ISSN 1798-4769
Journal of Language Teaching and Research, Vol. 3, No. 1, pp. 84-92, January 2012

doi:10.4304/jltr.3.1.84-92

Differential Item Functioning: Implications for Test Validation

Mohammad Salehi
Sharif University of Technology, Iran
Email: m_salehi@sharif.ir

Alireza Tayebi
Sharif University of Technology, Iran
Email: alireza_tayebi2008@yahoo.com

***Abstract*—This paper attempts to recapitulate the concept of validity, namely construct validity (i.e., its definition and its approaches and role in language testing and assessment). Validation process is then elaborated on and proved to be integral enterprise in the process of making tests, namely English language proficiency tests. Then come the related concept of test fairness and test bias and its sources (e.g., gender, field of study, age, nationality and L1, background knowledge, etc) and contributions and threads to the validity of tests in general and in high-stakes tests of English language proficiency in particular. Moreover, in the present study, different approaches to investigate the validity of tests will be reviewed. Differential Item Functioning (DIF), among the other methods to investigate the validity of tests is also explained along with the description and explanation of its different detection methods and approaches mentioning their advantages and disadvantages to conclude that logistic regression (LR) is among the best methods till now.**

***Index Terms*—validity, Differential Item Functioning, item bias, test fairness, logistic regression, IRT**

I. Introduction

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 the overall English language knowledge of a person. 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 such 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, it is the quality of interpretations made out of test scores (Bachman, 1995). To prevent such inappropriate consequences, Bachman believes, bias must be detected and removed, a complex procedure. It can be detected through various methods and procedures and present study mainly focuses on one of the most important and currently used procedures of bias detection, known as Differential Item Functioning (DIF, here after) and its different detection procedures and methods.

There is no single, absolute measure of validity to establish validity; however, various kinds of evidence can help to support it (Brown, 2004) so different facets of validity manifest involving content-related, criterion-related, construct-related, consequentional, and face validity. Among these aspects, chiefly construct-related validity is scrutinized, and to some extents criterion-related and consequentional validity are also investigated concerning one of the high-stakes tests administered in Iran the description of which is provide in chapter three. Therefore, the last two types are defined before turning to the first kind, construct-validity, and dealing with it in detail. Brown (2004) defines criterion-related validity which is divided into two kinds of concurrent and predictive validity as "the extent to which the 'criterion' of the test has actually been reached" (p. 24), and regarding consequentional validity he declares "consequentional validity encompasses all the consequences of a test, including such considerations as its accuracy in measuring intended criteria, its impact on the preparation of test takers, its effect on the learner, and the (intended and unintended) social consequences of a test's interpretation and use" (p. 26).

II. Related Literature Review

A. Validity and Validation Process

For validity, one of the most complex criterion of tests and the most fundamental in psychometrics (Angoff, 1988), various definitions have been proposed yet expressing the same central idea to which results of the tests must conform in order for them to be regarded as an effective and valid test. Before relation the concept of validity and approaches to

validation process to the purpose of the study and revealing how it is related to test bias and DIF providing definition of this concept and its related aspects seems helpful. As an example, one of the traditional definitions of validity is the correlation of test scores with "some other objective measure of that which the test is used to measure: (Bingham, 1937, p. 214; cited in Angoff, 1988). As another instance, one of the conventional definitions is suggested by Gronlund (1998, p. 226; cited in Brown H. D; 2004) who says "the extent to which inferences made from assessment results are appropriate, meaningful, and useful in terms of the purpose of the assessment". Anastasi (1986, as cited in Angoff, 1988) 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. Brown (2005) too defines validity as "the degree to which a test measures what it claims, or purports, to be measuring" (p. 220) and gains especial importance when involved in making decisions about students; therefore, after taking into account issues of practicality and reliability, validity should be concerned.

Construct validity is a chief issue concerning validating 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).

Brown (2004) defines construct as "A construct is any theory, hypothesis, or model that attempts to explain observed phenomena in our universe of perceptions. Constructs may or may not be directly or empirically measured-their verification often requires inferential data. "Proficiency" and "communicative competence" are linguistic constructs; "self-esteem" and "motivation" are psychological constructs". (p. 25).

Zumbo (1999) also points out that, the traditional view of validity is expanded in this regard. Principally, the current view of validity makes validation a central issue which is not simply evaluated by computing correlations with another measure, for explicit statistical studies which examine bias and concept-focused and policy studies of value implication for decision making are needed. Therefore, in validation process, we begin with the construct definition stage followed by writing the items or selecting a measure, and we continue through item analysis while using a measure.

Research in investigating the validity, namely construct validity, is abundant in the field. For instance, Salehi and Rezaee (2008) used the design of multitrait-multimethod (MTMM) to investigate 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. As they state, this test is a high-stakes and the results of this test have a kind of life-changing implications for the test takers. It was found that the test possessed both convergent and discriminant validity. As they argue, MTMM designs have two advantages. Firstly, by using this design it is possible to examine both convergent and discriminant validity, and secondly, by making use of this design the effect of measurement method can be distinguished from the effect of the trait method.

In another study done by Rezaee and Salehi (2009) 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) and the subject of their study was grammar section. 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.

B. Current View of Validity

McNamara and Roever (2006) point out that contemporary discussions of validity takes into account such issues as test fairness by developing procedures that supports the rationality of decisions based on tests. They refer to Cronbach as "father" of construct validity a term coined by Meehl and Challaman. As they explain, in construct validation, the validity of inferences are concerned rather than the validity of instruments. They state that Cronbach (1989) argued that such a thing as a "valid test" does not exist: "one validates not a test, but an interpretation of data arising from a specified procedure" (Cronbach, 1971, p. 447; cited in McNamara and Roever, 2006). Cronbach and Meehl (1995; cited in McNamara and Roever, 2006) make a distinction between a weak and a strong program for construct validation with weak program being concerned with any available means to verify interpretation which is mostly correlational and the strong program being based on the falsification idea advanced by Popperian philosophy.

III. Test Validation

One of the most important considerations in the process of making tests is test validation which can refer to any attempt to eliminating irrelevant factors and sources of bias from any kind in order for a test to yield valid results. According to Bachman (1995), the purpose of validation is highly in line with the specific groups of test takers-as well as test users. The members of test takers may differ in some aspects, other than language ability, such as gender, age, field of study, nationality, etc. Sometimes, each of these points of difference may cause some fluctuations in the language ability level of examinees; therefore, these points of dissimilarities may be regarded as sources of bias (Zumbo, 1999). Consequently, 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. This phenomenon which is referred to as test bias may contribute to misinterpretation of test scores; sexist or racist content unequal prediction of criterion performance; unfair content with respect to the experience of test takers; inappropriate selection procedure; inadequate criterion measures; and threatening conditions of testing (Nitco 1983: 43-7; as cited in Bachman, 1995).

In terms of 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). The distinction between validations of measure versus validation of inferences has significant contributions for assessment. Besides, all empirical measures have a need for validation of inferences.

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, and 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 approach is factor analysis, and finally the last one is test bias or assessing the impact of gender, field of study, age, background knowledge, etc.

Recently validity has been dealt with under the light of new considerations. For instance, in terms of Zumbo (1999) the current view of validity makes it so central that computing it simply by correlation with another measure would not be an appropriate method. It also highlights that explicit statistical studies examining bias and concept-focused and policy studies of value implications in decision making. 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. As noted earlier, the process of validation must examine the relationship between test performance and the test itself. Moreover, the process of validation is addressed to specific uses of the test, and the specific examinees group taking the test.

Brown (2005) also refers to the change in thinking of validity in the field of testing and assessment, current conceptions of validity. He makes a distinction between traditional and current view of validity. The traditional process of validation involves picking the most suitable type of validity (i.e., content, criterion-related or construct validity) and then conducting statistical analysis. Whereas, the current view of validity expands the conceptual framework of traditional view. If one aims at using the measure for decision-making purposes, research should be conducted to ensure that there is no bias in the measure. In fact, in the current view of validity, validation is a central issue which is not resolved by computing a correlation with another measure. That is, explicit statistical studies which examine test bias are needed. Such a need is due to the fact that validation process is never entirely complete.

A. Test Fairness

As explained above, test bias removing which contributes to test fairness is an important building block in the process of test validation. In the last two decades, the issue of test fairness and test bias has become increasingly important and it has been the subject of great deal of recent research focusing on the use of any psychological and/or educational tests. Tests and measures are mostly used for the purpose of decision-making. One of the important considerations in selection and use of any test is that test must not be biased, that is test must be fair to all candidates. 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 group, for instance, in terms of gender, age, minority status, race and so forth (Zumbo, 1999).

As to the definition of a fair test one can refer to Roever (2005; as cited in Perrone 2006) who points out that a fair test is one being 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). Fairness, according to Brown (2005), in addition, is defined as the degree tests' impartiality 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 can lead to systematic errors distorting the inferences made in selection and classification. In terms of 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). In other words, test items are biased if they contain sources of difficulty which is not relevant to the construct measured, in this particular case, the performance of examinees on proficiency test. Hence, items containing sources of difficulty beyond those of interest which results in a discrimination against particular groups are regarded as bias. Recently, a technique called DIF has been largely used in researched as new standard in psychometric bias analysis (Zumbo, 1999).

Accordingly, if an examination item is biased it functions differentially for a specific subgroups of test takers depriving testees of their equal chance of success (Zumbo, 1999); because, a biased item measuring irrelevant attributes to the tested construct which impact test takers' performance (Williams, 1997, Zumbo, 1999; as cited in Perrone, 2006).

In addition, if an item contains language or content differentially difficult for different subgroups of test-takers, it is biased and it might also demonstrate item structure and format bias in the case of involving ambiguities or inadequacies in the item stem, test instructions, or distracters (Hambleton & Rodgers, 1995; cited in Perrone, 2006).

Researchers in the field of second language assessment like users of psychological tests, policy makers, and personnel selection officers should be aware of the current thinking in test bias analysis, one of the standards or techniques of which is DIF, as systematic errors distorting the inferences made of tests may appear (Zumbo, 1999). Various aspects of fairness including fairness with respect to standardization, test consequences/score use, and item bias (Kunnan, 2000; Shohamy, 2000; as cited in Perrone, 2006) have been the center of attention in the literature; however as Roever (2005; as cited in Perrone, 2006) declares DIF developed by the Educational Testing Service (ETS) in 1986, has been known as the standard of psychometric bias analysis. Accordingly, 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).

In the context of the off-hand essay test, for instance, the issue of bias gains importance especially when a prompt or topic is biased against certain group(s) of examinees resulting in distorting the meaning of the essay score for different examinees subgroups jeopardizing the validity of score interpretations (Sheppard, 1982; cited in Park, 2006). Therefore, devising essay prompts that are not biased and are fair to all examinees so that no test taker will be unfairly disadvantaged and could demonstrate their true ability. As another example, an intelligence test is biased if it contains items which assess and tap specific knowledge and abilities not intended to be measured by the test (Noortgate & Boeck, 2005).

In sum, as Hambleton and Rogers (1995; cited in Perrone, 2006) declare when a dimension on the examination is not related to the construct being measured the result of which is favoring one group of examinees and placing the other group of examinees at a disadvantage in taking the examination, then test items will be regarded as biased items. Consequently, if DIF is not evident for an item, then the item is not biased. However, DIF is required but not sufficient for declaring item bias. In other words, an item might show DIF but the difference of performance on a test and responding to the item is due to the fact that one group of test-takers is at a high 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, Schumacher, 2005; cited in Perrone 2006). Therefore, after seeing evidence for the occurrence of DIF application of subsequent item bias analysis (e.g., empirical evaluation or content analysis) would be needed in order to prove the presence of item bias as DIF can be considered bias if there are construct irrelevant factors in a test resulting in differences in a group's ability to respond to a test item (Zumbo, 1999).

B. Differential Item Functioning (DIF)

This part provides a brief explanation on perspective and foundation of DIF, a review of statistical techniques to conduct DIF as well as summaries of some highly related practical research in this regard. In order to detect item bias and remove that, several methods have been proposed and used so far.

Previously, as Subkoviak et al (1984; as cited in Perrone) points out, a variety of methods such as the transformed item difficulty method, the Chi-square method, and the three-parameter item characteristic curve had been proposed to detect item bias. However, DIF procedures are considered as the new dominant psychometric methods to address fairness in standardized, achievement, aptitude, certification and license testing. That is to say, one of the most important considerations in selection and use of any test is that it must not be biased, so it should be fair to all applicants. 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.

This technique, DIF, has been mostly used in research as a rather new standard in psychometric bias analysis. DIF procedures are in fact a response to the legal and ethical need to ascertain that comparable test applicants are treated equally (Jodin and Gierl, 1999).There have been several definitions of DIF in the literature and the exact and to the point definition of DIF, according to Teresi (2004) varies across methods and the fact that whether binary or polytomous items (usually ordinal) items are to be examined adds to this diversity; however, broadly defined DIF is "conditional probabilities or conditional expected item scores that vary across groups" (p. 2).

Zumbo (1999) notes that for the measurement practitioners, DIF often means that there exists a type of systematic but construct irrelevant variance that is being tapped by the test or measure. Moreover, the source of construct irrelevant variance is related to group membership. In other words, presence of multidimensionality as well as its pervasiveness depends on group membership. In conducting DIF and detecting test and/or item bias, as Roever (2005; cited in Perrone, 2006) declares, logically is locating examination items on which one group of test-takers performs significantly better than the other group(s).

In literature, there is a clear distinction between "item impact" and "DIF" (Clause & Mazor, 1998; Penfield & Lam, 2000; as cited in Park, 2006) so that the former may be present when examinees from different groups show different probabilities of success on an item due to their difference in the ability measured. That is to say, in such cases "true" differences between the groups in the underlying ability being measured by the item results in differences in examinee performance on the item. Zumbo (1999), in addition, elucidates the matter explaining that when examinees from

different groups endorse an item differently item impact is evident since true differences exist between the groups regarding the underlying ability measured by the item ;and, item bias occurs due to the fact that some characteristics of the test item or testing situation that is not relevant to the test purpose differentiates performance of group of examinees making one group less likely to answer the item correctly and the other group more likely do so. As he further elaborates, the difference between item impact and item bias pertains to the fact that group differences are due to relevant or irrelevant characteristics (respectively) of the test. Teresi (2004) also points out that impact implies that there are group differences and that item impact is typically examined in terms of effect sizes. DIF is necessary condition for item bias, but it is not sufficient. In other words, if an item does not show DIF, then no item bias is present. Nevertheless, in cases that DIF is apparent, subsequent item bias analyses (e.g., content analysis, empirical evaluation) are needed to provide evidence to declare item bias.

According to Holland and Wainer (1993; cited in Monahan et al, 2007), in DIF analyses after adjusting groups for overall performance with regard to measured trait, they are compared on item performance. In other words, in assessing 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). Putting it in other words, Noortgate and Boeck (2005) explain that DIF analysis are often used to substantiate the fact that in a standardized test items do not favor the reference group or a majority group (e.g., males, white people, etc) compared with one or more focal or minority groups (e.g., females, people of color, etc). This helps researchers or test developers determine whether item responses are equally valid for distinct groups of test takers (Zumbo, 1999).

As to 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 uniformly is favored by one group over another across the ability continuum) and (b) nonuiform (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). Putting it in other words, Teresi (2004) clarifies that "Uniform DIF indicates that the DIF is in the same direction across the entire spectrum of disability, while nonuniform DIF means that an item favors one group at certain disability levels, and other groups at other levels" (p. 2). 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. In item response theory terms, in addition, nonparallel item characteristic curves indicates nonuniform DIF. Moreover, nonuniform DIF is much more difficult to interpret and due to the existence of the interaction between ability level of examinees' and their group membership (Park, 2006) Also, the identification of nonuniform DIF in polytomous items may become more important than that of nonuniform DIF in dichotomous items (Spray & Miller, 1994; cited in Park, 2006). Such methods as the Generalized Mantel-Haenszel procedure and the standardization method cannot detect nonuniform DIF (Miller & Spray, 1993; as cited in Park).

C. 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. In addition, 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, proficiency test is not limited o any single course, curriculum, or skill in the language. That is to say it tests overall language ability.

The investigation of DIF is crucial in language proficiency tests, where examinees with various backgrounds are involved, since DIF items pose a considerable threat to the validity of the test (Kim, 2001). It is notable that one can study measurement bias investigating external or internal relationships. DIF, however, is a matter of internal item relationships of the items to another. Studies of DIF are one of the primary methodological devices to address standardized assessment programs whether in second/foreign language assessment programs or in the field of psychology. What follows is the review and summary of the research done by different scholars and researchers in this regard.

As an instant, 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, logistic regression, is useful and powerful to detect most forms of DIF; however, large amounts of data manipulation was required and this, sometimes, makes interpretation of the results difficult. In another study, Jodoin and Gierl (1999) focus on the logistic regression procedure for DIF detection a model-based approach designed to identify both uniform and non-uniform DIF. they have developed a new classification method based on which established simultaneous item bias test. They also examined whether the effect size measure affects type I error and power rates for the logistic regression DIF procedure. The conclude that an inclusive view of the variable associated with statistical inferences is required in DIF. besides sample size, type I error, rate power and effect sizes are interrelated and must be considered with careful attention to the inferences from a statistical test.

Using tow 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. They also discuss the pros and cons of effect sizes, as they declare that, previous research using binary logistic

regression (LR) for detecting DIF in dichotomously scored items did not report an effect size while these LR effect sizes are valuable to practitioners especially for avoiding flagging unimportant DIF in large samples. Concerning how to use the effect sizes the authors recommend that, firstly, when deciding if items show DIF effect size a statistical test be used; secondly, practitioners decide on the values of the effect size for the intended purpose (i.e., negligible, moderate and large magnitudes); third, one take some steps to facilitate interpretations (e.g., calculating the reciprocal of odds ratios less than one; using scatter, line, and bar graphs aiding in discerning relative distances between DIF magnitudes; and sorting items according to direction and magnitude of DIF in tables; and forth, comparisons of DIF procedures can be facilitated using effect sizes.

Other DIF studies include Geranpayeh and Kunnan (2007) who used DIF procedure to investigate whether the test items on the listening section on the Certificate in Advanced English examination function differently for test takers among three different age groups. DIF analysis in this study identified six items exhibiting DIF. However, the findings of the study did not clearly show item bias toward any of the age groups examined. Different academic backgrounds also can be regarded as one of the factors causing bias in any test, a source of bias. Tae-II Pae (2004) does 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. In this study, DIF was detected using both Mental-Haenzel procedure and the IRT likelihood ratio-approach.the result of the research indicates that there were 18 DIF items with 28 DIF parameters.

Scherbaum and Gold stein (2008) studied the relationship between race-based DIF and item difficulty. They examined relationship between DIF and Item Response Theory. They, actually, replicated Freedles' findings by using alternative DIF techniques. They found a substantial correlation between item difficulty and DIF using different DIF techniques and a different source of data in comparison to Freedles' (2003) research. The results of their study indicates that there was a small correlation between item difficulty and DIF values.

D. DIF as a Validation Technique in Social and Psychological Measures

The field of psychology too, as indicated above, is concerned with issues of fairness and test bias. DIF analysis is an effective tool to explore personality constructs, too (Smith, 2002; as cited in Sheppard, 2006). Test bias reveals psychometric inequalities among different subgroups and, according to Drasgow, 1984; cited in Sheppard, 2006) is divided into two forms of relationship bias (i.e., concerning with the association between an external criterion measure and a test score) and measurement bias (i.e., concerning with properties of test items). Nevertheless, studies of test bias in employment-oriented personality inventories still remain unexplored and this is in contrast to the exhaustive research on bias in ability tests (Sackett & Willk, 1994; Jensen, 1980; Thissen, Steinberg, & Gerrard, 1986; as cited in Sheppard et al, 2006). Here also differential hiring rates would be regarded unfair and biased if test bias rather than true differences causes group differences. (Sheppard, et al 2006). Measurement bias was examined investigating differential item functioning across sex and two racial groups (Caucasian and Black) in the Hogan Personality Inventory to explore the themes of the potentially biased items by Sheppard et al (2006). They found that, 81 out of 138 items in the HPI exhibited DIF among which 38 percent of the total items were shown to be potentially biased by sex and 38 percent of them exhibited DIF by race. They conclude that, considerable degree of measurement bias exists in an employment-oriented personality test.

IV. Different DIF Detection Methods and Techniques

A. General Classification of Various Methods and Approaches

There does not exist any single "best method" of DIF analysis which is effective and useful for all purposes (Anastas & Urbina, 1997; as cited in Lai, Teresi, & Gershon, 2005). Various methods exist to examine DIF to estimate the level of disability, disease, capability, etc., among which some methods assume the existence of a latent variable which is estimated by using marginal maximum likelihood, some other methods assume a “valid” target dimension distinct from secondary “nuisance” factors, and finally other methods assume the existence of an external “gold standard” diagnostic variable (King, Murray, Salomon, and Tandon, 2004; as cited in Teresi, 2004). However, according to McNamara and Roever (2006) the following four broad categories of methods are used for detecting DIF: (1) analysis based on item difficulty (comparing item difficulty estimates); (2) nonparametric approaches (procedures using contingency tables, chi-square, and odd ratios); (3) item-response-theory-based approaches (approaches including 1, 2, and 3-parameter analyses which frequently compare the fit of statistical models); and (4) 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 dtect both uniform and nonuniform DIF; (f) can examine polytomous responses; (g) can include covriates in the model, and they (h) must use a categorical studied (group variable).” (p. 5).

It is also possible to classify matching approaches where testees with the same level of ability are matched versus those that do not: in this case, DIF is surly present if test takers of the same ability level but belonging to different groups have different likelihood of correctly answering an item as item are functioning differentially between test takers not because of underlying ability but because of group membership (e.g., males vs. females).

B. Item Response Theory (IRT)

In addition, IRT methods, are according to French and Miller (1996), theoretically preferred as to detect DIF in dichotomous items and recently research has examined these methods to investigate whether they are also useful for polytomous case as well (cf. Swaminathan and Rogers, 1990; Wainer, Sireci, and Thissen, 1991). Item Response Theory 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). However, 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. Item response theory procedures (Shepard et al, 1981; Hambleton & Swaminathan, 1985; cited in Swaminathan and Rogers, 1990) in comparing the performance of groups of examinees takes it into account the continuous nature of ability. However, IRT-based procedures' shortcomings are that they are sensitive to sample size and model-data fit and 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 Item Response Theory (IRT), 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). IRT-related approaches, according to Lai et al (2005), involves comparison of the Item Characteristic Curves (ICCs), or comparison of the item parameters and unlike the MH and LR methods do not rely on observed scores. As they explain, an ICC describes the relationship between a respondent's location on a latent trait continuum and the probability of respondent's giving specific response to a particular item on that trait continuum.

C. Mantel-Haenszel (MH)

Another widely accepted and probably once the most popular statistic in use for dichotomous DIF detection is the Mantel-Haenszel (MH) technique (Mantel & Haenszel, 1995) especially when our sample size is small (Holland and Thayer, 1986; cited in French and Miller, 1996). Swaminatha and Rogers (1990) state that, the Mantel-Haenszel procedure is particularly attractive concerning its 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, nonuniform DIF may not be detected (Swaminathan and Rogers, 1990; cited in French and Miller, 1996). MH technique can be adapted to be used to analyze polytomous data as well. In MH procedure, the subjects are divided into two groups (i.e., focal and reference group) which are matched on a conditional variable which is directly pertinent to the construct being measured. MH procedure, then, assumes that the ratio of answering a particular item correctly is equal between reference and focal groups across all ability levels (Lai, Teresi, Gershon, 2005). They indicate that the major feature of this procedure that distinguishes it from the other methods and procedures is that "instead of testing against a general alternative hypothesis of any difference in correct response rates between groups, this statistic tests against a particular alternative of a common odds-ratio across all blocking, or matching, categories." (p. 284).

D. Logistic Regression (LR)

This part presents a brief explanation of logistic regression as a statistical method for computing DIF. Logistic regression, 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. As Monahan et al (2007) state that 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). In addition, Logistic regression is useful technique for detecting both kinds of DIF, uniform and nonuniform DIF, in dichotomously scored items (Swaminathan and Rogers, 1990). Logistic Regression approaches (LR), in a predictive context, use regression of the external criterion on test score (Lai et al, 2005).

According to Noortgate and Boeck (2005) there exist another variety of logistic models called logistic mixed model in which "the main effect of group membership is modeled by means of one or more additional dummy covariates for the groups" (p. 445). As they explain the mixed model perspective it is suggested that random group effects in item response models be used if people belong to groups regarded as a random sample of groups. In such a case, the specific group effects does not concern researcher primarily, but rather the distribution of these effects is important. Noortgate and Boeck (2005) point out that there are three major strengths of the logistic mixed models approach for modeling DIF: "First, logistic mixed models are easy to understand and very flexible.…Second, the logistic mixed model framework allows considering items and/or groups to be random, resulting in more economical models.…Third, using random DIF effects, a hypothesized explanation for DIF can be included in the model through the effects of covariates, without requiring that the explanation is perfect. The covariates can relate to items, persons, or (higher-level) groups of items or persons" (p. 461).

There are a number of advantages attributable to this technique, logistic regression. For example, French and Miller (1996) point out that, "The logistic regression technique is attractive because it can model both uniform and nonuiform

DIF within the same equation and can test coefficients for significant uniform and nonuiform 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).

In addition, McNamara and Roever (2006) point out that:" Logistic regression is useful because it allows modeling of uniform and nonunifom DIF is nonparametric, can be applied to dichotomous and rate items, and requires less complicated computing than IRT-based analyses. In fact, Zumbo gives several examples that only require SPSS" (p.116). Referring to some researchers such as Lee, Breland, and Muraki (2004) who used logistic regression for a study of DIF in writing prompts for the computer-base TOEFL, they state that logistic regression is similar to the other DIF detection techniques in that it is mostly applied to dichotomous items and it focuses on DIF at the item level. Lee, Breland, Muraki (2002; 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 sacale; and secondly, it does not require a linear relationship between the dependant and independent variables. Mellenbergh (1982) used log-linear model of LR approaches to predict item responses from group membership, disability level, and the interaction of these factors where strong interaction term proves the presence of nonuniform DIF.

Logistic regression procedures are, according to Swaminathan and Rogers (1990) as powerful as the MH technique to detect uniform DIF and it is more powerful than the MH technique concerning detecting nonuniform DIF in dichotomous items. However, they also discovered that logistic regression procedures are much more expensive than MH procedures regarding both time and computer running costs. They argue that, such factors as sample size, test length, and the nature of the DIF are likely to highly affect the power of these two procedures (i.e., logistic regression and Mantel-Haenszel) in that the power of logistic regression procedure can be affected by sample size through its effect on estimation, subsequently, in small samples the asymptotic results may not hold rendering the test statistic unlikely to be valid indicator of the presence of DIF; and in Mantel-Haenszel procedure, the stability of the estimates of odds ratio in each score group may be affected by small samples.

E. Other Methods

Several methods are also based on examination of likelihood ratios connected to nested models which examine the group differences in log-liklihoods associated with compact and augmented models where the augmented models contain additional terms or parameters, and the impact model is more economical. In order to test the difference between modelsa likelihood ratio test, distributed as a chi-square, is examined to provide evidence for indicating DIF (Teresi, 2004).

The DFIT framework is also used in addition to IRT and LR to examine the effect of item-level DIF on the scale (Lai et al, 2005). In this framework, items that show DIF using a single approach might be further analyzed and items that do not exhibit significant DIF or ignorable impact at the scale level may be retained.

V. Conclusion

To summarize what was presented in the present paper, it can be said that validity is an important and essential property of any test, whether tests of second and/or foreign language (e.g., English proficiency test) or tests of social and psychological measures in order for their results to be dependable, generalizable and trustworthy. In order to achieve validity a test must undergo the process of validation to produce results and inferences which exactly mirrors the actual knowledge and ability of the test takers in question. Test bias and item impact were also explained and it was stated that these must be removed from a test so that a test can be regarded as a fair test. To achieve fairness, DIF is used which is a powerful technique in the analysis of bias not only for tests of language but also for tests of social and psychological and heal-related measures. In addition, it was mentioned that, there are different methods and approaches to detect DIF, among Item Response Theory (IRT) methods, Mantel-Haenszel (MH) technique and logistic regression are good examples. Finally, with regard to the advantages and disadvantages attributed to any of these methods and techniques it can be concluded that logistic regression (LR) technique is the one of the best methods developed and used from the early emergence of DIF detection methods up till now.

References

[1] Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & Braun, H. (EDs.) *Test validity* (p. 19-32). Hillsdale, NJ : Erbaum.
[2] Alderson, C., Clapham, C., & Wall, D. (1995). Language test construction and evaluation. NY: CUP.
[3] Bachman, L. F. (1990). Fundamental considerations in language testing. Oxford: Oxford University Press.
[4] Brown, H. D. (2004). Language assessment: Principles and classroom practices. London: Longman.
[5] Brown, J. D. (2005). Testing in language programs: A comprehensive guide to English language assessment. New York: McGraw-Hill.
[6] 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.
[7] 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.

[8] 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. University of Alberta.
[9] Kim, M. (2001). Detecting DIF across the different language groups in a speaking test. *Language Testing*, 18, 89-114.
[10] 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.
[11] McNamara, T., & Roever, C. (2006). Language testing: The social dimension. New York: Blackwell publishing.
[12] 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.
[13] Pae, T. (2004). DIF for examinees with different academic backgrounds. *Language Testing*; *21*, 53-73.
[14] 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.
[15] 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.
[16] Rezaee, A., & Salehi, M. (2008). The construct validity of a language proficiency test: a multitrait multimethod approach. *TELL, 2* (8), 93-110.
[17] 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.
[18] 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.
[19] Sheppard, R., Han, K., Colarelli, S. M., Dai, G., & King, D. W. (2006). Differential item functioning by sex and race in the Hogan personality inventory. *Assessment. 13* (4), 442-453.
[20] 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.
[21] Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement. 27* (4), 361-370.
[22] Teresi, J. (2004). Differential item functioning and health assessment. *Columbia University Stroud Center and faculty of Medicine.* New York State Psychiatric Institute, Research Division, Hebrw Home for the Aged at Riverdale. 1-24.
[23] 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.

Mohammad Salehi is currently a faculty member of the Languages & Linguistics center at Sharif University of Technology and holds a PhD degree in TEFL from the University of Tehran. He got his MA degree from the University of Allameh Tabatabai in TEFL. He earned a BA degree in English literature from Shiraz University. He has taught in University of Kashan, University of Teacher Training, University of Amirkabir, Azad University of Karaj, University of Tehran, and University of applied sciences. He has presented articles in Turkey, Cyprus, Dubai, Armenia, Australia and Iran. His research interests include language testing and second language acquisition research. He has also written books on language testing and vocabulary.

Alireza Tayebi was born in Tehran, Iran on Feb. seventh, 1986. He got his B.A in English Literature from Sheikh Bahayee University (SHBU) in Isfahan, Iran, in 2009. Now he is an M.A student of Applied Linguistics, working on his M.A thesis in the field of language testing. Vocabulary acquisition and language assessment are his main areas of interest so far. He has published one paper on the topic of vocabulary acquisition and strategy use in this regard. He has taught English at different language institutes to different levels for a couple of years and simultaneously had been working as Teacher Assistant (T.A.) at Sharif University of Technology, Tehran, Iran since 2009 up to 2010. Currently, he is working as Administrative Assistant (A.A.) at the same university.