A confirmatory study of Differential Item Functioning on EFL reading comprehension

Alireza Ahmadi*
(Assistant Professor, Shiraz University, Iran)
*Corresponding author email: arahmadi@shirazu.ac.ir

Touraj Jalili
(MA Graduate, Shiraz University; PhD Candidate, University of Isfahan, Iran)

(Received: 2014/04/23, Accepted: 2014/06/30)

Abstract
The present study aimed at investigating DIF sources on an EFL reading comprehension test. Accordingly, 2 DIF detection methods, logistic regression (LR) and item response theory (IRT), were used to flag emergent DIF of 203 (110 females & 93 males) Iranian EFL examinees' performance on a reading comprehension test. Seven hypothetical DIF sources were examined in this regard: text familiarity, gender, topic/text interest, guessing, and the social variables of location, income, and educational status. Only LR, for gender and text familiarity, could preempt DIF with gender supporting the gendered-text effect while text familiarity benefiting, inversely, the participants with low level of text familiarity. For interest in topic, LR found a single item favoring the group with higher levels of interest and the IRT model detected DIF in either extreme. Regarding guessing and income, the LR indicated DIF supporting the low guessers and high-income group whereas IRT, conversely, showed DIF favoring the high guessers and low-income group. For location and education both methods, correspondingly, demonstrated DIF for the expensive location and educated groups. Finally, the differential test functioning result made it clear that only three sources of DIF (gender, income, & interests) were transferred to the test level. The findings could support a proportional effect of DIF sources.

Keywords: DIF, LR, IRT, reading comprehension

Introduction

Differential item functioning (DIF) occurs when equally knowledgeable individuals from different subgroups are of different likelihood of correctly answering (or endorsing) an item (Shepard, Camilli, & Averill, 1981). DIF is a necessary but not sufficient condition for bias (Zumbo, 1999). In other words, bias only exists when the groups illegitimately differ in their performance on an item. Because items indicating DIF can function as a threat to the validity of a test, DIF analysis has become an essential step in the validation of a test (Camilli & Shepard, 1994; Zumbo, 2007; Zumbo, & Rupp, 2004) especially in high-stakes testing situations (Pae & Park, 2006).

The present study was an attempt to focus on several hypothetical factors that may cause DIF on a reading comprehension test. As pinpointed by McNamara and Roever, (2006), the literature lacks research on social or contextual sources of DIF. As such, several hypothetical social sources of DIF were also brought into focus in this study.

Background

Many studies have been conducted on DIF. The majority of these studies have focused on gender or ethnicity and only some have shed light on other sources of DIF. Among the studies conducted, mention could be made of those focusing on gender (e.g., Li & Suen, 2012; Pae, 2012; Ryan & Bachman, 1992; Takala & Kaftandjieva, 2000), ethnicity (e.g., Hammer, Pennock-Roman, Rzasa, & Tomblin, 2002; Stoneberg, 2004), language background (e.g. Harding, 2011; Li & Suen, 2012), topic familiarity or content knowledge (Pae, 2004), and field of study (Barati, Ketabi, & Ahmadi, 2006; Näsström (2004)).

Such studies have indicated that for example east Asian test takers outperform test takers from European-language backgrounds on Michigan English Language Assessment Battery (Li & Suen, 2012); that females tend to outperform males in tests of verbal and written abilities, especially if constructed response items are included (Willingham & Cole, 1997); they also outperform males on aesthetics, human affairs, and contextualized reading items (Carlton & Harris, 1992) on the mood, impression, or tone items of a reading passage (Pae, 2012), on items related to human relations, human rights, aesthetic and on items referring to stereotypical female activities (O'Neill & McPeek, 1993; Sadker & Sadker, 1994), whereas males outperform females on antonyms, and analogies (Carlton & Harris, 1992), on reading items of inferencing type (Carlton & Harris, 1992; Lawrence & Curley, 1989), on items involving visualization and items calling upon knowledge and experiences acquired outside of school (Hamilton & Snow, 1998), and on items related to science and on items referring to stereotypical male activities (O'Neill & McPeek, 1993; Sadker & Sadker, 1994).

The results of studies on DIF have sometimes been contradictory and far from being conclusive. Furthermore, the studies have been limited to few sources of DIF and lack of research is felt on social/contextual sources of DIF (McNamara & Roever, 2006). Very few studies have focused on how social factors may create DIF on different tests. The present study tried to fill this research gap by focusing on a number of social factors that may explain DIF on a reading comprehension test. It also tried to shed more light on a number of non-social factors mentioned as sources of DIF in the literature. As such, the following questions were put forward in the current study:

1. Can text familiarity, gender, interest, guessing, and contextual variables (location, income, and educational status) function as sources of DIF on EFL reading comprehension items?
2. To what extent do the results of DIF detection methods (IRT and LR) converge?
3. Can item-level differential performance manifest itself at the scale-level analysis?

Method

Design of the study

Unlike the DIF studies in the literature which select an exploratory approach toward the analysis of DIF sources, the present study benefited from a mixed exploratory-confirmatory approach in this regard; that is, in dealing with social sources of DIF, an exploratory approach was adopted in which the test was administered to the test takers who were divided into groups based on the social variables (location, income, & educational level) and then their performance on the test was analyzed for instances of DIF. This approach is the one adopted in DIF studies in the literature. The present study, however, adopted a confirmatory approach in studying sources of DIF mentioned in the literature; that is, instead of running the test and then searching for the probable sources of DIF, the study, following McNamara and Roever's (2006) suggestion, selected the most cited sources of DIF in the literature and intentionally included them in the reading test to see whether they really manifest themselves in the DIF study. The assumption was that if such claimed variables are really sources of DIF, then an item which embraces such variables should necessarily indicate DIF. A counterargument for this will hold that such sources may be context and test specific and hence do not necessarily manifest themselves in DIF analysis of any test in any context.

Participants

The participants of this study included 203 English as a Foreign Language (EFL) learners (93 males and 110 females) selected from among around 300 intermediate-level students at Shiraz University Language Center (SULC) in the spring semester in 2011. Only those participants who based on their final scores in previous semesters were roughly identified as intermediate were selected for the study. Due to the administrative limitations, only the participants' scores on the regular tests administered by the institute at the end of each term were taken into account and no specific proficiency test was employed. However, it should be mentioned that this was quite enough for the purpose of the present study because the DIF software (BILOG MG) automatically estimates the participants' proficiency based on the test used for DIF analysis and does not need a separate test to be used for proficiency. The reason why we tried to roughly homogenize the participants based on their proficiency was only to make sure that the test used for DIF analysis was not too far from their level of proficiency and hence more realistic answers would be given by the participants; otherwise, DIF studies rely on a single test.

Instrumentation

Reading comprehension test

A reading comprehension test was developed for the purpose of the study. The passages of the reading test were originated from Anderson's (2007, 2008) books entitled "ACTIVE Skills for Reading: Book 2" and "ACTIVE Skills for Reading: Book 3". The books were selected from among a series of four books to fit the intermediate proficiency level. Furthermore, as these books are not usually taught in the Iranian language institutes or universities, there was a very low chance for the students to have seen the texts before. Attempt was made to include the hypothetical sources of DIF. To this end, first of all, on the basis of gender-based familiarity with text topic, 6 short passages were chosen. Following the literature (Bugel & Buunk, 1996; Newman, Groom, Handelman, & Pennebaker, 2008; O'Neill & McPeek, 1993) some topics were hypothesized to be gender-stereotypical; that is, such topics as practical affairs, money, mechanical tools, soccer, occupation were considered male favorable, whereas topics related to humanities, social sciences, aesthetics and human relations, wedding, and household chores were considered female favorable. Therefore, two passages were taken as male-friendly (*What does a Million Dollars Buy? & Meet Freddy Adu, Soccer Sensation*), two passages as female-friendly (*Wedding Customs & The Right Job for Your Personality*), and two others as gender-neutral (*Are Human Beings Getting Smarter?* & A *Different Kind of Spring Break*).

To develop the test, each of the 6 passages was followed by five commonly used questions in reading comprehension tests (Broukal, 2007). They were items asking for details (facts), reference to a word or phrase, vocabulary knowledge, main idea (theme), and inference (logical conclusion). Therefore, the final test was a 30-item reading comprehension test. Two faculty members who were experts in L2 language testing and had worked on DIF checked the reading texts and questions in line with the suggestions made in the literature to make sure that the final test served the purpose of the study. The reliability of the test was also estimated through KR21which turned out to be .52. The moderate reliability observed was not far from logic as due to the confirmatory nature of the study the sources of error variance (DIF sources) were intentionally included in the test which could noticeably decrease the reliable variance proportion in the measurement equation.

DIF questionnaire

The second instrument was a very short questionnaire the first part of which was a set of bio-data questions asking the test takers to provide information about their gender, residency neighborhood, families' monthly income, and parents' educational level. The second part, however, was attached to the end of each passage and comprised these questions:

a) How familiar was the text topic or content to you?
b) How much were you interested in the reading topic?
c) Which item did you answer by guessing?

The purpose of this section was to collect information on the participants' topic interest, topic familiarity and guessing. They were expected to indicate their level of familiarity with and interest in the text by selecting one of the five options *very much, much, to some extent, little,* and *very little/ none*.

Data collection procedure

The factors included in DIF analysis were literature-based variables hypothesized to contribute to test takers' differential performance on EFL reading tests. They were gender, text familiarity, text interest, guessing, location in two levels (down=living in less expensive neighborhoods with a housing price of $700-$1500 per squared meter; up=living in the expensive neighborhoods with a housing price of $2200-$2800 per squared meter), income in two levels (low=less than $1000 per month, high= more than $1000 per month), and educational level in two levels (academically educated, academically uneducated).

Data analysis

It is recommended that more than one method of DIF analysis be employed in DIF studies to come to more dependable results (Camilli, 2006; Camilli & Shepard, 1994; Pae, 2012, Uiterwijk & Vallen, 2005). In line with this suggestion, the present study employed two methods: (a) a classical method: logistic regression (LR), (b) 1-p item response theory (IRT) model. This could add to the dependability of the results and made it possible to compare the degree of correspondence between the results of the two methods.

Results and discussion

DIF analysis based on Logistic Regression

Overall, about 47% (14 items) of the whole test displayed DIF through the use of LR. Only three of these items indicated large DIF and the majority indicated moderate DIF based on the criteria recommended by Hidalgo and Lopez-Pina (2004); that is, negligible DIF: $\Delta R^2 < 0.13$, moderate DIF: $0.13 \leq \Delta R^2 \leq 0.26$), and large DIF: $\Delta R^2 > 0.26$. In what follows the results of DIF analysis based on LR are presented in detail for each of the hypothetical sources.

DIF based on gender

As previously mentioned, the reading texts were selected with an eye toward the gender differences in topic familiarity of a reading test (e.g., Brantmeier, 2003; Pae, 2012). Overall, five items (17% of the whole test) were flagged for DIF based on gender. The results of LR indicated that in the female-friendly passages 13.33% of the items; that is, four items (two inference items, one vocabulary item, & one reference item,) favored females. In the male-friendly passages only a single item

(3.33% of the items) which was a vocabulary item favored males over females; and in the neutral passages no item indicated DIF in favor of a gender. The results of DIF overall, functioned in line with the literature on gender-based DIF that there exists DIF or differential performance on gendered texts (e.g., Carlton & Harris, 1992; Lawrence & Curley, 1989). In other words, the females got advantage of the female-friendly texts and the males were favored by male-friendly texts. Neither group was reported to gain benefit from the gender-neutral texts. More specifically, the results partially echoed those of the study by Newman et al. (2008) that most of the differences between men and women are related to the application of function words (e.g., reference-type item in the present study) and lexical words (e.g., vocabulary-type items in the present study).

DIF based on the familiarity with text topic or content

One part of familiarity was discussed under the rubric of gender-based DIF. However, mention was made that that particular analysis did not find any DIF for gender-neutral texts. Thus, familiarity was also considered separately with the intention of its effect on the items dispensed with the gender influence. The idea was that although two passages were claimed to be female-friendly, two male-friendly and two neutral based on the literature, still individual differences were possible in terms of their familiarity with the texts regardless of their gender. For example, saying that a text is more female-friendly based on the topic does not eliminate the chance that some males could be familiar with such texts. As such, the study focused on familiarity with the text topic as a factor for differential performance regardless of the gender. Therefore, the participants were divided into three groups based on their familiarity with each text (highly-familiar, moderately-familiar, and slightly-familiar). This was done based on the participants' answers to the questionnaire items indicating their level of familiarity with each text. Therefore, those selecting *very much* and *much* options were considered as the highly-familiar group, those selecting *to some extent* as the moderately-familiar and those selecting *little* and *very little* as the slightly-familiar group. Then DIF analysis was performed. Unlike the results of DIF based on gender, the results of DIF based on familiarity found some traces of DIF in the gender-neutral texts. Overall, topic familiarity was found to be the source of DIF for three items, two of which were related to the gender-neutral texts. Therefore, the results were not in line with the results of gender-based DIF. It is therefore manifest that, with three items (10% of the whole test) indicating DIF, the familiarity was not supported to have a leading role in DIF occurrence since only one inference-type item (3.33% of the whole test) favored the moderately-familiar group and two items of details- and vocabulary-type (6.66% of the whole test) benefited the slightly-familiar group.

This finding is against the literature (e.g., Sadker & Sadker, 1994) and does not support the assumption that higher familiarity may lead to higher chances of endorsing an item correctly. Therefore, taken with the results of the gender-based DIF together, it seems that gender (focusing on gender-friendly or gender-stereotypical texts) will provide us with a better explanation of DIF. In other words, familiarity is better to be considered together with the gender effect to explain the DIF on a test.

DIF based on the interest in text topic

To avoid the possible bias, test developers may want to select unfamiliar text topics, but by doing so the texts may not be so interesting and relevant to the test taker (Bachman, 1990). The challenging task of a test writer is to avoid either extreme, i.e. to develop topics that are very general and innocuous and at the same time interesting and relevant. Familiarity and interest pose difficulty in the design, development, and analysis of reading tests. Thus, we can hypothesize that the familiar texts would be interesting to a particular group often to the detriment of the other. To study the effect of interest as a source of DIF, like what we did for the familiarity, at first we divided the participants into three groups based on their interest in each text; that is, highly-interested, moderately-interested, and slightly-interested. This was done for each text separately and then DIF was checked for each item. Contrary to our expectations, the notion of interest in text topic did not turn out to be an influential source of DIF because only one item (3.33% of the whole items) which was a vocabulary item favored the highly-interested group over the others.

DIF based on guessing

It goes without saying that guessing is a hypothetical source of DIF in an MC test. In this study, the test takers were asked to inform the researcher whether for each passage with five items, they answered a particular item by chance. That is, they were asked to say if they guessed at a particular item or not based on which they were divided into three groups of low, mid, and high guessers. The results of DIF analysis flagged five items (16.66% of the whole test) based on guessing which favored low-guessers. In other words, the results indicated that being a member of the non-guesser (or low-guesser) group increased the probability of endorsing these items correctly. It seems that those who had guessed in this study were mostly wild guessers and thus their guessing did not help them. Guessing, therefore, was not supported to play a significant role in DIF results as low/non-guessers were more successful.

DIF based on location

Only one item was found to support the idea that the test takers' neighborhood (location) can function as a source of DIF. This item which was a vocabulary item functioned to the favor of those living in rich neighborhoods. Thus, living in different neighborhoods (and by generalization in different cultural settings) was corroborated by a single item (3.33 % of the whole test) to function significantly in the test takers' differential performance on the vocabulary-type reading item. This finding, though based on a single item, supports Zumbo & Gelin's (2005) idea that by ignoring socio-geographic differences one may lose the whole picture of DIF.

DIF based on income

Some social groups, as a result of high income, may have access to more high-quality educational opportunities which, in turn, can lead to their gradually better performance on language tests. Income, in this study, emerged as a source of DIF in two items. One of the items was an inference-type item and the other one a main-idea item. Both items indicated much higher probability of endorsing the items correctly for those from higher-income families. Thus, income was supported only by two items (6.66% of the whole test) to distinguish between test takers in performing on the main idea- and inference-type reading items.

DIF based on the educational level

Unlike the location and income variables that reverberate the community-level contextual factors, the family (or parental) educational level is less susceptible to the community at large and compares test takers at the individual level (Zumbo & Gelin, 2005). To take note of this contextual factor, the test takers were asked to report whether their families or parents were academically educated or uneducated. The results of DIF analysis in this regard indicated that only one item (inference-type item) presented this contextual factor as a significant predictor of DIF in favor of the test takers coming from educated families. Thus, a small portion of the test (3.33%) supported the advantage of the educated group.

IRT analysis

Overall, 33.33% (10 items) of the whole test displayed DIF through the use of IRT. The results of IRT analysis for gender and familiarity variables did not show any meaningful and statistically significant DIF. As such, in what follows the remaining DIF sources (interest, guessing, location, income, & educational level) are taken into consideration. In each section the difficulty differences between the contrasting groups, called group threshold differences, and the standard error of measurement are provided for each item. Items for which the threshold difference is roughly twice (1.96) or more the size of the standard error display DIF at the p = .05 level (Thissen, Steinberg, & Wainer, 1993). The only parameter to be attended to in this program was the difficulty value (b) and therefore the lower threshold value for a particular group means that the item was easier for them. That is, the negative or positive direction of the threshold differences indicates which particular subgroup was favored.

DIF based on the interest in text topic

The IRT analysis revealed the threshold differences output for the interest in texts. As Table 1 indicates (see Appendix), items 1, 3, 5, 20, and 27 displayed DIF. The table indicates that in both mid-low and high-low comparisons, Item 1 functioned as a DIF item (Thissen et al., 1993). The negative threshold values in both cases indicate that the item (details type) was more difficult for the low group than the mid and high groups. Thus, those with high and moderate interest in Text 1 were favored by item 1. For item 3 the results demonstrate that in both cases (high-low, high-mid) the item was more difficult for the high group. Thus, it is concluded that the low and mid groups were favored by Item 3. By the same token, the results indicate that for Item 5, the low and mid groups benefited from the item. In item 20, the negative threshold difference (-0.792) indicates that for the low group the item was more difficult; and hence the high-interest group was favored by this inference-type item. Finally, for item 27 it is clear that the high group was favored because their threshold was lower than that of the mid group. The detection of 16.66 % of the whole items (five out of 30) as including DIF demonstrates a greater proportion of interest effect in the IRT analysis in comparison with the other factors.

DIF based on guessing

The negative threshold differences for items 11 and 14 and the positive difference for item 15 indicate that all the high guessers had more likelihood to get the details- (11), main idea- (14), and inference-type (15) items right. Thus, guessing as a speculative source of DIF indeed made the MC items function in favor of the high guessers.

DIF based on location

Regarding the residential neighborhoods of the respondents (*up* vs. *down*) only item 4 (3.33% of the whole test) revealed significant differential performance between the groups (1.417). Thus, the u*p* group was reported to take advantage of this main idea-type item.

DIF based on income

Out of the 30 items, only one item (3.33% of the whole test) displayed DIF and supported the hypothetical function of income in the IRT analysis. Only item 19 was flagged for DIF. The negative threshold difference for this main idea-type item makes it clear that the low-income group found the item easier to endorse. In fact, those test takers coming from low-income families were favored by this particular item. As you will see in the DTF section, income was one of the sources transferred from the item level to the test level and had a proportional effect on the test bias.

DIF based on educational level

The comparison between the difficulty parameters of the educated- and uneducated-family test takers indicated that only item 5 (inference type) displayed significant DIF. The positive threshold difference reveals that upon answering item 5, the educated group was favored because for them the item was easier to endorse. Therefore, only 3.33% of the whole items supported the speculative effect of the parental educational level on differential performance of the test takers on reading comprehension.

Comparing the LR and IRT results

The overall results of DIF analyses are summarized in Table 2. The first column represents the hypothetical DIF sources; the second and the third illustrate DIF identified through LR and IRT respectively. The last column shows the items identified by both methods and therefore indicates the agreement between the two methods.

Table 2: The significant DIF sources identified by LR and IRT

DIF Sources	Items flagged as DIF through LR	Items flagged as DIF through IRT	Items flagged as DIF through both LR and IRT
Gender	2, 5, 13, 15, 18		
Familiarity	10, 21, 23		
Interest	18	1, 3, 5, 20, 27	
Guessing	8, 14, 17, 27, 28	11, 14, 15	14
Location	8	4	
Income	15, 29	19	
Education	5	5	5

As the above table indicates, only items 5 and 14 were identified as DIF by both DIF detection methods. This indicates a low level of correspondence between the results of the two methods. The IRT framework found no DIF related to the gender and familiarity variables. With respect to the interest variable both techniques could flag some DIF items. Through a small portion of correspondence between the models, interest as a source of DIF, which acts in a balanced correspondence with familiarity in reading comprehension (Bachman, 1990), was reported to function in a mixed direction, i.e. leading some items to favor interested test takers and some others to benefit the uninterested ones. Guessing, as a source of DIF in MC reading items was identified by both methods but with different functions. That is, the LR method indicated that low guessers were more successful, but the IRT model reported conversely that the DIF items were in favor of those test takers who guessed highly at the answers, hence a total divergence between the two methods. The results of both methods for location, in a complete correspondence, indicated that those living in the *up* neighborhoods were favored. In a similar vein, both frameworks identified a single item (item 5), in a complete convergence, as displaying education-oriented DIF in favor of those coming from educated families. However, for income-based DIF, the methods stood in a sharp contrast. In fact, unlike LR, IRT found that those coming from the low-income families were favored.

Let us interpret what we have come up with so far in two parts. First, as the above table depicts the two methods overlap in the preemption of only two DIF items. Why is there so much variance between them? Why are their results not supportive of each other? This finding could be explained in terms of the sample size. It is mentioned that for binary items at least 200 people per group is required for LR to function well and smaller sample sizes could deteriorate the results (Zumbo, 1999). Further explanation of this finding, may come from the significance level used for DIF analysis. Usually IRT is more accurate and does much better with a smaller sample size than LR in flagging items for DIF. So a Bonferroni correction test could be useful to apply to the LR analysis results for more careful analysis (Alavi & Karami, 2010; Runnels, 2013; Thompson, 2006). This can lead to a lower number of items to be identified as DIF. However, this procedure was not of much help in the present study to resolve the differences in the results of IRT and LR DIF analyses because the differences did not lie in the number of items rather in the type of items.

However, overall the results of the two methods should be considered together, as there would be no method which is foolproof with all samples and for all contexts. Both methods, likewise, indicated DIF related to the contextual factors despite the low reputation of the factors in the literature. Furthermore, the presence of more DIF related to guessing rendered it as a very important source of DIF.

Second, of all the sources in the study guessing caused the most DIF followed by interest, gender, familiarity, and contextual factors. The study centered around reading comprehension tests and that is why the emergence of guessing DIF for MC items, interest in topics, and gender DIF for gendered passages was not unconceivable. Those sources that could be manipulatively included in the research yielded the results much more in keeping with the literature than those factors over which, in an ex post facto manner, the researchers had no control, i.e. the contextual (social) factors.

Analyzing differential test functioning (DTF)

The presence of DIF items is considered to be a threat to the test validity to the extent that it manifests itself in the scale-level differential performance (Pae & Park, 2006; Zumbo, 2003). The results of a multiple-regression analysis indicated that although the effect of the variables on the total score was significant, only three variables (gender, income, and interest) were transferred from DIF to DTF level and added to the complexity of the DTF analysis (Pae and Park, 2006; Takala & Kaftandjieva, 2000). The remaining DIF sources, like the results of Zumbo’s (2003) study, did not emerge in the scale-level results.

Conclusion

The primary purpose of this study was to move beyond the hypothetical interpretation of DIF by following McNamara and Roever's (2006) suggestion, conducting an empirical confirmatory study. Its outlook was toward taking advantage of the falsification philosophy in a way that rival hypothetical DIF sources were proposed and none was presumed to predict the why of DIF occurrence. Two DIF sources were included in the reading test and the others were checked through a questionnaire attached to the end of each subtest (passage). The study could not support the notion of familiarity with text topic as a major source of DIF. For gender only the LR model could detect some male- and female-friendly DIF items. With the small number of DIF items for the interest, guessing, and income-related DIF, the results of the two DIF detection methods stood in a noticeable contrast to each other. However, the methods were found to be in a complete convergence in regard to the location and family educational status. Two other findings were related to the item type and bias in the test. The results indicated that gender could have an impact on reference- and vocabulary-type reading items. Furthermore, the DTF result made it clear that of all the sources, only three factors (gender, income, & interests) were transferred to the test level.

The findings of the study overall may warn language teachers and test developers about failing to pay due attention to DIF on reading comprehension tests. This would be double problematic in criterion-referenced tests where the number and size of DIF items can greatly affect the mastery and nonmastery decisions made based on a single cutoff score. Even more problematic would be ignoring DIF on high stakes tests which are usually used for gate-keeping purposes. The stigma of failing in such tests may distort the educational life (or even the entire personal life) of the test takers.

References

Alavi, S. M., & Karami, H. (2010). Differential Item Functioning and ad hoc interpretations. *TELL, 4*(1), 1-18.

Anderson, N. J. (2007). *ACTIVE skills for reading: Book 2.* Boston: Thomson ELT.

Anderson, N. J. (2008). *ACTIVE skills for reading: Book 3.* Boston: Thomson ELT.

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

Barati, H., Ketabi, S., Ahmadi, A. (2006). Differential item functioning in high-stakes tests: the effect of field of study. *IJAL*, *19*(2), 27-42.

Brantmeier, C. (2003). Beyond linguistics knowledge: Individual differences in second language reading. *Foreign Language Annals, 36 (1),* 33-43.

Broukal, M. (2007). *TOEFL reading flash*. NJ: Peterson's.

Bugel, K. & Buunk, B. P. (1996). Sex differences in foreign language text comprehension: The role of interests and prior knowledge. *The Modern Language Journal, 80 (1),* 15-31.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (pp. 221-256). Westport: American Council on Education & Praeger Publishers.

Camilli, G. & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Carlton, S. T. & Harris, A. M. (1992). *Characteristics associated with differential item functioning on the Scholastic Aptitude Test: Gender and majority/minority group comparisons.* ETS Research Report. Princeton, NJ: ETS.

Hamilton, L. S., & Snow, R. E. (1998). *Exploring differential item functioning on science achievement tests* (CSE Technical report No. 483). Los Angeles: Center for Research on Evaluation, Standards, and Student testing.

Hammer, C. S., Pennock-Roman, M., Rzasa, S. & Tomblin, J. B. (2002). An analysis of the test of language development-primary for item bias. *American Journal of Speech-Language Pathology*, *11*(3), 274-284.

Harding, L. (2011). Accent, listening assessment and the potential for a shared-L1 advantage: A DIF perspective. *Language Testing, 29 (2),* 163-180.

Hidalgo, M. H., & Lopez-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Hanszel procedures. *Educational and Psychological Measurement, 64,* 903-915.

Lawrence, I. M. & Curley, W. E. (1989). *Differential item function for males and females on SAT-Verbal Reading subscore items: Follow-up study.* Educational Testing Service

Research Report. Princeton, NJ: ETS.

Li, H. & Suen, H. K. (2012). Detecting native language group differences at the subskills level of reading: A differential skill functioning approach. *Language Testing, 30 (2),* 273-298.

McNamara, T. & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA & Oxford: Blackwell.

Näsström, G. (2004). *Differential item functioning for items in the Swedish national test in mathematics course*. Retrieved from http://www. Vxu.se/msi/picme1o/L2ng.PDF.

Newman, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 1400 text samples. *Discourse Processes, 45,* 211-236.

O'Neill, K. A. & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (255-276). Hillsdale, NJ: Lawrence Erlbaum Associates.

Pae, T. (2004). Gender effect on reading comprehension with Korean EFL learners. *System, 32 (3),* 265-281.

Pae, T. (2012). Causes of gender DIF on an EFL language test: A multiple-data analysis over nine years. *Language Testing, 29 (4),* 533-554.

Pae, T.-I. & Park, G.-P. (2006). Examining the relationship between differential item functioning and differential test functioning. *Language Testing, 23 (4),* 475-496.

Runnels, J. (2013). Measuring differential item and test functioning across academic disciplines. *Language Testing in Asia*, 3:9, doi:10.1186/2229-0443-3-9.

Ryan, K. & Bachman, L. F. (1992). Differential item functioning on two tests of EFL proficiency. *Language Testing, 9 (1),* 12-29.

Sadker, M. & Sadker, D. (1994). *Failing at fairness. How our schools cheat girls*. Toronto, ON: Simon & Schuster.

Shepard, L. A., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics, 6,* 317-375.

Stoneberg, B. D. (2004). *A study of gender-based and ethnic-based differential item functioning (DIF) in the spring 2003 Idaho Standards achievement Tests. Applying the Simultaneous Bias Test (SIBTEST) and the Mantel-Haenszel Chi Square Test*. Paper for EDMS 889 Measurement-Statistics Practicum, University of Maryland, College Park. Retrieved from http://files.eric.ed.gov/fulltext/ED483777.pdf

Takala, S. & Kaftandjieva, F. (2000). Test fairness: A DIF analysis of an L2 vocabulary test. *Language Testing, 17 (3),* 323-340.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thompson, B (2006). *Foundations of behavioral statistics: An insight-based approach*. London: The Guilford Press.

Uiterwijk, H. & Vallen, T. (2005). Linguistic sources of item bias for second generation immigrants in Dutch tests. *Language Testing, 22(2),* 211-234.

Willingham, W. W. & Cole, N. S. (1997). Fairness issues in test design and use. In W.W. Willingham & N. S. Cole (Eds.), *Gender and fair assessment* (pp. 227 - 346). Hillsdale, NJ: Lawrence Erlbaum.

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, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language Testing, 20 (2),* 136-147.

Zumbo, B. D. (2007). Three generations of DIF analysis: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4(2)*, 223-233.

Zumbo, B. D. & Gelin, M. N. (2005). A matter of test bias in educational policy research: bringing the context into picture by investigating sociological/community moderated (or mediated) test and item bias. *Journal of Educational Research and Policy Studies, 5,* 1-23.

Zumbo, B. D. & Rupp, A. A. (2004). Responsible modeling of measurement data for appropriate inferences: Important advances in reliability and validity theory. In D. Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences* (PP. 73-92). Thousand Oaks, CA: Sage.

Appendix

Table 1: Group threshold differences for location, educational level, guessing, income and interest

	Location	Educational Level	**Guessing**			Income	**Interest**		
Item	**Down-Up**	Uneducated-Educated	**Low-Mid**	**High-Mid**	**High-Low**	Low-High	**Mid-Low**	**High-Low**	**High-Mid**
1							-5.548	-6.996	-1.059
							1.970	2.761	2.175
3							0.510	4.545	2.953
							1.509	1.567	1.032
4	1.417								
	0.547								
5		1.328					1.558	5.023	2.536
		0.521					1.423	1.441	0.935
11			1.545	-1.545	-1.940				
			1.314	1.111	0.857				
14			-2.845	-1.521	0.838				
			1.538	0.741	0.967				
15			2.012	1.006	-0.642				
			0.912	0.637	0.548				
19						-1.243			
						0.485			
20							0.096	-0.792	-0.619
							0.374	0.380	0.314
27							0.681	-0.059	-0.781
							0.359	0.363	0.370