Language Testing in Asia

RESEARCH Open Access

Impact of test format on vocabulary test performance of EFL learners: the role of gender

Ghazal Akhavan Masoumi and Karim Sadeghi*

* Correspondence: k.sadeghi@urmia.ac.ir; kerimsadeghi@gmail.com
Urmia University, Urmia, Iran

Abstract

This study aimed to examine the effect of test format on test performance by comparing Multiple Choice (MC) and Constructed Response (CR) vocabulary tests in an EFL setting. Also, this paper investigated the function of gender in MC and CR vocabulary measures. To this end, five 20-item stem-equivalent vocabulary tests (CR, and 3-, 4-, 5-, and 6-option MC) were administered to 243 (132 male and 111 female) pre-intermediate students. Results of the study revealed that MC tests were easier than CR. Results indicated a gender bias, in that, males scored better than females in all versions of MC tests while females outperformed males in CR. The findings implied that testers should consider the effect of test format while assessing vocabulary knowledge and use a combination of test formats (MC and CR) in vocabulary assessment to reduce gender bias and format effect.

Keywords: Multiple choice, Constructed response, Vocabulary, Gender bias, Format effect

Introduction

Assessment is an integral part of any educational system, and it plays an important role in English as a foreign language (EFL) context. One of the components of language that is challenging for test constructors is vocabulary, a skill without which a learner cannot understand or communicate in a foreign language. Word knowledge or vocabulary repertoire is a fundamental component of language proficiency and also an essential component of communicative competence and acts as a vital element for production and comprehension in a second language (Coady & Huckin, 1997; Richards & Renandya, 2002).

Numerous researchers have highlighted the importance of vocabulary knowledge in their writings. Wilkins (1972), for example, asserted that "while without grammar very little can be conveyed; without vocabulary nothing can be conveyed" (pp. 111–112). His assertion highlighted the importance of vocabulary and placed it above grammar. According to Luppescu and Day (1993), it is necessary for students to build a large repertoire of vocabulary when learning a language because people with large

vocabularies are more proficient and competent than those with limited vocabularies. In other words, students with a good knowledge of vocabulary can communicate in a foreign language much better than low proficient students. Baker, Simmons, and Kame'enui (1998) emphasized that learning a language is mainly dependent on vocabulary and word knowledge; in other words, a high repertoire of word knowledge is required to be competent in a foreign language. Huyen and Nga (2003) and Zhao (2009) also highlighted that vocabulary (and its mastery) plays an important role in learning a second language, and without mastery over vocabulary, none of the other skills (listening, reading, speaking, and writing) are attainable.

Vocabulary is accordingly an indispensable component of language and any language education program. Indeed most researchers (Read, 2000; Zhao, 2009) agree that its mastery plays an important role in the process of language learning. Consequently, the practice and manner of its measurement becomes vital. In other words, examining the vocabulary knowledge of students correctly and providing accurate and relevant information about the process of language learning and teaching are imperative. Vocabulary knowledge is considered as a psychological attribute or mental ability which cannot be measured or observed directly, and special techniques are needed to measure it. Psychological attributes or constructs are "hypothetical concepts - products of the informed scientific imagination of scientist who attempts to develop theories for explaining human behavior" (Crocker & Algina, 2006, p. 4). Constructs are not visible, and indirect methods are used to measure them; the process of observing constructs through indirect methods is called operationalization. In other words, it is a way of moving from the abstract level to empirical level (Lewis-Beck, Bryman, & Liao, 2004).

A test is considered as an apparent example of the indirect method mentioned above, and it refers to the processes and procedures used by the tester to obtain information about the optimal performance of stakeholders or typical performance of individuals (Crocker & Algina, 2006). Different kinds of tests, from multiple-choice to gap filling, have been used to assess vocabulary knowledge in different levels of proficiency; however, the current study will mainly focus on selected response or Multiple Choice (MC) format and Constructed Response (CR) format.

Literature review

As mentioned above, the present study aimed to investigate the impact of format of the test, MC and CR, on the performance of students in vocabulary tests in an Iranian context. According to Bachman (1990), test format or test method should not interfere with the construct being measured. We believe this issue is of utmost importance for test developers since a slight change in the performance of test takers due to format of test will color the results of test in either a negative or a positive way, and the results will not be the real measures of students' abilities anymore. So, the statement "one test fits all" may not be true in language testing.

In this regard, reviewing some studies conducted in this field would be valuable. The literature on MC and CR tests from a quantitative perspective has mainly focused on two issues: "(a) differences in construct or trait measured using multiple-choice and open-ended formats and (b) differences between test scores in multiple-choice and open-ended formats (i.e., the relative difficulty of test formats)" (In'nami & Koizumi, 2009, p. 221). Also, according to Nixon and Kennedy (2002), there are three major

streams in comparisons of MC and CR test formats. The first and most prominent stream addresses and investigates whether MC and CR tests measure the same construct or not. The second stream deals with "how to scale or link MC and CR scores so as to create a single total score from one type of exam with those of another" (p. 959). The last stream, on which this study focuses, addresses whether tests in one format are more difficult than their equivalents in other formats.

Nixon and Kennedy (2002) compared the scores of students in stem-equivalent MC and CR tests of economics and showed that students do indeed score much better in MC. Gender in this study did not have an effect on the performance of students. Hastedt and Sibberns (2005) compared MC and CR test formats in Trends in International Mathematics and Science Study (TIMSS, a series of assessments of mathematics and science knowledge of students around the world) for 1995 and 1999. They observed only small differences between MC and CR scores, and based on such an observation, researchers suggested that "using MC and CR items in international studies, because it guarantees that test takers are treated equally and fairly" (Hastedt & Sibberns, 2005, p. 159). Gender analysis suggested that females performed better in CR while males outperformed in MC item format. Famularo (2007) revealed that there were significant differences between MC and CR items and MC tests were found to be much easier than their CR counterparts.

Liu, Lee, and Linn (2011) explored the function of explanation multiple choice (EMC) and showed that EMC and MC items were easier than CR, but EMC items were harder than MC items. Hickson, Reed, and Sander (2012) used a data set composed of thousands of observations on individual students in economics classes at a public university. They found that instructors paid too much attention to writing CR questions that assess higher level learning, but actually all these efforts were in vain, since little difference was found between MC and CR scores.

Shaibah and van der Vleuten (2013) compared the scores of students in MC and Free Response Format (FRF) (a version of CR) in a gross anatomy course. A Rasch model was utilized, and analysis revealed a strong correlation between MC and FRF scores. Shaibah and van der Vleuten suggested that MC test is a valid method which can be used as an alternative to FRF items. Moreover, their results showed that students scored better in recall MC tests than FRF understanding items.

Sangwin and Jones (2017) compared the performance of students in MC and CR tests in an online test and found that the overall score of students were higher in MC compared with CR. Gamer and Engelhard (1999) examined gender differences in the performance of students on MC and CR items. Researchers found that there were gender differences in doing MC and CR items. They explained that females performed relatively better than males in CR format while males outperformed females in MC item format.

Weaver and Raptis (2001) investigated the performance of male and female students in nine introductory atmospheric and oceanic science exams over 7 years. The analysis of performance of 295 male and 194 female students who participated in the study showed that there were no significant differences between the performance of male and female students. In a study conducted by Bacon (2003), gender differences in MC and short answer tests were studied. Results of t test suggested that there were no significant differences between the performance of male and female students. Taylor and Lee

(2012) compared the performance of male and female students in reading and math test items selected from state criterion-referenced tests with MC and CR items. Results of the study revealed that in both reading and math tests females did better in CR while males performed better in MC. Reardon, Kalogrides, Fahle, Podolsky, and Zárate (2018) investigated the association between test item format and gender achievement gaps on math and English language arts tests in fourth and eighth grades. They found that MC and CR tests measuring the same underlying constructs may rank the performance of males and females differently. In other words, gender gaps were sensitive to item format, in that males did better on MC tests and females performed better in CR.

Although there are numerous studies addressing the effect of test format on performance of test takers in different disciplines such as mathematics, psychology, and economics (e.g., Birenbaum & Tatsuoka, 1987; Hastedt & Sibberns, 2005; Simkin & Kuechler, 2005), few studies have been conducted in applied linguistics (AL) on the impact of test format on the performance of test takers in language tests in general and vocabulary tests in particular. For example, In'nami and Koizumi (2009) conducted a meta-analysis of the effect of test format on L1 and L2 reading, and L2 listening performance. Results of the study suggested that MC test was easier than CR in L1 reading and L2 listening. Although multiple-choice formats were found to be easier than open-ended formats, format effect in L2 reading was not observed.

Currie and Chiramanee (2010) compared the effectiveness of MC against CR items in the context of English language education in Thailand. Results of the study suggested that students achieved significantly higher scores in MC test than their stem-equivalent CR based on which Currie and Chiramanee mentioned that MC and CR tests were not measuring the same constructs. In other words, "a more realistic implication to draw from these results would be that the M/C format had the effect of distorting the measurement of language based abilities which were used by the participants in answering C/R items" (p. 485). The researchers, however, do not provide further explanation or any justification on why MC items had a distorting effect on measurement but CR items did not.

Performance of test takers on language tests is affected by different sources of variance which Bachman (1990) grouped into four broad categories: Communicative language ability (CLA), test method facets (TMF), personal attributes (PA), and random factors (RF). While the ultimate aim of language testing is to measure CLA, the outcome is often compromised by the other three (TMS, PA, and RF). Test makers have little or no control on random factors; however, they can exercise some control over test method facets and test-taker attributes. Test method facet or test format refers to the characteristics of tests or test tasks which are used to elicit information about test takers' knowledge about a matter (Bachman, 1990). In other words, Bachman believed that the performance of test takers on language tests was mainly the product of both an individual's language ability and the facets of test method employed (Weir, Vidaković, & Galaczi, 2013). This study draws on Guttman's (1980) facet theory and Bachman's (1990) test methods framework and aims to investigate the "type of response" facet (i.e., CR vs. MC), which is a characteristic of the "format" of the "expected response".

This study mainly investigates the effect of test method facet (test format) and personal characteristics (gender) on vocabulary test performance. For this purpose, we

decided to study the most important state wide high-stakes examination in Iran. This Iranian national university entrance exam is called "Konkur". Students in Iran sit for this MC exam (including a vocabulary test) and based on the results of the test, students are admitted to public universities. This test is a 4.5-h MC exam that covers all subjects taught in Iranian high schools, from math and science to Islamic studies and foreign languages. The exam is so high-stakes that students normally spend a whole year preparing for it. For the purpose of this study, we selected the English language section of this national test which consists of MC vocabulary and grammar items, a cloze test and a reading passage, followed with MC comprehension items. Also, more specifically, we selected only the vocabulary section (with four choices). For the parallel CR versions, we omitted the choices presented (as well as adding further choices to MC or dropping a choice when required). Since this exam is the most important examination in Iran affecting the future lives and of more than a million students and their families, we decided to examine one part of it (vocabulary section) in light of the facets proposed by Bachman. It should be mentioned that to the best of the researchers' knowledge, no similar study has been conducted based on this specific exam in Iran.

This study was accordingly an attempt to bridge a gap in the field of language testing by delving into the effect of test method effect on performance in vocabulary tests in an EFL setting in Iran. Furthermore, since research findings on the relationship between test format and gender are contradictory and inconclusive, the current project was aimed at investigating the differential function of gender (if any) across CR and MC test formats. In this regard, the following research questions were proposed:

Q1: Is there any significant difference between the performance of EFL students in MC and CR vocabulary tests?
Q2: Is there any significant difference between the performance of male and female students in MC versus CR vocabulary tests?

Methods

Participants and setting

The participants in the current study were 258 (140 male and 118 female) fourth year high school students within the age range of 17-18. The participants attended public high schools in Iran, where as part of their compulsory education, they received two hours of English education every week.

Instruments

The following four instruments were utilized in this research study:

Proficiency test

In order to guarantee the homogeneity of the participants in terms of language proficiency, an adapted version of Key English Test (KET) for schools (updated in 2009) was utilized. Before being given to the main study students, the adapted KET was administered to 25 students similar to the target group and the KR-21 reliability of test was calculated to be 0.78.

Vocabulary Pre-test

This vocabulary test was selected from Cambridge Key English Test 4 Self Study Pack (KET Practice Tests) (2006) by Cambridge ESOL. The test consisted of 24 MC vocabulary items with 3-options which was piloted with 24 students and enjoyed a KR-21 reliability of 0.83.

MC tests

Four versions of MC tests (constructed based on a CR test, see below) were used in the study. The CR and MC tests utilized in this study were stem-equivalent; in other words, all of them shared the same stems. The contents of these tests were based on the materials that the students studied and covered at school and matched their level of proficiency. Different versions of MC tests utilized in this study only differed in the number of options (which were based on the incorrect responses of students in the CR test). A 20-item 6-option MC test was constructed and piloted with 30 students similar to the target group and its KR-21 was estimated to be 0.87. The frequencies of all words which acted as answers and distractors were checked against Collins COUBUILD Advanced Learners' English Dictionary (2006), and were found to have similar frequency. After the 6-option test was administered to the pilot group, the least chosen distractors were omitted and 3-, 4-, and 5-option MC tests were constructed accordingly. These tests were piloted with 23, 25, and 30 students and their KR-21 was found to be 0.79, 0.82, and 0.85, respectively. These four versions of MC tests were also reviewed by two English language teaching (ELT) professionals, two high school English teachers, and two native speakers before being used in the main study. They approved of the appropriateness of the tests, and some minor revisions (on wording) were applied to a few items based on their suggestions.

CR tests

The fourth instrument used in this study was a 20-item CR test which was adapted from a 4-option vocabulary MC test in entrance examinations in Iran. For preparing this test, different entrance examination tests for Bachelor of Arts (BA) and Bachelor of Science (BS) from different years (2008-2016) were reviewed by the researchers, and faulty and problematic items were changed, revised or substituted after revision by an expert and after piloting. The final version of the test was reviewed by two ELT professionals, two high school English teachers, and two native speakers. It was piloted with 21 high school students similar to the target group and its Cronbach's Alpha was estimated to be 0.85. Exact-word scoring was utilized for scoring the test when it was used in the main study. Incorrect responses provided by test takers during the piloting of CR were used to construct appropriate distractors for the 6-option MC test.

Data collection

This study sought to determine the possible impact of test format on the performance of Iranian pre-intermediate fourth year high school students across gender. In order to conduct this study, ten intact classes (five male and five female) were chosen from different high schools in Urmia, Iran. To make sure that all the participants were homogeneous and of the same proficiency level, the researchers first administered an

Table 1 Design of the study

Step 1	Sampling	Proficiency test (KET)
Step 2	Pre-test	Vocabulary test
Step 3	Tests	CR
		3-option
		4-option
		5-option
		6-option

adapted KET for schools test (KR-21 reliability of 0.83). Based on the results of KET, six students were regarded as outliers and were excluded from the study, and later nine other students who did not participate in the vocabulary pre-test were omitted; consequently, the number of participants was reduced from 258 to 243. In this study, there were five parallel groups (based on their proficiency and vocabulary performance) for each gender: Each parallel group (male and its counterpart female group) received one test (either a CR format or one of four MC formats) on a random basis. That is, the first group of female participants and a parallel group of male students received one test in CR format. All the participants in these two groups had 15 min to answer 20 CR vocabulary items. The second parallel group of female and male students received MC vocabulary test with 3-options and the students in both groups had 10 min. to answer the stem-equivalent 3-option MC test. The CR group was given more time since they had to write their responses. The third, fourth, and fifth parallel groups of female and male students received MC vocabulary tests with 4-, 5-, and 6-options, respectively, through the same procedures (Table 1).

As shown in Table 2, after excluding the outliers and the students who did not participate in the vocabulary pre-test, the number of participants reduced from 258 to 243 (132 male and 111 female).

Results

A series of independent samples *t* test were conducted to compare the mean score of groups that took MC (3-, 4-, 5-, and 6-option) and CR test formats. Table 2 shows the descriptive statistics of the learners' vocabulary performance in 3-, 4-, 5-, and 6-option MC and CR tests.

As the mean and standard deviation scores in Table 3 shows, there are differences between EFL learners' performance in 3-option MC and CR tests. However, in order to get more accurate and reliable results, an independent samples *t* test was run, the results of which are displayed in Table 4.

So, in response to the first question about the differences in the performance of test takers in different test formats (3-, 4-, 5-, 6-option MC, and CR), it can be concluded

Table 2 Descriptive statistics: participants' profile

	CR	3-option	4-option	5-option	6-option	Total
Male	26	26	27	24	29	132
Female	23	23	20	23	22	111
Total	49	49	47	47	51	243

Table 3 Descriptive statistics for 3-, 4-, 5-, and 6-option MC and CR vocabulary tests

	Test	*N*	Mean	Std. deviation	Reliability
Vocabulary	CR	49	12.3878	3.63327	0.85
	3-option	49	17.1020	2.26610	0.79
	4-option	47	17.0000	2.17695	0.82
	5-option	47	15.5319	2.68531	0.85
	6-option	51	13.9608	2.93912	0.87

that the performance of test takers in all versions of MC tests were significantly better than their performance in stem-equivalent CR test (Table 5).

In order to examine the gender differences (second research question) in MC (3-, 4-, 5-, and 6-option) and CR test scores, a series of independent samples *t* test were conducted. Table 4 indicates the means and standard deviations for males and females in MC and CR tests.

Table 6 illustrates the results of *t* test statistics.

Overall, the results of a series of *t* tests indicate a significant difference between the performance of males and females in doing MC and CR formats. It can be stated that females performed better than males in CR format while males performed better than females in MC format.

Discussion

In this study, we have examined the effect of test format and gender on the vocabulary performance of fourth year high school (pre-university) students.

Results of the analysis (Table 4) suggested a significant difference between the performance of test takers in MC and CR tests; in other words, findings revealed that vocabulary performance of test takers varied based on test format (MC/CR), and that their performance was remarkably better in MC test formats. Findings suggest that test takers performed relatively better in selective response format than in constructed or productive format in the context of the current study.

The findings of the current study are in line and consistent with other response format studies that showed test takers perform relatively better in MC than their stem equivalent CR format (e.g., Currie & Chiramanee, 2010; Famularo, 2007; In'nami & Koizumi, 2009; Nixon & Kennedy, 2002). For example, Nixon and Kennedy (2002) carried out a study to compare the performance of test takers in stem-equivalent MC and CR tests. They found that test takers performed much better in MC items. Famularo (2007), also, compared the scores of test takers in MC and CR items, and their findings indicated that MC format was significantly easier than CR version of the same test. They found that test takers benefit from test taking strategies and corrective feedback

Table 4 Independent samples *t* test results for 3-, 4-, 5-, and 6-option MC and CR Differences

	F	Sig.	Mean difference	Std. error difference
Vocabulary 3-option and CR	25.373	0.000	-4.71429	0.61172
Vocabulary 4-option and CR	27.804	0.000	-4.61224	0.61454
Vocabulary 5-option and CR	13.510	0.000	-3.14416	0.65428
Vocabulary 6-option and CR	8.278	0.020	-1.57303	0.65962

Table 5 Descriptive statistics for gender in CR, 3-, 4-, 5-, and 6-option MC vocabulary test

	Gender	N	Mean	Std. deviation
CR	Male	26	9.3462	1.80980
	Female	23	15.8261	1.33662
3-option MC	Male	26	18.1538	1.59229
	Female	23	15.9130	2.35320
4-option MC	Male	27	17.7778	1.78311
	Female	20	15.9500	2.25890
5-option MC	Male	24	16.7500	2.06945
	Female	23	14.2609	2.70046
6-option MC	Male	29	15.0000	2.60494
	Female	22	12.5909	2.83950

while doing MC items; in other words, options in MC items provide some additional cues for test takers which facilitate answering MC items.

The findings of the current study also confirm those reported by Currie and Chiramanee (2010). They compared the performance of test takers in MC and CR English structure test. Their results indicated that scores of test takers were significantly better in MC than their stem-equivalent CR. Shaibah and van der Vleuten (2013) indicated that scores of test takers were better in MC. On the other hand, the studies done by Hastedt and Sibberns (2005) and Hickson et al. (2012) appear to contradict with the findings of this study. In their study, Hastedt and Sibberns (2005) observed only little differences between the scores and Hickson et al. (2012) found that item format did not affect the scores of test takers. The studies done by these researchers were related to different fields such as science, mathematics, and economics, and the differences in the characteristics, age, and knowledge of test takers could have led to different results.

The present study aimed to find the effect of test format on the performance of test takers. The comparison of mean scores suggested that test takers performed better in MC items. Based on the previous studies conducted (Cohen, 2012; Shohamy, 1984), the researchers believe that one of the reasons for this finding may be that test takers in doing MC and CR items make use of different skills and processes. While answering MC items, they just comprehend and select the option; while in doing CR items, they comprehend and produce an answer, which requires more processing (Shohamy, 1984). Furthermore, options in MC items provide additional information and cues for test takers, and by looking at options test-takers can remember and deduce the answer. Test taking strategies (TTS) refer to the processes that test takers have consciously selected in order to address language issues and item-response demands (Cohen, 2012). Test takers utilize a variety of strategies (such as facilitation and problem solving) to

Table 6 Independent samples *t* test results for gender in CR and MC vocabulary test

	F	Sig.	Mean Difference	Std. Error Difference
CR	2.798	0.000	− 6.47993	0.45965
3-option MC	0.990	0.000	2.24080	0.56824
4-option MC	0.791	0.003	1.82778	0.58941
5-option MC	2.934	0.001	2.48913	0.69995
6-option MC	0.660	0.003	2.40909	0.76563

improve their scores in an exam. Facilitation strategies help test takers facilitate doing a process (test tasks) and problem solving strategies are utilized when a problem comes in.

Also, test takers in doing MC items make use of a wide range of strategies such as elimination and test wiseness, which help them to answer and select the option. Test wiseness strategies are a sub category of construct irrelevant strategies (Cohen, 2012) which assist test takers in MC items to find possibly the correct answer among several distractors: stem-option cues, grammatical options, similar options, and item giveaway (Allan, 1992) are examples of test wiseness strategies utilized in MC items. All such skills, strategies, and cues facilitate in answering MC items, increasing the chance for a better score. Overall, the findings of this study suggest that employing MC and CR format in tests of vocabulary is likely to create format related noise or effect, and teachers are recommended to consider the cost and effectiveness of each format while choosing an appropriate format for measuring an intended construct like vocabulary. For example, in using MC format because of its practicality, teachers must weigh it against the risk that the measurement of construct is likely to be contaminated or affected by constructing irrelevant factors such as item format or guessing. It is recommended that teachers use a wide variety of formats such as CR, MC, matching, or cloze test while measuring a construct like vocabulary to decrease the effect of format-related factors.

Findings of this study showed that males performed better in all version of MC while females outperformed males in CR test. In other words, we found that MC vocabulary items may bias male students while CR items would bias female students. It should be noted that the issue of gender and its effects on educational tests have long been a concern for testers (Brown & McNamara, 2004).

However, there are mixed results in the literature. As a prime instance, Mauldin (2009); Simkin and Kuechler (2005) and Weaver and Raptis (2001) examined performance of students in MC and CR tests and found no significant differences between male and female test takers. However, findings of this study suggested that males perform better in MC items while females do better in CR items. Similar to our results, Taylor and Lee (2012) found that males performed better in items that asked them to identify interpretations (answers) while females did better in items that asked them to write their own answers and interpretations.

The results of the current study are consistent with findings of earlier studies (Bolger & Kellaghan, 1990; Gamer & Engelhard, 1999; Hellekant, 1994; Taylor & Lee, 2012). Findings of the current study also corroborate with those of DeMars (2000); Hastedt and Sibberns (2005)and Reardon et al. (2018) in that females perform better in CR while males outperform in MC items. Moreover, the findings of this research support those of Taylor and Lee (2012). Their study indicated that in reading and math tests, females did better in CR while males performed better in MC items. Furthermore, the results discussed above show that the findings of the present research are in line with claims of Elder (1998). She noted that females mostly perform better in CR rather than in MC items, which could be the result of their superior verbal ability, and also other factors which are unrelated to language abilities (e.g., handwriting). On the other hand, males perform better in forced choice test format or MC.

We cannot determine the exact reasons for the difference in the measured gender gaps on tests with different item formats, but we feel that the differences are large

enough to have meaningful consequences for students especially in Iranian context where the study was conducted. We hypothesize that one of the reasons for this gender difference in MC item may be related to different levels of confidence in males and females. Males are believed to be more confident and risk taking than females in doing MC items (Biria & Bahadoran Baghbaderani, 2015) while males trust the option they choose, females tend to doubt and change their answers several times which results in losing time and getting confused and more stressed in doing other MC items; on the other hand, in CR format unlike MC format females do not doubt or do not get confused and perplexed, so they rely on the first answer that they believe is right.

The current study is subject to a number of limitations. One limitation relates to the fact that the current study only investigated the performance of test takers in MC and CR test formats and due to operationalizability issues the researchers could not include other item formats. Also, the number of the participants and items utilized were limited (243 participants and 20 items). The researchers only analyzed the mean score of the students on the tests and further studies with other psychometric analyses that need to be conducted. Also, test taking strategies that the students exactly used for answering the MC and CR questions were not studied.

Considering the limitations of this study, further studies need to be conducted with more item formats, more participants so as to be able to recommend more generalizable findings and better alternatives for vocabulary test/formats. Furthermore, more research is needed to investigate whether male and female students differ in their answering strategy for MC and CR questions. The results of the current study were mainly based on the adapted MC (and CR) vocabulary questions used in a state wide entrance examination in Iran, and the findings may only be context specific. However, we believe that teachers and policy makers should consider the related format and gender-related differences in constructing MC and CR formats and develop test formats that are less biased toward a specific gender or use a wide range of formats to compensate for gender related differences.

Conclusions

The results of this study suggested that the performance of students were much better in MC vocabulary test than CR, and that they would have less difficulty in answering stem equivalent MC items The results of the study also showed gender differences in the performance of test takers in MC and CR formats; the findings suggested that females performed better in CR while males performed relatively better in MC test format. Based on these findings, another implication for teachers, testers, and policy makers would be taking into account individual and more specifically gender differences while constructing and developing test items for stakeholders. As mentioned earlier, for reducing bias against any gender, test developers are recommended to use different formats.

Abbreviations

BA: Bachelor of Arts; BS: Bachelor of Science; CR: Constructed response; ELT: English language teaching; KET: Key English test; MC: Multiple choice; TTS: Test taking strategies

Acknowledgements

We are thankful to the students who participated in this study and also the teachers that cooperated with us.

Authors' contributions
GAM and KS collected the data, did the statistical analysis, and wrote the manuscript. Both authors read and approved the final manuscript.

Funding
The authors received no specific funding for this work.

Availability of data and materials
The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests
The authors declare that they have no competing interests.

Received: 11 November 2019 Accepted: 17 February 2020
Published online: 10 March 2020

References

Allan, A. (1992). Development and validation of a scale to measure test-wiseness in EFL/ESL reading test-takers. *Language Testing, 9*(2), 101–122.
Bachman, L.F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
Bacon, D. R. (2003). Assessing learning outcomes: a comparison of multiple-choice and short-answer questions in a marketing context. *Jounal of Marketing Education, 25*(1), 31–36.
Baker, S., Simmons, D., & Kame'enui, E. (1998). *Vocabulary acquisition: synthesis of the research*. Washington: Educational Resources Information Center.
Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats– it does make a difference for diagnostic purposes. *Applied Psychological Measurement, 11*(4), 385–395.
Biria, R., & Bahadoran Baghbaderani, A. (2015). Exploring the role of risk-taking propensity and gender differences in EFL students' multiple-choice test performance. *Canadian Journal of Basic and Applied Sciences, 5*(3), 144–154.
Bolger, N., & Kellaghan, T. (1990). Method of measurement and gender diferences in scholastic achievement. *Journal of Educational Measurement, 27*, 165–174.
Brwon, A., & McNamara, T. 2004. The devil is in the detail: Resaerching gender issues in language assessment. *TESOL Quarterly, 38,* 524–538.
Coady, J., & Huckin, T. (1997). *Second language vocabulary acquisition*. New York: Cambridge University Press.
Cohen, A. (2012). Test taker strategies and task design. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 262–279). Abingdon: Routledge.
Crocker, L. M., & Algina, J. (2006). *Introduction to classical and modern test theory*. Mason: Cengage Learning.
Currie, M., & Chiramanee, T. (2010). The effect of the multiple-choice item format on the measurement of knowledge of language structure. *Language Testing, 27*(4), 471–491. https://doi.org/10.1177/0265532209356790.
DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education, 13*(1), 55–77. https://doi.org/10.1207/s15324818ame1301_3.
Elder, C. (1998). What counts as bias in language testing? *Melbourne Papers in Language Testing, 7*(1), 1–42.
Famularo, L. (2007). *The effect of response format and test taking strategies on item difficulty: a comparison of stem-equivalent multiple-choice and constructed-response test items*. Ann Arbor: Boston College Retrieved from http://search.proquest.com/docview/304897354?accountid=14645 ProQuest Dissertations & Theses Full Text database. (3283877 Ph.D.),.
Gamer, M., & Engelhard, J. G. (1999). Gender differences in performance on multiple-choice and constructed response mathematics items. *Applied Measurement in Education, 12*(1), 29–51. https://doi.org/10.1207/s15324818ame1201_3.
Guttman, L. (1980). Integration of test design and anlysis: Staus in 1979. *Directions for Testing and Measurement, 5,* 93–98.
Hastedt, D., & Sibberns, H. (2005). Differences between multiple choice items and constructed response items in the IEA TIMSS surveys. *Studies in Educational Evaluation, 31*(2), 145–161.
Hellekant, J. (1994). Are multiple choice tests unfair to girls? *System, 22*, 348–352.
Hickson, S., Reed, W. R., & Sander, N. (2012). Estimating the effect on grades of using multiple-choice versus constructive-response questions: data from the classroom. *Educational Assessment, 17*(4), 200–213. https://doi.org/10.1080/10627197.2012.735915.
Huyen, N. T. T., & Nga, K. T. T. (2003). The effectiveness of learning vocabulary through games. *Asian EFL Journal, 5*, 90-105. Retrieved from http://asian-efl-journal.com/quarterly-journal/2003/12/31/learning-vocabulary-through-games-the-effectiveness-of-learning-vocabulary-through-games/.
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. https://doi.org/10.1177/0265532208101006.
Lewis-Beck, M. S., Bryman, A., & Liao, T. F. (2004). *The Sage encyclopedia of social science research methods*. Thousand Oaks: Sage.
Liu, O. L., Lee, H., & Linn, M. C. (2011). An investigation of explanation multiple-choice items in science assessment. *Educational Assessment, 16*(3), 164–184. https://doi.org/10.1080/10627197.2011.611702.
Luppescu, S., & Day, R. R. (1993). Reading, dictionaries, and vocabulary learning. *Language Learning, 43*(2), 263–279. https://doi.org/10.1111/j.1467-1770.1992.tb00717.x.
Mauldin, R. K. (2009). Gendered perceptions of learning and fairness when choice between exam types is offered. *Active Learning in Higher Education, 10*(3), 253–264.
Nixon, C., & Kennedy, P. E. (2002). Are multiple-choice exams easier for economics students? A comparison of multiple-choice and "equivalent" constructed-response exam questions. *Southern Economic Journal, 68*(4), 957–971. https://doi.org/10.2307/1061503.
Read, J. A. S. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Reardon, S. F., Kalogrides, D., Fahle, E. M., Podolsky, A., & Zárate, R. C. (2018). The relationship between test item format and gender achievement gaps on Math and ELA tests in fourth and eighth grades. *Educational Researcher, 47*(5), 284–294.
Richards, J. C., & Renandya, W. A. (2002). *Methodology in language teaching: an anthology of current practice*. Cambridge: Cambridge University Press.
Sangwin, C. J., & Jones, I. (2017). Asymmetry in student achievement on multiple-choice and constructed-response items in reversible mathematics processes. *Educational Studies in Mathematics, 94*(2), 205–222.
Shaibah, H., & van der Vleuten, C. (2013). The validity of multiple choice practical examinations as an alternative to traditional free response examination formats in gross anatomy. *Anatomical Sciences Education, 6*(3), 149–156. https://doi.org/10.1002/ase.1325.
Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing, 147*(1), 147–170. https://doi.org/10.1177/026553228400100203.
Simkin, M. G., & Kuechler, W. L. (2005). How well do multiple choice tests evaluate student understanding in computer programming classes? *Journal of Information Systems Education, 14*(4), 389–399.
Taylor, C. S., & Lee, Y. (2012). Gender DIF in reading and mathematics tests with mixed item formats. *Applied Measurement in Education, 25*(3), 246–280. https://doi.org/10.1080/08957347.2012.687650.
Weaver, A. J., & Raptis, H. (2001). Gender differences in introductory atmospheric and oceanic science exams: multiple choice versus constructed response questions. *Journal of Science Education and Technology, 10*(2), 115–126.
Weir, C. J., Vidaković, I., & Galaczi, E. D. (2013). *Measured constructs: a history of Cambridge English language examinations 1913-2012*. Cambridge: Cambridge University Press.
Wilkins, D. A. (1972). *Linguistics in language teaching*. Cambridge: MIT Press.
Zhao, N. (2009). Metacognitive strategy training and vocabulary learning of Chinese college students. *English Language Teaching, 2*(4), 123–129.

Publisher's Note