**University of Shiraz**

**Faculty of Literature and Humanities**

**M.A. Thesis in** Teaching English as a Foreign Language - TEFL

# AN EXPLORATORY STUDY OF DIFFERENTIAL ITEM FUNCTIONING (DIF) IN EFL READING COMPREHENSION

By
**Touraj Jalili**

Supervised by
**Dr. Ali Reza Ahmadi**

**July 2012**

*IN THE NAME OF GOD*

**In the Name of God**

# DECLARATION

Hereby, Touraj Jalili (880227) student of English teaching at the Faculty of Literature and Humanities certify that this thesis results from my own research and whenever I have utilized other sources, I have clearly reference them. I also declare that the research and the title of my thesis are novel and I promise, without the permission from the university, the results never be published or bring to someone else. The copyright of this thesis is the property of Shiraz University.

**Name and Surname: Touraj Jalili**

**Date and Signature: 18/7/2012**

# Dedicated to

# *MY DEAR WIFE*

**Acknowledgement**

I would like to express my heart-felt gratitude to Dr. Ahmadi for his considerate and attentive reading of my thesis and also his constructive comments on the drafts of the thesis; I owe him his great help throughout the thesis, without his support I could not manage to fill the important gaps in the work and timely complete the thesis. Thanks also go to Dr. Razmjoo and Dr. Rahimi, my first and second readers, for their extremely useful and insightful comments on the thesis. I am really thankful to the staff, teachers (my colleagues), and students at Shiraz University Language Center (SULC) who sincerely let me give the reading test to the participants and gather the required data. I am also indebted to my dear wife, my soul, who was always there to help me get rid of the stress during the work and focus my attention on the thesis. Finally, I would like to express my warmest thanks to my dear parents and all those who helped me to peacefully complete the job.

# Abstract


## AN EXPLORATORY STUDY OF DIFFERENTIAL ITEM FUNCTIONING (DIF) IN EFL READING COMPREHENSION

By
**Touraj Jalili**

The literature on the sources of differential item functioning (DIF) in reading comprehension is replete with a host of speculative variables with gender, familiarity with text topic, interest in text topic or content, guessing, and contextual factors being some of the prominent ones (Pae, 2004; Zumbo & Gelin 2005). The present study, drawing on Popperian falsification philosophy, concentrated on the above factors and attempted to empirically investigate whether the speculative sources of DIF in EFL reading comprehension could stand on a firm verifiability/falsifiability ground. In this study an account was given to the reading performance of 203 Iranian test takers, 110 females and 93 males, to help the researcher to find the DIF items by means of two statistical models, logistic regression (LR) and item response theory (IRT). To this end, a reading test with six passages (2 passages with female-friendly topics, 2 with male-friendly topics, and 2 with neutral topics) was developed and each passage was followed by five item types, knowledge-, reference-, vocabulary-, main idea-, and inference-type to determine whether gender could predict differential performance on the reading items in general and on the item types in particular. Furthermore, attached to each passage was a questionnaire asking the test takers to make it clear whether they guessed at the answers, to what extent the text topic was familiar to them, and how interested they were in the text topic. The test takers were also asked to inform the researcher about their residential location, monthly income, and families' educational level to account for the contextual sources of DIF. The results of the study made it clear that regarding the gender and familiarity only the LR model flagged DIF items. The gender DIF confirmed the literature-based speculations but the familiarity DIF worked in a reverse manner and benefited the low familiar test takers. The other sources of DIF were identified by both models. For the interest DIF the results were mixed. That is, while the LR found a single item favoring the interested group, the IRT model detected some items in favor of the interested and some benefiting the uninterested group. With respect to the guessing and income DIF the results of the models were contradictory. The LR indicated DIF in favor of the low guessers and high-income group whereas the IRT model showed DIF favoring the high guessers and low-income group. The results of location and educational level DIF

by both methods were in a complete correspondence with both sources leading to DIF in favor of the test takers living in the expensive neighborhoods and with academically educated families. The findings of this research could support, albeit with a small sample, the effect on the EFL reading items of gender, interest, location, and educational level. However, more empirical research is required to confirm, with large samples and in different cultural settings, the hypothetical functioning of guessing, familiarity, and income as sources of DIF in EFL reading comprehension.

**Keywords:** DIF, LR and IRT models, Reading comprehension, Item type, Gender, Familiarity, Interest, Guessing, Contextual factors

## Table of Contents

**Content** **Page**

# List of Tables

**Content** **Page**

# List of Figures

**Content** **Page**

## List of Abbreviations

ACTIVE: Activate prior knowledge, Cultivate vocabulary, Think about meaning, Increase reading fluency, Verify strategies, Evaluate progress
ANOVA: analysis of variance
CR: criterion referenced
CTT: classical test theory
DDF: differential domain functioning
DIF: differential item functioning
$DR^2$ = augmented $R^2$– null $R^2$
DTF: differential test functioning
EFL: English as a foreign language
ESL: English as a second language
ESLPE: UCLA's ESL placement test
ETS: Educational Testing Service
FCE: First Certificate in English
ICC: item characteristic curve
IQ: intelligence quotient
IRT: item response theory
IRT LRT: IRT likelihood ratio test
LR: logistic regression
L2: second language
MH: Mantel-Haenszel
MHD: Mantel-Haenszel delta
NS: native speaker
NNS: non native speaker
OR: odds ratio
SIBTEST: simultaneous item bias test
SPEAK: Speaking Proficiency English Assessment Kit
SULC: Shiraz University Language Center
TCC: test characteristic curve
TID: transformed item difficulty

TOEFL: Test of English as a Foreign Language
1PL IRT: one-parameter logistic IRT
2PL IRT: two-parameter logistic IRT
3PL IRT: three-parameter logistic IRT

# Chapter I

# Introduction

## 1.1. Overview

This section introduces the general concepts and ideas of DIF and bias and the relationships among them. It starts with a brief overview of the reading skill and the effects on it of various variables. Then, the discussion will be focused on the differential item functioning as the preliminary stage of bias analysis. The socially-oriented concepts of bias, fairness, and validity and their relations with each other will be discussed with an eye toward the implications of having DIF items for the validity and fairness. Finally, the significance and purpose of the study will be explained and some key terms will be defined.

## 1.2. Reading

Reading skill has always been indispensable for academic success (Grabe & Stoller, 2002; Levine, Ferenz, & Reves, 2000). Researchers have, thus, attempted a great deal to identify the critical components that have the most effect on reading performance. Gender, prior knowledge, interest, and language ability are among the factors that affect reading comprehension performance (Brantmeier, 2001, 2003; Grabe & Stoller, 2002; and Pae, 2004). Studies on the role of prior knowledge (or familiarity of content) and gender as key variables in reading comprehension and the interaction between them abound (e.g. Brantmeier, 2001, 2003; Keshavarz & Ashtarian, 2008; Newman, Groom, Handelman, & Pennebaker, 2008; Pae, 2004; & Shumaimeri, 2005). According to Keshavarz and Ashtarian (2008) there are two crucial factors that influence the process of reading comprehension.

They are reader and text variables. The former is related to the ways readers affect the process and includes the strategies employed by them, their background knowledge, motivation, attitude, age, personality, and gender (see also Brantmeier, 2003, 2004, & 2007). The latter, on the other hand, may include the difficulty level of the text with respect to vocabulary, grammar, organization, discourse, meaning, and situational or contextual use.

Due to its sociocultural nature, gender has been given utmost importance in the literature and has led researchers to investigate the social and behavioral function of men and women. In recent years there has been extensive theorizing about the nature and existence of differences between men and women (Newman, et al., 2008). Gender differences are more conspicuous in the social aspects of the real life. Language could be conceived of as the primary realization of the social communication and thereby, through meticulous analysis, reveals the gender differences in its clearest form. Thus, language is dealt with as an inherently social phenomenon. Analyzing it thoroughly, researchers get insight into whether, how, and why men and women make different use of this social phenomenon. However, gender differences in reading comprehension are intensified by the level of instruction such that the higher the level, "the wider the gap becomes between male and female students" (Brantmeier, 2003, p. 1).

## 1.3. Differential Item Functioning

Differential item functioning (DIF) occurs when equally knowledgeable individuals from different subgroups are of different likelihood of correctly answering (or endorsing) an item (Angoff, 1993; Camilli & Shepard, 1994; Holland & Thayer, 1988; & Shepard, Cmilli, & Averill, 1981). It is a statistical technique that is applied to uncover the differential item response patterns between groups of test takers and thereby helps detect the potentially biased items (Zumbo & Gelin, 2005). To define it from the item response theory (IRT) point

of view, DIF occurs "when a test item does not have the same relationship to a latent variable across two or more examinee groups" (Embretson & Reise, 2000, p. 251). Mathematically, DIF may be defined as follows:

$P ( x / g_1 , a ) \# P ( x / g_2 , a )$

where p is the probability of endorsing an item, and x, g, and a stand for a (dichotomous) response, a group membership, and an ability level, respectively (Millsap & Everson, 1993). That is, the group membership would act as the distinguishing factor between the groups' performance.

DIF is a statistical method used to flag potentially biased and problematic items. DIF is a necessary but not sufficient condition for bias (Zumbo, 1999). In fact, a biased item will certainly reveal DIF. The analogy is rather like the relationship between reliability and validity. If a test, which comprises different items, is valid, then it will certainly be reliable. Thus, since reliability and validity are psychometrically- and socially-oriented, respectively, the presence of DIF makes a test less reliable and when DIF item turns out to be construct-irrelevant, the biased item makes the interpretation and use of test scores less valid. In fact, test bias can equal invalidity. What is often neglected is the absence of a clear-cut boundary between ability domain and test method facet (Bachman, 1990; Popham, 1978) which, in turn, results in the difficulty to distinguish reliability from validity. Differential item functioning might occur by either (communicative) language ability (called item impact) or by construct-irrelevant variables (called item bias). Bachman (1990) speaks of random factors and test method facets as those parts of test scores that reflect variables other than the intended language abilities and considers them to be the sources of measurement error. Although DIF is said to be the necessary but insufficient condition for bias (Roever, 2005), attention should be paid on the extent to which DIF is related to test method facet. If DIF turns out to be part of the test method facet, then we can certainly conclude that DIF equals bias. Bachman (1990) exemplifies