Differential Item Functioning (DIF) in Terms of Gender in the Reading Comprehension Subtest of a High-Stakes Test

Mohammad Salehi
Assisstant Professor, Sharif University of Technology
m_salehi@sharif.ir

Alireza Tayebi
M.A., Sharif University of Technology
tayebialireza@ymail.com

Abstract

Validation is an important enterprise especially when a test is a high stakes one. Demographic variables like gender and field of study can affect test results and interpretations. Differential Item Functioning (DIF) is a way to make sure that a test does not favor one group of test takers over the others. This study investigated DIF in terms of gender in the reading comprehension subtest (35 items) of a high stakes test using a three-step logistic regression procedure (Zumbo, 1999). The participants of the study were 3,398 test takers, both males and females, who took the test in question (the UTEPT) as a partial requirement for entering a PhD program at the University of Tehran. To show whether the 35 items of the reading comprehension part exhibited DIF or not, logistic regression using a three step procedure (Zumbo, 1999) was employed. Three sets of criteria of Cohen's (1988), Zumbo's (1999), and Jodin and Girel's (2001) were selected. It was revealed that, though the 35 items show "small" effect sizes according to Cohen's classification, they do not display DIF based on the other two criteria. Therefore, it can be concluded that the reading comprehension subtest of the UTEPT favors neither males nor females.

Keywords: Validity, Test Validation, Test Fairness, Differential Item Functioning (DIF), Logistic Regression (LR), Item Response Theory (IRT)

Received: January 2011; Accepted: November 2011

1. Introduction and Theoretical Background

Testing language is always done for a particular purpose in a specific context. One of the most important tests nowadays is the test of English language proficiency used worldwide as an indicator of a person's overall English language knowledge. As Kim (2001) states, English as a second or foreign language proficiency tests are used mainly to measure the English language ability of language learners whose L1 is not English. In other words, proficiency tests usually assess the extent to which an examinee is able to cope with real-life language use situations. These tests are used mostly to become aware of the level of language ability of examinees to make some hopefully correct and logical decisions.

To reach right judgments, it is necessary that a test be valid, since validity is one of the essential features of tests' interpretation and use. That is to say, validity is the quality of interpretations made out of test scores (Bachman, 1990). To prevent inappropriate consequences, Bachman (ibid) believes that bias must be detected and removed, which is a complex procedure. It can be detected through various methods and procedures and the present study mainly focuses on one of the most important and currently used procedures of bias detection, known as Differential Item Functioning (DIF) and its different detection procedures and methods.

Takala and Kaftandjieva (2000) believe that language use might not be seen as a fully uniform phenomenon as there are a number of variants which can affect it depending on the context of use and the language user’s characteristics (e.g., social and geographical origin, education, age, gender, etc.). Among these variables they put more emphasis on the difference in terms of gender and establish a connection between gender difference and DIF. They argue that the relationship between the two (i.e., gender and DIF) is two-way.

In other words, as they state gender differences observed in some research might be due to the biased estimation of the observed variable; however, true gender difference may result in gender DIF.

According to Zumbo (1999), there are two approaches to examine potential measurement bias namely judgmental and statistical. Judgmental methods only rely on one or more expert judges' opinions regarding the selection of potentially biased items, hence it is an impressionistic methodology. It is recommended that, in a high-stakes context statistical techniques be used to investigate potential bias due to this methods defensibility. The present study has used statistical methods to examine potential bias of a high-stakes test.

A proficiency test, according to Brown (2004) is one which aims at testing global competence in a language. Traditionally consisting of standardized multiple-choice items on grammar, vocabulary, reading comprehension, aural comprehension, and sometimes writing skill and oral production performance, a proficiency test is not limited to any single course, curriculum, or skill in the language. In point of fact, it tests overall language ability. The investigation of Differential Item Functioning (DIF) is crucial, therefore, in language proficiency tests, where examinees with various backgrounds are involved, since DIF-exhibiting items pose a considerable threat to the validity of the test (Kim, 2001).

The University of Tehran English Proficiency Test (i.e., the UTEPT) is a high stakes test with almost 9000 testees taking it since according to Roever (2001) the results of such a test bring about life-changing implications for the candidates (e.g., admission tests for universities or other professional programs, certification exams, etc).

1.1. Validity and Validation Process

According to Brown (2004), there is no single, absolute measure of establishing validity; however, various kinds of evidence can help to support it (e.g., evidence from content-related, criterion-related, construct-related, consequential, and face validity). For validity, one of the most complex criteria of tests and the most fundamental in psychometrics (Angoff, 1988), various definitions have been proposed which express the same central idea (i.e., results of the tests must conform to these definitions to be regarded as an effective and valid test). As an example, one of the traditional definitions of validity is the correlation of test scores with some other related objective measure (Bingham, 1937, p. 214; cited in Angoff, 1988). Brown (2005) also defines validity as “the degree to which a test measures what it claims, or purports, to be measuring” (p. 220) and gains special importance when involved in making decisions about students; therefore, after taking into account issues of practicality and reliability, validity should also be considered.

Anastasi (1986) believes that validity should be considered from the very beginning steps of test construction as opposed to traditional criterion-related validation where validity is only limited to the final stages of test development. Construct validity is a chief issue concerning validation of large-scale standardized tests of English language proficiency (Brown, 2004). According to Angoff (1988), “Construct validation is a process, not a procedure; and it requires many lines of evidence, not all of them quantitative” (p. 26).

Research on investigating validity in general, an construct validity in particular, is abundant in the literature. For instance, using a multitrait-multimethod (MTMM) design, Salehi and Rezaee (2009) investigated the construct validity of a high-stakes test (i.e., the University of Tehran English Proficiency test, the UTEPT) where two traits-grammar and vocabulary- and two methods-multiple choice and contextualization- were used. In another study conducted by Rezaee and Salehi (2008) factor analysis was done to determine the construct validity of a high-stakes test. The researchers used exploratory factor analysis (EFA) through principal component analysis (PCA) with grammar section being the subject of their research. Varimax rotation yielded distinct factors so that in one sub-section eight distinct factors and in the other sub-section six distinct factors were extracted.

1.2. The Current View of Validity

McNamara and Roever (2006) point out that contemporary discussions of validity take into account such issues as test fairness by developing procedures that supports the rationality of decisions based on tests. According to Zumbo (1999) the current view of validity makes it so central that computing it simply by correlation with another measure is not taken to be an appropriate method. For example, in item bias studies, validation process involves construct definition as the first step before writing the items or selecting a measure and is followed by item analysis processes. Moreover, the process of validation is addressed to specific uses of the test in addition to the specific examinees group taking the test.

Brown (2005) also refers to the change in conceptualization of validity in the field of testing and assessment and makes a distinction between traditional and current view of validity. In fact, in the current view of validity, validation is a central issue which is not confined to computing a correlation with another measure. That is, explicit statistical studies which examine test bias are in demand. Such a need is due to the fact that validation process is never entirely complete.

1.3. Test Validation

According to Bachman (1990), the purpose of validation is highly in line with the specific characteristics of groups of test takers and the needs of test users. Due to the presence of the sources of bias, which is the result of individual characteristics, systematic differences in test performance are caused, hence the validity of our judgments or interpretations may be jeopardized as well.

For Angoff (1988) neither a test nor even the scores produced by the test are validated; rather, "the interpretations and inferences that the user draws from the test scores, and the decisions and actions that flow from those inferences" are to be validated (p. 24). Zumbo (1999) also notes that it is not the measure that is being validated; rather, the inferences made from a measure must be validated. Brown (2005) in the same line of argument points out that "validity is not about the test itself so much as it is about the test when the scores are interpreted for some specific purpose. In fact, it is much more accurate to refer to the validity of the scores and interpretations that result from a test than to think of the test itself as being valid" (p. 221).

Therefore, any inference made from a measure will be meaningless without validation. Test validation is an important consideration and it gains more importance when the test to be validated is a high-stakes one (Rezaee & Salehi, 2008). The approaches to test validation are many. Alderson, Clapham, and Wall (1995) mention the following approaches to construct validation. The first approach is the correspondence with the theory, the second approach is internal correlations, the third one is factor analysis, and the last one is test bias or assessing the impact of gender, field of study, age, background knowledge, etc, among which the last approach is the concern of the present study.

1.4. Test Fairness

In the last two decades, the issue of test fairness and test bias has gained momentum and has been extensively investigated. One of the important considerations in the selection and use of any test is that it must not be biased. If we want to use the results of tests and measures to make decisions, then, we have to conduct research to ensure that our measure is not biased. That is, we need to have organizationally and socially relevant comparison groups, for instance, in terms of gender, age, minority status, race and so forth (Zumbo, 1999).

As for the definition of a fair test one can refer to Roever (2005) who defines a fair test as the one which is valid for all groups and individuals providing each person with an equal opportunity of demonstrating his/her skills and knowledge relevant to the purpose of the test. In other words, test takers with similar knowledge of material on a test (based on their total scores) must logically perform similarly on individual examination items irrespective of their gender, culture, ethnicity, or race, otherwise it is biased (Subkoviak, Mack, Ironson, & Craig, 1984; as cited in Perrone, 2006). Also, according to Brown (2005), fairness is defined as the degree of impartiality of tests and treating every student the same which leads teachers and testers "to find test questions, administration procedures, scoring methods, and reporting policies that optimize the chances that each student will receive equal and fair treatment" (p. 26). Bias also refers to any factor within a test that systematically prevents valid estimates or interpretation of candidates' ability (Mousavi, 2009). Bias can lead to systematic errors distorting the inferences made in selection and classification (Park, 2006).

Specifically, item bias occurs when test takers of one group are less likely to answer an item correctly than examinees of another group because of some characteristics of test item or testing situation that is not relevant to test purpose. According to Teresi (2004) “item bias implies that a sustentative review has been undertaken, and that the cumulative body of evidence suggests that the item performs differently, may have different meaning or may be measuring an unwanted nuisance factor for one group as contrasted with another” (p. 3). Differential Item Functioning (DIF) has been largely used in research as a new standard in psychometric bias analysis (Zumbo, 1999).

Various aspects of fairness including fairness with respect to standardization, test consequences/score use, and item bias (Shohamy, 2000) have been the focus of attention in the literature; however, as Roever (2005) maintains Differential Item Functioning (DIF) developed by the Educational Testing Service (ETS) in 1986, has been known as the standard of psychometric bias analysis. Accordingly, Differential Item Functioning (DIF) which may reflect measurement bias has received a great deal of attention in educational measurement (Millsap & Everson, 1993; as cited in Noortgate & Boeck, 2005).

1.5. Differential Item Functioning (DIF)

To facilitate systematic investigation of potential sources of DIF analytic techniques are required (O’Neill & McPeek, 1993). It is beneficial to use statistical techniques (i.e., DIF procedures) in order to investigate potential bias, especially in high-stakes tests (Park, 2006).

Differential Item Functioning (DIF), as Schumacker (2005; as cited in Perrone, 2006), explains is a collection of statistical methods used to determine the fairness and appropriateness of examination items with regard to different groups (e.g., male and female, etc) of test takers, hence aiding in the identification of biased test items. DIF by investigating performances of groups of interest-after they are matched on some criterion like gender-focuses on the issue of differential validity across groups (Dorans & Holland, 1993). According to O'Neill and McPeek (1993) "the fundamental principle of DIF is simple: Examinees who know the same amount about a topic should perform equally well on an item testing that topic regardless of their sex, race, or ethnicity" (p. 256).

DIF procedures are in fact a response to the legal and ethical need to ascertain that comparable test applicants are treated equally (Jodin & Gierl, 1999). There have been several definitions of DIF in the literature. Teresi (2004), however, broadly defined DIF as "conditional probabilities or conditional expected item scores that vary across groups" (p. 2).

However, DIF is a required but not a sufficient tool for detecting item bias. In other words, an item might show DIF but the difference of performance on a test and responding to the item might be due to the fact that one group of test-takers is at a higher level of ability and the other group in a lower level of ability the item must not be considered biased because this difference in the performance of groups of examinees is not indicative of test bias, but of item impact (Roever, 2005).

1.6. Uniform and Non-uniform DIF

As for the various types of DIF, one can classify it according to different factors. For example, there are, as French and Miller (1996) state, two possible types of DIF: (a) uniform (i.e., occurring when an item is uniformly favored by one group over another along the ability continuum) and (b) non-uniform (i.e., when there is an interaction between test-takers' ability level and their performance on an item contributing to change in the direction of DIF along the ability scale). Concerning group type, they explain that, there are again two distinct type of groups: focal and reference group, with the first one being of primary interest in DIF analysis and the second being taken as the standard.

1.7. Significance and Purpose of the Study

High stakes tests are administered in Iran on a yearly basis. Often some of traditional analyses are conducted at the neglect of sources of bias (i.e., gender, field of study, nationality, age, etc). Among these high stakes tests one can refer to English proficiency tests of universities in Iran administered as an entrance exam for those candidates aiming at pursuing PhD programs. Though some parts of universally-used tests are included in these high stakes tests, there seem to be some traces of sources of bias in these tests, namely difference in gender. Therefore, it is deemed necessary to scrutinize these tests to locate the source(s) of bias and eliminate them in order for the tests to achieve more validity.

Having a highly valid set of scores, as explained above, is of great importance in that the ultimate goal of any evaluation is to have accurate interpretation of scores. Thus, identification of factors which may jeopardize validity is a very important step in reaching a valid interpretation. Test bias, one of the factors affecting test validity, can be diagnosed by performing DIF procedures. To investigate whether or not English proficiency test of Tehran university (i.e., the UTEPT), as a high stakes test, exhibit DIF in terms of gender, as one of the sources of test bias, can be of great help in having a more accurate interpretation of the participants' performance on that test. The results of the study can be beneficial in revising or writing similar tests.

Therefore, individual test takers' gender may be a cause of test bias, which can, in turn, result in misinterpretation of test scores. The purpose of the present study is to investigate whether the English proficiency test of Tehran

University exhibits DIF among test takers in terms of gender. For example, if it is found out that some items exhibit DIF, care must be exercised to scrutinize the items in light of the degree of DIF based on the various criteria existing in the literature so far (e.g., Cohen, 1988; Zumbo, 1999; Jodin & Gierl, 1999).

1.8. The Research Question

With regard to the very nature of the study, the following research question is put forward:

- Do the items in the reading comprehension section of University of Tehran English Proficiency Test (the UTEPT) exhibit DIF with regard to the gender of the participants?

2. Review of the Related Literature

2.1. Different DIF Detection Methods and Techniques

There is not any single "best method" of DIF analysis which is effective and useful for all purposes (Anastasi & Urbina, 1997; as cited in Lai, Teresi, & Gershon, 2005). However, according to McNamara and Roever (2006), the following four broad categories of methods are used for detecting DIF: (a) analysis based on item difficulty (comparing item difficulty estimates); (b) nonparametric approaches (procedures using contingency tables, Chi-square, and odd ratios); (c) item-response-theory-based approaches (approaches including one, two, and three-parameter analyses which frequently compare the fit of statistical models); and (d) other approaches (including logistic regression, generalizability theory, and multifaceted measurement). Teresi (2004), in addition, classifies different DIF detection methods according to whether they "(a) are parametric or non-parametric; (b) are based on latent or observed variables; (c) treat the disability dimension as continuous; (d) can model multiple traits; (e) can detect both uniform and non-uniform DIF; (f) can examine polytomous responses; (g) can include covariates in the model, and they (h) must use a categorical studied (group variable)" (p. 5). Following techniques and methods have been used so far to detect DIF.

2.1.1. Item Response Theory (IRT)

Item Response Theory (IRT) methods are, according to French and Miller (1996), theoretically preferred as to detect DIF in dichotomous items and recent research has examined these methods to investigate whether they are also useful for polytomous cases as well (cf. Swaminathan & Rogers, 1990). IRT models are an interesting and useful tool to understand and model DIF, though the most popular techniques to detect DIF are not IRT-based (Lord, 1980; Thissen, Steinberg, & Wainer, 1993; cited in Noortgate & Boeck, 2005). IRT methods are generally constrained by sample size requirements, model fit assumptions, and software to calibrate the items and all of these problems are aggravated in the polytomous case. The shortcomings of the IRT-based procedures are that they are sensitive to sample size and model-data fit, are time consuming and that indexes as the area between item characteristic curves have no associated tests of significance.

As Noortgate and Boeck (2005) explain "in IRT models, the probability of a correct response is related to person and item covariates. These covariates often are person and item indicators (dummy covariates), weighted with parameters that are called ability and difficulty, respectively" (p. 443). In IRT model, DIF occurs "when a test tem does not have the same relationship to a latent variable across two or more examinee groups" (Embreston & Reise, 2000, p. 251; cited in Lai, Teresi, & Gershon, 2005).

2.1.2. Mantel-Haenszel (MH)

Another widely accepted and probably once the most popular statistic in use for dichotomous DIF detection is the Mantel-Haenszel especially when our sample size is small (Holland & Thayer, 1986; cited in French & Miller, 1996). Swaminathan and Rogers (1990) state that, the Mantel-Haenszel (MH) procedure is particularly attracting in terms of implementation and having an associated test of significance. Nevertheless, MH statistic is sensitive to the direction of DIF, meaning that if in the middle of the matching score distribution the direction of DIF changes, non-uniform DIF may not be detected (Swaminathan & Rogers, 1990).

2.1.3. Logistic Regression (LR)

Logistic Regression (LR), one of the DIF detection techniques, according to Zumbo (1999), "is based on statistical modeling of the probability of responding correctly to an item by group membership and a conditioning variable which is usually the scale or sub-scale total score" (p. 22). As Monahan, McHorney, Stump, and Perkins (2007) state binary Logistic Regression (LR) procedure has become increasingly popular for detecting DIF in dichotomous test items ever since Swaminathan and Rogers (1990) used it for this purpose (i.e., the detection of DIF in dichotomous test items). Logistic Regression is a useful technique for detecting both kinds of DIF, uniform and non-uniform DIF, in dichotomously scored items (Swaminathan & Rogers, 1990). Logistic Regression approaches (LR), in a predictive context, use regression of the external criteria on test score (Lai et al., 2005).

There are a number of advantages attributable to the LR technique. For example, French and Miller (1996) point out that, "the logistic regression technique is attractive because it can model both uniform and non-uniform DIF within the same equation and can test coefficients for significant uniform and non-uniform DIF separately. Specifically, this procedure models the probability of observing each dichotomous item response as a function of two explanatory variables: observed test score and a group indicator variable." (p. 317). Lee, Breland, and Muraki (2002; as cited in Park, 2006) point out that two advantages of logistic regression over linear regression are that, firstly, the dependant variable does not have to be continuous, unbounded, and measured on an interval or ratio scale; and secondly, it does not require a linear relationship between the dependant and independent variables.

2.2. Review of Studies Using Different DIF Techniques and Methods in Language Assessment

Differential-groups studies as Brown (2005) points out are those studies comparing the performances of two groups on a test aiming at demonstrating that the test scores differentiate between groups with one group having the construct being measured and the other group not lacking it. Few gender-related DIF studies have been done for tests which are developed for English as a Foreign Language (EFL) learners the majority of which have utilized U.S samples; therefore, whether or not DIF findings may be generalized across nationalities is not clear (Tae, 2004).

In addition, Tae (2004) examined the effect of gender on English reading comprehension subtest (i.e., 38 items) of the 1998 Korean National Entrance Exam for Colleges and Universities for Korean EFL learners using a DIF methodology. The results of this study indicated that those items which are classified as Mood/Impression/Tone favored females, whereas those classified as Logical Inference tended to be easier for males irrespective of item content.

Content analysis revealed that passage content is not a reliable factor that can predict interaction between examinees' gender and their performance in reading comprehension.

As an instance, French and Miller (1996) conducted a computer simulation study to determine whether it is feasible to use logistic regression procedures to detect DIF in polytomous items. They found that this technique is useful and powerful to detect most forms of DIF; however, large amount of data manipulation was required and this, sometimes, makes interpretation of the results difficult. In another study, Jodoin and Gierl (1999) focused on the Logistic Regression (LR) procedure for DIF detection a model-based approach designed to identify both uniform and non-uniform DIF. They conclude that an inclusive view of the variable associated with statistical inferences is required in DIF.

Using two large data sets, Monahan, McHorney, Stump, and Perkins (2007) present the equations for obtaining useful effect sizes for the logistic regression procedure, explain them and demonstrate their application for uniform DIF.

Other DIF studies include Geranpayeh and Kunnan (2007) who employed DIF procedure in terms of age to investigate whether the test items on the listening section of the Certificate in Advanced English examination function differently for test takers in different age groups. DIF analysis in this study identified six items exhibiting DIF. Pae (2004) did a research to investigate DIF on the English subset of the 1998 Korean National Entrance Exam for examinees with different academic backgrounds (humanities Vs science) using Item Response Theory (IRT).

Kim (2001) also investigated DIF across two different broad language groupings, Asian and European, in a speaking test. Logistic Regression (LR) and likelihood ratio procedure were used for DIF analysis. The results showed that, 'grammar' and 'pronunciation' functioned differentially across the two groups. Moreover, the content analysis of the study shows that the type and number of scoring scales may influence test validity. Mellenberg (1982; cited in Swaminathan & Rogers, 1990) used the log-linear model in order to predict item responses from group membership, ability level and the interaction between the two so that the presence of non-uniform DIF is indicated by a nonzero interaction term.

In addition, Park (2006), used a three-step logistic regression procedure for ordinal items to investigate DIF of ten writing prompts from the writing subtest of Michigan English Language Assessment Battery (MELAB) and found that the effect sizes were far too small for few prompts (i.e., those which were initially flagged due to statistically significant uniform/or non-uniform group effects) to be classified as having an important group effect.

Scherman and Goldstein (2008) investigated the relationship between race-based DIF and item difficulty and found a substantial correlation between item difficulty and DIF using different DIF techniques and a different source of data. The results of their study indicated that there was a small correlation between item difficulty and DIF values.

3. Methodology

3.1. Participants

The present study was conducted with 3,398 participants, both males and females, who took an English language proficiency test as a prerequisite for entering PhD programs at the University of Tehran (See section 3.2.).. The participants were selected out of a pool of 8,964 test takers. The participants had different fields of study, both humanities and non-humanities (e.g., philosophy, management, physics, chemistry, etc.).

3.2. Instrumentation

University of Tehran administers an English language proficiency test on a yearly basis known as the UTEPT, University of Tehran English Proficiency Test, as a partial requirement for those who intend to enter PhD programs at this university. The UTEPT is a 100-item test consisting of three sections of grammar, vocabulary and reading comprehension. The grammar section includes 35 items with the first 20 items being multiple-choice completion items and the second 15 items being error identification type among which 10 items (items 36 to 45) deal with grammar and vocabulary tested in context. 10 items also test grammar and vocabulary in context. The next section dealing with vocabulary is divided into two parts: part one having 10 items (items 46 to 55) and part two having 10 items (items 56 to 65) as well. The last section of the test also consists of thirty five items of reading comprehension tested in six passages. Table 1 illustrates the different sections of the test.

Table 1. Different Sections and Sub-Sections of The UTEPT

Section	Grammar	Vocabulary	Reading
Number of total items	45 (item 1 up to 45)	20	35
Sub-category item numbers	1 to 20: multiple-choice completion items of structure. 26 to 35: error identification of written expression items. 36 to 45: grammar and vocabulary in context.	46 to 55: vocabulary in sentence. 56 to 65: fill in the blanks type items.	66 to 100: 6 passages with 35 comprehension items.

3.3. Procedure

Though regarded as only a partial requirement for entering PhD programs, candidates from various fields of study involving both males and females, annually sit for the University of Tehran English Proficiency test (i.e., the UTEPT) to show his/her mastery of English language.

3.4. Data Collection

The UTEPT was administered by the faculty of foreign languages and literature. The data was obtained from the Information Center of University of Tehran with the permission to deal with and conduct this study. The data was fed into computer by university staff in the form of Excel. It was then converted into SPSS version.

3.5. The Design

This project is an ex post facto design. According to Hatch and Farhady (1982) ex post facto designs are mostly used when there is no selection and manipulation of the independent variables; hence, in such a design researchers are concerned with the type and/or degree of relationship between dependent and independent variables rather than their relationship in terms of cause-and-effect. In addition, correlational designs where there are two sets of data on two different variables are the most common subset of ex post facto designs. Also, a design that compares two groups of subjects on one measure is another type of ex post facto design called criterion group design. Hatch and Lazaraton (1997), moreover, point out that ex post facto designs are feasible to use in cases that there is not possibility of conducting true experimental designs (i.e., having random selection and assignment, controlling preexisting differences, and using control groups). They also assert that ex post facto design is the most commonly used design in applied linguistics since it discovers “what is going on” rather than “what caused this” and investigating “what is going on” is one of the first step in planning for instructional innovation. Hence, it is so much applicable to many projects which do not involve any kind of treatment.

3.6. Data Analysis

Logistic Regression (LR) is the method used in the present study due to the very nature of this research and several advantages attributable to LR as mentioned above.

Concerning this analysis, as Zumbo (1999) states, two most commonly used scoring formats for tests and measures are: (a) binary scores (also known as dichotomous item responses) and, (b) ordinal item responses (also referred to as graded response, likert, likert type or polytomous). He also notes that the question format is not important here but it is the scoring format which is highly important. The items scored in a binary format include items scored as correct/incorrect in aptitude or achievement test as well as item dichotomously scored according to a scoring key in a personality scale (as true/false questions). Items scored based on ordinal scale might include likert type scales like a 5-point strongly agree to strongly disagree scale on personality or attitude measures. The study at hand used ordinal scoring format.

According to Holland and Wainer (1993; as cited in Monahan et al., 2007), in DIF analyses after adjusting groups for overall performance with regard to the measured trait, they are compared on item performance. In other words, in assessing the test-takers’ response patterns to specific test items, or doing DIF, the comparison groups (e.g., males vs. females) are initially matched on the underlying construct of interest (e.g., verbal ability or mathematics achievement). This helps researchers or test developers determine whether item responses are equally valid for distinct groups of test takers (Zumbo, 1999).

Testing for statistical significance of DIF follows a natural hierarchy of entering variables into the model which, as Zumbo (1999) mentions, are:

Step # 1: Entering the conditioning variable (i.e., the total score),

Step # 2: Entering the group variable, and finally

Step #3: Entering the interaction term into the equation.

Having this information and the Chi-square test for logistic regression helps one compute the statistical tests for DIF. In other words, one must firstly obtain the Chi-square value for Step #3 and then subtract from it the Chi-square value for Step #1. The produced Chi-square value with 2 degrees of freedom (i.e., the model Chi-square statistic as step #3 is three and the model Chi-square statistic at step # 1 is one, therefore, difference will be 2 degrees of freedom) can be compared to its distribution function. Finally, the resultant two-degree of freedom Chi-square test is a simultaneous test of both uniform and non-uniform DIF (Swaminathan & Rogers, 1990).

This study also used this three-step model, as the main method of analysis, to conduct the DIF analysis. That is, logistic regression analysis was conducted in the following three steps: step 1, entering the matching or the conditional variable only (i.e., the total score); step 2: the group membership variable is entered into the regression equation; and step 3 (i.e., the full model), the interaction term (i.e., English language ability-by-group) is finally added to the regression equation.

In order to mark the amount of the group difference, *p*-values for the Chi-square test were used in addition to R^2 effect size estimates, which according to Zumbo (1999), provides information about the practical significance of DIF to interpret the results. However, there is a lack of consensus regarding what constitutes small or negligible, moderate or medium, or large effect sizes (Park, 2006). For instance, Cohen (1988) states that R^2 effect sizes of 0.02, 0.13, and 0.26 are to be considered as "small", "medium", and "large" effect sizes. Zumbo (1999), moreover, suggests that, the 2-degree of freedom Chi-square test between steps 1 and 3 have to have a *p*-value less than or equal to 0.01 for an item to be classified as displaying DIF, and the R^2 difference between them should be at least 0.13. In addition, Jodin and Gierl (1999) propose that R^2 differences of 0.035, 0.035 to 0.070, and greater that 0.070 be considered as "negligible", "moderate", and "large" effects, respectively.

In a similar vein, in terms of sample size, the literature gives credence to the fact that for binary items at least 200 people per group is adequate; however, more people in each group with no missing data yield better results (Zumbo, 1999).

One of the most important challenges in conducting DIF is to find an appropriate variable to be used to match test takers of different groups on their overall ability. In the DIF procedure, this overall matching must be done before between-group comparisons can be accomplished for individual items. In the case of standardized multiple-choice measures, the tests' total score will serve this function (Lee, Breland, & Muraki, 2002, as cited in Park, 2006). In this study, accordingly, this matching variable is created by summing up the examinees' scores on three sections of the University of Tehran English Proficiency Test (i.e., the UTEPT) as their total score.

SPSS version 15 was used to conduct all the statistical analysis of the present study (e.g., descriptive statistics, frequency statistics, independent sample t-test, and logistic regression).

4. Results

4.1. Descriptive Statistics

To give a general overview of the score information of the participants (e.g., females, males, ad all participants respectively) descriptive statistics are provided below. Tables 2, 3, and 4 present overall means and standard deviations of females and males separately and for all candidates, too. As Table 2 shows, the mean of the female group is 44.57. The mean of the male group, on the other hand, as shown in Table 3 is 44.65. Therefore, it can be concluded that these two groups have performed similarly in terms of their mean differences (i.e., mean difference is 0.07 only), hence they are comparable and DIF analysis can be done for them.

Table 2. Descriptive Statistics for Female Candidates

	N	Range	Minimum	Maximum	Mean		Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Total score Valid N (list wise)	2655 2655	67.00	14.00	81.00	44.5778	.21750	11.20696	125.596	.233	.048	-.334	.095

Table 3. Descriptive Statistics for Male Candidates

	N	Range	Minimum	Maximum	Mean		Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Total score Valid N (list wise)	2513 2513	67.00	16.00	83.00	44.6566	.22584	11.32132	128.172	.290	.049	-.176	.098

Table 4. Descriptive Statistics for all Candidates

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
Gender	6172	1.00	.00	1.00	.6416	.47957	.230
Reading total	5882	25.00	2.00	27.00	12.6284	4.03318	16.267
Total score	5168	69.00	14.00	83.00	44.6161	11.26169	126.826
Valid N (list wise)	3670						

4.2. T-test

An independent sample t-test was also performed, as a preliminary analysis, to compare the means of the comparison groups (e.g., males and females). The results are shown in Tables 5 and 6. According to table 6, while there is a difference in the mean of the two groups (i.e., 0.07) this difference is not statistically significant (p=0.802, t=0.251, df=5199).

Table 5. Group Statistics

Group	N	Mean	Std. Deviation	Std. Error Mean
Total score 1.00	2655	44.5778	11.20696	.21750
2.00	2513	44.6566	11.32132	.22584

Table 6. Independent Samples Test Results

	Levene’s Test for Equality of Variances		t-test for Equality of Means						
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower
Total score Equal variances assumed	.011	.917	-.251	5166	.802	-.07881	.31346	-.69331	.53570
Equal variances not assumed			-.251	5144.179	.802	-.07881	.31354	-.69349	.53587

4.3. Logistic Regression DIF

As described in the method section, Zumbo’s (1999) three-step modeling process was used as the main method of analysis in the current study. Table 7 summarizes the results of the logistic regression DIF analysis for the two groups in terms of gender.

Table 7. The Reading Section DIF Analysis

Item numbers: Reading comprehension	R-squared values at each step in the sequential hierarchical regression			DIF χ2 (2) Test	DIF R-squared
	Step#1 Total score in the model	Step #2 Total score, and uniform DIF variable in the model	Step ≠3 Total score, uniform, and non-uniform DIF variables in the model		
Item 66	.283	.000	.279	387.356 P=.0000	.004
Item 67	.282	.000	.280	303.408 P=.0000	.002
Item 68	.200	.000	.208	105.738 P=.0000	.008
Item 69	.138	.000	.132	175.921 P=.0000	.006
Item 70	.322	.002	.340	378.229 P=.0000	.018
Item 71	.156	.000	.157	173.183 P=.0000	.001
Item 72	.238	.001	.240	268.12 P=.0000	.002
Item 73	.172	.001	.172	201.816 P=.0000	0
Item 74	.168	.001	.162	222.886 P=.0000	.0006
Item 75	.000	.000	.000	.477 P=.639	0
Item 76	.031	.001	.031	22.126 P=.000	0
Item 77	.007	.000	.011	1.607 P=.000	.004
Item 78	.147	.000	.164	121.308 P=.000	.017
Item 79	.037	.000	.040	31.261 P=.000	.003
Item 79	.037	.000	.040	31.261 P=.000	.003
Item 80	.058	.000	.062	63.342 P=.000	.004
Item 81	.013	.000	.015	6.984 P=.000	.002
Item 82	.019	.000	.014	32.576 P=.000	.005
Item 83	.128	.000	.148	82.396 P=.000	.02
Item 84	.001	.000	.001	2.216 P=.563	0
Item 85	.072	.000	.064	79.704 P=.000	0
Item 86	.170	.007	.187	150.401 P=.0000	.017
Item 87	.001	.000	.001	.709 P=.332	0
Item 88	.005	.002	.012	8.419 P=.000	.007
Item 89	.087	.000	.086	97.376 P=.000	.001
Item 90	.215	.000	.209	276.335 P=.000	.006
Item 91	.132	.000	.135	118.169 P=.000	.003
Item 92	.141	.001	.142	155.144 P=.000	.001
Item 93	.015	.000	.015	19.532 P=.000	0
Item 94	.172	.002	.179	181.149 P=.000	.007
Item 95	.083	.001	.090	76.964 P=.000	.007
Item 96	.01	.000	.001	1.649 P=.164	.009
Item 97	.022	.000	.025	15.79 P=.000	.003
Item 98	.002	.001	.004	2.393 P=.003	.002
Item 99	.003	.000	.005	.992 P=.009	.003
Item 100	.015	.001	.018	6.048 P=.000	.003

For each single item of the reading comprehension subtest of the UTEPT, R^2 effect sizes were computed to be checked against the three sets of classifications mentioned earlier. Table 8 shows the final results of the DIF analysis on the subtest in question according to the Cohen's (1988), Zumbo's (1999), and Jodin and Gierl's (2001) criteria.

Table 8. The Results of the Reading Section DIF Analysis

Item Numbers: Reading Comprehension questions.	**R-squared effect sizes' criteria**			**DIF R-squared**
	Cohen (1988)[a]	**Zumbo (1999)[b]**	**Jodin and Gierl (2001)[c]**	
Item 66	Small	No DIF	Negligible	.004
Item 67	Small	No DIF	Negligible	.002
Item 68	Small	No DIF	Negligible	.008
Item 69	Small	No DIF	Negligible	.006
Item 70	Small	No DIF	Negligible	.018
Item 71	Small	No DIF	Negligible	.001
Item 72	Small	No DIF	Negligible	.002
Item 73	Small	No DIF	Negligible	0
Item 74	Small	No DIF	Negligible	.0006
Item 75	Small	No DIF	Negligible	0
Item 76	Small	No DIF	Negligible	0
Item 77	Small	No DIF	Negligible	.004
Item 78	Small	No DIF	Negligible	.017
Item 79	Small	No DIF	Negligible	.003
Item 80	Small	No DIF	Negligible	.004
Item 81	Small	No DIF	Negligible	.002
Item 82	Small	No DIF	Negligible	.005
Item 83	Small	No DIF	Negligible	.02
Item 84	Small	No DIF	Negligible	0
Item 85	Small	No DIF	Negligible	0
Item 86	small	No DIF	Negligible	.017
Item 87	Small	No DIF	Negligible	0
Item 88	Small	No DIF	Negligible	.007
Item 89	Small	No DIF	Negligible	.001
Item 90	Small	No DIF	Negligible	.006
Item 91	Small	No DIF	Negligible	.003
Item 92	Small	No DIF	Negligible	.001
Item 93	Small	No DIF	Negligible	0
Item 94	Small	No DIF	Negligible	.007
Item 95	Small	No DIF	Negligible	.007
Item 96	Small	No DIF	Negligible	.009
Item 97	Small	No DIF	Negligible	.003
Item 98	Small	No DIF	Negligible	.002
Item 99	Small	No DIF	Negligible	.003
Item 100	Small	No DIF	Negligible	.003

[a] Cohen's criteria (1988): R^2effect sizes of 0.02, 0.13, and 0.26 as "small", "medium", and "large".

[b] Zumbo's (1999) criterion: R^2 should be at least 0.13 to display DIF

[c] Jodin and Gierl (2001): R^2differences of 0.035, 0.035 to 0.070, and greater than 0.070 are considered as "negligible", "moderate", and "large" effects.

As Table 8 above indicates, all of the items (i.e., 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, and 100) are considered to have "small" and "negligible" R^2 effect sizes according to Cohen (1988) and Jodin and Gierl (2001), hence these items do not display DIF according to Zumbo (1999) as well.

5. Discussion and Conclusion

As mentioned above in the introduction part, the main purpose of this study was to investigate DIF across gender groups in the reading comprehension subtest of the English Language Proficiency of Tehran University (i.e., the UTEPT). In so doing, the three-step modeling process based on logistic regression (Zumbo, 1999) was employed which is, as explained before, efficient in investigating both uniform and non-uniform group effects, simultaneously. The statistical significant tests of DIF with a two-degree-of-freedom Chi-square test were then supplemented with corresponding effect sizes via R-squared. These R^2 effect sizes were finally scrutinized to determine whether the 35 items of the reading comprehension test exhibit DIF or not using Cohen's (1988), Zumbo's (1999), and Jodin and Gierl's (2001) sets of criteria.

Contrary to the study of Tae (2004) whose results indicated gender DIF in reading comprehension of the Korean National Entrance Exam for Colleges and Universities, the results of the study showed that, on the whole, none of the 35 items of reading comprehension subtest displayed DIF taking into account their uniform and non-uniform group effects since their effect sizes were either too small or negligible to be classified as DIF-exhibiting reading items. That is to say, items of this section of the UTEPT do not favor any particular groups of examinees regarding their gender; hence, it can be concluded that this section can be considered fair to all male and female test takers.

There are some plausible explanations as to why such results were arrived at in the present study. For instance, that the items of reading comprehension have not exhibited DIF might be pertinent to some characteristics of test takers such as their being non-native speakers of English language, their language background and level of exposure to English language which seems to be nearly the same. That test maker(s) carefully and flawlessly wrote the reading items is also another justification as to why the items did not show DIF.

Roznowski and Reith (1999; as cited in Takala & Kaftandjieva, 2000) in their study show that if a test contains DIF items it is not therefore inevitably biased. Because, DIF is a necessary condition for item (and test) bias however it is not sufficient. One of the implications of their finding is that doing single DIF analysis cannot by itself ensure that a test is or is not biased. Therefore, DIF analysis should not be limited to the item level; rather it should further investigate how DIF items affect the total scores (Bolt & Stout, 1996; as cited in Takala & Kaftandjieva, 2000). Complementing the results with the other methods as IRT can also help substantiate the truth of the findings. So, another possible explanation for obtaining the results of this study is that maybe the 35 reading comprehension items are biased but DIF procedure was not able to show this due to the fact that the results were not supplemented with any other analysis.

It is entirely possible that there are genuine differences between males and females in the test taking process but these differences occur in their performances in the reading comprehension items. Perhaps other construct irrelevant factors such as field of study may be better candidates for DIF studies. But that does not necessarily mean that DIF studies in terms of gender should be abandoned in favor of other factors. Furthermore, differential bundle functioning should also be taken into account. On an item by item basis, the test did not exhibit differential functioning. Items need to be investigated in light of combinatory analyses.

Moreover, the current study makes a contribution to the DIF literature providing a good deal of information about DIF involving Iranian test takers taking into consideration the fact that, to date, very few DIF studies have utilized Iranian sample. The results of the current research bring about several potential implications for educational centers and different organizations that make decisions regarding the future academic and occupational lives of individuals and society at large. As an instance, findings of the present study indicate that substantial effort has been devoted to making the reading section of the UTEPT due to the fact that it does not have any adverse impact on gender groups. That is to say, it is proved by this study that the reading items of the UTEPT are totally fair to all candidates whether they are male or female.

(De) limitations

The present study is subject to some (de) limitations like any other research in general and applied linguistics research in particular. They are as follow:

1- The first one is that, the participants of the study came from various mother tongues (e.g., Persian, Turkish, etc) which could have affected their performance in the test under investigation due to some linguistic matters. It would have been better if it had been feasible to control this issue.

2- It would have been ideal if the results of the study had been complemented with other statistical analyses such as factor analysis and multi-trait multi-method (MTMM) designs. This would have shed more light on the findings of the current study.

3- Moreover, the study was conducted using a single set of data; the data collected from only University of Tehran. Though this set of data is not narrow in scope, other data sets, if added, might have yielded different results which would in turn have enhanced the generalizability of the results.

4- The present study investigated DIF only in the reading comprehension subtest of the UTEPT in terms of gender. The other two subtests of the test (e.g., grammar and vocabulary) could have been examined as well in terms of gender.

Suggestions for Further Research

Future researchers can benefit immensely from the following avenues of research:

1- In order to improve the accuracy of parameter estimates, according to scholars such as Fischer & Formann, 1982; Mislevy 1988; Mislevy et al., 1993; cited in Swanson et al., 2002, IRT models could be used which can generate information about hidden relationships between items. IRT software (e.g., BILOG/MG) might also be used to estimate DIF effect sizes for individual items along with associated standard errors and covariances among parameter estimates. Hierarchical modeling software could then be used to look for regularities in the DIF estimates which are systematically pertinent to item characteristics. The same approach might also be used to look for relationships between item characteristics and other effect size indices of DIF (e.g., M-H indices, SIBTEST) to improve the accuracy of the estimates of effect size (Swanson, et al., 2002).

2- It is also suggested that one investigate DIF of the test in question (i.e., the UTEPT) taking into account fields of study as a polythomous variable and investigate whether the test favors any specific groups of examinees with regard to their field of study or not (e.g., mathematics, management, chemistry, philosophy, etc).

3- There are still a host of other standardized tests administered internationally (e.g., TOEFL, IELTS, etc) and those that are nationally administered (e.g., TOLIMO) that can be the subject of similar investigations. They are wider in scope and if flawlessly executed can yield more generalizable results.

4- More specifically, Takala and Kaftandjeva (2000) point out that two-way interaction between gender difference and DIF, as explained earlier, is more questionable in the context of L2 vocabulary testing context due mainly to two reasons: firstly, that it has not been widely investigated, and secondly few available research studies report that there is no gender DIF and the items with significant gender DIF are not discussed from the point of view of content analysis. That is why it is also suggested that one examine gender difference and L2 vocabulary knowledge in the context of a high-stakes test.

References

Alderson, C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. New York: Cambridge University Press.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Reviews of Psychology, 37,* 1-15.

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & Braun, H. (Eds.) *Test validity* (pp. 19-32). Hillsdale, NJ: Erbaum.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Brown, H. D. (2004). *Language assessment: Principles and classroom practices.* London: Longman.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment.* New York: McGraw-Hill.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd Ed.). Hillsdale, NJ: Lawerence Erlbaum Associates, Inc.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds), *Differential item functioning* (pp. 137-166). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

French, A. A., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement, 33* (3), 315-332.

Geranpayeh, A., & Kunnan, A. J. (2007). Differential item functioning in terms of age in the certificate in advanced English examination.
Language Assessment Quarterly, 4 (2), 190-222.

Hatch, E., & Farhady, H. (1982). *Research design and statistics for applied linguistics.* Rowley, Massachusetts: Newbury House.

Hatch, E., & Lazaraton, A. (1997). *The research manual: Design and statistics for applied linguistics.* Boston, MA: Heinle &Heinle Publishers.

Jodin, M. G., & Gierl, M. J. (1999). Evaluating type I error and power using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14,* 329-349.

Kim, M. (2001). Detecting DIF across the different language groups in a speaking test. *Language Testing, 18,* 89-114.

Lai, J. S., Teresi, J., & Gershon, R. (2005). Procedures for the analysis of differential item functioning (DIF) for small sample sizes. *Evaluation & the Health Professions, 28* (3), 283-294.

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension.* New York: Blackwell publishing.

Monahan, P. O., McHorney, C. A., Stump, T. E., & Perkins, A. J. (2007).
Odds ratio, delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression. *Journal of Educational and Behavioral Statistics, 32* (1), 92-109.

Mousavi, S. A. (2009). *An encyclopedic dictionary of language testing.* Tehran: Rahnamma Press.

Noortgate, W. V. D., & Boeck, P. D. (2005). Assessing and examining differential item functioning using logistic mixed models. *Journal of Educational and Behavioral Statistics, 30* (40), 443-464.

O'Neill, K. A., & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland & H. Wainer (Eds), *Differential item functioning* (pp. 255-267). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Pae, T. (2004). Gender effect on reading comprehension with Korean EFL learners. *System, 32,* 265-281.

Park, T. (2006). Detecting DIF across different language and gender groups in the MELAB essay test using the logistic regression method. *Spaan Fellow Working Papers in Second or Foreign Language Assessment,* 4, 81-96.

Perrone, M. (2006). Differential item functioning and item bias: Critical considerations in test fairness. *Columbia University Working Papers in TESOL & Applied Linguistics, 6* (2), 1-3.

Rezaee, A., & Salehi, M. (2008). The construct validity of a language proficiency test: A multitrait multimethod approach. *TELL, 2* (8), 93-110.

Roever, C. (2001). Web-based language testing. *Language and Learning and Technology,* 5 (2), 84-94.

Roever, C. (2005). *"That's not fair!" Fairness, bias, and differential item functioning in language testing.* Retrieved November 18, 2006, from the University of Hawai'i System Web site: http://www2.hawaii.edu/~roever /brownbag.pdf

Salehi, M., & Rezaee, A. (2009). On the factor structure of the grammar section of university of Tehran English Proficiency Test (the UTEPT). *Indian Journal of Applied Linguistics, 35* (2), 169-187.

Scherman, C. A., & Goldstein, H. W. (2008). Examining the relationship between race-based Differential Item Functioning and Item Difficulty. *Educational and Psychological Measurement, 68*, 537-553.

Shoahmy, E. (2000). Fairness in language testing. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (pp. 15-19). Cambridge, UK: Cambridge University Press.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27* (4), 361-370.

Swanson, D. B., Clauser, B. E., Case, S. M., Nungester, R. J., Featherman, C. (2002). Analysis of Differential Item Functioning (DIF) Using Hierarchical Logistic Regression Models. *Journal of Educational and Behavioral Statistics, 27*(1), 53- 75.

Tae, P. (2004). Gender effect on reading comprehension with Korean EFL learners. *System*, 32, 265-281.

Takala, S., & Kaftandjieva. (2000). Teat fairness: A DIF analysis of an L2 vocabulary test. *Language Testing, 17,* 323-340.

Teresi, J. (2004). Differential item functioning and health assessment. *Columbia* University *Stroud Center and faculty of Medicine.* New York State Psychiatric Institute, Research Division, Hebrew Home for the Aged at Riverdale. 1-24.

Zumbo, B. D. (1999). *A Handbook on the theory and methods of Differential Item Functioning (DIF): Logistic regression modeling as a unitary framework for binary and likert-type (Ordinal) item scores.* Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.