JSLTE

Journal of Studies in Learning and Teaching English

Online ISSN: 2476-7727, Print ISSN: 2251-8541

https://jslte.shiraz.iau.ir/

13(1), 2024, pp. 1-14

Research Article

A Gender-Based Investigation into the Relationship between Test Method and Iranian EFL Test-Takers' Grammar Performance

Amir Reza Ali-Akbar [1], Shokouh Rashvand Semiyari [2]

1.2. Department of English Language Teaching, West Tehran Branch, Islamic Azad University, Tehran-Iran

* Corresponding author: **Shokouh Rashvand Semiyari,** Email: sh_rashvand@yahoo.com

ARTICLE INFO

Submission History

Received: 2023-03-11
Accepted: 2023-09-28

Keywords

Test Method
Grammar Performance
Gender

ABSTRACT

In the domain of educational assessment, comprehending the elements that shape test-takers' achievements is quite significant. This research delved into how test method and gender might affect grammar performance. To this end, 274 intermediate EFL learners in the 18-30 age range, studying in Qotb Ravandi Institute in Tehran took a grammar test in four different formats specified to comparatives, superlatives, and present perfect tenses. The results of the correlation analysis revealed that there was a positive correlation between total score (grammar performance) and error correction, word changing, word order, and completion scores. The results of regression analysis also indicated that gender was a significant predictor of grammar performance. There was a negative statistically significant correlation between gender and grammar performance, indicating that male students tended to score lower than their female counterparts. Furthermore, the predictor variable of grammar performance could accurately classify 63.6% of females and 31.3% of males in their groups, with the overall precision of the regression model being 50%. Therefore, it can be argued that there would be a statistically significant relationship between test-takers' gender and their grammar performance. The implications and suggestions for further studies were also highlighted.

Introduction

As far as a foreign language is concerned, it becomes essential to learn and understand the grammar of that language. Due to the debate that has been going on about the part grammar plays in language learning, some educators have decided to ignore teaching grammar, but those studies have proven to be inefficient. Every year millions of people join the massive group of language learners who fall into two main categories; the first group relates to volunteers who choose to learn English language for academic or occupational purposes and the second group has to do with those who are

studying English through the curriculum system and their textbooks (Chambers & Schilling, 2013). As Close (1982) defines, "English grammar is chiefly a system of syntax that decides the order and patterns in which words are arranged in sentences" (P. 13). However, definitions of grammar differ greatly depending upon one's knowledge or orientation. The question of "what is grammar?" is not the first thing that comes to our minds when we are studying a new language, yet it is the first thing we often learn in academic settings. When thinking about grammar's definition, we usually think of "a set of rules that govern a language" which is true; however, such a definition does not even scratch the surface of what grammar really is. Grammar is not unchangeable and there are factors that can direct such changes, including literature, culture, and time. It can be different from one language to another and even from one individual to another (Debata, 2013). It provides us with the information we need in order to measure the effectiveness of our teaching and context. Additionally, our society has been characterized by diversity, with gender representing a notable dimension of this variety. Grasping these subtleties can enhance students' ability to interact adeptly with native speakers and showcase their cultural proficiency. In essence, there is a direct relationship between gender and language learning (Mirzaei & Rahimi, 2016). As Sumami and Rachmawaty (2019) stated, males usually tend to use more analytical, while females prefer to use more communicative learning strategies. They believed there are different reasons for gender differences in language proficiency among which we can refer to social and cultural factors. By providing multiple methods of testing for students, targeting a particular aspect of language, and considering the moderating role of gender, the researchers aimed to see whether different types of test methods could help students with their performance in grammar.

Review of the Related Literature

The Significance of Teaching and Testing Grammar

Under no circumstances can grammar teaching be ignored since Grammar is the bedrock of achieving proficiency in a language. When the proper knowledge of grammar is introduced to students, they can improve their levels of English language proficiency. If teachers are reluctant about grammar teaching and unwilling to explore grammar teaching methods, they can hinder language learning process (Zhang, 2009). Testing has always been of great value to all educators. It is an integral part of teaching and they are somehow inseparable. However, test taking can be stressful especially when the outcome of a test is not similar to what a student has anticipated (Dikmen, 2023). Language tests can provide teachers with information they need in order to measure the effectiveness of their teaching. Moreover, language learners may assume that the only use of testing is to measure their skills. Therefore, some may adopt a negative attitude towards it; hence, it is important to help them realize testing is not used only to measure the students. In fact, students' assessment is not solely related to determining their language proficiency, but to see whether the testing methods applied or even the teaching of an educator have actually been influential or not.

Testing and Evaluation: Definitions and Backgrounds

The act of teaching involves imparting ideas from one or more individuals to a significant number of other individuals. A teacher's job is to guide learners and facilitate this process. Testing is utilized in order to measure the effectiveness of the teaching and the proficiency that learners have developed (Paudel, 2018). A common misconception is that only language teachers need to know and learn about assessment. However, it is critical for not only language teachers, but everyone to understand the principles of language assessment for multiple reasons (Olmezer-Ozturk & Aydin, 2018). Firstly, language tests are extremely important in people's lives. They can influence many key moments in our lives such as education, employment, and emigration. Secondly, language educators need the required information in order to assess and grade students on specific courses. Lastly, in order to conduct research in language study, students' language proficiency must be measured first (McNamara, 2000). As Douglas (2010) explains, "a test is a measuring device, no different in principle from a ruler, a weighing scale,

or a thermometer. A language test is an instrument for measuring language ability" (p. 2). By taking tests, we provide equal opportunities for everyone to demonstrate his progress. Tests also provide information that can help teachers confirm their assessments and gain more confidence in making decisions by giving them a 'second opinion' about students' progress. The basis of achievement tests is the materials that students learn or whether students have learned what they should have learned after finishing a course of study. These tests can show administrators that students are learning and therefore making progress, and teachers are accomplishing their duties (Ozan & Kincal, 2018). Proficiency tests, however, give us information that help predict students' performance in situations outside the classroom (Douglas, 2010). People's views on language and language use directly influence language testing development (Mao, 2022). Pedagogical and research functions which testing provides for teachers, have made it an inseparable part of language teaching. Testing has proven to play an effective role in increasing the quality of teaching. The data retrieved from testing can help shape the instruction and if implemented correctly, testing can be applied as a strong engine of change (Roediger et al., 2011). Nowadays, a significant amount of money and thousands of hours are spent on administrating the standardized tests.

Evaluation in General Concept

A common misconception about evaluation and testing is that they are the same, while testing is only a part of the process of evaluation. The first thing that comes to mind when we talk about evaluation might be schools, curriculum, or examinations. But evaluation is not limited to education and it is used in our everyday lives. When we are listening to a speech, radio, or an interview, we are constantly making judgments about the speakers. Whether consciously or unconsciously, we make evaluations on a daily basis. In education, however, the judgments we make must be based on explicit criteria and evaluation must be systematic, because the validity and reliability of the educational decisions that we make are dependent upon the process of evaluation.

There are two major evaluation categories, general and specific topic-related purposes. Three reasons can be also mentioned for general evaluation including accountability, curriculum design, and self-improvement. Evaluation for purposes of accountability is related to whether something has been efficient and effective. It views everything from an economical angle. The information gathered from the evaluation of accountability provides valuable information for sponsors and heads of institutes and it is not particularly helpful to curriculum development or classroom practice. Evaluation for purposes of curriculum and design is mostly concerned with teachers since they have key roles in contributing toward curriculum development and renewal. The information teachers have about the context and their evaluation of a classroom is far more comprehensive than a test taker or a test designer. Evaluation for purposes of self-improvement is not concerned with the measurement and mainly focuses on paying attention to the process rather than the product. It helps teachers understand what is actually happening in the classrooms rather than what is supposed to happen (Cordeiro, 2021).

Teachers and students are not the only components of a class. There are other factors such as textbooks, classroom settings, and available resources to a teacher. Rea-Dickins and Germaine, (1992) highlight in order to figure out whether teaching and learning programs are working, the evaluator needs to be very specific about what needs to be examined. They believe two factors are influential in making the best possible decisions; the decision maker's ability and the information on which his or her decision is based. This information can also be gathered by monitoring the students' performance and the overall impressions; thus, evaluation is not limited to testing (Bachman, 1990).

As mentioned before, evaluation is an important part of teaching and learning, but it is also an important part of testing. Test-makers typically analyze tests to remove the weak items and design different test formats even before students come across them. By doing so and having a positive

feeling toward the test, class attitude, students' motivation and even their performance will be enhanced (Madsen, 1983). Douglas (2010) points out that in order to evaluate the learners, we do not necessarily need measurement. He further adds a teacher's impression of a student's improvements and language skills is evaluation without measurement. Teachers sometimes give students some scores to show administrators and parents that they are making progress. These scores may be based on various activities such as class participation and homework, but there is no test involved. This is called measurement without a test. Tests come in different forms including selected-response, constructed-response, etc. to assess students' learning.

Selected vs. Constructed-Response Tests

In selected-response tests, several items are given to test-takers only one of which is correct. The test takers must find and choose the correct answer in order to gain the score specified to that test item (Onaiba & Jannat, 2019). According to McNamara (2000), although selected-response tests are very efficient to administer and score, they don't measure productive skills such as speaking and writing. These tests are typically designed to assess either a particular component of language (e.g., grammar and lexical range) or measure the students' general understanding such as listening and reading comprehension (McNamara, 2010). One of the most commonly known selected-response tests is multiple-choice test. In multiple-choice tests, the examiner provides several options for the test-takers and they have to choose the correct option (Douglas, 2010). An apparent advantage of selected response tests, is the scoring process which is quite rapid and simple (Hughes, 1998). Another advantage of multiple-choice items is the vast number of students it can measure. This is why multiple-choice items are very popular among educators.

There are, however, some downsides to using multiple choice items according to Bush (2001). It is educationally important to know whether a student is able to produce the required answer without having the correct one as an option. For instance, a student may not be able to find the error in a sentence if the error is not presented as an option. Producing more difficult questions and making the incorrect answer more likely to be correct will not make up for this limitation. It might mislead a student-who otherwise could answer the question without being exposed to possible answers- to choose the wrong answer that was very close to the correct option. As a result, a different category of testing items, known as constructed response test was introduced by Livingston (2009).

In constructed-response tests, test-takers are expected to construct the correct answer to the designed questions, instead of choosing the right answers. The response that is given to a constructed-response test may be more extended compared to a selected-response test. Constructed-response tests are also much more suitable in order to test writing skills and paragraph development. As advantageous as constructed-response tests can be, there are some shortcomings regarding these tests (Livingston, 2009). It is very difficult to score them in an accurate and reliable way (Onaiba & Jannat, 2019). Four different types of tests have been used in this study, all of which are constructed-response tests including completion, error correction, word changing, and word order items, each has been explained in brief as follows:

Completion Items; according to Kitao and Kitao (1996), in completion items function words such as prepositions and articles are left blank and it is on test-takers to fill them correctly. An advantage of using this type of test is encouraging production rather than recognition (Kitao & Kitao, 1996). According to Sireci and Zenisky (2016), this category comprises many specific item formats. They are used in various assessment contexts and need test-takers to generate their own answers and complete one or more blanks. One disadvantage of using such items is that they are marked by hand and sometimes require raters to make judgments (Sireci & Zenisky, 2016).

Error Correction Items; in this type of test, students are given a sentence in which there are some errors. In some cases, the errors are underlined and in others, the errors are not marked and test-takers must find them. Teachers may also ask students to correct the mistakes. One of very good sources of error correction items for teachers

are mistakes and errors students make in their writing (Kitao & Kitao, 1996). This type of test is mostly used in computer-based language tests, yet it comes to language testing as a variation of multiple-choice test or in forms of short answer items (Dolan et al., 2011).

Word Changing Items; in these types of grammar test items, in order to answer the question correctly, students must have knowledge of different word forms and the way they are used in different sentences. For example, the base form of a verb in a sentence is presented, and the correct form of the verb which completes the sentence is required. Word-form tests can be served as reliable and valid measures of word-formation knowledge, and that they are usually associated with overall language proficiency (Kitao, & Kitao, 1996).

Word Order Items; whether students are asked to write something or alternative answers are provided, word order items are dependable methods for testing grammar. Such items not only show students' knowledge of related grammar, but by giving students several sentences and asking them to put them in order, cohesive devices and knowledge of references can also be measured (Kitao & Kitao, 1996).

Language Learning and Gender

Gender has an important impact on the language learning strategies the learners employ (Aslan, 2009). He believes female students usually employ more cognitive and metacognitive strategies, such as summarizing and self-monitoring, while male students tend to use more memory and social techniques, such as repetition and group work. Moreover, female learners tend to use more language learning strategies and achieve better success in learning English grammar compared to their male counterparts (Otayf, 2019). Holmes (2007) proposed different viewpoints on gender and language, including the dominance model, the difference model, and the social constructionist model. The dominance model regards language as a means of oppression, while the difference model focuses on gender-related variances in language use. The social constructionist model considers gender as a social construct that is continuously negotiated through language. These differences might arise due to socialization, cultural factors, and individual learner preferences. It is therefore crucial to acknowledge such differences when designing language testing techniques.

Empirical Background

Many studies have investigated the role of gender on grammar to-date (e.g., Azizmohammadi & Barjesteh, 2020; Beller & Gafini, 2000; Mozaffari, et al., 2017; Zoghi, et al., 2013). Almost in all of the studies, it has been shown that female students outperformed male students in grammar performance. Results of the study carried out by Pope et al. (2006) showed that female students outperformed male students in language achievement tests while male students scored higher in mathematics. In another study conducted by Pomplun and Capps (1999), it was demonstrated that female students outperformed male students in constructed-response items regarding reading comprehension. Regarding gender gap in multiple-choice items and open-ended items, a study by Beller and Gafni (2000), highlighted that, male students performed much better in multiple choice items while female students had better performance regarding open-ended items.

Likewise, a plenty of studies have investigated the effect of different test methods on students' performance (e.g., Bensoussan, 1984; Bleske-Rechek et al., 2007; Bridgeman, 1992; Cheng, 2004; Haynie, 1994). Even though previous research shows male students perform better than female students in selected response items, nearly all of the studies demonstrate that all students generally perform much better in selected-response rather than constructed-response items. The effect of testing method on different language components has also been studied by many scholars. The study done by Shohamy (1984) on the effect of the testing method on measuring reading comprehension showed that the more difficult the test method was, the greater effect it had on students, specifically those with lower proficiency levels. In another case, Akhavan Masoumi and Sadeghi (2020) examined how different test formats, including multiple-choice and constructed response, impacted the

performance of students in vocabulary tests in Iran. The study showed that the test format did not affect the construct being measured, as even a small variation in test takers' performance could significantly impact the test results. DeKeyser (1993) investigated how test format (error correction) affected grammar performance. The findings revealed that test format was useful in evaluating grammar knowledge. These results indicate that test format selection can considerably influence the evaluation of grammar knowledge and that teachers and test developers should take this into account when creating grammar tests for EFL learners.

Research Questions

Testing is an integral part of teaching and they are somehow inseparable. Test taking can be stressful and when the outcome of a test is not similar to what a student has anticipated, it can cause dissatisfaction and discouragement. The outcomes of a test can be a great source to any teacher to evaluate his own performance. Not all teachers benefit from enough experience or constructive supervision to ensure that the teaching and testing methods they have applied were effective (Pienemann & Brindley, 1989). To address such gap empirically, this research intended to find out how test method would influence test takers' grammar performance by taking the moderating role of gender into account. Accordingly, the following research questions were formulated:

RQ1. Is there any significant relationship between test method and test-takers' grammar performance?

RQ2: Is there any significant relationship between test-takers' gender and their grammar performance?

Methodology

Participants

The sample consisted of 274 Iranian EFL students (141 female and 133 male students) who were studying English at Qotb Ravandi Institute in Tehran. Their age range was between 18 to 30 years old (their mean age was 23.54 (SD= 2.23)). Students' levels of language proficiency were intermediate as selected by an Oxford Placement Test (OPT) and administered by the researchers at the beginning of the study. Participants took the Grammar test in different formats while measuring the same content. Table 1 summarizes demographic data of the participants:

Table 1.
Demographic Background of the Participants

No. of Students	274
Gender	Males (133) and Females (141)
Proficiency Level	Intermediate
Native Language	Persian
Institute	Qotb Ravandi
Academic Year	2021-2022

Instruments

To meet the purposes of the research, the researchers used the following research instruments:

Oxford Placement Test (OPT)

To determine the students' levels of language proficiency, an OPT (version 1) was administered in the beginning of the study. This test is often used by researchers as the language proficiency test in which participants' scores are ranked according to the test norms from beginners to upper intermediate levels. The OPT consists of two parts with 60 items in the form of multiple-choice questions and cloze tests. The first part consists of 40 questions measuring learners' grammar knowledge and the second part consists of 20 questions assessing learners' vocabulary knowledge. The allocated time for this test was 60 minutes.

Grammar Test

In order to check the effect of different test methods on students' grammar performance, a grammar test was given to the students. Three

grammatical contents were selected to be tested in different methods. The selected grammatical contents included comparatives, superlatives, and present perfect. For each grammatical content, three questions were extracted in four different test method formats, coming to a total of 12 questions for each grammar content. There were three error correction, three completion, three word order, and three words changing items. Grammar test comprised 36 questions in total among which 10 were drawn from Topnotch 1B, 13 originated from Topnotch 2A, and an additional 13 were selected from the Touchstone 3.

Design

The design of the study was descriptive correlational design. To achieve the goals of the study, the correlational analysis was carried out to examine the relationships among the variables under investigation. The amount and degree of the relationships were also presented. The regression analysis was also conducted to indicate whether the moderating variable of the study could be used to predict any changes in the dependent variable. In this study, students' grammar score was the dependent variable, different test methods were the independent variable, and gender was the moderating variable.

Procedure

Data were collected in several phases. Firstly, the homogenization applied in order to evaluate learners' general English knowledge. The participants were totally 274 Iranian EFL students (141 female and 133 male students). All of the students were measured through the Oxford Placement Test (OPT). Due to the scores of this test (from 60 points), students with one SD above and below the mean were selected as the suitable proficient levels (intermediate). After doing the first part, the grammar test was given to the participants. Three grammatical contents including comparatives, superlatives, and present perfect were selected to be tested in different methods of error correction, completion, word order, and words changing items. Each testing method comprised 9 questions and the whole grammar test included 36 questions in total. Participants were assured that their personal information would be kept confidential and would be only used for research and not for any other purposes.

Data Analysis

After administering the grammar tests to participants, to analyze, interpret, and report the findings, Correlation Coefficient and Regression Analysis were applied. Correlation coefficient analyses are used in science to assess the degree of association between two variables while the logistic regression evaluates predictors of dichotomous outcomes, i.e., outcomes that either occurred or did not.

Results

Analysis Results of the First Research Question

The first research question of this study was as follows:

1. Is there any significant relationship between test method and test-takers' grammar performance?

To answer this research question, a Pearson Product Moment Correlation Coefficient was used to calculate the correlations between scores of different types of tests (i.e., error correction, word changing, word order, completion) and test-takers' grammar performance (i.e., the total score). Descriptive statistics of the variables in the correlation analysis are presented in Table 2.

Table 2.
Descriptive Statistics of Variables in the Correlation Analysis

	Minimum	Maximum	Mean	Std. Deviation	Skewness	
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error
Error Correction	0.00	9.00	6.88	2.32	-.98	.15
Word Changing	2.00	9.00	7.11	1.95	-.68	.15
Word Order	1.00	9.00	7.12	1.81	-.93	.15
Completion	2.00	9.00	6.70	1.69	-.30	.15

	Minimum	Maximum	Mean	Std. Deviation	Skewness	
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error
Grammar Performance	13.00	36.00	27.81	6.42	-.69	.15

Table 3 shows that there are positive statistically significant correlations between total score (grammar performance) and error correction score (r = .84, p =.00), word changing score (r = .87, p = .00), word order score (r = .76, p = .00), and completion score (r = .80, p = .00). These findings indicate that as the total score of the participants increase, their scores in different test methods should increase in turn (see the positive significant correlations between the scores of different test methods as well in the matrix). Therefore, it can be concluded that there is a positive statistically significant relationship between test method and test-takers' grammar performance, and the null hypothesis (i.e., there is no significant relationship between test method and test-takers' grammar performance) is rejected.

Table 3.
Correlation Matrix of Different Testing Methods and Grammar Performance

		Grammar Performance	Error Correction	Word Changing	Word Order	Completion
Pearson Correlation	**Grammar Performance**	**1.00***	**.84***	**.87***	**.76***	**.80***
	Error Correction	.84*	1.00*	.73*	.41*	.55*
	Word Changing	.87*	.73*	1.00*	.57*	.56*
	Word Order	.76*	.41*	.57*	1.00*	.60*
	Completion	.80*	.55*	.56*	.60*	1.00*
p value	Grammar Performance		.00	.00	.00	.00
	Error Correction	.00		.00	.00	.00
	Word Changing	.00	.00		.00	.00
	Word Order	.00	.00	.00		.00
	Completion	.00	.00	.00	.00	

* Significant at lower than 0.05

Analysis Results of the Second Research Question

The second research question of this study was as follows:

2. Is there any significant relationship between test-takers' gender and their grammar performance?

In order to answer this research questions, a binary logistics regression analysis was used. More specifically, this type of regression was utilized to shed light on the correlation between grammar performance and gender of the participants. It should be mentioned that in this regression model, gender, a criterion variable, was a binary variable (categorical) with two levels, that is, male and female, and grammar performance was the predictor variable (a continuously measured variable).

Table 4.
Descriptive Statistics of Variables in Regression Analysis across Gender Types

Female				
	Minimum	Maximum	Mean	Std. Deviation
Total	13.00	36.00	28.75	6.37

Male				
	Minimum	Maximum	Mean	Std. Deviation
Total	13.00	36.00	26.83	6.34

As can be seen in Table 5, gender was the significant predictor of grammar performance (B= -.04, p = 01). More specifically, as Wald test was showing, there was a negative statistically significant correlation between gender and grammar performance. Since we coded female as 1 and male as 2 in the analysis, this means that male students tended to score lower than their female counterparts in the test. Pertaining to the prediction power of the grammar performance score, as can be seen in Table 6, this predictor could accurately classify 63.6% of females and 31.3% males in their groups, with the overall precision of the regression model being 50%. In the main, it can be argued that there is statistically significant relationship between test-takers' gender and their grammar performance, and hence the related null hypothesis (i.e., there is no significant relationship between test-takers' gender and their grammar performance) is rejected.

Table 5.
Regression Weight and its Wald test

		B	S.E.	Wald	df	*p* value	Exp(B)
Regression	**Total**	**-.047**	**.019**	**6.062**	**1**	**.014**	**.954**
	Constant	1.278	.551	5.371	1	.020	3.589

Table 6.
Classification Table for Logistic Regression

			Predicted		
			Gender		
Observed			Female	Male	Percentage Correct
	Gender	Female	89	51	63.6
		Male	92	42	31.3
	Overall Percentage				47.8

Discussion

The purpose of this study was to investigate the relationship between test method, test-takers' gender, and their grammar performance among Iranian EFL learners. First research question sought to examine the relationship between test method and test-takers' grammar performance. In order to answer the first research question, a Pearson product moment correlation coefficient was run. The results of the analysis showed that there were positive statistically significant correlations between total score (grammar performance) and error correction, word changing, word order, and completion score. These findings suggest that different test methods can be used to measure different aspects of grammar proficiency, and that using a variety of test methods can provide a more comprehensive and accurate measure of learners' grammar knowledge. This is consistent with previous research that has shown the benefits of using multiple measures of language proficiency (e.g., Cheng, 2004; Shohamy, 1984). The results were also in line with a study done by Mozaffari et al. (2017), which investigated the performance of Iranian EFL learners on multiple-choice and open-ended grammar tests. The results showed that there was a significant difference between the two test formats, with the open-ended test format yielding higher scores. However, this is consistent with the results that different test methods can be used to measure different aspects of grammar proficiency. The results are also consistent with the research conducted by Akhavan Masoumi and Sadeghi (2020), which investigated the differences between multiple-choice items and constructed response

items, revealing a significant difference. The results are also in line with a study done by Birenbaum and Tatsouka (1987), which showed considerable differences between test format and performance. The findings of the study are in line with those of Bridgeman (1992), who also reported substantial differences between the open-ended and multiple-choice response formats. The results also show that students tend to perform almost equally across all test formats which is inconsistent with other studies that have shown that test-takers tend to perform better in selected-response than in constructed-response formats, such as those conducted by Currie and Chiramanee (2010), Famularo (2007), and In'nami and Koizumi (2009). For instance, Ackermann and Siegfried (2019), compared the performance of test-takers in stem-equivalent selected-response and constructed-response tests and found that selected-response items yielded much better results. Similarly, Famularo (2007) compared the scores of test-takers in selected-response and constructed-response items and found that the selected-response format was significantly easier than the constructed-response version of the same test.

The second research question investigated the relationship between test-takers' gender and their grammar performance. The results of the binary logistics regression analysis showed that gender was a significant predictor of grammar performance, with male students tending to score lower than their female counterparts on the test. These findings are consistent with previous research that has shown gender differences in language proficiency (e.g., Aslan, 2009; Holmes, 2007; Hyde & Linn, 1988), and may reflect broader social and cultural factors that influence the language learning experiences and outcomes of male and female students. Results are also in line with the study done by Azizmohammadi and Barjesteh (2020). According to their findings, there is a statistically significant difference in grammar performance between male and female learners, with female students demonstrating a tendency to perform better than their male peers. Nevertheless, the findings were not in line with those reported by Izadpanah et al., (2023). They investigated the impact of gender on Iranian EFL learners' grammar performance. The results showed that there were no significant differences in the mean scores of male and female students in grammar performance. This is inconsistent with the results that male students tended to score lower than their female counterparts in the test. The findings of this study indicate that female students perform better than male students across all test formats, which contrasts with the results of earlier research conducted by Mauldin (2009), Simkin and Kuechler (2005), and Weaver and Raptis (2001). These studies examined the performance of male and female test-takers in both multiple-choice and constructed-response tests but did not identify any significant differences between genders.

Conclusion

The primary objective of the present study was to investigate the correlation between test method and Iranian EFL learners' grammar performance with respect to the moderating role of gender. The evaluation and testing of language skills and competencies are crucial aspects of language instruction. Testing is an essential part of teaching because it offers valuable insights into learners' growth and achievement, as well as their learning difficulties, styles, and anxiety levels. Presenting effective instruction and utilizing different test methods are somehow interconnected and influential (Roediger et al., 2011). Tests in different formats assess not only the progress and achievement of learners but also the effectiveness of the teaching materials and methods employed. The new paradigm of learning assumes that all students are capable of learning at higher levels; however, it should not be viewed as a limitation, as other factors such as race, gender, and sex can also have an impact on students' performance (Hijazi & Naqvi, 2006). Testing and evaluation play significant roles in language instruction and acquisition. Various types of tests in different formats can be administered to assess the student's language skills. Through language skills evaluation in different forms, we can identify the specific areas where students may face difficulties in learning. Once these problems are identified, we can develop appropriate remedies to address them. Using multiple measures of grammar and considering the

diversity of learners' backgrounds and experiences are important for creating more effective and equitable language testing and assessment practices (Brown & Abeywickrama, 2018).

The findings of this study have important implications for language testing and assessment practices. The positive correlations between different test methods and grammar performance suggest that using multiple measures of grammar can provide a more accurate and comprehensive assessment of learners' grammar ability. The gender differences in grammar performance highlight the importance of considering the diversity of learners' backgrounds and experiences in language teaching and assessment, and the need to develop more inclusive and equitable language programs and assessments. Additionally, the outcomes suggest that test method and gender are important factors that influence test-takers' grammar performance. Future research should continue to explore these relationships and investigate other factors that may impact language performance. By doing so, language testers and educators can better support learners in achieving their learning goals.

There are also some limitations to this study that should be acknowledged. First, the study was conducted from a single institution, which may limit the generalizability of the findings. Second, the study focused only on grammar performance and did not consider other aspects of language proficiency, such as vocabulary, fluency, and pragmatics. Based on the findings of this study, there are several recommendations for future research. First, future studies should include larger and more diverse samples to increase the generalizability of the findings. This study was conducted with a relatively small sample of participants from a single institution and may not be representative of other populations. Second, as already mentioned, future research should consider other aspects of language proficiency. This study focused only on grammar performance, and it would be beneficial to investigate the relationships between test method, gender, and other aspects of language proficiency. Third, future studies should explore the potential impact of culture on language proficiency. This study did not consider cultural factors that may influence language learning and proficiency. Exploring the impact of culture on language proficiency could provide valuable insights into how language testers and educators can better support learners from diverse cultural backgrounds.

References

Ackermann, N., & Siegfried, C. (2019). Does a balanced test form regarding selected-response and constructed-response items overcome gender gap in test scores? An analysis of the format-gender relation in the test of economic-civic competence. *Citizenship, Social and Economics Education, 18*(3), 158-176. https://doi.org/10.1177/2047173419892531

Akhavan Masoumi, G., & Sadeghi, K. (2020). Impact of test format on vocabulary test performance of EFL learners: the role of gender. *Language Testing in Asia*, *10*(1), 1-13. https://doi.org/10.1186/s40468-020-00099-x

Aslan, O. (2009). *The role of gender and language learning strategies in learning English.* MA Thesis, Middle East Technical University. https://hdl.handle.net/11511/18929

Azizmohammadi, F., & Barjesteh, H. (2020). On the relationship between EFL learners' grammar learning strategy use and their grammar performance: Learners' gender in focus. *Journal of Language Teaching and Research*, *11*(4), 583-592. http://dx.doi.org/10.17507/jltr.1104.08

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

Beller, M., & Gafni, N. (2000). Can item format (multiple choice vs. open-ended) account for gender differences in mathematics achievement? *Sex Roles*, *42*(1), 1-21. https://doi.org/10.1023/A:1007051109754

Bensoussan, M. (1984). A comparison of cloze and multiple- choice reading Comprehension tests of English as a Foreign Language. *Language Testing, 1*(1), 101-104. https://doi.org/10.1177/026553228400100109

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. https://doi.org/10.1177/014662168701100404

Bleske-Rechek, A., Zeug, N., & Webb, R. M. (2007). Discrepant performance on multiple- choice and short answer assessments and the relation of performance to general scholastic aptitude.

Assessment & Evaluation in Higher Education, *32*(2), 89-105. https://doi.org/10.1080/02602930600800763

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement*, *29*(3), 253-271. https://doi.org/10.1111/j.1745-3984.1992.tb00377.x

Brown, H. D., & Abeywickrama, P. (2018). *Language assessment: Principles and classroom practices (3rd Ed.)*. Pearson Education. https://doi.org/10.3390/socsci7110227

Bush, M. (2001). A multiple choice test that rewards partial knowledge. *Journal of Further and Higher Education, 25*(2). 157-163. https://doi.org/10.1080/03098770123674

Cesur, K. (2008). *Students' and teachers' perceptions of the test techniques used to assess language skills at university level* (Unpublished master's thesis). Çanakkale Onsekiz Mart University, Institute of Social Sciences, Department of English Language Teaching.
https://doi.org/10.13140/RG.2.2.16501.63202

Chambers, J. K., & Schilling, N. (2013). *The handbook of language variation and change.* Wiley-Blackwell. https://doi.org/10.1002/9781118335598

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. https://doi.org/10.1111/j.1944-9720.2004.tb02421.x

Close, R.A. (1982). *English as a foreign language.* George Allen and Unwin. ISBN 10: 0044250258. ISBN 13: 9780044250258

Cordeiro, P. F. (2021). *Accountability evaluation in systems-of-information systems based on systems thinking*. Doctoral Thesis, PPGI/UNIRIO. https://doi.org/10.13140/RG.2.2.35828.22402.

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

Debata, P. K. (2013). The importance of grammar in English language teaching: A reassessment. *Language in India, 13*(5), 482-486. ISSN 1930-2940

DeKeyser, R. M. (1993). The effect of error correction on L2 grammar knowledge and oral proficiency. *The Modern Language Journal, 77*(4), 501-514. https://doi.org/10.1111/j.1540-4781.1993.tb01999.x

Dikmen, M. (2023). Test anxiety in online exams: scale development and validity. *Current Psychology, 42*(1), 30210–30222. https://doi.org/10.1007/s12144-022-04072-0

Dolan, R. P., Goodman, J., Strain-Seymour, E., Adams, J., & Sethuraman, S. (2011). *Cognitive lab evaluation of innovative items in mathematics and English language arts assessment of elementary, middle, and high school students*. Pearson. https://doi.org/10.13140/RG.2.2.21857.02407

Douglas, D. (2010). *Understanding language-testing* (pp.2). Routledge. ISBN 9780340983430

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*. Boston College ProQuest Dissertations Publishing.

Haynie, W. J. (1994). Effects of multiple-choice and short-answer tests on delayed retention learning. *Journal of Technology Education, 6*(1), 32-44. https://doi.org/10.21061/jte.v6i1.a.3

Hijazi, S. T., & Naqvi, S. M. M. (2006). Factors affecting students' performance: A case of private colleges. *Bangladesh e-Journal of Sociology*, *3*(1), 1-10.
https://api.semanticscholar.org/CorpusID:17496544

Holmes, J. (2007). Social constructionism, postmodernism and feminist sociolinguistics. *Gender & Language, 1*(1), 51-65. https://doi.org/10.1558/genl.2007.1.1.51

Hughes, R., & McCarthy, M. (1998). From sentence to discourse: Discourse grammar and English language teaching. *TESOL Quarterly*, *32*(2), 263. https://doi.org/10.2307/3587584

Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis.
Psychological Bulletin, 104(1), 53-69. https://doi.org/10.1037/0033-2909.104.1.53

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

Izadpanah, J., Sadighi, F., & Akbarpour, L. (2023). The effect of explicit corrective feedback on EFL learners' retention of grammar: Does the medium of feedback matter? *Journal of Studies in Learning and Teaching English*, *12*(1), 99-122. 20.1001.1.22518541.2023.12.1.6.4

Kang, S. H., McDermott, K. B., & Roediger III, H. L. (2007). Test format and corrective feedback modify

the effect of testing on long-term retention. *European Journal of Cognitive Psychology, 19*(4-5), 528-558. https://doi.org/10.1080/09541440601056620

Kitao, S. K., & Kitao, K. (1996). Testing grammar. *The Internet TESL Journal*. Retrieved from http://iteslj.org/Articles/Kitao-TestingGrammar.html

Kuechler, W. L., & Simkin, M. G. (2010). Why is performance on multiple-choice tests and constructed-response tests not more closely related? Theory and an empirical test. *Decision Sciences Journal of Innovative Education, 8*(1), 55-73. https://doi.org/10.1111/j.1540-4609.2009.00243.x

Livingston, S. A. (2009). Constructed-response test questions: Why we use them; how we score them. R&D Connections, (11). *Educational Testing Service.* http://www.ets.org/Media/Research/pdf/RD_Connections11.pdf

Madsen, H. S. (1983). *Techniques in testing*. Oxford University Press.
https://doi.org/10.1177/026553228500200109

Mao, A.M. (2022) Literature review of language testing theories and approaches. *Open Access Library Journal, 9*(5), 1-5. https://doi.org/10.4236/oalib.1108741.

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. https://doi.org/10.1177/1469787409343191

McNamara, F. (2000). *Language testing*. Oxford University Press. ISBN-10, 0194372227.

McNamara, D. S. (2010). Strategies to read and learn: Overcoming learning by consumption. *Medical education*, *44*(4), 340-346. https://doi.org/10.1111/j.1365-2923.2009.03550.x

Mozaffari, F., Alavi, S. M., & Rezaee, A. (2017). Investigating the impact of response format on the performance of Grammar tests: Selected and constructed. *Teaching English as a Second Language (Formerly Journal of Teaching Language Skills)*, *36*(2), 103-128. https://doi.org/10.22099/jtls.2017.23918.2154

Ölmezer-Öztürk, E., & Aydin, B. (2018). Toward measuring language teachers' assessment knowledge: development and validation of Language Assessment Knowledge Scale (LAKS). *Language Testing in Asia, 8*(20), 1-15. https://doi.org/10.1186/s40468-018-0075-2

Onaiba, A., & Jannat, F. (2019). Test method effect and test-takers' scores: a critical review of the pertinent literature. *Scientific Journal of Faculty of Education, Misurata University-Libya, 1*(14), 3-22. http://mdr.misuratau.edu.ly/handle/123456789/1084

Otayf, Y. A. (2019). The role of gender in language learning strategies among male and female students at Jazan secondary schools. *مجلة كلية التربية (أسيوط)*, *35*(9.2), 1-28. https://doi.org/10.21608/mfes.2019.102838

Ozan, C., & Kincal, R.Y. (2018). The effects of formative assessment on academic achievement, attitudes toward the lesson, and self-regulation skills. *Educational Sciences: Theory & Practice, 18*(1), 85-118. https://doi.org/10.12738/estp.2018.1.0216

Paudel, P. (2018). Use of test-teach-test method in English as a foreign language classes. *Journal of NELTA Surkhet, 5(15)*, 15–27. https://doi.org/10.3126/jns.v5i0.19482

Pienemann, M., Johnston, M., & Brindley, G. (1989). Constructing an acquisition-based procedure for second language assessment. *Studies in Second Language Acquisition*, *10*(2), 217–243. https://doi.org/10.1017/S0272263100007324

Pope, G. A., Wentzel, C., Braden, B., & Anderson, J. (2006). Relationships between gender and Alberta achievement test scores during a four-year period. *Alberta Journal of Educational Research*, *52*(1), 4-15. *https://api.semanticscholar.org/CorpusID:201024355*

Pomplun, M., & Capps, L. (1999). Gender differences for constructed-response mathematics items. *Educational and Psychological Measurement, 59*(4), 597-614. https://doi.org/10.1177/00131649921970044

Rea-Dickins, P., & Germaine, K. (1992). *Evaluation.* Oxford University Press. ISBN 0194371387

Roediger, H., Putnam, A., & Sumeracki, M. (2011). Ten benefits of testing and their applications to educational practice. *Psychology of Learning & Motivation: Cognition in Education, 55*(1), 1-36. https://doi.org/10.1016/B978-0-12-387691-1.00001-6

Schmitt, N., & Carter, R. (2004). *Formulaic sequences in action. Formulaic sequences:*
Acquisition, processing and use. John Benjamins. https://doi.org/10.1075/lllt.9.02sch

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension.

Language Testing, *1*(2), 147-170. https://sid.ir/paper/599667/en

Simkin, M. G., & Kuechler, W. L. (2005). Multiple-choice tests and student understanding: What is the connection? *Decision Sciences Journal of Innovative Education*, *3*(1), 73-97. http://dx.doi.org/10.1111/j.1540-4609.2005.00053.x

Sireci, G. S., & Zenisky, L. A. (2016). Computerized innovative item formats: Achievement and credentialing. In S. Lane, M. R. Raymond & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 313-334). Routledge. ISBN 9780415626026

Sumarni, S. & Rachmawaty, N. (2019). Gender differences in language learning strategies. Ethical Lingua: *Journal of Language Teaching and Literature*. *6*(1), 13-22. 10.30605/ethicallingua.v6i1.1169.

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. https://doi.org/10.1023/A:1009412929239

Zhang, J. (2009). Necessity of grammar teaching. *International Education Studies*, *2*(2), 78–81. https://doi.org/10.5539/ies.v2n2p184

Zoghi, M., Kazemi, S. A., & Kalani, A. (2013). The effect of gender on language learning. *Journal of Novel Applied Sciences*, *2*(4), 1124-1128. *https://api.semanticscholar.org/CorpusID:58666518*