Biannual Journal of Applications of Language Studies (JALS) *http:// jals.hmu.ac.ir/*
Vol. 1, No. 1, Autumn & Winter 2017-2018 *pp. 147 - 172*

A Bias Guideline Preempts Experimental DIF Detection

Mahmood Safari[*]

Assistant professor of TEFL, Department of English Translation, Hazrat-e Masoumeh University, Qom, Iran

(Received: October 24, 2017; Accepted: February 14, 2018)

Abstract

The present article reports on a judgmental analysis of the items in the English subtest of Iranian University Entrance Exam (2009) investigating possible biased items. Plenty of statistical techniques, such as Logistic Regression, Mantel-Haenszel method, and IRT approaches, are developed to detect differential item functioning (DIF) and biased items. They require a pilot study of the test with a sizeable number of subjects. However, sometimes pre-testing is not possible and only subjective and judgmental analysis can be used to detect potential biased items. Off course, judgments should be informed by research findings and experts' opinions. This study suggests that research findings and experts' opinion can be combined to create a *bias guideline* for language test development. If research shed enough light on the issue of biased items and tests, a bias guideline may preempt the experimental DIF detection. This study utilized previous research findings, experts' opinion on bias, and the author's intuition to propose a bias guideline and figure out the possible biased items or bundles (groups of items) in the Iranian University Entrance Exam.

Keywords

Bias, DIF (Differential Item Functioning), Language testing.

[*] **Author's Email:** m.safari@hmu.ac.ir

Introduction

If a test measures something else in addition to what it purports to measure and, hence, favors a group of test takers who have more of the second construct, it is considered as a biased test. Bias is an aspect of test validity: It can be seen as construct irrelevant variance that distorts the test results and makes inferences based on test scores less valid. Henning (1989, p 189) defines bias as "nonrandom distribution of measurement error. It usually results in an unfair advantage for one or more groups or categories of individuals over other groups taking the same test". McNamara and Roever (2007) suggest "test-inherent bias distorts measurement of the construct of interest by allowing other test-taker characteristics to influence scores systematically, thereby introducing multidimensionality into the measurement" (p. 81).

To avoid the societal disgrace of he term 'bias' and its association with 'discrimination', another term was coined for the more technical analysis of items: Differential Item Functioning (DIF). Richards and Schmidt define DIF as "a test item that functions differently either for or against a particular group of test takers (e.g. those with Koran as their L1 or those with French as their L1). A DIF item may be considered biased when a score difference between two or more groups is due to a factor (e.g. test takers' L1) that is not the construct being tested (e.g. L2 listening comprehension)" (p. 157). DIF occurs when two groups of equal ability level show a differential probability of a correct response to an item. Naturally, biased items lead to differential performance of various groups of test takers with the same amount of the construct being measured. Angoff (1993, cited in McNamara & Roever, 2007) argued for keeping the two terms (bias and DIF) separate and using DIF for statistical analysis of score differences between groups and using bias to talk about larger social issues caused by DIF.

When two or more groups of test takers perform differently on an item or a test, there are three possibilities:

1. There is a real difference in the ability between the groups and

the groups have different degrees of the construct being measured. The groups are not equal in terms of the intended construct in the first place.

2. "There are confounding variables within the test which systematically masks or distorts the ability being tested" (Elder, 1997). There is an unintended secondary dimension which is unrelated to the construct being measured and is considered as *nuisance*. The test takers have the same amount of the intended construct but perform differently on the test or item due to the nuisance dimension.
3. The secondary dimension is relevant to and an appropriate part of the construct measured by the test. Now the secondary dimension is called *auxiliary*. The test takers have the same degree of the construct of interest but the item is more relevant to one of the groups due to issues such as gender, language, and socioeconomic status.

Only the second category is considered as bias. The relevant terminology regarding the differential performance of various groups of test takers are *item impact*, *DIF*, and *item bias*. Zumbo (1999) defines them as following:

> Item impact is evident when examinees from different groups have differing probabilities of responding correctly to (or endorsing) an item because there are true differences between the groups in the underlying ability being measured by the item. DIF occurs when examinees from different groups show differing probabilities of success on (or endorsing) the item *after matching on the underlying ability* that the item is intended to measure. Item bias occurs when examinees of one group are less likely to answer an item correctly (or endorse an item) than examinees of another group because of some characteristic of the test item or testing situation that is not relevant to the test purpose. DIF is required, but not sufficient, for item bias (p. 12).

Zumbo (1999) adds "thus, if DIF is *not* apparent for an item, then no item bias is present. However, if DIF is apparent, then its presence is *not* sufficient to declare item bias; rather, one would have to apply follow-up item bias analyses (e.g., content analysis, empirical

evaluation) to determine the presence of item bias" (ibid.). Also Bachman (1990) points to the probability of DIF not being a sufficient criterion for the presence of bias and states "it is important to note that differences in group performance in themselves do not necessarily indicate the presence of bias, since differences may reflect genuine differences between the groups on the ability in question" (p. 271).

Differential test functioning (DTF) defines situation in which the whole test is in favor of or against a particular group of test takers. According to Pae and Park (2004):

> a test shows DTF (i.e. test bias) when the expected true score at the scale level is not the same for two groups of examinees (e.g. Drasgow, 1987; Ellis and Raju, 2003) or when measurement invariance of a test does not hold for two groups of examinees (e.g. Maller & Ferron, 1997; Raju et al., 2002; Zumbo, 2003). Page number?

Sources of test bias

McNamara and Roever (2007) refer to race, gender, socioeconomic status, and first language as background characteristics that are commonly investigated but they add that "theoretically, any background characteristic that some test takers possess and others do not (or not to the same degree) can introduce systematic construct-irrelevant variance and lead to bias" (p. 82). Bachman (1990) asserts "the topic of test bias is a complex one. It may take a wide range of forms, including the 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 procedures, inadequate criterion measures, and threatening atmosphere and conditions of testing" (p. 272). However, he refers to cultural background, prior knowledge of subject matter, field independence, ambiguity tolerance, native language, ethnicity, sex, and age as "characteristics that may cause our tests to be biased for or against various test takers" (p. 291). J. D. Brown (1996) points to race, gender, religion, and nationality as some sources of test bias which must be avoided at all costs. Nevertheless, in addition to test content,

test method has been investigated as a source of test bias. A study revealed that multiple-choice tests are unfair to female test takers and underestimate their real ability or knowledge.

Fairness and bias

Fairness and bias are closely related concepts dealing with the social dimensions of language testing; that is, justice and impartiality. Fair tests are those which bring about justice for all stake holders specially test takers and test users and biased tests are the ones which are partial towards some particular social groups. Fair tests need to be unbiased since impartiality distorts fairness. McNamara and Roever (2007) define the notion of test bias as "the fairness of tests for particular social groups" (p. 81).

Shohamy (1997) considered bias and fairness as two aspects of ethicality: Ethical tests are those which employ methods that are fair to all test takers (i.e. unbiased tests) and those which do not "aim to exercise control and manipulate stake holders rather than provide information regarding proficiency levels" (i.e. fair tests).

McNamara and Roever (2007) considers bias as a part of fairness that deals with "the functioning of test items in ways that advantage or disadvantage groups of test takers" (p. 81). They believe that:

> Test fairness is a broad area, encompassing quality management in test design, administration and scoring, adequate coverage of relevant content, sufficient construction validation work, equal learning opportunities and access to testing, and items measuring only the skill or ability under investigation without being unduly influenced by construct irrelevant variance introduced through test taker background factors (Kunnan, 2000; Savill, 2003, 2005; Shohamy, 2000)

The last point in the above quotation (the development of items which are not unduly influenced by construct irrelevant variance due to background characteristics) is what test bias focuses on.

Nevertheless, some authors suggest that in the purely psychometric sense bias is ethically neutral and quite separate from the issue of fairness. Jensen (1980, p. 375), cited in Elder (1997), articulated that "the assessment of bias is a purely objective, empirical, statistical and

quantitative matter entirely independent of subjective value judgments and ethical issues concerning fairness or unfairness of tests and the uses to which they are put".

Fairness mainly deals with the development and implementation of fairness reviews and cods of ethic which aid the test developers to reduce bias and unfairness from the early stages of test creation. Some organizations, such as Educational Testing Service (ETS) and Cambridge ESOL, apply 'sensitivity review' and 'fairness review' as a part of overall test development process. Fairness is seen as an overall characteristic of a well-constructed test and there isn't any specific discussion of DIF. Fairness review is a guideline for the test developer, during the item writing process, to avoid potentially offensive or upsetting item content and too specialized or technical material. The ETS Fairness Review Guidelines identify offensive content and construct irrelevant knowledge as the two major sources of variance. "Offensive content is interpreted as leading to construct irrelevant variance through test takers being distracted or responding emotionally rather than logically to test material" (McNamara & Roever, 2007, p. 132). The guideline has provided a lengthy list of topics to avoid in tests. Also, the guideline sets a standard that overly complex words, idioms, or syntactic structures should be avoided and military, sports-related, and religious topics and terminology should not be used. As mentioned above, there is no discussion of DIF and statistical DIF analysis in fairness reviews and guidelines.

Methods of bias detection

Various approaches have been adopted to investigate potential test bias or item bias. Broadly speaking, there are two approaches for test bias detection: Judgmental and statistical bias detection approaches.

Judgmental method relies on one or more expert judges' opinion to find potentially biased items. In many situations experimental investigation of bias is almost impossible and only judgmental bias detection approach can be used. J. D. Brown (1996) states:

> The only practical way to avoid bias in most situations is to examine the items carefully and have other language professionals

> also examine them. Preferably these colleagues will be both male and female and will be drawn from different racial, religious, nationality, and ethnic groupings. Since the potential for bias differs from situation to situation, individual teachers will have to determine what is appropriate for avoiding bias in the items administered to their particular populations of students. Statistical techniques can also help teachers to spot and avoid bias in items; however, these statistics are still controversial (p. 53)

However, Zumbo (1999) considers the judgmental method as an 'impressionistic methodology' and says "Instead of the sole reliance on expert (i.e., content area) judges, I recommend that in a high-stakes context faced by human resource organizations one rely on statistical techniques for investigating potential bias because of this method's defensibility" (p. 13).

Statistical bias detection works at two levels: item level and test level. At the level of item bias detection, analysis is done through test-internal bias detection; that is, procedures for identifying DIF.

Test level bias investigation is mainly through external bias detection procedures which typically take the form of "regression analyses of scores of one test against those obtained on another measure deemed to be eliciting the same ability" (Elder, 1997). However, there are other procedures for test bias detection. According to Pae and Park (2004) test bias has been typically investigated:

a) by studying the association between the test score and an external criterion, e.g. Cleary, 1968; Jensen, 1980; Petersen and Novick, 1976; Thorndike, 1971);
b) by computing expected true scores for two groups of examinees using the IRT Test Characteristic Curve, e.g. Drasgow, 1987; Pae, 2004; or
c) by comparing internal factor structures across identifiable subgroups of examinees, e.g. Reise *et al*., 1993; Maller & Ferron, 1997; Raju *et al*., 2002; Zumbo, 2003.

They elaborate the three procedures by adding:

> Most of the external criterion methods are based on the effect of test score regression on an external criterion (e.g. Grade Point Average). These assume that the criterion measure is unbiased and

> represents 'the external standard for evaluating the test' (Camilli & Shepard, 1994, p.9). .... The TCC (Test Characteristic Curve) method estimates true score difference between two groups for each ability (i.e. theta) level. …. Comparison of the resulting TCCs will reveal the cumulative effect of DIF at the test level, hence the presence of test bias. Differences in internal test structure (i.e. factor structure) across subgroups of examinees are best examined by the confirmatory factor analysis (CFA) procedure (ibid.).

Methods used for the investigation of bias at the item level (i.e. DIF detection) can be categorized as those which consider the real difference between test takers in the ability by matching test takers of the same ability level and those that ignore the probability of real differences between the test taker groups.

According to McNamara and Roever (2007, p. 93) methods used for DIF detection fall into four broad categories:

1. Analyses based on item difficulty. They compare item difficulty estimates.
2. Nonparametric approaches which use contingency tables, chi-square, and odds ratios.
3. IRT-based approaches including 1, 2, and 3-parameter IRT analyses.
4. Other approaches including logistic regression, generalizability theory, and multifaceted measurement.

The first category includes some early studies on DIF which took differences in item difficulty or total test scores as automatically indicating bias. One approach in this tradition is delta plot or transformed item difficulty (TID) developed by Angoff (1993). The basic idea behind the TID method is to compare the relative ordering of item difficulty indices across groups, and items which are outliers in terms of item difficulty are flagged for bias. Item difficulties (*p*-values) are computed separately for each group and then transformed to a standardized metric, such as z-score or ETS delta scale. Finally the standardized values are correlated and displayed in a scatter plot. The items which are too difficult or too easy for one group, in comparison to the other group, would be located far away from the

regression line and would be suspect of showing DIF. In this approach examines are not matched on the ability to be measured and the difference between the groups may be real and not due to a bias.

The second category includes nonparametric approaches of Mantel-Haenszel odds ratio (α MH) and the standardization procedure. In the Mantel-Haenszel procedure for every score level a 2 by 2 table is created which is used to estimate the relative odds of a correct response for the reference and the focal group (advantaged and disadvantaged groups respectively). The odds ratios for all score levels are summed and divided by the number of score levels to work out the average odds ratio. The resulting DIF index is the Mantel-Haenszel odds ratio (α MH). Sometimes this index is transformed to fit the delta metric, a scale commonly used by Educational Testing Service (ETS). The scale is centered at 0 (indicating no DIF), stretches from -13 to + 13 and has a standard deviation of 4. Then the index is reported as Mantel-Haenszel delta (MH Δ). DIF items are classified into three types: *negligible DIF* (type A) if the item's MH Δ is nonsignificant at 5% level or its absolute value is less than 1, *large DIF* (type C) if MH Δ exceeds 1.5 and is larger than 1 at a 5% significant level, and other items are classified as *intermediate DIF* (type B).

The second nonparametric DIF detection approach is the standardization procedure also known as the conditional *p*-value, which "compares the proportion of test takers who answered the item correctly for the reference and focal groups at each score level" (McNamara & Roever, 2007, p. 101). More weigh is given to score levels with more test takers. The result is a value between -1 and +1. A difference of 0.1 (10%) indicates that one group has a correct response rate which is, on average, 10% higher than that of the other group at all ability levels. Although this procedure is not mathematically and conceptually complex, it requires large samples of test takers to be implemented productively.

The IRT approaches to DIF detection are based on the comparison of item parameters (difficulty, discrimination, and guessability)

between the reference and focal group. IRT plots test taker's ability level against the probability of answering the item correctly. This can be displayed as a curve which is known as Item Characteristic Curve (ICC). The presence of DIF is signaled by two different ICCs across the groups. "From an IRT perspective, an items functions differentially if test takers at the same ability level but belonging to different groups have a different likelihood of getting the response correct" McNamara and Roever (2007, p. 107).

However, there is a challenge for IRT models of DIF: The calculation of the actual amount of DIF and its significance. Cammili and Shepard (1994), cited in McNamara and Roever (2007), discuss a range of formulas and approaches for calculating the amount of DIF, the most prominent of which is SPD-θ index. To estimate this index the probability of a correct response for focal group members at each ability level is subtracted from the reference group members' likelihood.

Among the last category is the logistic regression, a DIF detection technique that has recently gained increased attention. Logistic regression is useful since it is nonparametric, can be applied to dichotomous and rated items, needs less complex computation than IRT-based techniques, and allows modeling of uniform and non-uniform DIF; that is, when an item shows probability difference in item difficulty (i.e. uniform DIF) or in item discrimination (i.e. non-uniform DIF). "Logistic regression assesses to what extent item scores can be predicted from total scores alone, from total scores and group membership, or from the total scores, group membership, and the interaction between total scores and group membership" McNamara and Roever (2007, p. 116). Interested readers can refer to Zumbo (1999) for a detailed discussion.

Another approach to DIF analysis is generalizability theory (G theory). It is not a common method for DIF detection; however, it is sometimes used for DIF detection since it shows interactions between facets, such as persons, items, ratings, and topics, and it can provide some indication of possible bias. G theory utilizes analysis of variance (ANOVA) to distribute the variance associated with the facets.

The last approach to be mentioned is multifaceted Rash measurement. It applies 1-parameter IRT (Rash modeling) to locate various facets (typically including task difficulty, rater harshness/leniency, and test taker ability) on a common scale. This approach is mainly used for speaking and writing tests with small subject populations. Multifaceted Rash measurement models the relationship among all of the facets in a study and takes task difficulty and rater strictness into consideration when estimating test taker ability.

Problems with DIF/DTF detection

There are some problems with DIF and DTF analysis which should be mentioned. First, they are very complex and not easy for interpretation. Second, they require quite large number of subjects. The third problem deals with the choice of the validity of control measures of ability or criteria used as benchmark for comparing different groups of test takers. Establishing the validity of the concurrent or predictive validity in external bias detection procedures is the limitation of this statistical analysis. In internal bias analysis (i.e. DIF detection) the validity of the total test score as the measure used to compare the performance of different groups of test takers is questionable as it "may itself be suspect since it is the aggregate of a possible biased set of test items" (Elder, 1997). Dorans and Holland (1993), cited in Elder (1997), recommend the purification procedure as a first step in DIF analysis; that is, to remove items with extreme DIF values from the test and then use the refined criterion (i.e. a purified total score) for another DIF analysis of the remaining test items. The fourth problem is related to categorization or how the grouping of the test takers (focal and reference groups) is determined. Many categories are crude dichotomous classifications, for instance native speakers versus nonnative speakers (NNSs versus NSs), because "for many constructs and operational tests, many NNSs can be just as able as NSs" (McNamara & Roever, 2007, p. 123). There is a host of other factors, such as socioeconomic status, motivational and educational factors, and societal expectations, associated with a

category (e.g. gender or L1), rather than mere being a member of a category, which brings about the differential performance of the groups of test takers. For instance, much of the behavior of males and females are accounted for by environmental factors, such as societal expectations from boys and girls, rather than biological and cognitive factors. These societal expectations may differ from culture to culture, and research findings regarding gender may not be generalizable to other contexts. Also differential performance attributed to L1 may be better explained by motivational factors. In a study, German test takers outperformed the Philippine examinees in RC subtest while the Filipinos did better on the grammar subtest. This differential test performance may be better explained by the Filipinos' desire to enter the American universities and the need to pass tests such as TOEFL, than their L1.

Finally, Gipps (1995) points to several limitations of psychometric approach to bias detection claiming that it does not deal with the problem of equity properly. She states:

> The limitation of this approach is that it does not look at the way in which the subject is defined (i.e. the overall domain from which test items are to be chosen), nor at the initial choice of items from the thus-defined pool, nor does it question what counts as achievement. It simply 'tinkers' with an established selection of items. Focusing on bias in tests, and statistical techniques for eliminating 'biased' items, not only confounds the construct being assessed, but has distracted attention from wider equity issues such as actual equality of access to learning, 'biased' curriculum, and inhibiting classroom practices. Page number??

Research on bias

For a century researchers have been exploring the effect of construct irrelevant test taker characteristics on test scores. Binet and his associates published five studies on the impact of socioeconomic status on subjects' performance on intelligence tests from 1905 to 1911. According to McNamara and Roever (2007, p. 85) in 1914, Stern conducted the first analysis of differences in item functioning,

when he considered which items were easier and more difficult for test takers of lower and higher socioeconomic status. Weintrob and Weintrob (1912) were the first to use race as an explanatory variable in their findings, which would later become a central issue in bias analyses in the United States.

The early work on bias and group related differences in the first half of the twentieth century was mainly related to fairness and the goal of avoiding construct-irrelevant variance and concerned IQ tests. However, in the second half of the century bias studies followed the notion of social equity and worked on tests that provided access to educational and job opportunities. The concern of the studies also shifted away from socioeconomic influences on scores to test taker background characteristics such as race, gender, and native speaker status. Golden Rule Settlement in 1970s highlighted the impotence of developing fair tests with little DIF. In 1976 The Golden Rule Insurance Company took legal action against Illinois Department of Insurance and Educational Testing Service (ETS) for alleged racial bias in the licensing test for insurance agents. The test was "for all practical purposes excluding Blacks entirely from the occupation of insurance agent" (Rooney, 1987, p. 2), cited in McNamara and Roever (2007, p. 86).

Concern for bias was very early addressed in the field of language testing. Bachman (1990) refers to the work of Briere (1973) and Briere and Brown (1971) in developing language tests for use with American Indians in the 1960s as "illustrative of an early concern with cultural differences as a potential source of bias in language tests" (p. 273). Bachman (1990) mentions Plaister (1967) and Condon (1975) as other studies in language testing which addressed the problem of cultural content as a possible source of bias.

Zeinder (1986) investigated the validity and cross-cultural generalizability of the test bias contention regarding English Language Aptitude tests which were employed for student selection and placement in Israeli colleges and universities. He applied internal criteria (factor structure and reliability) and external criteria (predictive validity) in his study and concluded that:

> English Language Aptitude Test scores may be equally applicable for aptitude assessment and prediction functions among varying cultural groups in the Israeli setting and provides little evidence suggesting that psychometric aptitude tests are meaningfully biased or unfair with respect to Israeli Arab minority group members applying to the university under consideration. (p. 94)

Another possible source of bias which has been explored is background knowledge or topical knowledge. Bachman (1990) suggests that the studies by Chacevych et al. (1982), Erickson and Molly (1983), Alderson and Urquhart (1983, 1985), Chavanachart (1984), and Hale (1988) "have provided quite convincing evidence of an interaction between test takers' familiarity with content area and performance on tests of listening and reading comprehension" and cloze test (p. 273).

Pae (2004) applied both Mantel-Haenszel procedure and IRT likelihood ratio approaches to examine DIF on the English subset of the 1988 Koran National Entrance Exam for Colleges and Universities, examining the performance of test takers with different academic backgrounds (i.e. Humanities Versus Sciences). The subtest consisted of Listening Comprehension and Reading Comprehension tests containing 55 items. The study identified 18 DIF items with 28 DIF parameters. Preliminary content analysis of those items flagged for DIF further suggested that DIF may be associated with content characteristics specific to group membership (e.g., science-related topics or human issues).

However, there has been a controversy on whether content knowledge is part of the construct being measured or a separate irrelevant construct, especially in the context of English for Specific Purposes (ESP). According to Bachman and Palmer (1996) "if language test tasks are authentic and interactional, and elicit instances of language use, test takers' topical knowledge will always be a factor in their test performance". Nevertheless, "historically language testers have viewed topical knowledge almost exclusively as a potential source of test bias or invalidity" (p. 120). They add that "although topical knowledge may, in many situations, be a potential source of

bias, there are other situations in which it may, in fact, be part of the construct the test developer wants to measure" (p. 121).

The impact of cognitive characteristics (field independence and ambiguity tolerance) on test performance of examinees, on certain types of language tests, has also been investigated. Witkin et al. (1977), cited in Bachman (1990), define field dependence/independence as "the extent to which a person perceives part of part of a field as discrete from the surrounding field as a whole, rather than embedded, or … the extent to which a person perceives analytically" (p. 7). Hansen and Stanfield (1981) and Stanfield and Hansen (1983) studied the relationship between field independence and language test performance of American college students of Spanish as a second language. They found a significant correlation between the scores of GEFT (Group Embedded Figures Test), the most commonly used measure of field independence and cloze test scores. The results suggest that field independent individuals perform better on cloze tests than field dependent people. Hansen's (1984) research findings corroborated those of Stanfield and Hansen by indicating significant positive correlations between GEFT scores and measures of language proficiency, the highest of which was the correlation between the cloze and the GEFT. However, in Chapelle and Roberts' (1986) study, the highest correlations were between the GEFT and multiple choice tests of structure, although the correlation between the GEFT and the cloze were nearly as high.

There have, also, been some studies on the impact of L1 on test performance. Chen and Henning (1985), using IRT approach, examined the English as a Second Language Placement Examination (ESLPE), employed at the university of California, and identified some vocabulary items which were biased in favor of the Spanish test takers and against the Chinese examinees. The items tested the English words for which close cognate forms exist in Spanish. Swinton and Powers (1980) and Alderson and Holland (1981), cited in Bachman (1990, p. 277), found differential performance on TOEFL across different L1 groups. The latter compared the performance of

test takers of the same ability level (with comparable total TOEFL scores) on individual TOEFL items (Bachman, 1990, p. 278).

Brown and Iwashita (1996) explored native language as a source of differential performance on grammar test items on the basis of the hypothesis that L1 influences acquisition of grammar strongly. 1400 students in Australia, China, and Japan (all of them having studied Japanese for 150 to 500 hours) each took a Japanese grammar test of 50 items. Item difficulties were shown to be quite different for the three groups of examinees. Elder (1996) examined the performance of Australian school-age language learners from three different native languages (Chinese, Greek, and Italian) on the reading and listening comprehension tests of Australian Language Certificate (ALC). She used Mann Whitney and Mantel-Haenszel procedures to investigate test score differences and the differential item functioning between those with a home background in each of the three target languages and those without it. The results indicated a strong relationship between home exposure to the language and level of performance the listening and, to a less extent, the reading comprehension of tests.

Gender bias is explored by Takala and Kaftandjieva (2000) who used the IRT 1-paramterr approach (separate calibration t-method) to analyze the performance of 475 Finnish test takers on a 40-item vocabulary section from the Finnish Foreign Language Certificate Examination, investigating gender based DIF. 27.5% of the items (11 items) were shown to b functioning differently (six in favor of men and five favoring women). They hypothesized that DIF items were explainable the different life experiences conditioned by gender roles. The English words *grease*, *rust*, and *rookie* were biased towards male examinees and the words *ache*, *turn gray*, and *jelly* were in favor of female test takers. However, gender roles were not accountable for the words *plot*, *ward*, *association*, and *estate*, which caused DIF. Ryan and Bachman (1992) investigated gender and native language based DIF through the analysis of the performance of 1426 test takers on the TOEFL test and two sections of the FCE (First Certificate in English). Native language was categorized into Indo-European and non-Indo-European test takers. In gender analysis no type C items (i.e. large

DIF) were found and only 4-5% of the items were type B (i.e. medium DIF). In native language analysis the results were completely different. 27% of the items, on both TOEFL and FCE, were classified as type C. 12% and 17% of the items were type B on FCE and TOEFL respectively. However, type B and type C items cancelled out ach other at the scale level. Items having more America-specific content in the areas of culture, academia, and technology were biased towards the non-Indo-European examinees. This could be due to the educational and immigration desire of theses test takers, although the authors did not hypothesize this or any other explanation. Kunnan (1990) used delta plot analysis of DIF to investigate the performance of male and female test takers from four native languages (Japanese, Chinese, Spanish, and Korean) on the ESLPE, containing 150 items. Intended categorizations were age and native language. 13 items (9%) in the native language analysis and 23 items (15%) in the gender analysis were found DIF. Four of the DIF items in L1 analysis were English words with cloze cognate forms in Spanish which favored the Spanish examinees. 11 of the DIF items in the gender analysis were in Listening and Reading Comprehension subtests and were in favor of the male test takers. Their topics were related to business, anthropology, and aerospace engineering. 61% of the DIF items were explainable by the background characteristics hypotheses (e.g. the hypothesis that cognate forms advantage the relevant native languages) and 39% were not.

Some studies have investigated the impact of multiple background characteristics simultaneously on language test performance. Farhady (1982) studied the relationship between academic major, nationality, university status, and sex, and performance on several tests of language ability. The research findings supported the hypothesis that performance on various ESL measures (cloze, dictation, listening comprehension, reading comprehension, grammar, and functional tests) is closely related to test takers' educational and language backgrounds. He suggested that in order to "decrease test bias, the theoretical definition of language proficiency should be modified" (Farhady, 1982).

Spurling and Ilyin (1985) explored how the learner variables of age, sex, language background, high school graduation status, and length of stay in the US influence performance on different language tests (two cloze tests, two listening comprehension tests, a reading comprehension test, and a structure test). Applying regression approach, posed by analysis of variance, they found that age, language background, and high school graduation status have significant effects on tests. Zeinder (1987) investigated age, sex, and ethnicity as sources of bias in the English Aptitude Tests for selection and placement of students in Israeli colleges and universities. The study revealed significant differences between ethnic groups, sexes, and age groups in Language Aptitude Test scores.

Bias is more intricate in performance assessment as in tests of writing and speaking. The small number of tasks included in performance assessment makes it more difficult to achieve a balance of content, context and response mode which is crucial in minimizing the effect of group differences in prior knowledge. Moreover, "the judgmental nature of the scoring is particularly vulnerable aspect of performance assessment in relation to fairness: the perceptions and biases of the rater must not be reflected in the student's score" (Gipps, 1995, p. 103). However, there is ample amount of evidence indicating that markers are influenced by the characteristics of the examinees and of the piece of work. Neatness of presentation and clear handwriting will affect marks in an upwards direction (Wood, 1991, cited in Gipps, 1995, p. 69). Gipps points to a study by Goddard-Spear (1983) in which the same pieces of science writing received lower marks when assigned to girls than when assigned to boys. In short, it seems that when markers can infer gender or ethnic group from the test taker's name, "stereotypes come in to play, with curriculum area interacting" (Gipps, 1995, p. 70).

There have also been statistical investigations of bias in writing and speaking tests. Lee et al. (2004) investigated native language DIF in 81 writing prompts for the computer-based TOEFL. The participants were 254, 435 test takers with Indo-European and Asian native languages, who took the TOEFL computer-based test between 1998

and 2000. Three-step logistic regression procedure for ordinal items was used in this study. They found 27 items with significant group membership effect but the effect size was extremely small in all cases and irrelevant for practical purposes. Kim (2001) used two dimension IRT to analyze the performance of 1038 test takers, from Indo-European and Asian first language background, on the SPEAK (Speaking Proficiency English Assessment Kit) test which focused on grammar, pronunciation, and fluency. The results indicated significant DIF in the grammar and pronunciation scores but not the fluency scores. Grammar and pronunciation favored European test takers at low ability level but at high ability levels Asian test takers performed much better on them.

The effect of test method on test takers' performance is explored by Hellekant (1994), who investigated differential performance of male and female examinees on two different types of test format (multiple choice and open ended questions). She compared the scores of around four thousand Swedish boy and girl upper-secondary school students at the English subset of the National test for the years 1986-1993. Half of the items were multiple choice questions and the second half was free response format. Boys did better than girls at the multiple choice test up to 9% while girls outperformed boys at the open ended questions around 2%. The study suggests that multiple-choice format is unfair towards girls and underestimate their true ability.

Finally, some few studies have examined the effect of DIF on test level bias; that is, whether a test with DIF items lads to DTF or test bias. The relation of DIF to DTF is important because decisions are based on total test performance than performance on an individual item. Some studies have indicated that DIF does not lead to DTF since DIF items cancel out the effect of each other. In Takala and Kaftandjieva's (2000) study the 40-itm vocabulary test did not show gender bias at the test level, possibly since the number of DIF items in favor of each group was almost the same. Zumbo (2003) investigated whether DIF manifested itself in scale level analyses. He applied multi-group confirmatory factor analysis using both covariance and tetrachoric correlation matrix to measure the effect of item level DIF

on test level (i.e. DTF). The results indicated that DIF may not affect test level invariance regardless of its size and magnitude. However, Pae and Park's (2006) research revealed that item level DIF may be carried to the test level bias regardless of the DIF directions, thereby showing mixed evidence to the previous findings reported in the literature.

Bias guideline

Based on the studies on DIF and bias, we can develop a bias guideline which may preempt experimental bias detection. In many cases DIF items or item composites reveal real differences between test taker groups and should not be eliminated as biased items. In some cases elimination of DIF items may reduce the construct being measured and bring about construct under representation. Moreover, in many situations test developers can involve equal number of DIF items favoring each group (e.g. male and female) so that they neutralize each other's effect.

Here a sketchy bias guideline is offered based on research findings, expert opinions in the literature, and the author's intuition. It surely requires adjustments and modifications and need to be improved.

Bias guideline:

1. Test developers should take into consideration the intended test takers and the different groupings of them so that possible biases for these groups are detected. Some items may be biased for or against certain groups of test takers but these groups may not be among the test takers.
2. When DIF items measure auxiliary constructs (e.g. cognate words) they should be proportionate to their frequency in real language use.
3. Vocabulary items should not be technical and specific to certain fields or registers which may disadvantage test takers of other fields.
4. Vocabulary items should not be specifically relevant to certain gender roles and disadvantage the members of the opposite sex. In cases where this leads to construct under representation,

there should be an almost equal number of supposedly biased items favoring each gender.

5. Topics in Listening Comprehension and Reading Comprehension tests should be as general as possible, and interesting and engaging at the same time. Larger number of passages with various topics can be used so that the total test does not advantage or disadvantage any particular group.
6. Various sub-skills and strategies of reading and listening comprehension (such as scanning for details, identifying synonyms and paraphrases, drawing an inference, and recognizing the main idea) should be involved in tests to avoid bias for or against examinees with different reading strategy tendencies.
7. In speaking and writing tests numerous prompts with various topics should be used so that test takers can select the ones which they are more familiar with and competent at.
8. In writing tests, test takers' personal background characteristics should remain anonymous for the raters and more raters examine the writings.
9. Where possible, test takers' writing pieces should be in electronic format so that the hand writing and neatness do not affect the raters and their ratings.
10. Different test formats should be used so that the test method does not advantage or disadvantage certain groups of test takers.

Off course more work and much research is required to complete this guideline as many areas and topics have been ignored in the literature.

Bias in Iranian University Exam

Research findings in the literature, expert opinions regarding bias, and the author's intuition were employed to figure out the possible biased items in the Iranian University Entrance Exam (2009).

There are five groups of test takers taking this test: Sciences, Humanity, Mathematic and Physics, Arts, and Languages students.

Each group takes a separate English test consisting of 25 items. It includes ten multiple choice vocabulary and structure items, a cloze passage with five blanks which measures vocabulary and structure in context, and two reading comprehension passages each containing five questions.

Fist, The Fog index of readability was used to compare the readability of the texts in the tests for different groups. Readability of the texts was dissimilar in the tests for different groups. Table 1 displays the average readability figures for each group.

Table 1. readability figures for each group

Arts	Humanities	Sciences	Math & Physics
22.33	24.4	27.96	32.45

Although test takers struggle with other examinees within their own group for their favorite university fields, there are many fields for which test takers of all groups compete. Readability of texts for the Languages group was not compared to other groups as the test takers in this group are supposed to have a higher reading ability than others since they ascribe for foreign language majors.

Then judgmental bias detection was implemented to work out the possible biased items and item composites.

In the English test for the Sciences group, the topic of the cloze passage and one of the reading comprehension passages (out of two) were almost the same (calendar). In reading tests topic familiarity is a strong source of invariance and large number of passages with various topics is proposed to minimize bias. Different topics should have been used.

In the English test for Arts group, item 83 (a vocabulary item) deals with food and meals and would possibly favor female test takers. Also, the topic of the cloze passage is military and army and possibly would be unfair towards female examinees; additionally, fairness guidelines inhibit military topics in the tests.

In the English test for Languages group, item 112 (a vocabulary item) is relevant to email services. It seems that it would favor

examinees with higher socioeconomic status. They have personal computers and access to the Internet and are familiar with email services. The correct choice is 'forward' which test takers of high socioeconomic status may have seen repeatedly on their email webpage and could easily match it with the word 'email' in the stem without knowing its meaning. One of the reading passages was about a religious story from the bible which may lead to a bias for or against certain religion groups; even, some test takers may know the replies for some questions without the need to read the passage. The English subtests for all the groups can be retrieved from www.sanjsh.org.

Implications and further research

Fair and impartial tests are an urgent requirement especially for high-stakes testing situations such university entrance exams. Test developers need to be well-trained and on the alert for biased items. Provision of a *bias guideline* will equip test writers to create unbiased and fair tests. So far much research has been done on test bias and numerous findings are achieved; although a vast amount of research on bias is still required. Gathering research findings and combining them with experts' opinions on bias would create a productive guideline for unbiased test development. Also, there should be a leeway for the intuition of informed test writers for local and novel situations and bias sources unpredicted in the literature. The present study is a fledgling attempt to provide such a guideline; however, the guideline is far way from complete and requires much improvement.

To complete the guideline, research is required in many new areas and on more sources of bias. Probably, there is a need for 'call for research' to bridge the gap in the neglected and controversial areas. Much of the research on bias needs to be replicated in new contexts; some features of categories (e.g. gender) vary from culture to culture. Moreover, as McNamara and Roever (2007) pinpoints, there should be a "principled theory about the effect of test taker background characteristics on scores" (p. 123). Majority of the explanations for DIF is ad hoc and justificatory. In a study, the researchers totally

changed their explanations after they recognized that they had identified the wrong items as DIF. There is a need for studies which examine the functioning of the items which are developed to favor or harm a particular group. Also, researchers can compare judgmental bias analyses of items with the statistical bias detections of the same items to see how much of DIF and bias could be recognized by judgment. Interested researchers can statistically examine the items which are considered as potentially biased in this study.

References

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

Bachman, L., & Palmer, A. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Brown, A., & Iwashita, N. (1996). Language background and item difficulty: The development of a computer adaptive test of Japanese. *System*, *24*, 199-206.

Brown, J. D. (1996). *Testing in language programs*. Prentice Hall Regents.

Chapelle, C., & Roberts, C. (1986). Ambiguity tolerance and field dependence as predictors of proficiency in English as a second language. *Language Learning, 36,* 27–45.

Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing, 2,* 155–63.

Elder, C. (1996). The effect of language background on foreign language test performance: the case of Chinese, Italian, and modern Greek. *Language Learning, 46,* 233–82.

Elder, C. (1997). What does test bias have to do with fairness? *Language Testing*, *14*, 261–277.

Farhady, H. (1982). Measures of language proficiency from the learner's perspective. *TESOL Quarterly, 16*, 43–59.

Gipps, C. V. (1995). *Beyond Testing*. London: The Falmer Press.

Hale, G.A. (1988). Student major field and text content: interactive effects on reading comprehension in the Test of English as a Foreign Language. *Language Testing, 5,* 49-61.

Hansen, J., & Stanfield, C. (1981). The relationship between field dependence- independence cognitive styles and foreign language achievement. *Language Learning, 31,* 349–367.

Hansen, L. (1984). Field dependence-independence and language testing: evidence from six Pacific island cultures. *TESOL Quarterly, 18*, 311–324.

Hellekant, J. (1994). Are multiple-choice tests unfair to girls? *System, 22*, 349-352.

Henning, G. (1987). *A Guide to Language Testing*. Newbury House Publishers.

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

Kunnan, A.J. (1990). DIF in native language and gender groups in an ESL placement test. *TESOL Quarterly*, *24*, 741–746.

Lee, Y. W., Breland, H., & Muraki, E. (2004). *Comparability of TOEFL CBT writing prompts for different native language groups (TOEFL Research Reports No. RR–77).* Princeton, NJ: Educational Testing Service. Retrieved February 22, 2006, from http://www.ets.org/Media/Research/pdf/RR-04-24.pdf

McNamara, T.F., & Roever, C. (2007). *Language testing: the social dimension*. Blackwell Publication.

Pae, T.-I. (2004). DIF for examinees with different academic backgrounds. *Language Testing 21, 53–73.*

Pae, T.-I., & Park, G-P. (2006). Examining the relationship between differential item functioning and differential test functioning. *Language Testing, 23*, 475– 496.

Richards, J. C., & Schmidt, R. (2002). *Dictionary of Language teaching and Applied Linguistics.* Pearson Education Limited.

Ryan, K., & Bachman, L-F. (1992) Differential item functioning on two tests of EFL proficiency. *Language Testing, 9*, 12–29.

Shohamy, E. (1997). Testing methods, testing consequences: are they ethical? Are they fair? *Language Testing*, *14*, 340–349.

Spurling, S., & Ilyin, D. (1985). The impact of learner variables on language test performance. *TESOL Quarterly*, *19*, 283–301.

Stanfield, C., & Hansen, J. (1983). Field dependence-independence as a variable in second language cloze test performance. *TESOL Quarterly*, *17*, 29–38.

Takala, S., & Kaftandjieva, F. (2000). Test fairness: A DIF analysis of an L2 vocabulary test. *Language Testing, 17*, 323–40.

Zeinder, M. (1986). Are English language aptitude test biased towards culturally different minority groups? Some Israeli findings. *Language Testing*, *3*, 80–95.

Zeinder, M. (1987).A comparison of ethnic, sex, and age biases in the predictive validity of English language aptitude tests: Some Israeli data. *Language Testing, 4*, 55–71.

Zumbo, B. (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. Retrieved February 22, 2006, from http:/www.educ.ubc.ca./faculty/zumbo/DIF/index.html.

Zumbo, B. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language Testing*, *20*, 136–47.