Scientific Journal of Faculty of Education, Misurata University-Libya, Vol. 1, No. 14, Sep. 2019

Published online in September

المجلة العلمية لكلية التربية، جامعة مصراتة، ليبيا، المجلد الأول ـ العدد الرابع عشر، سبتمبر 2019م

# Test Method Effect and Test-Takers' Scores: A Critical Review of the Pertinent Literature

**Abdulhamid Elmurabet Onaiba**
**Department of English, Faculty of Education, Misurata University**
**hamidmo@edu.misuratau.edu.ly**

**Fareehah Bashir Jannat**
**Department of English, Faculty of Arts, Misurata University**
**f.jannat@art.misuratau.edu.ly**

## Abstract

Language tests are introduced to test-takers in different formats and designs. They can be administered by means of selected-response and/or constructed-response items. The type of the designed questions of a test may have some effects on the type of score a candidate gains, and this is known as "test method effect". In line with this, literature shows that several empirical research studies investigated the extent to which the type of test format and construction may impact upon the scores that test-takers gain in a test. Thus, this study aims at shedding more light on the effect of using different test methods on students' performance. So, this paper attempts to provide a clear explanation of the notion "test method effect" with reference to the pertinent literature, journal articles and books. As an integrative literature review of several studies, findings of this study highlighted the ideal existence of the notion "test method effect". However, this effect may or may not have a linear relationship with test-takers performance in a test, i.e. other test facets may come into play. The study may also suggest the need for deeper inquiries to be carried out in order to

better understand this notion, particularly in contexts where such issue has not attracted a research focus yet.

*Keywords:* test method effect, test-taker, scores, selected and constructed response items.

## تأثير طرق الاختبار وعلاقتها بنتائج الممتحنين

### ملخص البحث

تقدم الاختبارات اللغوية للممتحنين باستخدام صيغ وأشكال مختلفة، حيث يتم إجراء الاختبارات اللغوية باستخدام الأسئلة الموضوعية أو الأسئلة المقالية. إن تنوع هذه الأسئلة بين مقالي وموضوعي قد يؤثر على الدرجات المستخلصة من الاختبارات، أو ما يعرف بـ " تأثير طرق الاختبار"، هذا ما دفع المختصين في هذا المجال لإجراء العديد من البحوث التجريبية التي تدرس مدى تأثير نماذج وأشكال الاختبار على الدرجات التي يتحصل عليها الممتحنون. و عليه فإن هذه الدراسة هدفت إلى جذب أنظار المعنيين بهذا المجال لهذه التأثيرات؛ وذلك عن طريق توضيح وشرح مفهوم" تأثير طرق الاختبار"، وتقديم نماذج وأمثلة من الدراسات السابقة في هذا المجال. كما اهتمت هذه الورقة البحثية أيضا بتسليط الضوء على حتمية تأثير طرق الاختبار على درجات الممتحنين. استخدم الباحثان طريقة التحليل النصي، متمثلة في مراجعة نقدية وتحليله لدراسات سابقة نشرت في مجلات علمية محكمة وكتب مطبوعة. خلصت الدراسة إلى أن عملية تأثير طرق إعداد الامتحانات قد جذبت العديد من البحاث في هذا المجال، و لكن النتائج أوضحت أنها قد لا تكون ذات سلطة أو تأثير مباشر على أداء الممتحنين في الاختبارات، وذلك لاحتمالية تدخل غيرها من المؤثرات الاختبارية. و من هذا المنطلق توصي هذه الورقة البحثية بضرورة الشروع في دراسات أعمق لإثراء هذا المجال، خاصة في البيئة الليبية، وذلك لخلو مكتباتها من دراسات تهتم بهذا الموضوع.

*الكلمات المفتاحية:* تأثير طريقة الاختبار ، الممتحنون و درجاتهم ، الامتحانات الموضوعية، الامتحانات المقالية.

## Introduction

This section, firstly, describes the two types of test format, and briefly mention how each type can be scored. The second sub-section deals with the notion "test method effect". The sub-section that follows highlights the rationale and significance of the study, followed by the raised research objectives and questions. Then, how the raised research questions to be addressed is discussed in the methodology section.

### Types of Test Format

"Test item format fall into two broad types: selected-response and constructed-response"(Osterlind, 2002, p.30). Selected-response test items require the test-taker to respond to a question item by selecting the given correct answer from a number of other available options. As demonstrated by Douglas (2010), multiple-choice task is the most common type of selected response items. In a multiple-choice item, examinees are instructed to choose the correct option or alternatives from those presented by the examiner. Khodadady (1999) illustrates that the alternatives used in multiple-choice items could be categorized as keyed response and distracters. Further, keyed responses are of different types; they are one correct response, the best answer and multiple responses. To clarify, alternatives are typically straightforward and distinct from each other in the item of one correct response. The best answer is the situation when the most appropriate answer is to be selected by the examinees. Nevertheless, multiple response items can be quite different as they sometimes involve more than one correct response; besides, test takers are not penalized if they choose only one option for the reason that the examiners are primarily interested in test-takers' strategies in solving problems.

Furthermore, other types of selected response item formats include matching tasks, dichotomous tasks, ordering tasks, information transfers and editing tasks (Douglas, 2010). Selected response items are used to assess abilities including mastering vocabulary, grammatical structures and other aspects of comprehension and subtle knowledge of languages. Furthermore, selected response items are economical in scoring though they are quite demanding in their construction and development (Douglas, 2010).

Constructed response items, on the other hand, require test takers' to create or construct a response. In order to respond to constructed response items, test takers have to construct a word or a short sentence perceived to be the correct answer; further, constructed response items may include items that require more extended responses (Osterlind, 2002). Short response items can be presented in different item types; for instance, gap filling which requires examinees to provide a word or a phrase to complete missing information in a sentence. Other item types of short response include short-response questions, cloze test and c-tests (Douglas, 2010).

Regarding extended response items, the discourse required to accomplish the task is longer than a single sentence, for example, it might be in the form of an essay or a composition. According to Kubiszyn and Borich (2003), essays can be longer than one page taking the form of open-ended essays, which are referred to as ‘extended-response essays’; on the other hand, the given responses can be restricted to one page or less, which are referred to as ‘restricted-response essays’. Downing (2009) explicates that constructed response items have many strengths, one of which is that they are more useful when testing writing skills such as adequacy of sentences and paragraph development. Moreover, the responses to these items are non-cued

and essay questions are easier to construct than selected-response format. Further, constructed response items have the advantage of facilitating partial credit scoring.

Notwithstanding, it is contended that constructed response items are difficult to be scored accurately and reliably (Downing, 2009). Besides, when there are a large number of examinees, the scoring is time consuming and costly. Other problems related to validity might be encountered when using constructed response items; for instance, potential threats on validity may occur as a result of the subjective nature of scoring and biases associated with raters. Furthermore, content validity is limited since sampling is restricted because of testing time constraints (Downing, 2009).

Scoring selected response items and constructed response items is of different methods and procedures. When a test is designed using closed response items, the scoring typically involves counting the number of correct responses before raw scores being transformed into more meaningful scales. This type of scoring does not require raters to make judgments or personal opinions about a response. "Cloze-test procedure test items are dichotomously scored" (Osterlind, 2002, p.31). That is, two predetermined categories are possible for test takers' responses which are correct and incorrect. Such type of scoring can be mechanical, using scoring keys or computer programs.

Moreover, according to Downing (2009), there are two basic formulas to score selected response tests: they are 'number-correct' score and 'correction-for-guessing' score. The former is manipulated by counting the number of correct responses and the latter attempts to mitigate the perceived effects of random guessing. Correction-for-guessing score formula either rewards test

takers for not trying guessing or penalize guessing. However, even though reliable, those formulas are criticized. Downing (2009) claims that such formulas may not achieve the goal for which it was stated. The author further explicates that correction-for-guessing formula reduces validity evidence because they count a personality trait, which is construct-irrelevant variance.

Scoring constructed response items typically require judgments on the part of test raters and all alternative responses are to be considered when scoring. Douglas (2010) emphasizes the importance of providing explicit detailed criteria for scoring constructed response items. This criteria would reflect the abilities to be measured, such as syntactic forms and vocabulary. Moreover, raters are needed to be well-trained in order to insure consistency in results. An example of constructed response scoring is essay scoring. There are two approaches to essay scoring, analytical and holistic (Downing, 2009). Analytical scoring methods rate one or more specific features and characteristics of an essay, while holistic scoring rate the overall quality of an essay.

Notwithstanding, "restricted-response-questions are difficult to score consistently across individuals" (Kubiszyn & Borich, 2003, p.135). That is, when scoring essay questions different scorers may give different ratings to the same answer. However, to enhance scoring reliability, well-written items and explicit rating criteria are to be employed when scoring constructed response items.

**Test Method Effect**

Test method effect is a term used when the method used for testing a language ability affects students' scores. That is, students' performance and

scores are likely to be affected by the features of the test format; such effect should be reduced as much as possible (Bachman & Palmer, 1996).

"Recent research has demonstrated that the wording and format of test items can greatly influence the psychological perspectives that the examinee brings when considering a response" (Osterlind,2002, p.2). As illustrated by Osterlind (2002), these psychological perspectives such as anxiety, motivation and performance are important to be attended when constructing good items. It was further explained that other technical considerations are also to be attended such as the level of vocabulary and determining the optimal number of response alternatives.

Moreover, Bachman and Palmer (1996) stress the importance of understanding the fundamental issues, approaches and methods used in measurement and evaluation when preparing a test. Further, research in language testing demonstrated that a test might affect test-takers with different characteristics in ways that are irrelevant to the abilities being tested (Bachman, 1990). The degree to which test-takers could perform well in a given test might vary according to the method in which the test is presented. For instance, some individuals might perform better on a multiple-choice test than on a constructed response test as a consequence of the different individual attributes. Bachman (1990) highlights the importance of test method facets since "performance on language tests varies as a function both of an individual's language ability and of the characteristics of the test method" (Bachman, 1990, p. 113). There are five sets of test method facets as presented by Bachman (1990). These include: the testing environment, the test rubric, the nature of the input the test-taker receives, the nature of the expected response to that input, and the relationship between input and

response. Towards a deeper understanding of the test method effect, "more research looks at how students actually respond to a particular test method" (Alderson, Clapham & Wall, 1995, p.44).

## Rationale and Significance of the Study

This study is important for several reasons. First and foremost, more interest for carrying out other studies in different contexts can be attracted based on the findings of this study, particularly in the context of the current study, i.e. Libyan English classrooms. This is because, to the current researcher's best of knowledge, there has been no research carried out to find out the extent to which the type of test methods may affect positively or negatively students' obtained scores in tests in the context of this study. Secondly, conclusions drawn from this paper, hopefully, would have important implications for both test-takers and test-makers when English language tests are administered in language classrooms

Thirdly, this study is significant in terms of its methodology. The study deploys a documentary research method. "This [kind of] method has had little attention compared to other methods … and often marginalized or even used, it only acts as a supplement to the other general social research methods" (Ahmed, 2010, pp. 1 - 2). Also, it was contended by Bowen (2009) that "there is some indication that document analysis has not always been used effectively in the research process, even by experienced researcher" (p.27). So, methodologically, this study will add insights to the literature pertinent to research methods used in social and educational researches. That is, document analysis can be used as a valuable instrument to collect data in order to carry our research studies.

## Objectives and Research Questions:

The general objective of this paper is to clarify the concerns related to the notion of test method effect based on the analysis of previously carried out research. It aims at describing the process researchers used to examine the existence of test method effect. Further, it intends to explain the test method effect and its relationship with candidates' performance by reviewing and analyzing pertinent literature.

So, this research attempts to investigate the scope of the following questions:

1. How are the investigations regarding test method effect carried out?
2. Which methods and systems are used in this field?
3. What kind of relationship has literature documented between the type of test methods and test-takers' gained scores in a test?

## Methodology

This integrative literature review research is descriptive in nature utilizing qualitative research design by making use of document analysis of journal articles and books. By doing so, an objective critique and drawing conclusions about the issue investigated can be reached with the help of the analysis of the previously carried out research studies (Christmals , Dela & Gross, 2017), and such type of data collection method "can be treated as a source of data in their own right" (Denscombe, 2003 ,p. 212).

So, this paper is a narrative review of related literature as the current researchers conducted an online search process databases from journal articles and books to obtain selectively related papers. Papers went through screening process to determine their usefulness to be included in this paper. Although the reviewed studies were evaluated from different angles, this

study focused primarily on the issues related to the notion "test method effect" endeavoring to shed light on the topic studied. Then, the study information was extracted and reported, and conclusions were drawn.

**Empirical Research on Test Formats Affecting Examinees Performance**

The different test formats discussed above have been questioned by several researchers who have investigated whether they have some impact on students' performance or not (Bachman & Palmer, 1996). However, there is considerable research in language testing that demonstrates this effect.

Currie and Chiramanee (2010) examined the test method effect of two test formats, MC (multiple-choice) and CR (constructed response), in English language tests. The study was conducted as an attempt to raise awareness of the effectiveness and effect of the use of MC item. Two tests with the two formats were administered to correlate the results and the distracters on the MC test were based on incorrect answers from the CR test. The argument presented in the above mentioned study totally criticizes MC format and challenges it as being inadequate to assess dimensions of cognitive performance. The study claims that MC formats emphasize recall and test taking strategies rather than generation of answers. To illustrate, Currie and Chiramanee (2010) demonstrated that 'format-related noise' is sought through MC format rather than language related responses that would be reflected in other test formats. They support this claim by pointing to a finding that the options selected in MC items by test-takers do not correspond with their own previous answers on CR items, particularly when giving incorrect answers. Thus, such findings questioned the ability of MC questions to be a valid method of measuring language knowledge and proficiency.

In contrast with previous research, Lissitz and Hou (2007) claim that elimination of CR items from test comprising of MC items and CR items does not form critical harm to the conclusions derived from the tests. This study manipulated correlating and observing test results on three situations of eliminating either MC or CR item formats from a test and a mixture of both formats. The study was conducted on different subjects including English language and on different groups categorized in accordance with their race and gender. The results demonstrated that the order of the ethnic groups depending on their scores remained the same for the three situations with similar patterns of standard deviation. Further, females performed better in the English test regardless of whether CR items were removed or not.

In addition, the results of the above study were also reported in terms of reliability which decreased when removing CR items. Nevertheless, hypothetically adding MC items equal to the number of points lost from dropping the CR items is believed to counter the effect of reliability being decreased as a result of removing CR items. However, the researchers did not refer to the fact that the unchanging order of the groups across different format might infer other dimensions of test method effect. Thus, further investigations were needed to correlate results and find out whether there was test bias resulting from using different test methods.

Based on the results of Lissitz and Hou (2007), Lissitz and Hou (2012) further investigated the previous results concluding that there is a multidimensionality of mixed format tests because of the interrelated correlations presented between MC and CR item formats. It was assumed to be due to the characteristics of the tests. To illustrate, the examined interaction of the item formats of ethnic groups elicited a performance gap

between groups that was smaller in CR items than MC items in English language test. Moreover, interactions were also found between item format and gender. As deduced by the researchers, the relatively higher level of performance of females to males on CR items was presumably a support of the theory that females possess greater verbal abilities which consequently resulted in higher scores on CR items. However, Lissitz and Hou (2012) concluded that skills are not equally assessed using CR and MC item formats.

Another research on the effect of test method on students' performance was conducted by Hassani and Maasum (2012). They examined students' reading comprehension on summary writing and open-ended question formats. For deeper investigation, the researchers grouped test takers in accordance to their proficiency using TOEFL reading comprehension test into intermediate and low achievers. The researchers concluded that both groups showed better performance in summary test regardless of their level of proficiency. The researchers justified this diversity in results by referring to the fact that students were unfamiliar with the open-question types, for example, literal, application, inferential and evaluation questions. Nevertheless, it might be inferred that students being aware of the methods used in a test would have yielded more practical and applicable data tackling the focus of the research.

A further research carried out by Shahivand, Paziresh, and Raeeszadeh (2014) investigated whether test methods have impact on Iranian EFL test-takers' performance in terms of their proficiency level. It was concluded that test format had impact on examinees' performance variously and at different proficiency levels. Intermediate and upper-intermediate levels were reported to show poorest performance on MC test format compared to a much better

performance on C-test, T/F and cloze test. According to the researchers, the reason for this relatively low performance is hesitation due to the learners vast knowledge of grammar and vocabulary, which formed some sort of misgiving because they had to choose the most correct option among the correct options presented. Such performance was claimed to be a result of examinees' either lack of attention or over-attention to the clues provided. On the contrary, the intermediate and low level students had their poorest performance on cloze test. It was assumed to be as a result of their unfamiliarity with this format. On other formats, intermediate and low level students' would recall and choose the correct response when they read some clues in the options or first part of the stem presented. This study concludes that using different item formats results in better judgment and enhances reliability of scores and test validity.

Shin (2008) examined the possibility of different response formats affecting the hierarchical level of information in a constructed response task comprising incomplete outline, open-ended questions and summary responses. Listening comprehension of main ideas performance was best on summary task and worst on open-ended questions. It was explained that this variance is "due to the fact that open-ended questions ask for more straightforward options than summary tasks do" (Shin, 2008, p. 119). On the other hand, when major ideas are being examined, test-takers were observed to perform better on incomplete task than on open-ended task. It was explained that this result was due to the more cues provided by incomplete outline tasks. However, it was reported that summary tasks were easier in this respect because incomplete outline task involves understanding the predefined text structure as presented. On the contrary, when supported

details were examined, the easiest task was the incomplete outline whereas the summary was the hardest. It was claimed that the mean score differences in main, major and supporting detailed ideas between test formats reflect the difficulty associated with eliciting information from the text.

Salehi and Sanjareh (2013) investigated the impact of response format on learners' test performance of grammatical judgment tests. Comparing dichotomous and multiple choice types, the results were drawn to conclude that response format could affect reliability and consequently validity of a test to differing degrees. Thus, the researchers pointed out the problem of response format as a systematic measurement error on test-takers' performance.

Kindergarteners' expressions of depth of knowledge of vocabulary were examined in relation to in-context and out-of-context format by Christ, Chiu, Currie, and Cipielewski (2014). Eliminating any other variance, the researchers concluded that most differences in students' answers were a result of different test format. To clarify, it was deduced that out-of-context test format is more sensitive for detecting child's deep vocabulary knowledge while in-context format is more sensitive for detecting low levels of vocabulary knowledge, or decontextualized knowledge of a word because students might be deriving vocabulary meaning rather than having the response derived from knowledge in their lexicon.

MC item, CR item and CRE item formats in English language reading comprehension test of two groups of students were investigated by Zheng, Cheng, and Klinger (2007). The researchers concluded that higher achievement was achieved in MC questions, lower in CR questions and lowest in CRE questions. It was demonstrated that MC questions are

generally easier to be answered correctly as compared to the other two formats. This was explained to be the result of MC items require comprehension and selection while CR items require comprehension and production. However, the higher results on MC items might be the result of strategies such as 'test-wiseness' (Zheng et al., 2007).

Ward, Dupree and Carlson (1987) examined students' performance on MC item format and CR item format in a reading comprehension test. The items were balanced in accordance with the kind of information processing needed to answer them. However, different applications proved the existence of slight systematic differences associated with test format. On the other hand, evidence of systematic differences for individuals by format interactions was detected. It was explained that individual performance between the two formats differed amongst students.

With relation to scoring rules and procedures, correlations and distinctions between MC and CR item format were investigated. Kastener and Stangl (2011) alerted that using different scoring rules when comparing test formats would yield deferent results. It was claimed that CR tests would be equal to MC tests with *NC* scoring- number correct-. Other scoring rules like *WU* –University-specific- and *AN* –All-or-nothing- scoring rules cannot be used interchangeably with CR tests as they are regarded to be stricter as *WU* and *AN* scoring methods do not reward partial knowledge and penalize guessing.

In the above surveyed research, discussions and conclusions towards test method effect were mainly held by manipulating correlations between the scores of tests with different item formats. For example, Currie and Chiramanee (2010) correlated the results of MC and CR test formats to reach

their conclusions. Further, researchers attempted to spot any interactions between test scores and the target groups referring to influences such as gender, race or proficiency level of the candidates. Referring to such interrelations would permit the researchers to raise assumptions rather than merely explain figures. To exemplify, examining such relations of interactions among the groups investigated by Lissitz and Hou (2012) brought the researchers to the conclusion of supporting the theory stating the greater verbal ability that females possess when compared to males. Shahivand et al. (2014) could deduce individual attributes related to the proficiency level of the candidates such as hesitation.

Other research investigated the correlations of test item formats referring to the relationship between the input and the output in given tests. It was also manipulated that there has been a correspondence between test item formats and the required information processing to achieve correct answers (Shin, 2018; Zheng et al., 2007). A different dimension was tackled by Kastener and Stangl (2011) who brought up the assumption that scoring methods might interfere with reliability of candidates scores; thus, the investigation was held not only by correlating two different item formats but also by correlating different results of different scoring methods that could be used in a test.

Different researchers have concluded divergent assumptions in the field of understanding the existence and the degree to which test methods affect students' performance. "Unfortunately, our understanding of test method effect is still so rudimentary that it is not possible to recommend particular methods for testing particular language abilities" (Alderson et al., 1995, p.45).

## Conclusion

Language tests can be presented in different formats, and scored by different scoring rules and procedures. They can be administered using constructed response or selected response means. This diversity in formats raise varying questions regarding the impact of test formats on students' performance; thus, test method effect was investigated through considerable research. Arguments regarding this issue range from supporting the existence of test method effect to partially denying its direct relation to students' performance. To exemplify, MC items were at times questioned for not being reliable and adequate tool of measuring language abilities as asserted by Currie and Chiramanee (2010). At other times, MC items were accused of measuring constructs irrelevant to language abilities. Further, using different test formats would result in unequal assessment of the skills as pointed out by Lissitz and Hou (2012). However, some research claims that test method effect is not highly harmful to be critically considered. On the one hand, researchers may not accuse the test item format used to be responsible for candidates different performance, but they refer to the scoring methods and procedures as an affecting factor of test scores, such as Kastener and stangl (2011).

To conclude, using different research processes and investigating different angles of relationships that could be encountered when assessing language, whether language related or not, resulted in reaching divergent conclusions regarding the notion of 'test method effect'. This multiplicity of suppositions require deeper investigations to be implemented to better specify the impact of test method effect. Thus, this review may attract a

further research to be conducted, particularly in the context of the current study.

## References

Ahmed, J. U. (2010). Documentary research method: New dimensions. *Indus Journal of Management & Social Sciences*, *4*(1), 1-14.

Alderson, C., C. Clapham & D. Wall. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press

Bachman, L. & Palmer, A. (1996). *Language testing in practice*. NY: Oxford University Press

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

Bowen, G. A. (2009). Document analysis as a qualitative research method. *Qualitative research journal*, *9*(2), 27-40.

Christ, T., Chiu, M., Currie, A & Cipielewski, J. (2014). *The Relation Between Test Formats and Kindergarteners' Expressions of Vocabulary Knowledge.* In *Reading Psychology,* vol. 35 (6). (pp. 499-528). NY: Routledge. Retrieved from: http://dx.doi.org/10.1080/02702711.2012.746249

Christmals, C. Dela, & Gross, J. J. (2017). An Integratıve Lıterat ure Revıew Framework For Postgraduate Nursıng Research R eviews. *European Journal of Research in Medical Sciences*, *5*(1) , 7–15

Currie, M. & Chiramanee, T. (2010). *The effect of the multiple-choice item format on the measurement of knowledge of language structure.* In *Language Testing,* vol. 27(4). (pp. 471–491). DOI: 10.1177/0265532209356790

Denscombe, M. (2003) *The good research guide : for small-scale social research projects*, 2nd ed. Buckingham: Open University Press.

Douglas, D. (2010). *Understanding language testing*. NY: Routledge.

Downing, S. (2009). *Written Tests: Constructed-response and Selected-response Formats.* In *Assessment in Health Professions Education,* (pp. 169-204)

Hassani, L. & Maasum, T. (2012). *A Study of Students' Reading Performance in Two Test Formats of Summary Writing and Open-ended Questions.* In *International Conference on Education and Educational Psychology,* vol. 69. (pp. 915-923). Universiti Kebangsaan Malaysia. Retrieved from: https://doi.org/10.1016/j.sbspro

Kastner, M. & Stangl, B. (2011). *Multiple Choice and Constructed Response Tests: Do Test Format and Scoring Matter?*. In *Procedia Social and Behavioral Sciences*, vol. 12 (pp. 263–273). Retrieved from: https://doi.org/10.1016/j.sbspro.2011.02.035.

Khodadady, E. (1999). *Multiple Choice Items in Testing: Practice and Theory.* Iran: Rahnama Publications.

Kubiszyn, T. & Borich, G. (2003). *Educational Testing and Measurement: Classroom Application and practice.* US: John Wily & Sons, INC.

Lissitz, R. & Hou, X. (2007). *Multiple Choice Items and Constructed Response Items: Does It Matter?.* University of Maryland

Lissitz, R. & Hou, X. (2012). *The Contribution of Constructed Response Items to Large Scale Assessment: Measuring and Understanding their Impact.* In *Journal of Applied Testing Technology,* vol.13 (3).

University of Maryland. Retrieved from: http://www.jattjournal.com/index.php/atp/article/view/48366

Osterlind, S. (2002).*Constructing Test Items: Multiple-Choice, Constructed-Response, Performance and Other Formats.* NY: Kluwer Academic Publishers

Salehi, M. & Sanjareh, H. (2013). *The Impact of Response Format on Learners' Test Performance of Grammaticality Judgment Tests.* In *Journal of Basic and Applied Scientific Research,* vol. 3.(2). (pp. 1335-1345). Retrieved from: https://www.researchgate.net/publication/283122456

Shahivand,Z., Paziresh, A. & Raeeszadeh, A. (2014*). The Effects of Test Formats on the Performances of Iranian EFL Students.* In *Theory and Practice in Language Studies,* vol. 4. (2) (pp. 366-373). Academy Bublisher: Finland.doi:10.4304/tpls.4.2.366-373

Shin, S. (2008). *Examining the Construct Validity of a Web-Based Academic Listening Test: An Investigation of the Effects of Response Format.* In *Spaan Fellow Working Papers in Second or Foreign Language Assessment,* vol. 6. (pp. 95–129). English Language Institute. University of Michigan.

Ward, W., Dupree, D. & Carlson, S. (1987). *A Comparison of Free-Response and Multiple-Choice Questions in the Assessment of Reading Comprehension.* In *Research Project.* New Jersey: Educational Testing Service Princeton.

Zheng, Y., Cheng, L. & Klinger, D. (2007). *Do Test Formats in Reading Comprehension Affect Second Language Students' Test Performance Differently?.* In *Queen's University.* Retrieved from: https://www.researchgate.net/publication/236682610