The Journal of Teaching Language Skills (JTLS)
Vol. 2, No. 3, Fall 2010, Ser. 61/4
(Previously *Journal of Social Sciences & Humanities)*

Gender-based DIF across the Subject Area: A Study of the Iranian National University Entrance Exam

H. Barati
Assistant professor
Isfahan University, Isfahan
email: h.barati@gmail.com

A. R. Ahmadi *
Assisstant professor
Shiraz University, Shiraz
email: ar.ahmadi@yahoo.com

Abstract

This study aimed at investigating differential item functioning (DIF) on the Special English Test of the Iranian National University Entrance Exam (INUEE). The effect of gender and subject area was taken into account. The study utilized one-parameter IRT model with a sample of 36000 students who sat for the INUEE Special English Test in 2004 and/or 2005. The findings confirmed the presence of DIF on this test. The cloze test indicated the lowest DIF whereas language function indicated the highest DIF. The results also revealed some general gender DIF patterns across the subject area. Females were favored on the three sections of grammar, language function, and the cloze test, whereas males were favored on the vocabulary and word order sections. The reading comprehension section favored males and females equally. It was also concluded that the item format alone could not explain DIF. In other words, it is the subject area or the interaction of the subject area and item format that determines the degree and direction of DIF.

Received: 6/6/2010 Accepted: 10/26/2010
* Corresponding author

Keywords: 1. IRT Models 2. Differential Item Functioning 3. High-stakes Tests 4. Iranian National University Entrance Exam 5. Gender, Subject Area

1. Introduction

Constructing a test which is fair enough to different groups of the same population has been a major concern for testing practitioners. This needs attempts at analyzing test items to make sure that individual items are not biased toward a particular group; in other words, all the test takers who are of the same language proficiency level should have the same probability of getting the item correct (Camilli & Shepard, 1994). Such a concern has received particular attention by policy-makers, administrators, and testing professionals in recent years (Perrone, 2006). Researchers have tried to pinpoint such bias and its possible sources. The Standards for Educational and psychological Testing (APA, AERA, NCME, 1999) have also underlined the importance of ruling out potential biasing factors in a test and have considered this as critical to a sound testing practice. In concert with that, investigating test scores for Differential Item Functioning (DIF) has changed to a quite popular practice in the field of language testing. DIF occurs when respondents with the same underlying level of knowledge have a different probability of answering the item correctly (Thissen, Steinberg, & Wainer, 1993). DIF is not equal to item bias; however, it's a necessary condition for item bias; that is, if an item is biased certainly there exists DIF whereas the existence of DIF does not necessarily mean that the item is biased.

> DIF methods allow one to judge whether items (and ultimately the test they constitute) are functioning in the same manner in various groups of examinees. In broad terms, this is a matter of measurement invariance; that is, is the test performing in the same manner for each group of examinees? (Zumbo, 2007, p. 1).

DIF items may function as a serious threat to the validity of the measuring instruments. Instruments containing such items may have reduced validity for between-group comparisons, as their scores may be indicative of a variety of attributes other than those the scale is intended to measure (Thissen, Steinberg, & Wainer, 1993). Thus, DIF analysis is an essential step in the validation of educational and psychological tests (Camilli & Shepard, 1994). Moreover, it becomes "intimately tied" to test validation when establishing the inferential limits of the test; that is, establishing for whom the test or item score inferences are valid (Zumbo, 2007; & Zumbo, & Rupp, 2004).

DIF studies especially deserve due attention in high-stakes test contexts (Pae & Park, 2006). This has been the motive for the present study to check the potential DIF on a high-stakes test (National University Entrance Exam) in Iran. This test is probably the most important test (at least in terms of the number of its applicants) in Iran, which is administered annually to screen the candidates for admission to different fields of study at universities. Considering the large number of students who sit for this test, and the fact that their future is highly affected by it, even small amounts of DIF deserve due attention to avoid unfair and unintended consequences. The present study focuses on this and attempts to scrutinize DIF in this test in terms of gender across different subject areas.

2. Literature Review

The differences observed between the performance of males and females on various standardized tests have been the subject of much research. A number of studies conducted especially in the US have confirmed the presence of gender-based differences in certain tests especially the Scholastic Aptitude Test (SAT). There is a good bulk of studies investigating gender DIF on SAT. These studies indicate that males outscore females on the SAT-math section (College Entrance Examination Board, 2001; Kanarek, 1988); women typically get equal or better grades in college math compared with their male classmates with

higher SAT-math scores (Bridgeman, & Wendler, 1991); and that females score lower on the SAT-verbal section (Kanarek, 1988; College Entrance Examination Board, 2001).

Similar studies have found that high school males generally outperform females in standardized tests of science, mathematics, history, and social studies although females generally have similar or higher school awarded marks (Halpern, 1992; Wightman, 1998). Conversely, females have generally been reported to outperform males in tests of verbal and written abilities, especially if constructed response items are included (e.g. Mazzeo, Schmitt, & Bleistein, 1993; Tyler, 1965, cited in Gokiert & Ricker, 2002; Willingham & Cole, 1997). Other studies of this kind found that males outperform females on antonyms, and analogies (Carlton & Harris, 1992); and on all types of problem solving activities (Gallagher, et al. 2000; Willingham & Cole, 1997).

Some studies have shown that gender DIF is related to item content (e.g. Bolger & Kellaghan, 1990; Lane, Wang, & Magon, 1996; Linn, De Benedictis, Delucchi, Harris, & Stage, 1987; Mazzeo, Schmitt, & Bleistein, 1993; Willingham & Cole, 1997). For example, the content-based DIF studies have shown that males outperform females on items related to practical affairs and science (Donlon, 1973); items involving visualization and those eliciting information about the real life (Hamilton and Snow, 1998); items related to science and those referring to stereotypical male activities (Mazzeo, Schmitt, & Bleistein, 1993; O'Neill & McPeek, 1993) whereas females outperform males on items related to human relations, human rights, aesthetic and on items referring to stereotypical female activities (Donlon, 1973; Mazzeo, Schmitt, & Bleistein, 1993; O'Neill & McPeek, 1993). Similar studies have also indicated that males tend to perform better than females on items related to geometry, ratio, proportions, and those containing tables, graphs, or figures, whereas females tend to perform better on items related to computation and items containing symbols (Harris & Carlton, 1993; O'Neill & McPeek, 1993).

Still some other studies have focused on the effect of test format on differential item functioning. Evidence indicates that males generally perform better than females on multiple choice items while females perform better on essay-type (constructed) items (Bolger & Kellaghan, 1990; Linn, Delucchi, and Stage, 1987; Mazzeo, Schmitt, & Bleistein, 1993; 1990). In addition, in measures of quantitative abilities (vocabulary, grammar, etc.), females tend to perform better than males when constructed response items are included (Lane et al., 1996). This difference has been attributed to the stronger writing skills and neater and more comprehensive answers that are provided by females (Lane et al., 1996; Mazzeo et al., 1993; Willingham & Cole, 1997). Also it has been suggested that girls are more reluctant to guess on multiple-choice questions than boys; boys overestimate their likelihood of success and hence take risks unknowingly, for which they are rewarded (Linn, De Benedictis, Delucchi, Harris, & Stage, 1987); and males tend to guess more on multiple-choice exams whereas girls tend to omit the items they are not sure about (Bolger & Kellaghan, 1990).

Less focus has, however, been on gender DIF in EFL tests. Studies of this type generally suggest that reading comprehension is differentially easier for females, whereas males are better performers on antonyms and analogies and that women are likely to perform better on items with more contexts (Carlton & Harris, 1992); that females are much superior in writing and speaking though in reading and vocabulary the difference is not significant, and that males are superior in listening comprehension (Cole, 1997); that females are favored in grammar and vocabulary whereas males in cloze tests (Lin & Wu, 2003); and that items classified as Mood/Impression/Tone are easier for females whereas items classified as Logical Inference are more likely to favor males regardless of item content (Pae, 2004). However, some others like Hyde and Linn (1988) maintain that no specific difference exists between females and males.

An analysis of the DIF literature indicates that most of these studies suffer from a number of shortcomings including the following:

1. Many DIF studies rely on an internal criterion to match the test takers. This criterion is usually defined as the total score on a test. This could be awkward especially when the test embraces different sections which are seemingly related. In this case selecting the overall score as the matching criterion would be quite problematic (Hamilton, 1997). This is because the test takers who are matched according to their total score on the test may differ in their knowledge of the test sections. As such it's better to match the test takers based on their score on each section and do the analysis for each section separately. To avoid such a problem in the present study the English Subtest of the INUEE was broken down into its subtests and the scores on the subtests were used as the matching criterion. The assumption was that such a procedure would yield more dependable results (Clauser, Mazor, & Hambleton, 1991; Donlon, Hicks, & Wallmark, 1980).

2. The majority of DIF studies have been carried out in a European or American context (Pae, 2004). This would cast doubts on the generelizability of the findings. Lack of the DIF studies is felt in other contexts (especially in Asian contexts). Of course, this does not necessarily mean that DIF differs from culture to culture or that it is culture-specific. It, however, means that since the educational system of the European and American countries and also their tests are basically different, probably some explanations for DIF could be found in the culture of a country. This needs further studies before any claims can be made. It is also likely that cross-cultural studies of DIF give us a clearer picture of the role of culture in DIF. The present study was conducted in an Iranian context on two successive versions of the INUEE English Subtest to help fill this gap.

3. A number of studies suffer from small samples and short tests. They have usually employed samples below 500 with tests of lower than 50 items (e.g. Roever, 2005). This study employed a large sample of participants (36000) through disproportionate stratified sampling (Ary, Jacobs, and Razavieh, 1996). The test included 70 items and was assumed a long test compared to those used in previous studies.

Not a single study, to the best of authors' knowledge, has been conducted so far in Iran to examine the presence of DIF on different language tests or in different contexts. This together with the vital role that the INUEE plays in the life of Iranian students in terms of their higher education served as an incentive for the present study to investigate the fairness of this test to male and female test takers using a DIF methodology.

It goes without saying that although it is rather fashionable these days to criticize DIF analyses for not providing the reason for differential test performance, it is clear that this criticism is somewhat misplaced because not all DIF studies are aimed at finding the reason for DIF. For example, one could only be interested in flagging DIF items in an operational language test and hence the reason for DIF is secondary to guaranteeing the adequacy of the inferences made from the test scores and reducing test bias against sub-groups of test takers (Shimizu, & Zumbo, 2005). The present study was also basically an observational one focusing on the specification of the number and type of DIF items in different sections of a high-stakes test in Iran (INUEE). It, however, tried to provide some explanations for the observed DIF patterns.

3. Methodology

3.1 Participants

The participants of this study were randomly selected from a population of about 500,000 high school graduates who sat for the Special English Test of the INUEE in 2004 and 2005. Disproportionate stratified sampling (Ary, Jacobs, & Razavieh, 1996) was used to select the participants from the two groups of male and female students. Eighteen thousand students were randomly selected from those who took part in the Test 2004 and 18000 from those taking part in the Test 2005. Half of the students were female and half male. Overall a total sample of 36000 test takers was selected for the present investigation. The assumption was that a large sample, unlike a small one which could hide interesting statistical effects, may point to statistically significant findings

where the effect is seemingly small and meaningless (Camilli, & Shepard, 1994).

3.2 Instruments

Iranian National University Entrance Exam

The Iranian National University Entrance Exam (INUEE) is designed to measure applicants' general ability and select the best for studying at higher education. This test is designed in five forms to screen candidates for admission to universities. The Special English Test is specifically designed to screen candidates for the English majors. This Test consists of 70 MC items in six areas of structure (10-12 items), vocabulary (20 items), word order (4-5 items), language function (4-5 items), cloze test (15 items), and reading comprehension (15 items). Two versions of this test (2004, and 2005) were chosen for the present study. Table 1 depicts the reliability indices measured through Cronbach's α for different sections as well as the total test of both versions.

Table 1. Reliability estimates for the Special English Test

Test 2004						
Structure	Vocabulary	Word order	Language function	Cloze test	Reading	Total
.66	.77	.76	.65	.86	.89	.95
Test 2005						
Structure	Vocabulary	Word order	Language function	Cloze test	Reading	Total
.70	.81	.69	.63	.84	.89	.94

As indicated, the language function section had the lowest reliability probably because of the small number of items in this section and the reading comprehension section had the highest reliability in both versions. The reliability of the total test was quite high for both versions.

3.3 Data collection procedures

The data for this study were collected from the Iranian Sanjesh Organization in Tehran. This organization is in charge of preparing and administering many of the important examinations held in the country

such as the university entrance exam for high school graduates, university entrance exam for MA candidates, etc. The organization provided the anonymous answer sheets of all the applicants for the English field, who had taken the English Subtest in 2004 (more than 275000 applicants) or 2005 (more than 220000 applicants).

3.4 Data analysis

Since the study was to be carried out through the IRT models of DIF detection, checking the assumptions behind the models was necessary to make sure of the suitability of a model. The first assumption was that of uni-dimensionality. This assumption holds that the items in a test should measure a single dominant trait. Nevertheless, it seems vital to mention that uni-dimensionality is not a strict concept since strict uni-dimensionality will result in a narrow construct (Mc Namara, 1996) that may not adequately represent the original "content map of expert reviewers" (Teresi, 2006, p. 20). As a result, a good model fit requires only a reasonably good approximation to the unidimensionality assumption (Mc Namara, 1996). Furthermore, research designed to assess the impact of violations of the unidimensionality assumption (e.g., Cooke & Michie, 1997; Hulin et al., 1983) has suggested that the unidimensional IRT models are relatively robust with respect to moderate violations of strict unidimensionality, and that the most important issue concerns the relative degree to which the item pool is dominated by a single latent trait. In concert with that, some researchers have proposed that a test can be considered essentially uni-dimensional if the major domain of its latent space contains only a single trait (Reise & Waller, 1990; & Stout, 1987). Similarly, Reckase (1979) suggests that if the first factor accounts for roughly 20 percent or more of the variance in addition to being several times larger than the second factor, the test can be considered as uni-dimensional and appropriate results could be obtained by using IRT models to such a data. This has been shown to be the least stringent criterion for unidimensionality (Choi & Backman, 1992).

Therefore, at first the TESTFACT program was applied to the data to check for the unidimensionality assumption. Tetrachoric correlation matrices were obtained. The results indicated that the two versions of the English Subtest were multidimensional since even the Reckase's criterion for uni-dimensionality was not met, and therefore had to be broken into the components which were determined through tetrachoric correlation matrices. This along with the fact that separating the sections of a test for DIF analysis could lead to more accurate results and therefore reduce the probability of type 1 error which mistakenly highlights items for DIF (Clauser, & Mazor, 1998; Donlon, Hicks, & Wallmark, 1980; Le, 2000 and Reeve 2003) led the researchers to break the test into different sections. These were basically the same sections categorized by the original examination board of the INUEE (i.e. grammar, vocabulary, language function, word order, cloze test, and reading comprehension).

As for the second assumption, local independence, BILOG MG assumes that local independence is met and provides no test of this assumption. However, based on the results of uni-dimensionality analysis which resulted in the English Subtest to be divided into 6 parts, the local independence assumption was also partly taken into account. This is because "the assumptions of uni-dimensionality and local independence are related in that items found to be locally dependent will also appear as a separate dimension in a factor analysis" (Reeve 2003, p. 12). Therefore when the assumption of uni-dimensionality is met, the local independence will also be met (Reeve 2003). Nevertheless, the relationship between uni-dimensionality and local independence does not mean that one can be ignored if the other is met. Thus, it seems the findings of this study should be cautiously generalized since the local independence assumption was only considered based on the results of the uni-dimensionality analysis and no specific test was employed to check it. Violation of the local independence assumption is more likely observed in the cloze test than the other sections of the English Subtest.

That may be because of the interconnectivity which is usually present in the cloze test.

Finally the data were analyzed for the presence of DIF using the one-parameter IRT model. The fit of the one-parameter IRT model with the data was estimated through BILOG MG. This software has been introduced as the steadiest and most accurate software for the estimation of item parameters (Liu, Shu, & Jeng, 1998).

Moreover, 1-parameter IRT model has been shown to be quite robust and not to result in high level of error even when guessing and discrimination are significant factors in the performance of candidates (Mc Namara, 1991 & 1996). In case of the Special English Test, guessing may not be an influential factor because the test takers are penalized for their wrong answers (1/3 of a score for any wrong answer). As such, utilization of the 1-parameter model in the present study seemed to be a good option. In the next section of this paper the results of the data analysis are explored in detail.

4. Results and Discussion

4.1 Gender DIF

The results of DIF analysis for the two versions of the Special English Test revealed no exclusive pattern **(**favoring only one group**)** for different test sections (subject areas). In fact, in all the subtests there were some items indicating DIF in favor of female students and some in favor of male students, hence the patterns were not exclusive but overall each subtest favored a particular group more than the other.

As indicated in Table 2. below 43.56 % of all the items in the two versions of the test were marked for DIF. The table presents the number of items indicating DIF in favor of each gender.

Table 2. Gender DIF in the English test: Number of items favoring each group

	Grammar	Vocabulary	Word order	Language function	Cloze	Reading	Total
Female	8	7	1	4	4	8	32
Male	5	10	2	2	2	8	29
Total	13	17	3	6	6	16	61
Percent	59.09%	42.5%	33.33%	66.66%	20%	53.33%	43.57%

The above table indicates that the cloze section carried the lowest amount of DIF whereas the language function carried the highest. Further, it is shown that, most of the test sections were in favor of females; that is, grammar, language function, and cloze sections were in favor of females whereas vocabulary and word order favored males. Reading comprehension favored males and females equally in terms of the number of items indicating DIF.

The first section in the Special English Test was grammar. This part basically embraces de-contextualized language items. In other words, this part is usage rather than use-oriented. It basically includes items on prepositions, articles, verb forms, relative pronouns, etc. It seems that females are favored on items of this type.

The "language function" section provided short dialogs with some blanks or dialogs followed by some comprehension questions. This section also favored females.

Similarly the cloze section favored females. Cloze tests usually depict contextualized language items. In fact, the cloze test represents integrative test of language by providing a real context of language use (Oller, 1979). Such tests have also been indicated to correlate well with measures of EFL proficiency (Fotos, 1991; Irvine, Atai, & Oller, 1974; & Oller & Conrad, 1971). DIF in this section could not be justified by the notion of "topic familiarity". According to "topic familiarity" particular groups will have better performance on a text they are familiar with. Topic familiarity has been indicated to be a significant factor in gender DIF (Brantmeier, 2003; Bugel, & Buunk 1996; Floyd & Carrell, 1987; Hyde & Lynn 1988). However, the results of this study in relation to the cloze test cannot be explained through "topic familiarity". The text used in the Cloze test (version 2005) was about the boxing champion "Muhammad Ali Clay". It seems logical in terms of topic familiarity to expect better performance on the part of male students since they are usually more interested in or have more information about boxing. However, exactly the reverse came true; that is, females outperformed males on this subtest.

The cloze passage in the English Subtest 2004 which indicated no gender differential performance was related to the "industrial revolution" and how it affected the working hours. Probably this topic could be considered a "gender-neutral" topic leading to more or less the same performance by the two groups. Similar performance is usually found on a text which is approximately of similar familiarity to different sex groups (Brantmeier, 2003; & Bugel & Buunk, 1996).

In contrast to the above three sections, the vocabulary and word order sections favored males. The results of the vocabulary section were in line with Lin and Wu (2003). They were, however, in contradiction with Carlton and Harris (1992) and O'Neill and Mc Peek (1993) who concluded that females would outperform males on abstract concepts. In the present study, the items that favored each group included both abstract and concrete items. In fact, most of the items in favor of females were testing concrete words such as food, stretcher, etc which was against the findings of Carlton and Harris (1992) & O'Neill & Mc Peek (1993).

In line with the vocabulary section, the word order section favored male students. This section basically included de-contextualized items. In other words, they were mechanical rather than meaningful or communicative items (Paulston & Brudes, 1976). In both versions of the Special English Test, word order provided students with four options one of which presented the correct structural order of a sentence. The advantage found here for male students is mostly in contrast to Carlton & Harris (1992) & O'Neill *et al.* (1993) who found that females outperformed males on sentence correction items and items in which the best written sentence had to be selected from the given options.

As for the reading comprehension section, none of the groups was favored in terms of the number of items indicating DIF. In fact, this section indicated four DIF items in favor of each gender group. Thus, we may state that the same number of items indicating DIF in favor of different groups may cancel the effect of one another; that is, DIF at the level of individual items may be canceled at the test level (Drasgow,

1987; Roznowki & Reith, 1999; Zumbo, 2003). Nevertheless, some other researchers have indicated that the DIF detected at the item level may be transferred to the test level bias regardless of the directions of DIF. In other words, there would be no DIF cancellation at the test level (Pae and Park, 2006; & Takala, & Kaftandjieva, 2000), and hence all the items that indicate DIF must be carefully dealt with. The reading comprehension texts (5 texts) were considered separately and more accurately for their DIF. This ended in some interesting points to turn up. Females were favored on three reading texts (East End Mall, Tour de France, & Edgar Allan Poe) whereas males were favored on two reading texts (tooth decay and physical barriers to the growth of animals and plants).

The notion of topic familiarity could only partly explain gender differential performance in this section. For example, it could explain why females were favored on the reading text about the East End Mall. This text is about a shopping center and we may claim that females have more information in this regard. But topic familiarity could not explain why females were also favored on the reading text about Tour de France which is an international bike race (mostly of interest to males) and Iranian male students are usually expected to be more interested or have more information in relation to such a text.

The idea that females score higher than males on humanities-oriented reading passages but lower than males on science-related passages (Curely & Schmitt, 1993; Lawrence et al. 1988; and Maller, 2001) could also account for some of the results of the reading section. It could explain why females were favored on the text related to “Edgar Allan Poe” (a humanities-oriented text) and why males were favored on the texts “physical barriers to the growth of animals and plants”, and “tooth decay” which are mostly scientific texts. But the fact that the reading section generally favored males and females equally in terms of the number of items indicating DIF could probably be explained in two ways:

First, reading skill is the most emphasized and practiced skill in Iranian secondary schools. In the same way, reading tests are familiar to students from the very first year of learning English at secondary schools. Therefore, it seems, that adequate practice in reading comprehension as well as the familiarity of the students with reading comprehension tests may have reduced the effect of gender as far as DIF is concerned.

Second, Special English Test may be considered as a speed test, since many students cannot complete the reading section which comes at the end of the test. This may lead to less variation in the performance of the different groups of test takers on the reading comprehension test. Eventual scores of the test takers may then seem to be highly affected by the speededness of the test and not by the real language abilities. Of course, this may lead to some inaccuracy in DIF studies and needs due attention. The present study didn't try to exclude such students from the study since the sample was randomly selected from a population of about 500000 candidates and many of the test takers appeared to have incomplete performance on the reading comprehension test. The assumption there was that excluding such students would mean disregarding many of the candidates. This could ,in turn, make the final sample not a true representation of the population.

4.2 DIF across MC item format

Investigating DIF across different formats was not a purpose of this study. In fact, this study has focused on the MC item format since the INUEE is only constructed in MC format. Thus nothing could be specifically stated regarding different item formats and their relation to DIF. However, a general conclusion could be made about item format based on the results of this study; that is, item format alone cannot explain DIF adequately. In fact, the idea that a specific format could be easier for males or females (e.g. Becker, 1990) is misleading if the subject area is not taken into account. In other words, it is the subject area (or probably the interaction of the subject area and item format) that determines the degree and direction of DIF not the item format alone.

This was quite clear in the present study. The study indicated that males were favored on the two sections of vocabulary and word order whereas females were favored on the three sections of grammar, language function, and cloze. Neither of the male or female groups was favored on the reading comprehension section. The fact that all of these sections were constructed in MC item format indicates that format alone cannot lead us to a safe and sound conclusion concerning DIF.

This conclusion may seem to be in contradiction with the previous research that indicated males to be better performers on MC items (Becker, 1990; Bolger & Kellaghan, 1990; Linn, Delucchi, & Stage, 1987; Mazzeo et al., 1993). This could probably be explained based on the following: First, previous studies have mostly been conducted without any special attention to the subject areas; that is, basically they have tried to find the differential performance due to item format regardless of the role that the subject area may have. However, the present study focused on MC item format across six different subject areas of grammar, vocabulary, word order, language function, cloze test, and reading comprehension and it's quite probable, as indicated by the results of the present study, that the subject area may also play a role, probably a more important role than the item format, in favoring one group over another. As such, DIF should not be judged only based on the item format, rather the influence of the subject area should also be taken into account.

Second, many of the studies that report superiority for males on MC items have been conducted basically on non-language tests, such as the SAT, IQ tests, mathematics tests, etc. (e.g. College Entrance Examination Board, 2001; Kanarek, 1988; Wainer & Steinberg, 1992). Few studies, however, have specifically focused on language tests (e.g. Cole, 1997; Lin & Wu, 2003; Pae, 2004).

The third point to be mentioned in this regard is the effect of "guessing". Some researchers (e.g. Bolger & Kellaghan, 1990) have concluded that males tend to guess more on multiple-choice exams; in contrast, girls tend to omit the items they are not sure about and are more

reluctant to guess on multiple-choice questions than are boys; boys overestimate their likelihood of success and so take risks unknowingly, for which they are rewarded in the multiple-choice format (Linn, et al 1987); therefore this risk taking on the part of males leads to their out-performance on MC item format. But it should be noted that the majority of the studies conducted in this regard have been concerned with the tests in which guessing was encouraged whereas in the Special English Test guessing is somehow suppressed since candidates are penalized for their wrong answers. Thus in such a testing context candidates may avoid guessing and try to answer the questions based on their knowledge. As such, the out-performance observed in the previous studies is not observed in the present study.

5. Conclusion

The present study aimed at investigating gender DIF in the Special English Test of the Iranian National University Entrance Exam. It confirmed the presence of DIF on this test and revealed some general DIF patterns across the subject area. It turned out that females were favored on three sections of the Special English Test; that is, grammar, language function, and cloze, whereas males were favored on the vocabulary and word order sections. The reading comprehension section favored both males and females equally. DIF in the cloze test and the reading comprehension section could partly be justified by the notion of "topic familiarity".

6. Implications

The findings of this study may bring about certain implications regarding gender DIF in the INUEE Special English Test:

1. The results indicated that different subject areas favored different groups based on their gender. At present, Iranian applicants of EFL courses are judged for their general English proficiency only on the basis of the results of the Special English Test. This test relies solely on MC item format and embraces only some of the language skills and

components. In fact currently, this test is administered in multiple-choice format in six subject areas of grammar, vocabulary, language function, word order, cloze test, and reading comprehension. This may give rise to validity and fairness questions. Equity concerns would probably dictate a mix of different types of assessment instruments as well as subject areas (Mazzeo, Schmitt, & Bleistein, 1993). This has been adapted in some tests since the mid-90's, during which "performance" assessment or "authentic" assessment was widely accepted as a better way of measuring student achievement than the MC format. However, in terms of eliminating test bias, it has so far shown rather disappointing results. In fact, in some cases, performance assessments have even shown wider achievement gaps than do multiple-choice formats. This may be because performance assessment relies heavily on expert judgment for its results and human judgment is notoriously difficult to standardize. On the other hand, multiple-choice formats may have understated the true extent of the achievement gaps, which are now revealed by the new assessments. From a practitioner's standpoint, performance assessment is very time-consuming and expensive to implement on a large scale. It has not yet shown its value as a tool to eliminate test bias, but has definitely expanded the practitioner's tool kit, (Schellenberg, 2004). Anyhow, great need is felt to do more research on DIF hopefully to find the best format and/or subject areas for different testing contexts. Certainly limiting testing instruments to MC format and also to some specific language skills and components (what is exactly done in the INUEE) could be far away from fair assessment. Standardized tests like IBT, IELTS, etc. could be considered as good models of language tests in that they are not limited to just some language skills and components and are not limited only to one item format.

2. The Cloze Test indicated the lowest degree of gender DIF. This is worth noticing since it may mean that cloze provides a fairer test of language ability and hence needs more attention in this regard. At present, 15 out of 70 items of the INUEE Special English Test are in cloze format.

3. The results may be beneficial to test developers by providing information concerning the effect of gender on the performance of test takers in the Iranian National University Entrance Exam (INUEE), and therefore highlighting the items that may unfairly work to the advantage or disadvantage of some examinees. Therefore eliminating and modifying the items which work to the advantage or disadvantage of some learners could be one of the objectives obtained.

4. Due to the absence of DIF studies in an Iranian context, the present research could be insightful to the practitioners in this field. It could function as a platform for further studies in this regard.

5. Finally, noting the great impact that high-stakes tests like the INUEE Special English Test have on teachers and their teaching, the findings of the present study could be helpful especially to Iranian English teachers and learners. Whatever that the Iranian English teachers do in their classes is heavily under the influence of the content of this test; that is, everything is streamlined toward the successful performance on this test rather than successful learning of the English language. For example, speaking and listening skills are paid but a lip service since these skills are not tested in the INUEE, whereas grammar is emphasized because it is one of the sections of the INUEE.

7. Further Research

1. DIF studies so far have basically paid attention to the statistical procedures for DIF detection. Few have noticed the causes of DIF (Schmitt et al, 1993; and Uiterwijk, and Vallen, 2005). More studies are needed to specifically focus on the causes of DIF by employing different qualitative techniques in line with the quantitative ones. In this line content analysis could be accompanied by verbal protocols, or think aloud procedures to yield better results. Content analysis alone cannot lead us far in this regard (Englhard et al. 1990; Nandakumar, 1993; Pae, 2004; Scheuneman & Gerritz, 1990).

2. We need to know what test-takers background variables interact with test items in what way.

3. We also need to search for the DIF detection methods that work well with small samples (Roever, 2005).

4. In particular, more studies are needed to tell us what to do with the items indicating DIF. Although Bridgeman, and Schmitt (1997) indicated that rewriting the DIF items would change them into good items, more studies are needed to possibly lead us to more vivid principles of what to do with DIF items: To rewrite them, eliminate them, ignore them? Or to include different DIF items in favor of different groups so that they would probably cancel each other out?

References

APA, AERA, & NCME (American Psychological Association, American Education Research Association, and National Council on Measurement in education). (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Ary, D., Jacobs, L. C., & Razavieh, A. (1996). *Introduction to research in education (5th ed.)*. Fort Worth: Harcourt Brace College Publishers.

Becker, B. J. (1990). Item characteristics and gender differences on the SAT-M for mathematically able youths. *American Educational Research Journal*, *27*(1), 65-87.

Bolger, N., & Kellaghan, T. (1990). Method of measurement and gender differences in scholastic achievement. *Journal of Educational Measurement*, *27*(2), 157-64.

Brantmeier, C. (2003). Second language reading research on passage content and gender: challenges for the intermediate level curriculum. *Foreign Language Annals*, *34*, 325-33.

Bridgeman, B. & Schmitt, A. (1997). Fairness issues in test development and administration. In Willingham, W. W & Cole, N. S. (Eds.), *Gender and fair assessment*, (pp. 185-226). Mahwah, NJ: Lawrence Erlbaum.

Bridgeman, B., & Wendler, C. (1991). Gender differences in predictors of college mathematics performance and in college mathematics course grades. *Journal of Educational Psychology*, *83*, 275-284.

Bugel, K. & Buunk, B. P. (1996). Sex differences in foreign language text comprehension: the role of interest and prior knowledge. *Modern Language Journal*, *60*, 15-31.

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

Carlton. S. T., & Harris, A. M. (1992). Female/male performance differences on the SAT: Causes and correlates. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Choi, I. C. & Bachman, L.F. (1992). An investigation into the adequacy of three IRT models for data from two EFL reading tests. *Language Testing*, *9* (1), 51-78.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: issues and practice*, *17*, 31-44.

Clauser, B. E., Mazor, K. M., & Hambleton, R. K. (1991). Influence of the criterion variable on the identification of differentially functioning test items using the Mantel-Haenszel statistic. *Applied Psychological Measurement*, *33*, 453-64.

Cole, N. S. (1997). *The ETS gender study: How males and females perform in educational settings.* Princeton, NJ: Educational Testing Service.

College Entrance Examination Board (2001). *A profile of SAT program test takers*. New York: author.

Cooke, D. J. & Michie, C. (1997). An item response theory analysis of the Hare Psychopathy Checklist-Revised. *Psychological assessment*, 9, humanities-sciences 4.

Curley, W. & Schmitt, A. P. (1993). Revisiting SAT verbal items to eliminate differential item functioning. *College Board Report 93-2*. New York: College Entrance Examination Board.

Donlon, A. (1973). Content factors in sex differences on test questions. NJ: Educational Testing Service, *Research Memorandum*, 73-28.

Donlon, T. F., Hicks, M. M, & Wallmark, M. M. (1980). Sex differences in item responses on the Graduate Record Exam. *Applied Psychological Measurement, 4*, 9-20.

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72*, 19-29.

Floyd, P., & Carrell, P. L. (1987). Effects on ESL reading of teaching content schemata. *Language Learning, 37*, 89-108.

Fotos, S. S. (1991). The cloze test as an integrative measure of EFL proficiency: a substitute for essay on college entrance examination? *Language Learning, 41*(3), 313-36.

Gallagher, A. M., De Lisi, R. & Holst, P. C., McGillicuddy-De Lisi, A. V., Morely, M., & Cahalan, C. (2000). Gender differences in advanced mathematical problem solving. *Journal of Experimental Child Psychology, 75*, 165-90.

Gokiert, R. J. & Ricker, K. L. (2002). Gender DIF and the WISC-III 1 Analysis of the Canadian Standardization Sample. *Center for Research in Applied Measurement and Evaluation, University of Alberta*.

Halpern, D. F. (1992). *Sex differences in cognitive abilities* (2 ed.), Hillsdale, NJ: Lawrence Erlbaum.

Hamilton, L. S. (1997). Identifying differential item functioning on science achievement tests. Paper presented at the annual meeting of the National Council on measurement in education, Chicago.

Hamilton, L. S. & Snow, R. E. (1998). Exploring differential item functioning on science achievement tests. *CSE Technical Report 483*.

Harris, A. M., & Carlton, S. T. (1993). Patterns of gender differences on mathematics items on the Scholastic Aptitude Test. *Applied Measurement in Education, 6*, 137-151.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: application to psychological measurement.* Homewood, IL: Dow Jones-Irwin.

Hyde, J. & Linn, M. (1988). Gender differences in verbal ability: a meta-analysis. *Psychological Bulletin, 104* (1), 53-69.

Irvine, P., Atai, P., & Oller, J. W., Jr. (1974). Cloze, dictation and the test of English as a foreign language. *Language Learning, 24*, 245-52.

Kanarek, E. A. (1988). Gender differences in freshman performance and their relationship to use of the SAT in admissions. Paper presented at the Northeast Association for Institutional Research Forum.

Lane, S., Wang, N., & Magon, M. (1996). Gender-related differential item functioning on a middle-school mathematics performance assessment. *Educational Researcher, 15,* 2 1-27,31.

Lawrence, I. M., Curley, W. E. & McHale, F. J. (1988). Differential item functioning for males and females on SAT verbal reading subscore items. *College Board Report, No: 88-4*. New York: College Entrance Examination Board.

Le, V. N. (2000). *Exploring gender differences on the NELS: 88 history achievement tests*. Unpublished doctoral dissertation, Stanford: Stanford University.

Lin, J. & Wu, F. (2003). Differential performance by gender in foreign language testing. Poster for the 2003 annual meeting of NCME in Chicago.

Linn, M. C., De Benedictis, T., Delucchi, K., Harris, A., & Stage, E. (1987). Gender differences in national assessment of educational progress science items: what does "don't know" really mean? *Journal of Research in Science Teaching*, 24(3), 267-278.

Liu, C. C., Shu, T. W., & Jeng, F. S. (1998). The comparison of accuracy for the software of IRT. *Journal of Educational Measurement and Statistics, 6*, 1-12.

Maller, S. J. (2001). Differential item functioning in the WISC-III: item parameters for boys and girls in the national standardization sample. *Educational and Psychological Measurement, 61*, 793-817.

Mazzeo, J., Schmitt, A. P., & Bleistein, C. A. (1993). Sex-related performance differences on constructed-response and multiple-

choice sections of Advanced Placement Examinations. *College Board Report*, 92-7.

McNamara, T. F. (1991). Test dimensionality: IRT analysis of an ESP listening test. *Language Testing, 8*(2), 139-59.

McNamara, T. F. (1996). *Measuring second language performance*. London: Addison Wesley Longman Limited.

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. *Journal of Educational Measurement, 19*, 73-90.

Oller, J. W. Jr. & Conrad, C. (1971). The cloze procedure and ESL proficiency. *Language Learning, 21*, 183-96.

Oller, J. W., Jr. (1979). *Language tests at schools: a pragmatic approach*. London: Longman.

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

Pae, T. & Park, G. P. (2006). Examining the relationship between differential item functioning and differential test functioning. *Language testing, 23*(4), 475-96.

Pae, T. (2004). Gender effect on reading comprehension with Korean EFL learners. *System, 32*, 265-81.

Paulston, C. B., & Brudes, M. N. (1976). *Teaching English as a second language: techniques and procedures*. Cambridge: Winthrop

Perrone, M. (2006). Differential Item Functioning and Item Bias: Critical Considerations in Test Fairness. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics, 6(2), the Forum*. Retrieved from http://www.tc.columbia.edu/tesolalwebjournal.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: results and implications. *Journal of Educational Statistics, 4*, 207-30.

Reeve, B. B. (2003). *An introduction to modern measurement theory*. Retrieved from http://appliedresearch.cancer.gov/areas/cognitive /immt.pdf. on 6 July 2006.

Reise, S. P. & Waller, N. G. (1990). Fitting the two parameter model to personality data: the parameterization of the multidimensional personality questionnaire. *Applied Psychological Measurement, 14*, 45-58.

Roever, C. (2005). That's not fair! Fairness, bias, and differential item functioning in language testing. *SLS Brownbag*. Retrieved 16/3/2006 from http://www2.hawaii.edu/~roever/brownbag.pdf.

Roznowski, M. & Reith, J. (1999). Examining the measurement quality of tests containing differentially functioning items: do biased items result in poor measurement? *Educational and Psychological Measurement*, *59*, 248-69.

Schellenberg, S. J. (2004). Bias or cultural bias: Have we really learned anything? *National Association of Test Directors, Annual Proceedings*.

Scheuneman, I. D., & Gerritz. K. (1990). Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. *Journal of Educational Measurement*, *27* (2), 109–31.

Schmitt, A. P., Holland, P. W., & Dorans, N. J. (1993). Evaluating hypotheses about differential item functioning. In Holland, P. W. & Wainer, H. W., (Eds.), *Differential item functioning*, (pp. 281-315). Hillsdale, NJ: Lawrence Erlbaum.

Seleshi Z. (2001). Gender differences in mathematics performance in the elementary grades: implications for women's participation in scientific and technical occupations. *Eastern Africa Social Science Research Review*, *7*(2), 107-27.

Shimizu, Y. & Zumbo, B. D. (2005). A logistic regression for differential item functioning primer. *Japan Language Testing Association Journal*, 7, 110-24.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589-617.

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

Teresi, J. A. (2006). Different approaches to differential item functioning in health applications: Advantages, disadvantages and some neglected topics. *Medical Care*, *44* (11), 15 math-sciences 70.

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

Uiterwijk, H. & Vallen, T. (2005). Linguistic sources of item bias for second generation immigrants in Dutch tests. *Language Testing*, *22*(2), 211-23.

Wightman, L. F. (1998). An examination of sex differences in LSAT scores from the prospective of social consequences. *Applied Measurement in Education*, *11*(3), 255-277.

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

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

Zumbo, B. D. (2007). Three generations of DIF analyses: considering where it has been, where it is now, and where it is going? *Language assessment quarterly*, *4* (2), 223-33.

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