*IN THE NAME OF GOD*

*THE COMPASSIONATE*

*AND*

*THE MERCIFUL*

**بسمه تعالی**

## تاییدیه اعضای هیات داوران حاضر در جلسه دفاع از رساله دکتری

خانم / آقای سارا جلالی رساله واحدی خود را با عنوان: مقایسه نظری و عملی دیدگاه‌های روان‌سنجی کلاسیک و سوال-پاسخ و طراحی مدل آزمون انطباقی رایانه‌ای براساس نظریه سوال-پاسخ

در تاریخ ارائه کردند.

اعضای هیات داوران نسخه نهایی این رساله را از نظر فرم و محتوا تایید کرده است و پذیرش آن‌را برای تکمیل درجه دکتری آموزش زبان انگلیسی پیشنهاد می‌کنند.

| اعضای هیات داوران | نام و نام خانوادگی | رتبه علمی | امضاء |
|---|---|---|---|
| 1- استاد راهنمای اصلی | غلامرضا کیانی | دانشیار | |
| 2- استاد راهنمای دوم | - | - | |
| 3- استاد مشاور اول | ولی‌الله فرزاد | استادیار | |
| 4- استاد مشاور دوم | - | - | |
| 5- استاد ناظر | سیده سوسن مرندی | استادیار | |
| 6- استاد ناظر | سید محمد علوی | دانشیار | |
| 7- استاد ناظر | پرویز بیرجندی | استاد | |
| 8- استاد ناظر | محمدرضا عنانی سراب | استادیار | |
| 9- نماینده شورای تحصیلات تکمیلی | رضا غفارثمر | استادیار | |

# آيين‌نامه حق مالكيت مادي و معنوي در مورد نتايج پژوهشهاي علمي دانشگاه تربيت مدرس

مقدمه: با عنايت به سياست‌هاي پژوهشي و فناوري دانشگاه در راستاي تحقق عدالت و كرامت انسانها كه لازمه شكوفايي علمي و فني است و رعايت حقوق مادي و معنوي دانشگاه و پژوهشگران، لازم است اعضاي هيأت علمي، دانشجويان، دانش‌آموختگان و ديگر همكاران طرح، در مورد نتايج پژوهشهاي علمي كه تحت عناوين پايان‌نامه، رساله و طرحهاي تحقيقاتي با هماهنگي دانشگاه انجام شده است، موارد زير را رعايت نمايند:

ماده 1- حق نشر و تكثير پايان نامه/ رساله و درآمدهاي حاصل از آنها متعلق به دانشگاه مي باشد ولي حقوق معنوي پديد آورندگان محفوظ خواهد بود.

ماده 2- انتشار مقاله يا مقالات مستخرج از پايان‌نامه/ رساله به صورت چاپ در نشريات علمي و يا ارائه در مجامع علمي بايد به نام دانشگاه بوده و با تاييد استاد راهنماي اصلي، يكي از اساتيد راهنما، مشاور و يا دانشجوي مسئول مكاتبات مقاله باشد. ولي مسئوليت علمي مقاله مستخرج از پايان نامه و رساله به عهده اساتيد راهنما و دانشجو مي باشد.

تبصره: در مقالاتي كه پس از دانش‌آموختگي بصورت تركيبي از اطلاعات جديد و نتايج حاصل از پايان‌نامه/ رساله نيز منتشر مي‌شود نيز بايد نام دانشگاه درج شود.

ماده 3- انتشار كتاب و يا نرم افزار و يا آثار ويژه حاصل از نتايج پايان‌نامه/ رساله و تمامي طرحهاي تحقيقاتي كليه واحدهاي دانشگاه اعم از دانشكده ها، مراكز تحقيقاتي، پژوهشكده ها، پارك علم و فناوري و ديگر واحدها بايد با مجوز كتبي صادره از معاونت پژوهشي دانشگاه و براساس آئين نامه هاي مصوب انجام شود.

ماده 4- ثبت اختراع و تدوين دانش فني و يا ارائه يافته ها در جشنواره هاي ملي، منطقه‌اي و بين‌المللي كه حاصل نتايج مستخرج از پايان‌نامه/ رساله و تمامي طرح‌هاي تحقيقاتي دانشگاه بايد با هماهنگي استاد راهنما يا مجري طرح از طريق معاونت پژوهشي دانشگاه انجام گيرد.

ماده 5- اين آيين‌نامه در 5 ماده و يك تبصره در تاريخ87/4/1 در شوراي پژوهشي و در تاريخ 87/4/23 در هيأت رئيسه دانشگاه به تاييد رسيد و در جلسه مورخ 87/7/15 شوراي دانشگاه به تصويب رسيده و از تاريخ تصويب در شوراي دانشگاه لازم‌الاجرا است.

**English Department**
**Faculty of Humanities**
**Tarbiat Modares University**

Thesis
Submitted in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy in
Teaching English as a Foreign Language

# Theoretical and Practical Comparisons of Two Psychometric Frameworks Classical Test Theory and Item Response Theory, and Designing the Computer Adaptive Test on the Basis of Item Response Theory

**By**
Sara Jalali

**Supervisor**
Dr. Gholam Reza Kiany

**Advisor**
Dr. Valiollah Farzad

**2010**

# DEDICATION

***To the loving memory of my father who was not able to see me complete my doctoral studies.***

***To my mother who supported me through this endeavor.***

***&***

***To my dearest sister, Dr. Pooneh Jalali, for her overwhelming support and encouragement.***

# Acknowledgements

I would like to express my profound gratitude to my supervisor, Dr. Kiany for his interest in the project and for his valuable comments and guidance from the selection of the topic up to the completion of the research. He has been there to answer questions and offer guidance. He has always asked the hard questions, and allowed me to grow as a researcher. I will be eternally grateful to him.

I am also indebted to my advisor, Dr. Farzad who kindly read and commented on the first draft of the dissertation and provided me with helpful guidelines.

I owe to my internal and external readers Dr. Ghafar Samar, Dr. Marandi, Dr. Alavi, Dr. Birjandi and Dr. Anani for their valuable advice. Their comments and discussions with me have made me think and reexamine my work. I should thank them for listening to me and helping guide my research ideas and questions.

I also owe special thanks to the staff of National Organization of Educational Testing (NOET) who helped me a lot in data gathering and handling the software.

# Abstract

There are two major theories of measurement in psychometrics: Classical Test Theory (CTT) and Item-Response Theory (IRT). Despite its widespread and long use, CTT has a number of shortcomings, which make it problematic to be used for practical and theoretical purposes. IRT tries to solve these shortcomings, and provide better and more dependable answers. One of the applications of IRT is the assessment of Differential Item Functioning (DIF). DIF tells the test developer whether the test item functions differently for different groups. Another important use of IRT is in the area of Computer Adaptive Tests (CAT). CAT is based on IRT, and the stepping-stone in preparing a CAT is the preparation of an item bank. Item banking is based on IRT. When IRT is ignored, item banking will not be applicable and consequently there will be no CAT.

The present study first provided a thorough comparison of CTT and IRT from both theoretical and practical perspectives. For this part of the study, the scores of 3000 testees were used. After that, IRT was utilized to estimate DIF between two gender groups and three fields of study i.e. mathematics, science and humanities in the specific English language part of the foreign language university entrance exam questions of the year 2006. For this part, the data of 15486 participants were used for finding gender DIF and the data of 3924 participants for field DIF. Then, IRT was used to prepare an item bank of the specific English language part of the mock foreign language entrance exam questions for the years 2006 and 2007. This mock exam is administered by an institute related to National Organization of Educational Testing (NOET). For preparing the item bank, specific new software i.e. FastTEST, was utilized. Finally, this item bank was utilized for preparing the CAT version of the English exam, which was the final goal of the dissertation.

The findings of this study showed that CTT- and IRT-based person statistics correlated highly across the three IRT models. Also, it was found that item difficulty and item discrimination indexes from CTT correlated highly with those from all IRT models. The DIF analysis showed that there were a number of DIF items in the exam and these items were analyzed in order to find the source of DIF. Finally, a suitable item bank along with the CAT version of the English exam was prepared. The findings of the present study can be of great importance for the educational system. The researcher proposed some suggestions as to the use of IRT and English CAT in Iran.

**Keywords:** Classical Test Theory (CTT); Item-Response Theory (IRT); Computer Adaptive Test (CAT); item bank; Differential Item Functioning (DIF); specific English part, foreign language university entrance exam

## Table of Contents

Title Page

**Chapter I: Introduction** ……………………………………….…………………………..1

1-1- Introduction ……………………………………………………………………..…2

1-2- Background ……………………………….………………………………….....2

1-2-1- Classical test theory (CTT) ………………………………………………..…..2

1-2-2- Item response theory (IRT) …………………………………………………....5

1-2-3- Differential item functioning (DIF) ………………………………………………7

1-2-4- Item bank ……………………………………………………………………………8

1-2-5- Computer adaptive tests (CATs) …………………………………………………10

1-3- Statement of the problem ……………………………….……………………11

1-4- Significance of the study …………………………………………………………..13

1-5- Research questions ……………………………….……………………………..17

1-6- Null hypotheses …………………….…………………..…………………………18

1-7- Definitions of the key terms.……………………………………………………......19

1-8- Delimitations and limitations of the study …………………………….……………...20

**Chapter II: Review of the literature** …………………………………………………....21

2-1- Introduction ………………………………………………………………………...22

2-2- Classical test theory (CTT) …………………………………………………………...22

2-3- Item response theory (IRT) ……………………………….……………….............27

2-3-1- IRT models ……………………………………………………………………28

2-3-2- Graphical information …………………………..…………………………...32

2-4- Differential item functioning (DIF) ………………………..………………………38

2-5- Item bank for EFL …………………..……………………………………………..48

2-6- English Computer adaptive tests (CATs) ……………………………………………51

**Title Page**

**Chapter III: Methodology** ……………………………………………………………58

3-1- Introduction …………………………………………………………………..59

3-2- Participants …………………………………………………………………...59

3-3- Instrumentation …….…………………………………………………………61

3-3-1- English tests ……………………………………………………………...61

3-3-2- Software ………………………………………………………………….62

3-4- Procedure ……………………………………………………………………...64

3-5- The design of the study ……………………………………………………….67

3-6- Data analysis ………………………………………………………………….68

**Chapter IV: Results and discussion** ……………………………………………….70

4-1- Restatement of the problem …………………………………………………...71

4-2- Introduction …………………………………………………………………...71

4-3- Research question one (Results & discussion) ………………………………..…...76

4-4- Research questions two and three (Results & discussion) ……………………………78

4-5- Research question four (Results & discussion) ………………………………………91

4-6- Research question five (Results & discussion) ………………………………………106

4-7- Research question six (Results & discussion) ……………………………………..…115

4-8- Research question seven (Results & discussion) …………………………………...120

4-9- Research question eight (Results & discussion) ………………………………….....125

**Chapter V: Conclusion** ……………………………………………………………...136

5-1- Introduction …………………………………………………………………....137

5-2- Summary ……………………………………………………………………....137

5-3- Pedagogical implications ……………………………………………………...140

5-4- Suggestions for further research ……………………………………………....142

**References** ……………………………………………………………………………...144

**Title** **Page**

**Appendices** ………………………………………………………………………………..156

Appendix I: Item statistics based on CTT and IRT ……………………………………..157

Appendix II: ICCs for DIF items ……………………………………………….………170

Appendix III: Items of the specific English part of the foreign language university entrance exam 2006) ……………………………………………………………………….....187

Appendix IV: Tree diagrams for the subparts of the test ……….…………………….....196

Appendix V: Preparing an item bank in FastTEST ………………………………….....209

**Persian abstract** ...........................................................................................................237

## List of abbreviations

| | |
|---|---|
| **ASVAB** | Armed Services Vocational Aptitude Battery |
| **a2PL** | Discrimination Parameter for Two-Parameter Logistic Model |
| **a3PL** | Discrimination Parameter for Three-Parameter Logistic Model |
| **b1PL** | Difficulty Parameter for One-Parameter Logistic Model |
| **b2PL** | Difficulty Parameter for Two-Parameter Logistic Model |
| **b3PL** | Difficulty Parameter for Three-Parameter Logistic Model |
| **BISER** | Biserial Correlation Coefficient |
| **CAST** | Computerized Adaptive Screening Test |
| **CAT** | Computer Adaptive Test |
| **CTT** | Classical Test Theory |
| **df** | degrees of freedom |
| **DIF** | Differential Item Functioning |
| **EFL** | English as a Foreign Language |
| **FCE** | First Certificate in English |
| **ICC** | Item Characteristic Curve |
| **IIF** | Item Information Function |
| **IRF** | Item Response Function |
| **IRT** | Item Response Theory |
| **LR** | Logistic Regression |
| **MH** | Mantel-Haenszel |
| **NOET** | National Organization of Educational Testing |
| **P & P** | Paper-and-Pencil |
| **PBISER** | Coint-Biserial Correlation Coefficient |
| **PCTT** | Proportion Correct based on Classical Test Theory |
| **1PLM/1PL** | One-Parameter Logistic Model |
| **2PLM/2PL** | Two-Parameter Logistic Model |
| **3PLM/3PL** | Three-Parameter Logistic Model |
| **SEE** | Standard Error of Estimation |

| | |
|---|---|
| **SEM** | Standard Error of Measurement |
| **SD** | Standard Deviation |
| **TEFL** | Teaching English as a Foreign Language |
| **TIF** | Test Information Function |
| **TOEFL** | Test of English as a Foreign Language |

## List of symbols

| **Symbol** | **Explanation** |
|---|---|
| **a** | discrimination parameter |
| **b** | difficulty parameter |
| $\mathbf{b_F}$ | item difficulty parameter estimate for females |
| $\mathbf{b_M}$ | item difficulty parameter estimate for males |
| **c** | guessing parameter |
| **D** | discrimination index |
| **e** | constant 2.718 |
| **P** | proportion correct |
| **p-value** | probability of success/significance |
| **p (θ)** | probability of a testee's success |
| **q** | proportion of test takers who get the item incorrect |
| **r** | Correlation |
| $\mathbf{r_{bis}}$ | biserial correlation coefficient |
| $\mathbf{r_{pbi}}$ | point-biserial correlation coefficient |
| $\mathbf{s_x}$ | standard deviation of test scores |
| $\mathbf{s^2_x}$ | observed score variance |
| $\mathbf{s^2_t}$ | true score variance |
| $\mathbf{s^2_e}$ | error score variance |
| $\mathbf{SE_F}$ | standard error of estimation for females |
| $\mathbf{SE_M}$ | standard error of estimation for males |
| **SQRT** | square root |
| **t** | t-value |
| **θ** | theta/latent trait |
| $\boldsymbol{\theta_T}$ | ability estimation based on the whole set of items |
| $\boldsymbol{\theta_N}$ | ability estimation based on a sub-set consisting of the items with no DIF |
| **x** | observed score |
| $\mathbf{x_t}$ | true score |

| | |
|---|---|
| $\mathbf{x_e}$ | error score |
| $\bar{x}$ | mean score |
| $\bar{x}_{\mathbf{p}}$ | mean score on the test for those who get the item correct |
| $\bar{x}_{\mathbf{q}}$ | mean score on the test for those who get the item incorrect |
| $\mathbf{z}$ | ordinate of normal curve corresponding to p |
| $\surd$ | square root |
| $\chi^2$ | Chi-square |
| $\phi$ | phi |

## List of tables

Table 2-1- The old rules in CTT …………………………………………………………..36

Table 2-2- The new rules in IRT ……………………………….………………………….37

Table 2-3- Methods for assessing DIF ………………………………………………………41

Table 4-1- Eigen values for all subparts of the test …………………………………….....72

Table 4-2- Residual correlations for the cloze test subpart …………………………………...74

Table 4-3- The reliability of all subparts of the test ……………………………..…………75

Table 4-4- CTT item statistics for the structure subpart ……………………………………...79

Table 4-5- IRT item statistics for the structure subpart (1PL) ………………………………..80

Table 4-6- IRT item statistics for the structure subpart (2PL) ………………………………..81

Table 4-7- IRT item statistics for the structure subpart (3PL) ………………………………..82

Table 4-8- Correlation between CTT and IRT for the difficulty parameter for all subparts of the test ………………………………………………………….……………...……………83

Table 4-9- Correlation between CTT and IRT for the discrimination parameter for all subparts ……………………………………………………………………………………..…....85

Table 4-10- Chi-square values of the IRT models for all subparts of the test ………….…...89

Table 4-11- Chi-square values of the final iteration for all subparts of the test ……..……...93

Table 4-12- Group differential item functioning for the structure subpart …………….………94

Table 4-13- Group differential item functioning for the vocabulary subpart ……………..…95

Table 4-14- Group differential item functioning for the word order subpart ……………..…96

Table 4-15- Group differential item functioning for the language function subpart ………..97

Table 4-16- Group differential item functioning for the cloze test subpart ………………....97

Table 4-17- Group differential item functioning for the reading comprehension subpart ….98

Table 4-18- Total number of group DIF items based on the IRT DIF procedure ………....100

Table 4-19- Comparison of the three DIF methods …………………………………………100

Table 4-20- Crosstabulation for the two variables …………………………………………..102

Table 4-21- Correlations for the three DIF procedures ……………………………………...102

Table 4-22- Chi-square values of the final iteration for all subparts of the test …..………...106

Table 4-23- Field differential item functioning for the structure subpart …………………...107

Table 4-24- Field differential item functioning for the vocabulary subpart ………………....108

Table 4-25- Field differential item functioning for the word order subpart …………………109

Table 4-26- Field differential item functioning for the language function subpart …………109

Table 4-27- Field differential item functioning for the cloze test subpart …………………...110

Table 4-28- Field differential item functioning for the reading comprehension subpart ……111

Table 4-29- Total number of field DIF items based on the IRT DIF procedure……………..113

Table 4-30- Comparison of the two DIF methods ……………………………………...…114

Table 4-31- Correlation between the two DIF procedures ………………..…………………115

Table 4-32- Comparison between ability parameter estimations for all subparts of the test (Males vs. Females) ……………………………………………………………………………..117

Table 4-33- Comparison between ability parameter estimations for all subparts of the test (Math vs. Humanities) ……………………………………………………………………….…118

Table 4-34- Comparison between ability parameter estimations for all subparts of the test (Math vs. Science) ……………………………………………………………………………...118

## List of figures

Figure 2-1- An ICC for the 3PLM ………………………………………………………..31

Figure 2-2- A typical Item characteristic curve (ICC) ……………………………………33

Figure 2-3- An example of an item information function (IIF) …………………………..35

Figure 2-4- An example of a test information function (TIF) …………………………….36

Figure 2-5- An example of an item that does not display DIF ……………………………...43

Figure 2-6- An example of an item that displays substantial uniform DIF …………………43

Figure 2-7- An example of an item that displays substantial non-uniform DIF …………….44

Figure 4-1- The relationship between unidimensionality and local independence …………75

Figure 4-2- Regression of ability on percentage correct for 1PL model ……………………77

Figure 4-3- Regression of ability on percentage correct for 2PL model ……………………77

Figure 4-4- Regression of ability on percentage correct for 3PL model ……………………78

Figure 4-5- Item bank for the structure subpart of the test ……………………………..…122

Figure 4-6- Item bank for the vocabulary subpart of the test ……………………..……....123

Figure 4-7- Item bank for the word order subpart of the test ……………………..…....123

Figure 4-8- Item bank for the language function subpart of the test ……………………..…124

Figure 4-9- Item bank for the cloze test subpart of the test ……………………..……..…124

Figure 4-10- Item bank for the reading comprehension subpart of the test ……………..…..125

Figure 4-11- Configuration for a test module ……………………………………………..…127

Figure 4-12- Sample CAT item for the structure subpart of the test ………...…………....128

Figure 4-13- Sample CAT item for the cloze subpart of the test …………………………….128

Figure 4-14- Sample CAT item for the reading comprehension subpart of the test …………129

Figure 4-15- The first Instruction Module of the test for testee's name and identification ………………………………………………………………………………………...129

Figure 4-16- The last Instruction Module for finishing the test session ……………………..130

Figure 4-17- Page 1 of an item-by-item report for the adaptive test of the structure subpart …………………………………………………………………………………..132

Figure 4-18- Page 2 of an item-by-item report for the adaptive test of the structure subpart …………………………………………………………………………………..133

# CHAPTER I

# INTRODUCTION

## 1-1- Introduction

In this chapter, first, a background to the study is presented. Then, an overview of five important concepts is provided i.e. classical test theory, item response theory, differential item functioning, item banks and computer adaptive tests. In the second section, the problems are stated. The third section describes the significance of this study. The remaining sections of the chapter state the research questions and hypotheses (sections four and five), elaborate on the definitions of pertinent terms (section six), and describe limitations of the study (section seven).

## 1-2- Background

In the theory of measurement, there are two major measurement frameworks: classical test theory (CTT) and item response theory (IRT). Differences are most evident in the statistical analysis underlying each theory.

### 1-2-1- Classical test theory (CTT)

Classical test theory (CTT) is best suited for traditional testing situations, either in group or in individual settings, in which all the members of a target population, e.g. persons seeking college admission, are administered the same or parallel sets of test items. CTT has four underlying assumptions (Bachman, 1990):