0	1,2025	61,0	,4735
16,0	,4799	62,0	,5082
17,0	,6839	63,0	,3026
20,0	,7257	64,0	,3544
21,0	,3968	66,0	,4538
22,0	,9545	69,0	,5003
23,0	,3647	71,0	1,1311
24,0	,8317	72,0	,6695
25,0	2,4846	73,0	,8667
33,0	,3568	74,0	,4172
34,0	,5480	75,0	1,0649
35,0	1,1553	76,0	,8360
36,0	,4571	77,0	,8167
37,0	,3728	78,0	,3386
38,0	,3694	81,0	,6813
40,0	,5319	84,0	,6477
41,0	,7736	86,0	1,3248
42,0	,3725	88,0	,7251
44,0	,6216	91,0	,3630
45,0	,5900	93,0	,3813
46,0	,3925	94,0	,5847
48,0	,5439	95,0	,3327
49,0	,4289	97,0	,3586
51,0	,4158	98,0	,3813
52,0	,4404	99,0	,5465
53,0	,6346	100,0	2,2947

APPENDIX G4
ITEM INFORMATION FUNCTIONS (IIF) OF THE HIGHEST 35 ITEMS IN THE 3 PARAMETER MODEL

Item Number	IIF
2,0	,5579
5,0	,6414
9,0	4,7365
11,0	,7129
14,0	1,2025
17,0	,6839
20,0	,7257
22,0	,9545
24,0	,8317
25,0	2,4846
34,0	,5480
35,0	1,1553
40,0	,5319
41,0	,7736
44,0	,6216
45,0	,5900
48,0	,5439
53,0	,6346
54,0	,5016
57,0	,7898
62,0	,5082
69,0	,5003
71,0	1,1311
72,0	,6695
73,0	,8667
75,0	1,0649
76,0	,8360
77,0	,8167
81,0	,6813
84,0	,6477
86,0	1,3248
88,0	,7251
94,0	,5847
99,0	,5465
100,0	2,2947