Score Variation in Multiple-Choice Tests of Grammar: On the Effect of Gender and Stem Type

Mohsen Shirazizadeh[*] [1], Gholam Reza Kiany[2]

[1] *English Department, Alzahra University, Tehran, Iran*

[2] *English Department, Tarbiat Modares University, Tehran, Iran*

*Corresponding author: m.shirazizadeh@alzahra.ac.ir

Received: 2016.2.20
Revisions received: 2016.4.4
Accepted: 2016.6.13

Abstract

This paper examines the effect of gender and type of stem on Iranian test takers' performance on the grammar section of a nation-wide language proficiency test. In so doing, the scores of 2931 examinees (1107 female and 1824 male) who sat Tarbiat Modares English proficiency test were obtained. The examinees' scores on three types of MC grammar items with different kinds of stems (i.e., blank filling, error recognition, and cloze) were compared to see if the type of stem has any effect on performance. Grammar scores of males and females were also compared to see if there is an effect for gender on the examinees' performance on the grammar test in general and on its three item types in particular. The results indicated that test takers performed better on cloze items than the other two types. It was also found that females outperformed males on both the whole test, and also on items with blank filling and cloze types of stems. Due to the particularity of the context and the small effect sizes found, the study calls for more research to be conducted on this topic.

***Keywords*:** Score variation, Gender, Stem type, Grammar tests

Introduction

Several factors have been recognized as affecting the performance of examinees on language tests. Many of such factors have been found to be part of the construct being investigated and thus necessary to be included. Other factors have, however, been considered as contaminating test scores with construct irrelevant variance, and thus, jeopardizing validity and reliability. Two of the factors being much discussed as varying test scores are test method or test format (Alderson, 2000; Buck, 2001; Kobayashi, 2002) and test takers' individual characteristics (Chen & Henning, 1985; Farhady, 1982; Lumley & O'Sullivan, 2005). The present study intends to examine *different types of stems* as well as *test takers' gender* as two possible sources of variation of multiple-choice grammar test scores.

Different aspects of how to ask the question are certainly influential on what the response is and how it is formulated. In language testing, this is pointed out by Bachman (1990) who emphasizes that "[p]erformance on language tests varies as a function both of an individual's language ability and *of the characteristics of the test method"* (p.113; italics not in the original). This influence of different aspects of the item on test scores has been the focus of investigation by many researchers. One of the pioneering studies investigating the effect of item format is that of Shohamy(1984). She compared multiple-choice and open ended items in reading comprehension tests and found that multiple choice items were consistently easier than open ended questions. Wolf (1993) compared three methods of testing reading comprehension and found that test takers performed much better on the multiple choice items than on the open-ended and cloze test items. Tsagari (1994) also reported that multiple-choice items are easier than constructed response items in a reading comprehension test. More recently, Cheng (2004) concluded that test takers perform significantly better on multiple choice items than constructed response items in a listening comprehension test. More interestingly, multiple-choice items were found to be easier than constructed response items even when the test takers were asked to construct the response in their native language.

In fact, a vast majority of the studies conducted on various aspects of test format have focused on the response part of the item and not the stem.

One of the few studies which examined the effect of different types of stems is that of Dávid (2007)who compared different types of multiple choice grammar items regarding the amount of information each can provide about students with different ability levels. The results of this study revealed that multi-track items (i.e., a type of item in which the test taker should choose the option which is grammatically wrong) provided more information about the candidates at different ability levels than did standard multiple choice items, multiple choice cloze test items and double blank multiple choice items. In fact, the focus on the response part of the item has resulted in losing sight of the importance of the stem in some studies investigating the effect of test format (In'nami & Koizumi, 2009; Rodriguez, 2003). In the present study, we will, therefore, focus on the effects of different types of stems on test takers' performance on multiple- choice grammar items.

Language test scores might vary due to not only test-related factors but also test taker-related factors. Bachman (1990), while emphasizing the role of test method as a source of performance variation, maintains that "the effects of different test methods themselves are likely to vary from one test taker to another" (p.113). These factors influencing test score include test takers' gender, age, native language, educational, cultural and ethnic background, background to name a few. Of these personal attributes affecting performance, examinees' gender has been one of the most widely studied.

A large number of gender studies conducted on proficiency tests have reported that females usually outperform males (James, 2010). For example, females were reported to score higher on International English Language Testing System (IELTS). This outperformance was observed in both the total score and the four sections of the test (University of Cambridge, 2006). Similarly, Educational Testing Service (2007) reported that females scored slightly higher on the Test of English as a Foreign Language Internet-Based Test (TOEFL iBT) administered between September 2005 and December 2006. However, this report indicated that females outperformed males in total test score and all the main sections except the reading comprehension part in which males performed better than females. The mean score of female test takers on Michigan English Language Assessment Battery

(MELAB) was also higher than that of men based on the data from 2007 (Johnson & Song, 2008). Female outperformance was also reported in all four skill parts of Canadian Academic English Language (CAEL) tests administered between 2002 to 2008 (Carleton University, 2009). James (2010) examined the effect of gender on performance on Accuplacer ESL tests. She found that females performed better than males on the whole test and on three of its five subtests. Kunan(1990), however, examining a multiple choice university placement test, contrary to many other studies, revealed that 20% of the listening, reading, grammar and vocabulary items were in favor of males.

On the other hand, some studies have investigated the effect of gender on language test performance while focusing on type and content of the question. Lawrence, Curley, and Hale (1988) and Lawrence and Curley (1989), for instance, studied the gender differences in the verbal section of scholastic aptitude test (SAT), and found that males outperformed females on items related to technical reading passages. Carlton and Harris (1992), in another study on SAT, found that females outperformed males on items categorized under the rubric of aesthetics and human relationships across antonyms, analogies, and sentence completion questions. Males however, performed significantly better on items whose content were related to science and the practical matters. In another study on reading comprehension, Pae (2004) reported that females performed better on items relating to mood, impression, and tone while men did better on passages requiring logical inference. The report by Breland, Bridgeman, and Fowles (1999) indicated that males outperformed females on multiple-choice parts of TOEFL while females did better on the essay portion of the test. Takala and Kaftandjieva (2000) in a DIF (i.e., Differential Item Functioning) study on an L2 vocabulary test concluded that many items were biased for either of the sexes while the test on the whole was found to be gender-neutral. Some other vocabulary studies have, however, found their tests to be either in favor of females (e.g. Cole, 1997) or in favor of males (Born & Lynn, 1994; Lynn & Dai, 1993).

Based on the studies reviewed above, the type of the question or item format seems to have a key effect on examinees' performance on various

aspects of communicative language ability. This is also implied that the effect of item format on performance is probably not the same for males and females. In other words, the literature supports the fact that how the question is formulated and whether the examinee is male or female affect the response. In view of the fact that tests of grammatical ability have not been sufficiently investigated in terms of the effect of item format and gender, this study breaks new ground by examining how the format of the stem as well as test takers' gender influence performance on a multiple-choice test of grammar. Thus, the present study set out to answer the following questions:

1. What is the effect of the stem type on test takers' performance on a multiple-choice test of grammar?
2. What is the difference between the performances of males and females on multiple-choice tests of grammar with different types of stems?

Method

Participants

The participants in this study all sat the December 2010 administration of Tarbiat Modares University English proficiency test (TMUEPT). They included 2931 examinees, 1107(37.8%) females and 1824 (62.2%) males. The examinees were all MA or MS students or graduates of various fields of studies, applying for a PhD position in the following academic year (i.e., 2011). They were from different ethnic and L1 backgrounds (e.g., Farsi, Turkish, Kurdish, Lori, etc.). All examinees were, however, fluent speakers of Farsi since it is the national language and the medium of instruction in Iranian universities.

Instrumentation

The instrument used in this study to collect the data was the grammar section of TMUEPT administered in December, 2010. The test is administered as preliminary requirement for taking PhD examination in different fields at Tarbiat Modares University, Tehran, Iran. It is held biannually by the examination center of the university and the scores are

valid for taking part in PhD examination at different universities in Iran. Test takers are required to gain a minimum score of 50 out of 100 on the whole test to be allowed to take part in field-specific PhD examination and interview. The whole test is composed of grammar (40 items), vocabulary (30 items) and reading comprehension (30 items) parts respectively. Each item is followed by four alternatives out of which examinees should choose the most appropriate one. The focus of the test is basically on assessing the participants' general academic reading ability. The scores of the grammar section are analyzed in the present study. These Items of grammar included three different types of stem as described below:

1) *Blank filling items*: In these items, the examinees were provided with one sentence with a single blank to be filled by choosing the appropriate alternative from among the four choices. The alternatives included various grammatical forms such as different tenses of a verb or different parts of speech of a root. There were a total of 15 items of this type in the grammar section. An example of blank filling item is:

   Neither the table nor the chairs……clean.

   A) is B) are C) being D) been

2) *Error recognition items*: In this type of items, the stem consisted of one or two sentences which contained four underlined parts. Of these four underlined parts, the one containing a grammatical error should be chosen by the test takers as the correct response. In fact, in this type of item, the alternatives were embedded in the stem itself. There were a total of 15 items of this type in the grammar section. An example of error recognition item is:

   <u>After the negotiations</u>, they made <u>little</u> changes in <u>their proposal</u>.

   A B C D

3) *Cloze-based items*: The stem of these items consisted of a short text which had ten mutilations. The examinees were asked to choose the option which best filled in the respective blank in the text. The cloze section was similar to blank filling items in that both had the same

form of alternatives. They, however, differed in the amount of context they provided for the test takers and the extent of local dependence among the items, with items of the cloze section being more dependent on each other. A sample of a cloze item is provided below:

It is predicted that policing in the future will be more different than it is today. Advances in technology, particularly in computers, televisions, and ***(1)…….****will assist the police* ***(2)……...****solving and preventing crimes……*

1. A) communicates B) communicating
C) communicative D) communications
2. A) in B) on C) to D) at

It should be noted, once again, that the difference among the above categories is related to the stem of the items. They, however, shared the same type of response, that is, four-option multiple-choice response.

Procedure

TMUEPT was administered in the morning and under the usual security precautions. The students were allowed to take only a pencil with them in the test session. The responses were recorded in an answer sheet on which the personal information of the examinees was already printed. The examinees were given 100 minutes to answer the 100 questions of the total test. No one was permitted to leave the session before the end of the time. The answer sheets were scored at the examination center of the university. The test takers' raw scores on each of the parts and subparts were calculated on the basis of the total number of correctly answered questions with no penalty for guessing. The examinees' scores on the grammar cloze part were multiplied by 1.5 to make all the three set of scores out of 15, and thus, comparable.

Results

Our analysis focused only on the grammar section of the test which included items with three different types of stem as described above. To compare thetest takers performance on the three types of grammar items, one-way repeated measures ANOVA was used. The difference between the means of males and females on the total grammar part was calculated by t-test. To compare the performance of males and females on each of the three item types, one-way MANOVA was applied. In all statistical analyses, $P<0.05$ was considered statistically significant. The analyses were carried out by SPSS version 18. Table 1 summarizes the descriptive statistics for the variables of the study.

Table 1

Descriptive Statistics of the sores

	Gender	Mean	SD	N
Blank filling	Female	7.12	2.52	1107
	Male	6.86	2.48	1824
	Total	6.96	2.50	2931
Error recognition	Female	7.11	2.57	1107
	Male	6.89	2.58	1824
	Total	6.97	2.58	2931
Cloze	Female	8.58	3.15	1107
	Male	8.21	3.26	1824
	Total	8.35	3.22	2931
Total grammar	Female	22.82	6.67	1107
	Male	21.97	6.77	1824
	Total	22.29	6.74	2931

One-way repeated measures ANOVA was conducted to compare the test takers' scores on the multiple-choice grammar items with different types of stems. As shown in Table 2, ANOVA with a Huynh-Feldt correction revealed that there was a significant effect for the type of stem on examinees' performance, $F (1.81, 5860) =459.38$, $p<.005$; Partial Eta Squared= .136.

Table 2
ANOVA Results: Multivariate Tests

	Effect	Value	F	Sig.	Partial Eta Squared
Item type	Pillai's Trace	.192	348.59	.000	.192
	Wilks' Lambda	.808	348.59	.000	.192
	Hotelling's Trace	.238	348.59	.000	.192
	Roy's Largest Root	.238	348.59	.000	.192

Post hoc tests using the Bonferroni correction revealed that test takers performed significantly better on items with cloze type of stem than the other two types ($p<.005$). No significant difference ($p>.05$) was found between the examinees' score on blank-filling and error recognition items (see Table 3).

Table 3
ANOVA Results: Pairwise Comparisons

(I) itemtype	(J) itemtype	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference	
					Lower Bound	Upper Bound
Error Recognition	Blank	.013	.044	1.000	-.092	.117
	Cloze	-1.380*	.058	.000	-1.519	-1.242
Blank	Error Recognition	-.013	.044	1.000	-.117	.092
	Cloze	-1.393*	.056	.000	-1.527	-1.259
Cloze	Error Recognition	1.380*	.058	.000	1.242	1.519
	Blank	1.393*	.056	.000	1.259	1.527

To investigate the difference between male and female test takers with regard to their total grammar score, an independent sample t-test was employed. As can be seen in Table 4, females ($\underline{M}$= 22.82, $\underline{SD}$=6.67) significantly outperformed males ($\underline{M}$= 21.97, $\underline{SD}$=6.77), $\underline{t}(2929)=3.31$, $p=.001$; Partial Eta Squared= .0037.

Table 4
Summary of the Results of Independent Samples T-Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Grammar Total	Equal variances assumed	.029	.864	3.31	2929	.001	.85	.256	.347	1.354
	Equal variances not assumed			3.32	2359.03	.001	.85	.255	.349	1.352

To further analyze the effect of gender on performance on each of the grammar item types, One-way MANOVA was conducted. The preliminary analysis of the data indicated that there was a statistically significant difference between males and females on the combined dependent variables, F(3, 2927)=3.79; p=.01; Wilks' Lambda= .99; Partial Eta Squared= .004. (see Table 5).

Table 5
MANOVA Results: Multivariate Tests

Effect		Value	F	Sig.	Partial Eta Squared
Intercept	Pillai's Trace	.913	10191.37	.000	.913
	Wilks' Lambda	.087	10191.37	.000	.913
	Hotelling's Trace	10.44	10191.37	.000	.913
	Roy's Largest Root	10.44	10191.37	.000	.913
Gender	Pillai's Trace	.004	3.79	.010	.004
	Wilks' Lambda	.996	3.79	.010	.004
	Hotelling's Trace	.004	3.79	.010	.004
	Roy's Largest Root	.004	3.79	.010	.004

The Post hoc separate analysis of each of the dependent variables (i.e., the three types of grammar items), with the Bonferroni adjusted alpha level of .017, showed that females significantly outperformed males on blank filling , F(1, 2929)=7.06, p=.008; Partial Eta Squared= .002, and cloze items, F(1, 2929)=9.12, p=.003; Partial Eta Squared= .003. (see Table 6).

Table 6
MANOVA Results: Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Gender	Blank	44.090	1	44.090	7.062	.008	.002
	Error Recognition	35.464	1	35.464	5.318	.021	.002
	Cloze	94.806	1	94.806	9.125	.003	.003

Note: P<0.017 is statistically significant; since there are three level for stem type, .05 should be divided by 3

Figure 1 illustrates the point vividly.

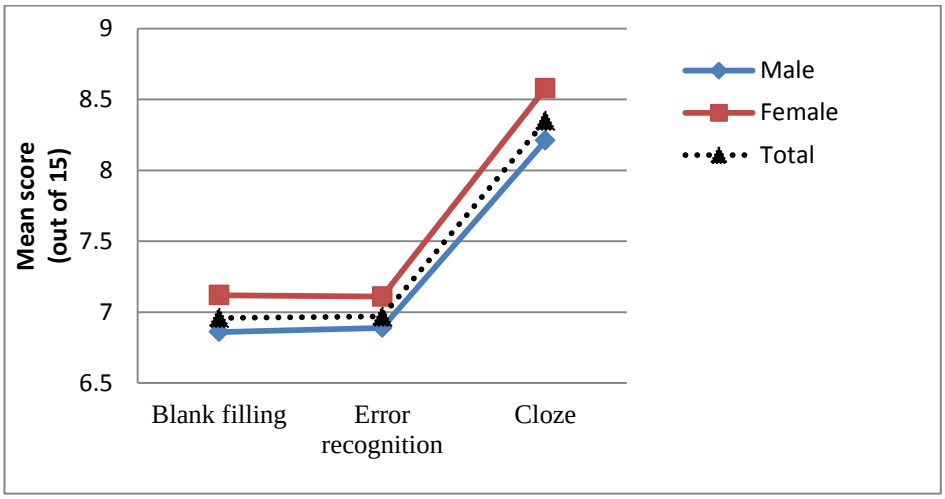

Figure 1
Test takers' performance on three types of grammar items

As indicated in Figure 1, females significantly did better than males on blank filling and cloze items.

Discussion

The findings of the analyses regarding the first research question showed that, in multiple choice items of grammar, type of stem had a significant effect on the test takers' scores. The examinees scored significantly higher on multiple choice cloze items than on the other two

types. No significant difference was found between error recognition and blank filling items. This finding shows that the features of the question, in this case different stems, can lead to noticeable variation in grammar test scores. This finding is quite in line with those of David (2007) who found that different types of grammar items would lead to variation in test scores, hence supporting the effect of item type. This variation in performance, which is due to aspects of the question, has been reported by many other researchers as well (e.g., Cheng, 2004; Wolf, 1993).

In our study, however, the focus was only on multiple choice items of grammar that differed in their stems. The test takers' scores were highest on items based on the cloze passage. This outperformance of the examinees on cloze based items can be justified by considering the fact that in these items, there is a much wider context available to the examinees. This availability of context can enable them to make more educated guesses as to the response of each of the questions. The context will, in fact, give them more clues as to, say, the dominant tense of the passage, the grammatical function of many words, the tone (i.e., passive vs. active) of the sentences etc., all of which are highly important in choosing the correct alternative that fills the blank parts of the text.

Further, our examination of the effect of gender on performance revealed that females performed consistently better on the three types of grammar items, hence, the total grammar section. This outperformance of females has also been frequently reported by many researchers (James, 2010; Johnson & Song, 2008). However, in only two of the item types, namely blank-filling and cloze-based items, were differences between males and females statistically significant. It should be noted that in both of these two types of items, the examinees should fill in the gap by choosing the grammatically correct alternative. They are, however, different from each other in the amount of context they provide. Both types of items, therefore, require a more or less holistic view of the stem. In error recognition items, on the other hand, examinees should have a more analytical ability to diagnose the erroneous part in the four underlined parts of the sentence. It can, therefore, be maintained that cloze-based and blank-filling items have more touch of reading ability within them, and that it is probably this

reading-based nature which makes them easier for females since some studies like that of James (2010) and the report by Carleton University (2009) have reported that reading comprehension is one of areas in which females considerably outperform males.

There are, however, some very important cautions to be taken into account when interpreting the findings of this study. Some of these points are discussed in the following section.

The results of this study indicated that both stem type and examinees' gender would affect performance on multiple-choice tests of grammar. These differences in performance were however too small to be considered practically and educationally significant. Given the large number of participants in this study, mean differences and corresponding P values are certainly less than definitive. This issue is very clear when we check the effect sizes, all of which are too small to attract any attention (Cohen, 1977).

In addition to small effect sizes, the findings are also limited by a number of other points which should be taken into account when making any interpretation thereof. First of all, the content of the questions for each of the three stem types were different. Differences in performance may, therefore, not be due to the effect of stem type but the content of the question. No content analysis at item level was done to control for such an effect. A DIF analysis of the items could shed light on the effect of item characteristics on performance. However, such an analysis required another rather different study from the present one. Many individual characteristics of the examinees were also not controlled in this study. Factors such as examinees' field of study, age, native language and many others are all possible confounding variables which could have exerted their influence upon gender differences found in this study. For example, it was not known in the study whether males and females were similar in proficiency level before taking the test. Although the large size of the sample might have neutralized any intervention of proficiency level as far as gender effects on performance is concerned, no definitive claim can yet be made that gender-based differences found here are not affected by proficiency level. In fact, the outperformance of females might be due to differences in language

ability (i.e., the construct being measured) and not gender (i.e. one of the possible construct irrelevant factors).

Another noteworthy issue is that the test was a criterion-referenced one. The examinees were only required to answer half of the questions correctly to pass the test. There was also no penalty for wrong answers. Therefore, many test takers might have entirely discarded some parts of the test to have more time for other parts or they might have answered many items by guessing. This fact could have substantial effect on participants' performance on different parts and subparts of the test. Another issue limiting the findings of the study is examinees' test-wiseness. Some of the participants were taking the test not for the first time as they had failed in previous administrations. As a result, a number of test takers were more familiar than others with the structure, timing and conditions of administration of the test. This test-wiseness might also have influenced their performance, hence the findings of the study.

Some general points can be concluded from the present study. First of all, significant differences found in large samples should be approached with much caution (Cohen, 1977). Second, any difference between males and females should be investigated from many different perspectives to make possible robust interpretations. Various types of analyses at the item and test level as well as controlling for possible confounding variables would make the findings much reliable and generalizable. This study could be considered a preliminary step in better investigating and understanding the effects of test method facets and individual characteristics on performance on language tests. Studies like the present one would provide us with information that could be utilized in "designing tests that are less susceptible to such effects, that provides the greatest opportunity for test takers to exhibit their 'best' performance, and which are hence better and fairer measures of the language abilities of interest" (Bachman, 1990, p. 156). Many more multi-faceted studies are yet required to gain a thorough understanding of the effects of such factors.

References

Alderson, C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

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

Born, M., & Lynn, R. (1994). Sex differences on the Dutch WISC-R: A comparison with the USA and Scotland *Educational Psychology, 14*(2), 249–255.

Breland, H. M., Bridgeman, B., & Fowles, M. E. (1999). Writing assessments in admission to higher education: Review and framework (Report No. 99-3). New York, NY: College Entrance Examination Board.

Buck, G. (2001). *Assessing listening*. Cambridge: Cambridge University Press.

Carleton University. (2009). CAEL test score and users' guide . Ottawa, Canada: Author. Retrieved from http://www.cael.ca/edu/testuserguide.shtml

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, 92–64. Princeton, NJ: ETS.

Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing, 2*(2), 155-163.

Cheng, H. f. (2004). A comparison of multiple-choice and open-ended response formats for the assessment of listening proficiency in English. *Foreign Language Annals, 37*(4), 544-553.

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

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

Dávid, G. (2007). Investigating the performance of alternative types of grammar items. *Language Testing, 24*(1), 65-97.

Educational Testing Service. (2007). Test and score data summary for TOEFL Internet-based test: September 2005-December 2006 test data. Princeton, NJ. Retrieved from www.ets.org/toefl.

Farhady, H. (1982). Measures of language proficiency from the learner's perspective. *TESOL Quarterly, 16*(1), 43-59.

In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing, 26*(2), 219-244.

James, C. L. (2010). Do language proficiency test scores differ by gender? *TESOL Quarterly, 44*(2), 387-398.

Johnson, J. S., & Song, T. (2008). MELAB 2007 descriptive statistics and reliability estimates. Ann Arbor, MI: English Language Institute, University of Michigan.

Kobayashi, M. (2002). Method effects on reading comprehension test performance: text organization and response format. *Language Testing, 19*(2), 193-220.

Kunnan, A. J. (1990). DIF in native language and gender groups in an esl placement test. *TESOL Quarterly, 24*(4), 741-746.

Lawrence, I. M., & Curley, W. E. (1989). *Differential Item Functioning for males and females on SAT-Verbal Reading subscore items: follow-up study.*: Princeton, NJ: Educational Testing Service.

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

Lumley, T., & O'Sullivan, B. (2005). The effect of test-taker gender, audience and topic on task performance in tape-mediated assessment of speaking. *Language Testing, 22*(4), 415-437.

Lynn, R., & Dai, X. Y. (1993). Sex differences on the Chinese standardiz-ation sample of the WAIS-R. *Journal of Genetic Psychology, 154* (4), 459–464.

Pae, T.-I. (2004). DIF for examinees with different academic backgrounds. *Language Testing, 21*(1), 53-73.

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: a random effects synthesis of correlations. *Journal of Educational Measurement, 40*(2), 163-184.

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing, 1*(2), 147-170.

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

Tsagari, C. (1994). *Method effects on testing reading comprehension: How far can we go?* . Unpublished MA thesis, University of Lancaster, UK.

Wolf, D. F. (1993). A comparison of assessment tasks used to measure fl reading comprehension. *The Modern Language Journal, 77*(4), 473-489.

Biodata

Mohsen Shirazizadeh is an assistant professor of Applied Linguistics at the English department of Alzahra University. His main areas of interest include corpus linguistics, phraseology, psycholinguistics and language testing.

Gholam Reza Kiany is an associate professor of Applied Linguistics at the English department of Tarbiat Modares University where he teaches postgraduate courses on Language testing and evaluation, and quantitative research methods. He is interested, in particular, in EFL program and teacher evaluation, large-scale testing and computer adaptive tests.