Tabaran Institute of Higher Education
Iranian Journal of Language Testing
Vol. ٤, No. ٢, October ٢٠١٤

ISSN ٢٢٥١-٧٣٢٤

Does the Type of Multiple-choice Item Make a Difference? The Case of Testing Grammar

Nasser Rashidi[١] , Faezeh Safari[٢]

Abstract

With the widespread use of multiple-choice (MC) tests, even if they were disapproved by many practitioners, investigating the performance of such tests and their consequent features is desirable. The focus of this study was on a modified version of multiple-choice test, known as multitrak. The study compared the multitrak test scores of about ٦٠ students against those of the standard MC and constructed-response (CR) tests. The tests employed in the study evaluated English language grammar while they all had identical worded stems. The results showed that multitrak items are at a higher level of difficulty in comparison to the other formats. The results suggest that these items can be used to test more advanced aspects of grammatical competence as the test taker requires going beyond mere syntactic knowledge to be competent in the range of alternatives being used in communication to find the unacceptable choice. Therefore, multitrak test is better geared for higher levels of proficiency and could provide better information about test takers who are more proficient. At the end, implications of the study for test constructors and test users, as well as implications for future research, are discussed.

Keywords: *Multitrak test, standard multiple-choice test, constructed-response test, testing grammar, test format*

١. Introduction

Among many factors which affect performance on language test, one of the crucial concerns is the influence of test formats on test performance. Test format constitutes the context of a language test and is comparable to the context of a language communication (Bachman, ١٩٩٠). The format of a test could influence the tester's performance and thereby enhance or impede the measurement of the construct. The format is not able to reflect the construct very well if it does not include certain construct elements or causes interference with such elements. Alternatively, the item format could improve the measurement of the construct by including some other elements (Dávid, ٢٠٠٧). Furthermore, the way these formats affect may vary on the part of the learners at different levels of proficiency. With regard to grammar tests, a variety of test formats

[١]*Department of Foreign Languages and Linguistics, College of Literature and Humanities, Shiraz University, Shiraz, Iran. Email: nrashidi@rose.shirazu.ac.ir*

[٢]*Department of Foreign Languages and Linguistics, College of Literature and Humanities, Shiraz University, Shiraz, Iran.*

ISSN ٢٢٥١-٧٣٢٤

have been developed, including gap-filling, matching, multiple-choice, constructed-response, ordering, etc. Since there is no format to function well in all conditions, the language tester must recognize the properties of each format and make the best selection based on the purpose of a test in each context (In'nami & Koizumi, ٢٠٠٩). This study examined the issues related to the effect of three different types of testing, namely standard MC, multitrak, and constructed-response test, on the measurement of English language grammar in terms of their difficulty level and discrimination power and their reliability in testing grammar.

١.١. Test of Grammar with the Focus on MC Test

An overview of teaching and testing of grammar reveals the changes happening over the years. Once, grammar had a pre-eminent position, along with the prescriptivism in linguistics. But, with the decline of prescriptivism, the importance of the concept of correctness also lost its former position (Dávid, ٢٠٠٧). Dávid (٢٠٠٧) noted a number of factors which impact on this changing fortune, including an increasing sociolinguistics awareness and the weakening of the high status dialects, much attention to spoken language, and a broadening notion of the grammar itself. At the end of the ١٩٨٠s, then again the role of the grammar in language teaching went through reconsideration and the form of language received much more attention while the importance of explicit grammar teaching and testing reappraised (Purpura, ٢٠٠٤). In spite of a reappraisal of the position of grammar, according to Purpura (٢٠٠٤), research on the grammar assessment still has not changed profoundly.

This lack of renewed research in grammar tests, consequently, seems to have an influence on the most-associated technique with it, i.e. multiple-choice. Dávid (٢٠٠٧) observed that despite the prevalence, research on multiple-choice test, specifically in testing grammar, is almost underdeveloped in recent literature. He, then, took into account a number of reasons for the relative unpopularity of the recent research on the MC format. MC tests are generally regarded as the partial, limited, and decontextualized form of the assessment and accordingly inadequate for many testing situations since they could not measure all parts of the construct. In addition, the MC format is unsatisfactory in connection with grammar testing because of the broadening concept of grammar developed over the years. Leech (١٩٨٣) defined this broadening concept by specifying three levels: syntactic level, semantic level, and pragmatic level, whereas grammar, traditionally specified narrowly at the syntactic level. Syntactic level embraces the rules for combination of elements into sentences and texts, the semantic level connects syntactically acceptable forms to meaning, and finally the pragmatic level involves the rules for the use of these syntactically and semantically correct elements in an appropriate way with regard to a specific situation. Another deficiency of MC formats is related to the difficulty of writing correct and acceptable MC items. For instance, an item may not lend itself to the construction of logical distractors. Further, an alternative considered as a wrong answer may become a plausible one in a certain discourse or special context, or in one of the varieties of English (Dávid, ٢٠٠٧).

Despite the shortcomings, many educational systems yet rely heavily on multiple-choice tests to fulfill their assessment demands due to the large number of test takers, the need for fast scoring, and the convenience and reliability of multiple-choice tests (Currie & Chiramanee, ٢٠١٠). However, regardless of the widespread use of multiple-choice, Dávid (٢٠٠٧) observed the unpopularity of recent literature in multiple-choice format as the focus of the research, especially

ISSN ٢٢٥١-٧٣٢٤

in the case of testing grammar. Therefore, it is both appropriate and desirable to research into the effect of different types of multiple-choice items in measuring the knowledge of language grammar.

١.٢. The Issue of Negative MC Items

Concerning Dávid's (٢٠٠٧, p.٦٨) classification, four types of MC items can be identified:

١. Four-choice sentence-based items, which are possibly the most widespread variety and are referred to as the *standard* type of multiple-choice (*standard MC).*
٢. Four-choice text-based items. The stem of such items is text-based and usually referred to as *MC cloze*.
٣. Four-choice sentence-based items, called *double-blank* as the stem encloses two blanks instead of the normal one. For example :

Teenagers _____ better _____ children.
(a) are // to be regarded as * (c) are // regarded to be
(b) are // regarded as (d) had // to regard themselves as

٤. A four-choice sentence-based item, like the standard one, but the most observable difference is that in this category, the test taker must select the unacceptable choice. They have been called *multitrak* items. For instance:

Look! Over there! That _____ be the man we're looking for.
(a) could (c) may (b) can * (d) might

A *multitrak* item, a term which seemingly was given by Dávid (٢٠٠٧), makes it possible that the test constructor writes all but one of the possible answers in a way that represent the communicative situations which are open to the speaker to choose among alternatives in a certain language use situation. Dávid (٢٠٠٧) evaluated the efficiency of multitrak items in relation to the other three MC item types in the Hungarian context. He identified that multitrak items provide more information about the test takers with the ability above the intermediate level. Further, he claimed that multitrak items allow a focus on more difficult content than the other MC item types.

Making use of an unacceptable option as the correct response, however, raises the issue of "testing negatively". A growing body of literature stands against "find the wrong answer" types of questions (e.g. Heaton, ١٩٨٨; Madsen, ١٩٨٣). Such guidelines argue that multitrak items are unacceptable and unfair since these items draw students' attention on errors. Tamir (١٩٩١) debated that such claims imply that standard MC items are better as they do not draw students' attention to what is wrong. Dávid (٢٠٠٧) criticized that such arguments are more based on methodology rather than research. He explained that such arguments are not persuasive as they beg "the question of why the one unacceptable choice in a multitrak item would be more of a focus on what is "wrong" than when the unacceptable choices outnumber the correct answer in standard items, usually by three or four" (p. ٧٠). Similarly, Tamir (١٩٩١) noted that since responding to a test is in itself a kind of learning situation, why not the students have been exposed to more correct information than to incorrect ones. Furthermore, in Structure and Written Expression section of the paper-based TOEFL and in some matriculation examinations, similar types of multitrak items, known as 'spot-the-error type' items, can be found in which the test takers should identify one of the four underlined parts contained an error. The challenge of

ISSN ٢٢٥١-٧٣٢٤

testing negatively can therefore be better understood if we consider that such items require test takers to involve an inverted cognitive-process (Dávid, ٢٠٠٧).

While multitrak test, as a modified format of MC, has been shown to have some potential benefits, it is of interest to conduct a study to bring together some evidence of its characteristics in comparison with other test formats. In this regard, this study aimed at seeking the answers to the following research questions:

Q١. Does test taker's performance on multitrak test differ from those of the standard MC and CR tests? Or more generally, does the test taker's score on grammar remain stable across the three test formats?
Q٢. Does the consistency between the test takers' responses to individual items vary depending on the test format?
Q٣. Does the item format affect the average level of item discrimination indices of the stem-equivalent tests?
Q٤. Do correlation coefficients between the scores from the three test formats support a conclusion that the tests measure essentially the same construct?
Q٥. Does multitrak test effectively distinguish learners of different proficiency levels?

٢. Methodology

٢.١. Participants

Two different groups participated in this study: First, for the purpose of pilot testing and estimating the reliability of two instruments, including multitrak test and standard MC test, ٤٠ freshmen students studying at Zand non-profit University took part in the study. In the second round, ٦٨ second-year students studying at Shiraz University made up the sample of the study. The first language of all the participants is Persian and English is their major. Their gender was not a controlled factor.

٢.٢. Instruments

Four instruments were employed in the study.

٢.٢.١. Proficiency test

A reduced form of TOEFL was used as the proficiency test. This is a ٦٠-item test constructed by Educational Testing Service (ETS, ١٩٩٨) and proved to have reliability and validity. The study has used the reduced form since the students would not have eagerly answered the whole test. This test was employed to classify students based on their proficiency levels in English. The basis for classification was the students' scores distribution. The top ٢٧٪ scores and the low ٢٧٪ scores were considered as the high and low levels, respectively. The rest ٤٦٪ scores were considered as the mid level.

٢.٢.٢. Multitrak Test

First, a ٥٧-item grammar test of English was written in multiple-choice format in which each item consisted of three acceptable choices and one unacceptable. These items were written based on the questions and grammatical tips provided in Hewings (١٩٩٩) and Murphy (١٩٩٤). Before being administered to the students, the test was examined by three native speakers. The primary

ISSN ٢٢٥١-٧٣٢٤

purpose was to confirm the researcher's judgments of which choices were acceptable and which were unacceptable for each item. The choices were accepted as correct based on a majority of two out of three native speakers and in other cases the item and/or choices were reviewed and revised. By doing so, after modification, ٥٠ items were accepted to be included in the test. The following item is given as an example:

Select the choice that does **NOT** fit the blank.
١. Can I have the newspaper when ……..?
a. you finish with it b. you've read it c. you will finish with it d. you've finished with it

٢.٢.٣. Standard MC Test
A ٥٠-item standard MC test then was written based on the multitrak items. Therefore, the stems of the items in both tests were equivalent but in the standard test, there were three unacceptable choices and one acceptable. The following item is given as an example:

Choose the best answer.
١. I was pleased …… the success of our money-raising efforts.
a. on b. in c. to d. about

In other words, both tests were totally the same except for two choices in each item. The ٥٠ items were distributed randomly throughout each test; by doing so, the order of the items in two tests was not the same. With respect to the instruction of the tests, it was expected that the candidates choose the best answer for *Standard MC Test* and choose the unacceptable choice in the *Multitrak Test.*

٢.٢.٤. Constructed-response (CR) Test
A ٥٠-item CR test also was constructed in this study. Each test item comprised a situation with a blank included in. The test taker should guess the word(s) appropriate for the blank based on the context of the item. The stems of this test were exactly similar to those of two other tests except when the context of the sentence was not so rich for the students to fill out the blank. In such cases one or two sentences were added to rich the context or a cue was provided right after each item in the parentheses. The following item is given as an example:
Read the situations and complete the sentences. You can make use of the words in parentheses in their most appropriate form given after some items to complete the sentences.

١. A friend is reading the newspaper. You'd like to read it after her. You ask:
Can I have the newspaper when ……..?

٢.٣. Data Collection Procedure
Following the construction of two types of MC, then, they were pilot tested using ٤٠ first-year students of Zand non-profit University. The pilot served to assess test reliability and to detect any problem associated with the items. In order to decrease the *practice* effect of the test, the participants were randomly divided into two groups. One group took the standard test and the other took the multitrak one. Each group then answered the other alternative, too. According to

ISSN ٢٢٥١-٧٣٢٤

the obtained results, both tests had good internal consistency, with the Cronbach Alpha coefficients reported of ٠.٧٥ and ٠.٨٠ for multitrak and standard MC tests, respectively.
Later, for collecting the data, the tests were taken by ٦٨ second-year students of Shiraz University. In order to decrease the effects of the order and the familiarity with the test, the participants were randomly divided into two groups. In the first round of testing, one group took the standard test and the other took Multitrak. Then, after ١٠ days, in the second round of testing, each group took the other alternative test. Among ٦٨ participants, however, only ٦٠ students took part in both sessions of testing and answered both tests.
Considering the specification of the participants' proficiency levels, the study relied on a reduced form of TOEFL. Among ٦٠ students who took part in both sessions, ٥١ students answered the proficiency test. Therefore, whenever the analysis requires the data of all three tests, the data of these ٥١ students have been employed, but in other cases, the study made use of other data, too.
One month later, the CR test was administered in which only ٢٥ students out of the ٦٨ participants took part. The scoring of the CR answers was done in relation to a response model which listed all the possible correct answers. Spelling errors were not counted. A number of papers (١٤) were re-scored by another rater to examine the interrater reliability and it was found to be over .٩٠.

It is worth mentioning that none of the tests were speeded; and in all test formats, each item was equally scored right (١ point) or wrong (٠ points).

٢.٤. Data Analysis and Results

Mean scores (Research Question ١): The means and standard deviations of the three tests are presented in Table ١. As displayed in the table, the multitrak test was the hardest (M = ٢٩.١٦, SD = ٤.٠٩٨), the CR test was in-between (M = ٣١.٥٨, SD = ٣.٤٣٧), and the standard MC test (M = ٣٤.٩٢, SD = ٤.٨٩٠) was the easiest. The results of the one-way repeated measure ANOVA indicated that the test format significantly affected the mean test scores, Wilks' Lambda = .١٣٩, F = ٥٢.٨٠, p < .٠٠٠٥. The effect size of the difference, expressed in partial-eta squared was .٨٦. According to Cohen (١٩٨٨), a value of partial-eta squared above .١٣٨ can be regarded as large. Therefore, the effect can be regarded as very large. Further, post-hoc comparisons with Bonferroni adjustments identified that all pairwise comparisons were significant at the .٠٠٥ level.

Table ١. Means and standard deviations for multitrak, standard MC, and CR tests

	Mean	N	Std. Deviation	Std. Error Mean
Multitrak	٢٩.١٦	٦٠	٤.٠٩٨	.٩٩٢
Standard MC	٣٤.٩٢	٦٠	٤.٨٩٠	.٨٣٣
CR	٣١.٥٨	٢٥	٣.٤٣٧	.٧٨٨

Test reliability (Research Question ٢): To investigate Research Question ٢ ('Does the consistency between the test takers' responses to individual items vary depending on the test format?'), Cronbach's alpha coefficients were examined. As is indicated in Table ٢, the reliability coefficient in the multitrak test was slightly higher than that of the standard MC while the reliability of the CR test was relatively lower than those of the other two tests.

ISSN ٢٢٥١-٧٣٢٤

Table ٢. Reliability coefficients for the three tests

	Cronbach's alpha
Multitrak	.٧٩
Standard MC	.٧٧
CR	.٦٩

Item discrimination (Research Question ٣): Another concern related to the efficiency of different test formats is that of the performance of the test items. It is of interest to explore how different item formats function in terms of discriminating among test takers. To explore this question, we computed point-biserial correlations (item to total score correlations). We used Fisher's *z* transformation (transformed the coefficient *r* to *z* value) to compute the mean item discrimination indices and to conduct significant test of the mean item discrimination indices among the three tests. The findings are summarized in Table ٣. As the table shows, the mean item discrimination in the multitrak test was relatively as large as that of the CR test while the mean item discrimination of the standard MC test is relatively lower than those of the other two tests. To investigate whether the differences among the means were statistically significant, a one-way analysis of variance (ANOVA) was used. Results of ANOVA indicated that there was no statistically significant difference at $p < ٠.٠٥$ in the means of item discrimination indices for the three forms ($F = ٠.٨٧$, $p = ٠.٤٢$). Hence, there was no evidence that the three test formats for grammar test items have any significant impact on item discrimination.

Table ٣. Mean item discriminations of the three tests

	No. of items	Mean	SD	Max	Min
Multitrak	٥٠	.٣١	.١٧	.٥٤	.٠٨
Standard MC	٥٠	.٢٦	.٢٠	.٥١	.٠٦
CR	٥٠	.٣٠	.١٩	.٤٦	.٠٣

Test score correlation (Research Question ٤): To see whether the three test formats essentially assess the same construct, we derived correlations between the test scores. Correlations are regarded as an indication that the construct measured by the tests are either the same or very closely related (see, e.g. Lissitz, ٢٠٠٩; Thorndike, ١٩٨٢). As indicated in Table ٤, the coefficients revealed that the test scores in multitrak and standard MC formats did not correlate highly with CR test scores. However, the magnitude of the correlation between multitrak and standard MC tests was relatively larger than those of the CR test.

Table ٤. Test score correlation coefficients (*r)*

	r	Sig.
Multitrak/standard MC test	.٦٠	.٠٠٠
Multitrak/CR test	.٣٥	.٢٩
Standard MC/CR test	.٤٨	.٠٨

Comparison based on proficiency groups (Research Question ٥): To address Research Question ٥, as explained earlier, the test takers were divided into high-, middle-, and low-proficiency subgroups based on their scores in the proficiency test. A summary of the size of the

ISSN ٢٢٥١-٧٣٢٤

subgroups and their mean scores is provided in Table ٥. It should be noted that for this question, the CR test scores were excluded from the analysis due to the small number of the test takers; thus only the test scores in the multitrak and standard MC were included for the analysis.

To see whether the tests can effectively distinguish learners of different proficiency levels, two ANOVA tests separately were conducted to compare the mean scores of the three proficiency groups, one for the multitrak test and the other for the standard MC test. Here, the assumption was that the high group would significantly outperform the other groups and the mid group scores much higher than the low group. The results of ANOVA tests are reported in Table ٦ and ٧ for multitrak and standard MC tests, respectively. It is evident from the table that in both tests there is a significant difference between the means of at least two groups. In order to see which mean differences were significant, a post-hoc analysis using Scheffe's test was run. The results appear in Table ٨ and ٩ for multitrak and standard MC tests, respectively. For standard MC test, the results reveal that the mean differences between the performances of all three proficiency groups are large and significant. Therefore, it can be concluded that the test is capable of distinguishing learners with different proficiency levels. The results of multitrak test, however, indicate that not all the mean differences are significant. It is evident from Table ٨ that highly proficient learners scored significantly different from the mid and the low groups but the mean differences between the performances of the mid and the low groups are not large and significant. Therefore, we can conclude that multitrak test is capable of distinguishing high proficient learners from other learners well; but there is no evidence that multitrak test is capable of distinguishing mid-proficiency learners from low-proficiency learners.

Table ٥. The test score means of the multitrak and standard MC tests split by proficiency groups

	N	Multitrak test mean	Standard MC test mean
High proficiency group	١٤	٣٤.٩٠٩	٣٨.٠٩١
Middle proficiency group	٢٣	٢٧.٨٩٥	٣٣.٢٦٣
Low proficiency group	١٤	٢٦.٠٠٠	٢٨.٢٧٣

Table ٦. ANOVA results for the multitrak test

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	٥٠٣.٣٥	٢	٢٥١.٦٧٥	٥.٦٥٧	.٠٠٧
Within Groups	١٦٩٠.٦٩٩	٤٨	٤٤.٤٩٢		
Total	٢١٤٩.٠٤٩	٥٠			

Table ٧. ANOVA results for the standard MC test

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	٥٣٠.٢٤٩	٢	٢٦٥.١٢٥	١١.٦٢٣	.٠٠٠
Within Groups	٨٦٦.٧٧٥	٤٨	٢٢.٨١٠		
Total	١٣٩٧.٠٢٤	٥٠			

Table ٨. Multiple comparisons for the multitrak test

(I) Group	(J) Group	Mean Difference (I-J)	Std. Error	Sig.	٩٥٪ Confidence Interval	
					Lower Bound	Upper Bound

ISSN ٢٢٥١-٧٣٢٤

low	high	-٨.٩٠٩٠٩*	٢.٨٤٤٢٠	.٠١٣	-١٦.١٥٤٦	-١.٦٦٣٦
	mid	-١.٨٩٤٧٤	٢.٥٢٧١٤	.٧٥٧	-٨.٣٣٢٦	٤.٥٤٣١
high	mid	٧.٠١٤٣٥*	٢.٥٢٧١٤	.٠٣٠	.٥٧٦٥	١٣.٤٥٢٢

*. The mean difference is significant at the ٠.٠٥ level.

Table ٩. Multiple comparisons for the standard MC test

(I) Group	(J) Group	Mean Difference (I-J)	Std. Error	Sig.	٩٥٪ Confidence Interval	
					Lower Bound	Upper Bound
low	high	-٩.٨١٨١٨*	٢.٠٣٦٤٨	.٠٠٠	-١٥.٠٠٦١	-٤.٦٣٠٣
	mid	-٤.٩٩٠٤٣*	١.٨٠٩٤٦	.٠٣١	-٩.٦٠٠٠	-.٣٨٠٩
high	mid	٤.٨٢٧٧٥*	١.٨٠٩٤٦	.٠٣٨	.٢١٨٢	٩.٤٣٧٣

*. The mean difference is significant at the ٠.٠٥ level.

٣. Discussion of the Findings

The paper reported the results of the study which examined the effect that various test format used for testing grammar have on the measurement of the trait. Specifically, we investigated the characteristics of a special type of MC test, called multitrak, in comparison to the properties of the standard MC test and the CR test.

First, the results indicated that each of the testing methods yielded different degrees of difficulty for the test takers, with the standard MC test being the easiest and the multitrak test being the most difficult. In other words, the study found that the multitrak test was easier than the CR test and more difficult than the standard MC test. The higher difficulty level of the multitrak test in comparison with the standard MC test was also found by Dávid (٢٠٠٧). He repeatedly commented on the rationale for multitrak items and believes that multitrak questions provide a good opportunity for the test constructors to evaluate the communicative competence of the learners by providing three acceptable response options for a certain language context. Therefore, this type of MC test could embrace a range of alternatives much like the system of alternatives usually used by the speakers in spoken discourse.

On the basis of the ATCFL Proficiency Guidelines (Chastain, ١٩٨٨), while a novice learner should be able to satisfy partially the requirements of basic communicative exchanges, towards the proficient end of the continuum, the learner should be able to satisfy the requirements of a broad variety of situations, including both formal and informal from a broad range of language. Bear this in mind and what has been discussed above, then, it seems reasonable that a multitrak test embraces a higher level of difficulty. Accordingly, in a multitrak test, the test taker gets involved more on the higher levels of grammar; so the focus of multitrak items goes beyond the syntactic level to come close to semantic and pragmatic levels, whereas a standard MC test mostly covers the narrow concept of grammar which is in syntactic level. Yet another explanation for the greater difficulty of the multitrak test might be a reflection of the fact that these items are the converse of standard items, or simply the fact that they were different from what students were used to.

Another finding related to the first research question is that multitrak test was found to be relatively more difficult than the CR test. This finding contradicts the earlier research with the

ISSN ٢٢٥١-٧٣٢٤

commonly held idea that multiple-choice tests are easier than constructed-response test as production probably is a higher-level and thus more difficult task than selection (see e.g. Cheng, ٢٠٠٤; Currie. & Chiramanee, ٢٠١٠; Shohamy, ١٩٨٤). Several explanations may contribute in interpreting the above result. First, the findings of the earlier research are almost based on the function of the standard multiple-choice test which seems to act differently than multitrak test. With respect to the multitrak items, it seems that although optional items can be clues, they can also be confusing as the testee needs to know several acceptable responses for a certain language context. This appears to be more difficult to handle on the part of the testee than the CR test in which the knowledge of just one acceptable response is sufficient. Obviously, this type of test could not evaluate the learner's competency in 'alternatives'. On the other hand, in the multitrak test, the test taker should be not only competent in the appropriate words acceptable for the blank but also competent in the alternatives open to the speakers in the flow of communication. In other words, multitrak items can measure a higher range of language competence. In the same vein, Dávid (٢٠٠٧) discussed that multitrak items require a different kind of thinking since the candidates should go through all response options carefully and draw on different kinds of grammatical knowledge to respond correctly.

Second, the findings reveal that there was no notable difference in test reliability values of the two multiple-choice formats, but there was much difference between the reliability coefficient of the CR test and the two multiple-choice formats. Therefore, it was concluded that the type of multiple choice test has no notable impact on test reliability but the type of required response (selection or production) affects the reliability of the test.

Third, though the means of item discrimination indices were not the same over the three formats, they were not significantly different among the three different formats. Hence, in the present study, there was no evidence that the three test formats for grammar test items have any significant impact on item discrimination. This could be because fundamentally, the items were the same across different formats. Or, viewing this from a completely different approach, it could be that statistical significant was difficult to reach due to the limited number of participants. Therefore, it would be of interest to see whether a study with more participants would produce similar results, or significant differences in discrimination indices would be obtained with a larger number of participants.

The issue of construct equivalence, as explained earlier, has often been considered on the basis of correlations between test scores. On that basis, the correlations shown in Table ٤ could be taken to suggest that the two multiple-choice formats measured the same construct and that these tests tapped into the same underlying skills or abilities but the suggestion that the CR test in comparison to the two MC formats were measuring the same construct was not supported by the data. Therefore, it could be argued that the CR test and the two MC formats are two methods of measuring different, though largely separate language-related constructs. In this regard, however, an alternative argument is provided by Currie and Chiramanee (٢٠١٠). They maintain that "a more realistic implication would be that the M/C format had the effect of distorting the measurement of the language based abilities which were used by the participants in answering the C/R items" (p. ٤٨٥). An important follow-up of this study then could be to carry out an investigation of the process that a test taker goes through in performing testing tasks with different formats. At the end, it should be noted that regarding the low correlations among the

ISSN ٢٢٥١-٧٣٢٤

formats, we need to carefully consider statistical factors that can affect low correlations, in this case, the measurement error and lack of score variability.

With respect to the function of the multitrak test across different proficiency groups, the results showed that the test was capable of segregating high proficiency test takers but failed to function well in distinguishing mid-level test takers from low-level ones. In contrast, it appeared that the standard MC test did well in discriminating the proficiency groups. This is not surprising considering that multitrak test was at a higher level of difficulty than standard MC test and the higher the difficulty level, the higher the ability of the test takers about whom the test provides the most information. Thus, one possible explanation for this finding is that multitrak test might be hard not only for the low proficiency group but also to some extent for the mid-level group as the test required being competent in the alternatives used in communicative contexts. Accordingly, the multitrak test was able to discriminate better between those who were more proficient. This is in line with Dávid's (٢٠٠٧) findings which revealed that multitrak items were better geared for testing of upper-intermediate level of candidates. Hence, it may be recommended to use multitrak test when the purpose of testing is to obtain better differentiation among high proficient learners.

٤. Implications, Limitations, and Further Directions

The results obtained in this study show that the type of the MC test can make a difference. On the whole, the findings suggest that though, for the lower levels of proficiency, a standard MC test could be helpful, for the higher level, a multitrak test might be a better informative indicator in terms of the information they deliver about the test takers. Dávid (٢٠٠٧) believes that this matching of the item difficulties and testee's abilities is a feature of good testing. In addition, multitrak items allow more difficult content focus to be measured. Another positive feature of multitrak tests is the lessening of what Loftus and Palmer (١٩٧٤) call 'misinformation effect'. According to Roediger and Marshall (٢٠٠٥) misinformation effect of multiple-choice tests is the exposure of distractor answers to the test takers. They clarify that the "students will sometimes come to believe that the distractor answers are correct and therefore leave the exam having acquired false knowledge" (p. ١١٥٨). This fact is obvious in standard MC items in which the distractors outnumber the correct option, usually by three or four. But, multitrak items could considerably lessen the misinformation effect by providing just one single distractor. Additionally, multitrak items can be used when it is quite difficult for test constructors to invent good distractors. Generally, it is recommended that test constructors employ different types of multiple choice items based on the intended purpose and the content focus to have a better picture of the test taker's proficiency.

The present study represents only a limited picture of the many contexts in which multiple choice tests are used. One of the major limitations of the study was the limited number of participants who eagerly participated in all testing sessions of the study. If the number of the participants was more, the study might come through stronger conclusions. The limited number of the participants should be considered in interpreting the results. Furthermore, this study investigated the difference among the tests within a limited context, i.e. the test of English language grammar at the upper intermediate level in an EFL context. Further studies might need to examine alternative types of multiple-choice in other contexts and in other language skill areas, with students at different levels. Further studies could examine the difficulty level attributed to

ISSN ٢٢٥١-٧٣٢٤

multitrak items to better distinguish between construct-relevant and –irrelevant sources of difficulty as it still remains to be seen whether this difficulty is related to item format or item focus. Finally, whereas a limited number of candidates took part in the CR test in the present study, it might be desirable to have a closer investigation on the different behaviors of multitrak and standard MC tests when they are compared with a CR test and to see which one(s) reflect the responses of the test takers more correctly.

٥. References

Bachman, L. F. (١٩٩٠). *Fundamental considerations in language testing*. Oxford, UK: Oxford University Press.

Chastain, K. (١٩٨٨). *Developing second-language skills: theory and practice*. Orlando, FL: Harcourt Brace Jovanovich.

Cheng, H. (٢٠٠٤). A comparison of multiple-choice and open-ended response formats for the assessment of listening proficiency in English. *Foreign Language Annals, ٣٧*(٤), ٥٤٤-٥٥٥.

Cohen, J. (١٩٨٨). *Statistical power analysis for the behavioral sciences* (٢nd ed.). Hillsdale, NJ: Erlbaum.

Currie, M. & Chiramanee, T. (٢٠١٠). The effect of the multiple-choice item format on the measurement of knowledge of language structure. *Language Testing,* ٢٧(٤), ٤٧١–٤٩١. doi: ١٠.١١٧٧/٠٢٦٥٥٣٢٢٠٩٣٥٦٧٩٠

Dávid, G. (٢٠٠٧). Investigating the performance of alternative types of grammar items. *Language Testing,* ٢٤ (١), ٦٥–٩٧. doi: ١٠.١١٧٧/٠٢٦٥٥٣٢٢٠٧٠٧١٥١٢

Heaton, J. B. (١٩٨٨). *Writing English language tests*. London: Longman.

Hewings, M. (١٩٩٩). *Advance grammar in use*. Cambridge: Cambridge University Press.

In'nami, Y. & Koizumi, R. (٢٠٠٩). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing,* ٢٦ (٢), ٢١٩–٢٤٤. doi: ١٠.١١٧٧/٠٢٦٥٥٣٢٢٠٨١٠١٠٠٦

Leech, G. N. (١٩٨٣). *The principles of pragmatics*. London: Longman.

Lissitz, R. W. (Ed.). (٢٠٠٩). *The concept of validity: Revisions, new directions, and applications*. Charlotte, NC: Information Age Publishing.

Loftus, E. F., & Palmer, J. C. (١٩٧٤). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior, ١٣,* ٥٨٥–٥٨٩.

Madsen, H. S. (١٩٨٣) *Techniques in testing*. New York: Oxford University Press.

Murphy, R. (١٩٩٤). *English grammar in use* (٢nd ed.). Cambridge: Cambridge University Press.

Purpura, J. E. (٢٠٠٤). *Assessing grammar*. Cambridge: Cambridge University Press.

Roediger, H. L. III. & Marsh, E. J. (٢٠٠٥). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, ٣١*(٥), ١١٥٥–١١٥٩.

Shohamy, E. (١٩٨٤). Does the testing method make a difference? The case of reading comprehension. *Language Testing* ١(٢), ١٤٧-١٦١.

Tamir, P. (١٩٩١). Multiple choice items: how to gain the most out of them. *Biochemical Education, ١٩*(٤), ١٨٨-١٩٢.

ISSN ٢٢٥١-٧٣٢٤

Thorndike, R. L. (١٩٨٢). *Applied psychometrics*. Boston: Houghton Mifflin.