**INDONESIAN JOURNAL OF APPLIED LINGUISTICS**
Vol. 9 No. 3, January 2020, pp. 536-544

Available online at:
https://ejournal.upi.edu/index.php/IJAL/article/view/23203

doi: 10.17509/ijal.v9i3.23203

# Washback effects of multiple-choice, cloze and metalinguistic tests on EFL students writing

**Danial Shirzadi* and Majid Amerian**
*Department of Literature and Foreign Languages, Faculty of Foreign Languages, Arak University, Iran*

**ABSTRACT**

The washback effects of different test formats on the writing performance of students have always been of great importance. However, this area of research has not fully touched upon by researchers of second language testing. Despite the importance of the issue, there is a dearth of empirical studies to unravel the effects of different types of tests on learning. To shed some light on the current issue, the present study intends to look into the washback effects of tests on students who are learning and using some special grammatical points in writing tasks. In order to fulfil this project, we made a set question in three formats of cloze, multiple-choice and metalinguistic on a grammatical form(i.e. present perfect and present perfect continuous)to use after each session of teaching (2 sessions of training) as an activity. The researchers devised and validated three tests on the target form; each test contained 20 questions and was in different formats of cloze, multiple-choice or metalinguistic. At the end of this two-session trainings, two focused writing tasks were implemented. The results indicated that supporting teaching grammatical points with metalinguistic tests yields the highest positive washback on students writing. Finally, some practical implications were suggested.

**Keywords:** Assessment; multiple choices; cloze; metalinguistic tests; writing

| **First Received:** | **Revised:** | **Accepted:** |
|---|---|---|
| 27 September 2019 | 20 November 2019 | 12 December 2019 |
| **Final Proof Received:** | | **Published:** |
| 18 January 2020 | | 31 January 2020 |

**How to cite (in APA style):**
Shirzadi, D. & Amerian, M. (2020). Washback effects of multiple-choice, cloze, and metalinguistic tests on EFL students writing. *Indonesian Journal of Applied Linguistics, 9*, 536-544. doi: 10.17509/ijal.v9i3.23203

## INTRODUCTION

It does not matter that in which context, school or university, the practice of language teaching is being conducted; teaching is always subdivided into four phases including planning, teaching and learning, and evaluation (Ellis, 2003). Teaching goals are set in the planning phase in order to help to find activities, which are capable of providing learners with meaningful learning processes. Then, when it comes to the teaching and learning phase, all that teachers must do is to engage their learners in suitable learning strategies (Biggs, 2003). Finally, teachers need to conduct an evaluation to find out about the efficiency of the utilized teaching and learning strategies for the accomplishment of the teaching goals. However, successful teaching cannot be implemented unless some kind of meaningful correspondence connects these ingredients. Furthermore, aligning learning activities and assessment strategies is a critical trait that needs to evolve in language teaching. Undoubtedly, such an alignment can be achieved when teaching goals, learning strategies, teaching strategies and evaluative tests all correspond to each other.

According to Ellis (2003), the educational purpose of assessment is to provide the language learners with feedback, motivation, guidance and learning support. To achieve a successful assessment, there should be a clear sense of what the course is designed to accomplish (Palomba & Banta, 1999). Once the learning outcomes have been clearly defined, the development of assessment methods for determining whether these outcomes have been met or not become more attainable. Teaching methods typically make general statements about the assessment methods (e.g., essay test, peer assessment, learning contract, oral examination). On the other hand,they should contain details regarding the

* Corresponding Author
Email: shirzadidanial@yahoo.com

assessment method alongside a concrete set of assessment resources (e.g. tests, test items, peer assessment forms).

As far as English language teaching is concerned, assessment seems to be unavoidable since there should be some method to measure a person's language ability (Brown, 2004). As it was previously mentioned, maintenance of correspondence or alignment among four phases of teaching is inevitable; therefore, tests must be closely associated with pedagogical purposes (Bachman & Palmer, 1996). Accordingly, a considerable portion of language testing literature refers to the effects of tests on teaching and learning known as the washback effect (Hughes, 2003). According to Hughes (2003), washback refers to the positive or negative influence that tests have on teaching and learning. Despite its relatively easy definition, the bulk of studies in this area suggest that it is an extremely complex phenomenon as there is no consensus on its effects(Green, 2006; Rea-Dickins & Scott, 2007; Spratt, 2005; Watanabe, 2004).Since studies have been conducted to examine the washback effect are scarce (Safa & Goodarzi, 2014) especially regarding writing skill, and to explore the test format which has the most washback effect on students' writing skills,the current study will be an attempt to fill such gap in the literature.

**Different ways of defining washback**

When it comes to applied linguistics, there are several ways to define the concept of 'washback'. In its most simplified version, it refers to the positive or negative effects that tests may have on teaching, learning processes, students, teachers, policymakers and other stakeholders(Alderson & Wall, 1993; Hughes, 2003). Today, there is a growing concern for such an influence among both theoreticians and practitioners in the realm of language teaching, and it also has been reflected in the curriculum, teaching materials, teaching methods, testing procedures and, in a nutshell, in the leaning process (Spratt, 2005). Despite having a seemingly straightforward definition, the literature suggests that washback is an extremely complex phenomenon as there is no consensus on the subject (Green, 2006; Rea-Dickins & Scott, 2007; Spratt, 2005). In order to come to a better understanding of this multidimensional phenomenon, scholars felt that the washback issue should be studied from various aspects such as different effects of it on different stakeholders.

One of the strongest determining factors that can enhance the washback influence of a test refers to the importance of that test in taking big decisions. Sometimes, tests have direct or indirect life-changing influences over careers' of the test takers that is they are high stake tests. A university entrance test is a good example of this notion from which the concept of measurement-driven instruction emerges (Pearson, 1988). Some scholars believe that this phenomenon could be beneficial for teaching and learning with the assumption of having properly constructed and implemented tests (Qi, 2005). On the other hand, there are other scholars who are criticizing washback due to its tendency to narrow down the curriculum (Madaus, 1988). They believe that test-driven instruction limits students' and teachers' creativity (Wall, 2000).

Although validity is a well-defined and properly inquired concept in testing, it is still one of the interesting areas for scholars who are interested in the washback issue. Morrow (1986) believes that a test's validity should be measured by the degree of its beneficial influence on learning and teaching. With this in mind, validity acquires a new educational purpose that could result in curricular and instructional changes (Pan, 2009). However, this perspective suffers from a serious weakness since scholars have not managed to introduce proper ways for the empirical establishment of test validity in this perspective. To confront this problem, Alderson and Wall (1993) tried to introduce a more unified concept of validity in which washback had been addressed as a part of the test validity:

Whereas validity is a property of a test, in relation to its use, we argue that washback, if it exists - which has yet to be established - is likely to be a complex phenomenon which cannot be related directly to a test's validity. (Alderson &Wall, 1993, p. 116)

Later, Messick (1996) utilized the term 'consequential validity' to propose a stronger argument and put this notion within a stronger theoretical framework. He suggests investigating "validity as a likely basis for washback" rather than "seeking washback as a sign of test validity" (p. 252). He believes that consequential validity entails facets such as:

Evidence and rationale for evaluating the intended and unintended consequences of score interpretation and use in both the short- and long-term, especially those associated with bias in scoring and interpretation, with unfairness in test use, and with positive or negative washback effects on teaching and learning. (Messick, 1996, p. 251)

**Intended vs unintended washback**

There is a common misconception that we one can differentiate good tests from bad ones based on their beneficial or detrimental washback effects (Heaton, 1988). However, a deeper look at the nature of tests and the washback phenomenon reveals that the correspondence between the quality of a test and positive washback is not always operational (Hughes, 2003). Of course, this new perspective has not led to the omission of utilizing positive and negative washback in the related literature of the field. Instead, the purpose is to highlight the fact that washback might be independent of the quality of the test and there may be other factors in the scene (Messick, 1996). In language testing, negative washback has been usually attributed to tests' limiting influence on content and creativity. As such, 'teaching to the test' is considered as an unholy byproduct of some tests that would result in lack of motivation and lack of knowledge. On the other hand, tests are usually able to enhance the learners' motivation and empower them with a sense of accomplishment (Pan, 2009).

**Empirical studies on washback**

The washback phenomenon had not received much attention from language testing researchers until the early 1990s. In 1993, Green wrote an article about the effects of established testing programs and introduced themselves as the pioneers of empirical research in the field (Green, 2013). Afterwards, many studies have been conducted to explore the washback effects of high-stakes tests with a focus on content, teaching and learning. In the following notes, some of these studies will be briefly reviewed.

Alderson and Hamp-Lyons (1996) studied the washback effects of TOEFL (international proficiency test) in the USA and found a widespread tendency for teaching to the test in TOEFL classes. A few years later, Andrews, Fullilove, and Wong (2002) inquired the washback influence of national Hong Kong advanced English oral examination required for admission into the university and concluded that due to high stakes of tests, linguistic knowledge and test-oriented skills were still the main focus of instructors, contrary to the intentions of test constructors.

In 2004, in New Zealand, Read and Hayes used interviews, questionnaires, classroom observations and tests scores in order to study the washback effects of IELTS (international proficiency test) for tertiary study and came to the conclusion that negative washback effects of such tests are more observable in intensive course (Read & Hayes, 2004).

One year later, Qi (2005) studied the washback effect of national matriculation English test (NMET) as part of university entrance test battery in middle schools of China and did not spot the presence of intended washback. However, Green (2006) whose research context's country was the same (China) found washback on course content in his study of IELTS academic writing for tertiary study. In 2009, Shih conducted an inquiry regarding the washback effects of GEPT (national English proficiency test) in Taiwan and found limited and teacher-specific washback on teaching practices in the context with GEPT requirement (Shih, 2009).

**Washback and writing tests**

It is possible for the washback effect to work for improving the learners' writing ability when the test design is in accordance with the identification of the ability which is supposed to be tested. Therefore, defining the construct – writing ability – is one of the most fundamental concerns in developing a test of writing. Writing is a very complex cognitive activity and to come up with a thorough understanding of this process we need to refer to previously established models (Bachman & Palmer, 1996; Grabe & Kaplan, 1996; Hayes, 1996; Hayes & Flower, 1980).

It is possible to translate the writing ability to two sets of features. The first set includes relevance and adequacy of content, compositional organization, cohesion, and adequacy of vocabulary, and altogether, they are labelled as communicative effectiveness. The second set includes grammar, punctuation, and spelling, and altogether, they are labelled as accuracy. Accordingly, the washback effect can be pedagogically beneficial in writing classrooms if two general results are achieved. First, we need to be able to collect, identify, describe and classify the errors of students through their performance in a writing test and statistically determine their level in writing ability. Second, we must be capable of exploring the effectiveness of adjusting the instructional program with the features of the second language which cause problems for the learners in developing the writing ability.

**Different types of tests and their washback**

Currie and Thanyapa (2010) studied the effect of the multiple-choice item format on the measurement of knowledge of language structure. They conducted their study with a sample of one hundred and fifty-two university undergraduates. These students took a test of English structure first in constructed-response format and later in three, stem equivalent multiple-choice formats. They found a significant and substantial increase in mean and generally in individual scores between the two tests. However, a direct comparison of the responses to the items in the two tests showed that only 26% of the responses were the same. This means that most of what the multiple-choice items measured was directly dependent on the item format.

In another study, Rauch and Hartig (2010) compared multiple-choice with open-ended response formats of reading test items. They focused on the dimensionality of a reading comprehension assessment with non-stem equivalent multiple-choice items and open-ended items with German test data of 8523 9$^{th}$ graders. Accordingly, they concluded that a two-dimensional item response theory model with within-item multi-dimensionality had a superior fit compared to a uni-dimensional model.

Mozaffari, Alavi and Rezaee (2017) investigated the impact of response format on the performance of grammar tests. They compared multiple-choice items with their constructed response stem-equivalent in a test of grammar using the Rasch model in order to compare item difficulties, fit statistics, ability estimates and reliabilities of the two tests. By means of two independent sample t-tests, they investigated whether the differences among the item difficulty estimates and ability estimates of the two tests were statistically significant.

There have been some studies addressing the issue of different test methods and their washback effect on language learning (e.g. Brame & Biel, 2015; Hemmati & Ghaderi, 2014; In'nami & Koizumi, 2009; Khoshsima & Pourjam, 2014; Ko, 2010; Kromann, Jensen & Ringsted, 2009; Mozaffari, Alavi & Rezaee, 2017; Rauch & Hartig, 2010; Safa & Goodarzi, 2014; Sze & Leung, 2014; Watanabe & Koyama, 2008; Wang & Wang, 2013; Zarei & Neya, 2014;) but most of them address a limited type of tests (i.e. they just investigate

 e-ISSN: 2502-6747, p-ISSN: 2301-9468

effects of single types of test like multiple-choice, and none of them addressed metalinguistic tests) or addressed the washback effects regarding to reading comprehension. As a result, the washback effects of different test format on writing performance of students have been rather neglected.

Despite the importance of the issue, there is a dearth of empirical studies to unravel the effects of different types of tests on learning. To shed some light on the current issue, the present study intends to look into washback effects of tests on students who are learning and using some special grammatical (i.e. present perfect and present perfect continuous) points in writing tasks. To pursue this goal, tests in three different formats had been provided including context embedded (cloze test), context reduced (multiple-choice items), and metalinguistic tests(i.e.tests that make students consciously ponder about the grammatical point taught). Afterwards, the study was carried out in three phases: first, grammatical points were taught to four different groups of students. Then, three groups received treatments by taking a test after the teaching phase, but the control group only received an extended time of teaching. At last, all groups took a focused writing task in which the target grammar forms are needed to be used.

In a nutshell, this study has been conducted in order to answer the following questions:Is there any washback effect regarding writing skill for the students who take tests as a learning activity?Which test format can have the most washback effect on students' writing skills when using as a learning activity?

## METHODS

### Subjects

The subjects of the current research were 120 upper-intermediate students, both male and female, studying English as their second language at two private language institutes in Mazandaran, Iran ranging from 17 to 23. To ensure the homogeneity of their proficiency, an Oxford Placement Test (OPT) (Allan, 2004) was administered to the students of four different classes, besides the fact that all of the participants were at the same level according to the institute's evaluation.The participants whose scores were one standard deviation above or below the mean were selected; the rest of the students were excluded from further analyses. Thus, the number of participants decreased to 108. Having eliminated outliers of the previous phase, the researchers measured writing ability of the students through writing section of TOEFL proficiency test from Longman Preparation Course for the TOEFL test (Phillips, 2004) prior to the beginning of the study. In the second phase, students' writings were measured in terms of their accuracy, fluency and syntactic complexity.

According to Kuiken and Vedder (2007), accuracy can be assessed as "the number of error-free T-units, error-free T-units per T-unit and the number of errors per T-unit" (p. 266). It was noted that since finding the first two criteria might be difficult in learners' production, the last one could provide more information about the general accuracy of L2 learners' writing. In the present study, the number of morphosyntactic, lexical, and spelling errors per T-units was counted to measure accuracy. Syntactic complexity was defined as "the number of clauses per T-unit, the number of dependent clauses per T-unit and the number of dependent clauses per total number of clauses" (Kuiken & Vedder, 2007, p. 266). In this study, the number of clauses per T-unit was considered to measure the syntactic complexity of participants' writing performance. Regarding fluency, a measure used by Ishikawa (2006) was adopted. Fluency was assessed in the TOEFL writing posttest as a measure of words per T-units.

Two raters,who were MA holders, scored more than 600 in TOEFL test and had more than ten years experience of teaching, analyzed the paragraphs and a coefficient correlation of 0.91 shows the reliability of assessment. Subsequently, the homogeneity of the respondents in their writing ability was proofed through the mentioned statistical method in the previous phase. Through the above-mentioned process, the total number of 80 upper-intermediate learners were chosen. Then, the participants were randomly assigned to three experimental groups and one control group. The number of participants in each group was 20.

### Instruments

In order to fulfil this project, an OPT test and writing section of a TOEFL test was used to ensure the homogeneity of the participants in terms of their general proficiency level and their writing capability. A set of researcher-made questions in three formats of cloze, multiple-choice and metalinguistic on a grammatical form (i.e. present perfect and present perfect continuous) was used after each session of teaching as an activity. The researchers devised and validated three tests on the target form; each test was in a different format of cloze, multiple-choice or metalinguistic and contained 20 questions(all in all 60 items). In the end, two focused writing tasks to guide participants toward using intended grammatical forms, which were extracted some textbooks (Ellis, 2003; Van Den Branden, 2006),were implemented to investigate the effects of different test formats accompanied by teaching on learners writing ability and their use of target forms.

To validate the three researcher-made sets of tests (cloze, multiple-choice and metalinguistic), the researchers piloted each format of the tests to a class of 30 learners and used a classical true score theory item analysis technique through which item facility, item discrimination and point-biserial correlation were computed for each item.Regarding items facility factor, following Tuckman (1978), items having the p-value of less than 0.33 or higher than 0.67 were considered misfit items for the present study. Tuckman (1978) believed that questions with the share of the right

 e-ISSN: 2502-6747, p-ISSN: 2301-9468

answer less than 0.33 or higher than 0.67 should be rejected. For an item discrimination index, both a point-biserial correlation and an item discrimination index were calculated for each item. According to Henning (1987), a minimum of 0.25 for point-biserial correlation and 0.40 for discrimination index are acceptable for an item to be included in the final version of a test. Accordingly, items with lower levels of correlation and discrimination were discarded. As a result of the above-mentioned process, a total number of 60 tests were chosen out of a pool of 90 items to make three types of test formats (i.e. cloze, multiple-choice and metalinguistic). Each test type encompassed 20 items.

**Procedure and design**

The present study was carried out in two sessions, and each session lasted for 45 minutes. 30 minutes of each session was devoted to teaching a target grammatical form (i.e. present perfect and present perfect continuous). Then, the three experimental groups were given a test of 10 questions (each group received a different type of test on the same subject) and 15 minutes to answer and work on it; while the control group only continued the routine process of teaching. In sum, the three experimental groups received 60 minutes of instruction plus 30 minutes of working with two sets of tests containing 20 items. In contrast, the control group received 90 minutes of teaching for two sessions per se. This course was taught through using one same method of teaching (i.e. inductive teaching) and three different approaches (i.e. context embedded, context reduced, and metalinguistic) of testing as a support to the language learning process.

After two sessions of the above-mentioned intervention, all four groups were asked to complete two grammar-focused writing tasks.The participants were told that task completion is a part of the research, but they were not informed about the purpose of the study until after it finished.

Two experienced raters (both PhD holders in TEFL with more than 15-year experience of teaching) analyzed the paragraphs in terms of their accuracy and syntactic complexity(or awareness of target grammatical form). Fluency was eliminated from the current research as the essence of our interventions is mainly grammatical.Cronbach' Alpha Coefficient correlation was used to ensure the inter-rater reliability and p-value of 0.96 shows an acceptable level of agreement between the two raters.

The design of this study was quasi-experimental, including experimental and control groups with pretest and posttest. Test type was considered as the independent variable (with three levels of context embedded, context reduced and metalinguistic)and writing task completion was considered as the dependent variable of the study. The learners' proficiency level was considered as a moderator variable. SPSS 19th (Statistical Package for the Social Sciences) software package was used for all the statistical analyses in this study. Significance of the observed differences in participants' posttest scores was investigated through ANOVA test.The results of this analysis are presented thoroughly in the following paragraphs.

**RESULTS**

This study aimed to analyze the effects of an independent variable in three levels (i.e. multiple choice, cloze, and metalinguistic testing methods) on a dependent variable (i.e. accuracy and syntactic complexity in writing ability). To reach this aim, a total number of 120 homogenized learners were divided into three experimental and one control groups and went under a two-session intervention. Each of the three experimental groups worked on a test of 10 items in each session after the teaching phase, while the control group only received teaching for all sessions. At last, all groups took part in a writing posttest in which two raters judged their writings in terms of their accuracy and syntactic complexity (or awareness of target grammatical form). The descriptive results of the posttest are summarized in Table 1.

Table 1. Descriptive statistics of the posttest

| | N | Mean | SD | Std. Error | Min. | Max. |
|---|---|---|---|---|---|---|
| Multiple choice | 30 | 12.50 | 2.62284 | .4788 | 7.00 | 17.00 |
| Cloze test | 30 | 15.26 | 2.16450 | .3951 | 11.00 | 18.00 |
| Metalinguistic | 30 | 17.33 | 1.82574 | .3333 | 13.00 | 20.00 |
| Control | 30 | 12.60 | 2.47191 | .4513 | 7.00 | 17.00 |
| Total | 120 | 14.42 | 3.03388 | .2769 | 7.00 | 20.00 |

Table 1 shows the results of the posttest. Participants who have been treated by metalinguistic tests after the teaching part outperformed other groups and the control group (M= 17.33 & Std. Deviation= 1.82). Participants who took the cloze tests achieved a mean score of 12.26 (Std. Deviation 2.16) following by those who took the multiple-choice items (M=12.50&Std. Deviation=2.62). As indicated by Table 1, the mean score of students who have been treated by multiple-choice tests is even lower than the control group who received mere teaching of intended grammatical point (i.e. present perfect and present perfect continuous). In order to answer the first research question and show the significance of observed differences, ANOVA test was run; the results of which are presented in Table 2.

A one-way between-subjects ANOVA was conducted to compare the effect of 3 different testing methods on learning and usage of grammatical forms in a writing task. As it is shown in Table 2, There was a

significant effect of testing methods on writing task at the p<.05 level for the three conditions [F(3, 116) = 30.851, p = .000]. Post hoc comparisons using the Tukey HSD test, which is summarized in Table 3, indicated that the mean score for the metalinguistic test condition (M= 17.33 & Std. Deviation= 1.82) and cloze test condition (M = 12.26 & Std. Deviation 2.16) was significantly different from the no tested treatment condition (control group). However, the multiple-choice test condition (M= 12.50 & std. Deviation= 2.62) did not significantly differ from the no test conditions (control group).

Table 2. ANOVA test to compare the effect of three different testing methods on writing task

| | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Between Groups | 486.092 | 3 | 162.031 | 30.851 | .000 |
| Within Groups | 609.233 | 116 | 5.252 | | |
| Total | 1095.325 | 119 | | | |

Table 3.Tukey HSD test

| | (I) tests | (J) tests | Mean Difference (I-J) | Std. Error | Sig. | 95% Confidence Interval | |
|---|---|---|---|---|---|---|---|
| | | | | | | Lower Bound | Upper Bound |
| Tukey HSD | multiple | cloze | -2.76667* | .5917 | .000 | -4.3091 | -1.2242 |
| | | meta | -4.83333* | .5917 | .000 | -6.3758 | -3.2909 |
| | | control | -.10000 | .5917 | .998 | -1.6424 | 1.4424 |
| | cloze | multiple | 2.76667* | .5917 | .000 | 1.2242 | 4.3091 |
| | | meta | -2.06667* | .5917 | .004 | -3.6091 | -.5242 |
| | | control | 2.66667* | .5917 | .000 | 1.1242 | 4.2091 |
| | meta | multiple | 4.83333* | .5917 | .000 | 3.2909 | 6.3758 |
| | | cloze | 2.06667* | .5917 | .004 | .5242 | 3.6091 |
| | | control | 4.73333* | .5917 | .000 | 3.1909 | 6.2758 |
| | control | multiple | .10000 | .5917 | .998 | -1.4424 | 1.6424 |
| | | cloze | -2.66667* | .5917 | .000 | -4.2091 | -1.1242 |
| | | meta | -4.73333* | .5917 | .000 | -6.2758 | -3.1909 |

*The mean difference is significant at the 0.05 level.

Taken together, these results suggest that taking metalinguistic and cloze tests as a learning activity really does have a significant effect on learning and using those forms in writing tasks. However, it should be noted that multiple-choice items were found not to have any significant effects on learners uptake and output of the intended forms. Accordingly, while MCs are an objective way of assessing students' mastery over a form in a context reduced situation but they are not a suggested method for assisting language learning based on the results of the current research, especially in boosting writing ability.

## DISCUSSION

Assisting language learning through testing is not a myth, but there is a consensus on the positive effects of testing on teaching and learning (Andrews et al. 2002; Chapman & Snyder, 2000). The best portrait of the issue may be pointed by Elton and Laurillard (1979) as they believe "the quickest way to change student learning is to change the assessment system". Most of the studies on different effects of assessment on learning have been carried out through the lenses of washback studies, and most of these washback studies have been concerned about teachers, learners or stakeholders' perspectives on the concept. Washback effects as a result of different practical assessment methods and techniques have been remained fairly obscured though they are of crucial importance to fully comprehend the concept (McNamara, 2001). Accordingly, the present study aimed to investigate the washback effects of different grammar-focused test techniques on learners writing task completion. The results of this study suggested that there is a positive and significant washback effect on students' writing performance as a result of assisting teaching through different testing techniques.

The mentioned finding is in line with Brame and Biel (2015), Chehrazad and Ajideh (2012), Ko, (2010), Kromann et al. (2009), Talebzadeh and Bagheri (2012), Zarei and Neya, (2014), but it is a rather sharp contrast with Loch (2010). Talebzadeh and Bagheri (2012) reported a positive washback effect of cloze tests on students' vocabulary learning. Brame and Biel (2015) declared that various testing format can enhance learning and they suggested that feedback on tests would enhance the beneficial positive washback effects of tests. Loch (2010) ,while accepting the joint effects of test format with other factors like text difficulty or test takers characteristics, mentioned that "task type and native language use as test method variables, rarely have a statistically significant affect separately" (Loch, 2010, p. 924). These rather opposing results could be partly due to "gender, language spoken at home, and school track" (Rauch & Hartig, 2010, p. 35). Test usefulness factors (i.e. reliability, construct validity, authenticity, interactiveness, impact, and practicality) may be in charge (Backman & Pulmer, 1996) which should be controlled in future studies.

As a post hoc test illustrated, metalinguistics items loaded the highest effect on students' writing performance, followed by cloze and multiple-choice tests. Furthermore, there was not any significant effect on multiple-choice items compared to the control group.

Khoshsima and Pourjam (2013) and Mozaffari and Alavi (2017) reported opposing results in favour of multiple-choice format tests but in these studies tests were the final goal and they do not relate tests to learning especially to skills such as writing. Alternatively, Mizumoto, Ikeda and Takeuchi (2016) accepted the significant positive effects of cloze tests on learning and proposed that "cloze tasks require greater cognitive processing than multiple-choice tasks in reading comprehension using brain imaging. Overall, brain imaging results supported this hypothesis, with greater mean cerebral activation for cloze tasks than for multiple-choice tasks and control tasks." (Mizumoto, Ikeda, & Takeuchi, 2016, p. 74)

The results indicated that supporting teaching grammatical points with metalinguistic tests would yield the highest positive washback on students writing. This is in line with the findings of Wang and Wang (2013) who found significant washback effects of explicit teaching and metacognitive awareness with academic writing and reading among English language learners. The superiority of metacognitive tasks to enforce grammaticality in writing could have happened due to some reasons. As Swain, Lapkin, Knouzi, Suzuki, and Brooks (2009) concluded that:

> It reflects how each test activity draws on different knowledge sources and abilities that vary across students, and it reflects the different language learning histories experienced by our learners. In the delayed posttest stage, whereas the written responses tap into the ability to produce the verb form required by the voice of the sentence, the languaging in the stimulated recall taps into the depth of understanding. (Swain et al., 2009, p. 22)

On top of that, Roehr (2006, 2007, 2008) in several studies emphasized the differences between linguistic and metalinguistic types of knowledge. He suggested that while linguistic knowledge is assumed to be "represented in terms of flexible and context-dependent categories which are subject to similarity-based processing", "explicit metalinguistic knowledge is characterized by stable and discrete Aristotelian categories which subserve conscious, rule-based processing" (Swain et al., 2009, p. 67). Likewise, the results of the current study show that tapping into students metalinguistic knowledge through test techniques would ideally suit a foreign language learning situation and more importantly, in supporting teaching grammaticality in writing tasks.As another possible explanation, Roehr (2006, 2008) found a significant positive correlation between learners metalinguistic knowledge and their proficiency; furthermore, it has been reported that learners with higher levels of metalinguistic awareness tend to show higher levels of learning gain over those with less (Mitchell, Myles, & Marsden, 2013).

## CONCLUSION

To put it in a nutshell, the present study addressed two research questions. Regarding the first one, we found a positive washback effect of tests on learning of grammatical points and producing those forms in writing tasks. We mentioned a plethora of agreeing on studies but a few opposing ones. Some explanations for the contrary findings may be gender, learners' mother tongue or other factors of test usefulness.

With respect to the second research question, the results of post hoc test indicated that experimental groups which were assisted by multiple-choice and metalinguistic tests significantly outperformed the control group in doing grammar-focused writing tasks while those who received multiple-choice tests did not show an improvement over the control group. This finding represents an update on former research. The results suggest that both metalinguistic and cloze tests are suitable activities to support the production of grammatically correct written forms, but there should be revisions about the effects of multiple-choice tests as some contrary evidence were probed. It is noted that cloze tests can induce higher loads of "cognitive processing" than multiple-choice tests so it could have the edge. The superiority of metalinguistic tests, which are neglected or even prohibited in most parts of language learning, could be explained by the difference in types of knowledge and its greater correlation with written modes of production. In addition, higher levels of metalinguistic knowledge cause higher levels of learning.

A number of implications are conceivable for the results of the current study. First and for most, all language teachers, students and material developers may need to reflect more on their perspectives on language testing and consider its possible negative and positive washback effects. Another point is if we assume that the main goal of every language assessment activity is to foster learning, and if we believe that assisting language learning through judicious type of test can lead to linguistic and metalinguistic development, then it is reasonable to call a coherent and extensive effort by all teachers, material developers, and stakeholders to develop nationally and internationally validated tests. Consequently, it is suggested to keep an eye on metalinguistic and cloze tests while teaching, studying or preparing course material for writing and grammar.

Needless to say, these proposals would benefit from further investigation. In particular, more controlled studies regarding test usefulness factors (i.e. reliability, construct validity, authenticity, interactiveness, impact, and practicality). Moreover, large scale longitudinal and qualitative studies are needed to fully document the underlying mental processes of these phenomena. Another point is the effects of cultural competence, schemata and background knowledge which should be investigated in relation to washback effects of different test formats.

## LIMITATIONS

One of the greatest limitations of this study was the limited number of treatment sessions that is two

sessions of testing cannot fully represent the washback of effects of testing. It is hoped that future studies address this limitation.

**REFERENCES**


Alderson, J. C., & Hamp-Lyons, L. (1996).TOEFL preparation courses: A study of washback. *Language Testing, 13*(3), 280-297. doi: 10.1177/026553229601300304

Alderson, J. C., & Wall, D. (1993). Does washback exist?. *Applied Linguistics, 14*(2), 115- 129. doi: 10.1093/applin/14.2.115

Allan, D. (2004). *Oxford placement test*. Oxford: Oxford University Press.

Andrews, S., Fullilove, J., & Wong, Y. (2002).Targeting washback: A case study. *System, 30*(2), 207–223. doi: 10.1016/s0346-251x(02)00005-2

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

Biggs, J. B. (2003). *Teaching for quality learning at university* (2nd ed.). Buckingham: Open University Press/Society for Research into Higher Education.

Brame, C. J., & Biel, R. (2015).Test-enhanced learning: the potential for testing to promote greater learning in undergraduate science course.*CBE-Life Sciences Education, 14*(2), 1-12. doi: 10.1187/cbe.14-11-0208

Brown, H. D. (2004). *Language assessment: Principles and classroom practices*.White Plains, NY: Pearson Education.

Chapman, D., & Snyder, C. (2000). Can high stakes national testing improve instruction: reexamining conventional wisdom. *International Journal of Educational Development, 20*(6), 457– 474. doi: 10.1016/s0738-0593(00)00020-1

Chehrazad, H., & Ajideh, P. (2012). Effects of different response types on Iranian EFLtest takers' performance. *Iranian Journal of Applied Language Studies, 5*(2), 29-50.

Currie, M., & Thanyapa C. (2010). The Effect of the multiple-choice item format on the measurement of knowledge of language structure. *Language Testing, 27*(4), 471-91. doi: 10.1177/0265532209356790

Ellis, R. (2003). *Task-based language learning and teaching*.Oxford: OUP.

Elton, L.R.B. & Laurillard, D.M. (1979) Trends in research on student learning. *Studies in Higher Education, 4*(1), 87-102. doi: 10.1080/03075077912331377131

Grabe, W., & Kaplan, R.B. (1996). *Theory and practice of writing*. Harlow: Longman.

Green, A. (2006). Washback to the learner: learner and teacher perspectives on IELTS preparation course expectations and outcomes. *Assessing Writing, 11*(2), 113-134. doi: 10.1016/j.asw.2006.07.002

Green, A. (2013). Washback in language assessment. *International Journal of English Studies, 13*(2), 39-51. doi: 10.6018/ijes.13.2.185891

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 1-27). Mahwah, NJ: Lawrence Erlbaum.

Hayes, J. R., & Flower, L. S. (1980).Identifying the organization of writing processes. In L. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3-30). Hillsdale, NJ: Lawrence Erlbaum.

Hemmati, F., &Ghaderi, E. (2014).The effect of four Formats of multiple-choice questions on the listening comprehension of EFL learners.*Procedia - Social and Behavioral Sciences, 98*, 637-644. doi: 10.1016/j.sbspro.2014.03.462

Henning, G. (1987). *A guide to language testing: development, evaluation, research*. Cambridge, MA: Newbury House.

Hughes, A. (2003). *Testing for language teachers* (2nd ed). Cambridge: Cambridge University Press.

Heaton, J. B. (1988). *Writing English language tests*. New York: Longman.

Ishikawa, T. (2006). The effects of task complexity and language proficiency on task-based language performance. *Journal of Asia TEFL, 3*(4), 193-225.

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. doi: 10.1177/0265532208101006

Khoshsima, H., & Pourjam, F. ( 2014). Comparative study on the effects of cloze tests and open-ended questions on reading comprehension of Iranian intermediate EFL learners. *International Journal on Studies in English Language and Literature, 2*(7), 17-27.

Ko, M. H. (2010). A comparison of reading comprehension tests: Multiple-choice vs open-ended. *English Teaching, 65*(1), 137-159. doi: 10.15858/engtea.65.1.201003.137

Kromann, C. B., Jensen, M. L., & Ringsted, C. (2009) The effect of testing on skills learning. *Medical Education, 43*(1), 21-7. doi: 10.1111/j.1365-2923.2008.03245.x

Kuiken, F., & Vedder, I. (2007). Task complexity and measures of linguistic performance in L2 writing. *International Review of Applied Linguistics, 45*(3), 261-284. doi: 10.1515/iral.2007.012

Loch, A. (2010). How do test methods affect reading comprehension test performance?. In Kovács, P., Szép, K. Katona, T. (Szerk.). *Proceedings of the Challenges for Analysis of the Economy, the Businesses, and Social Progress International Scientific Conference*. (pp. 924-935).

McNamara, T. (2001). Language assessment as social practice: Challenges for research. *Language Testing, 18*(4), 333-349. doi: 10.1177/026553220101800402

Madaus, G.F. (1988). The influence of testing on the curriculum. In Tanner, L.N. (Ed.), *Critical issues*

*in curriculum: eighty-seventh yearbook of the national society for the study of education*.(p.83-121). University of Chicago Press, Chicago.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*(3), 241-256. doi: 10.1177/026553229601300302

Mitchell, R., Myles, F., & Marsden, E. J. (2013). *Second language learning theories* (3rd ed.). Abingdon: Routledge.

Mizumoto, A., Ikeda, M., & Takeuchi, O. (2016). A comparison of cognitive processing during cloze and multiple-choice reading tests using brain activation. *Annual Review of English Language Education in Japan, 27*, 65-80.

Morrow, K. (1986). The evaluation of tests of communicative performance. In M. Portal (Ed.), *Innovations in Language Testing* (p.1-3). London: NFER/Nelson.

Mozaffari, F., Alavi, S., & Rezaee, A. (2017). Investigating the impact of response format on the performance of Grammar tests: Selected and constructed. *Journal of Teaching Language Skills, 32*(2), 103-128.

Palomba, C. A., & Banta, T. W. (1999). *Assessment essentials: Planning, implementing, and improving assessment in higher education*. San Francisco, CA: Jossey-Bass.

Pan, Y. (2009). A review of washback and its pedagogical implications. *VNU Journal of Science, Foreign Languages, 25*, 257-263.

Pearson, I. (1988). Tests as levers for change.In D. Chamberlain & R. J. Baumgardner (Eds.), *ESP in the classroom: Practice and evaluation* (p. 98-107). London: Modern English.

Phillips, D. (2004). *Longman introductory course for the TOEFL test: The paper test*. New York: Longman.

Qi, L. (2005). Stakeholders' conflicting aims undermine the washback function of a high- stakes test. *Language Testing, 22*(2), 142-173. doi: 10.1191/0265532205lt300oa

Rauch, D., & Hartig, J. (2010). Multiple-choice versus open-ended response formats of reading test items: A two-dimensional IRT analysis. *Psychological Test and Assessment Modeling, 52*(4), 354–379.

Rea-Dickins, P., & C. Scott (2007). Washback from language tests on teaching, learning and policy: Evidence from diverse settings. *Assessment in Education, Special Issue, 14*(1), 1-7. doi: 10.1080/09695940701272682

Read, J. & Hayes, B. (2004). IELTS Test Preparation in New Zealand: Preparing students for the IELTS academic module.In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (p. 97-112).New Jersey: Lawrence Erlbaum Associates.

Roehr, K. (2006). Metalinguistic knowledge in L2 task performance: A verbal protocol analysis. *Language Awareness, 15*(3), 180–198. doi: 10.2167/la403.0

Roehr, K. (2007). Metalinguistic knowledge and language ability in university-level L2 learners. *Applied Linguistics, 29*(2), 173-199. doi: 10.1093/applin/amm037

Roehr, K. (2008). Linguistic and metalinguistic categories in second language learning. *Cognitive linguistics, 19*(1), 67-106. doi: 10.1515/cog.2008.005

Safa, M. A., & Goodarzi, S. (2014). The washback effects of task-based assessment on the Iranian EFL learners' grammar development. *Procedia - Social and Behavioral Sciences, 98*, 90-99. doi: 10.1016/j.sbspro.2014.03.393

Shih, C. (2009). How tests change teaching: A model for reference. *English Teaching: Practice and Critique, 8*(2), 188-206

Spratt, M. (2005). Washback and the classroom: The implications for teaching and learning of studies of washback from exams. *Language Teaching Research, 9*(1), 5-29. doi: 10.1191/1362168805lr152oa

Swain, M., Lapkin, S., Knouzi, I., Suzuki, W., & Brooks, L. (2009). Languaging: University Students Learn the Grammatical Concept of Voice in French. *Modern Language Journal, 93*(1), 5-29. doi: 10.1111/j.1540-4781.2009.00825.x

Sze, P., & Leung, F.F.Y. (2014). Enhancing learners' metalinguistic awareness of language form: The use of eTutor resources. *Assessment and Learning, 3*, 79-96.

Tuckman, B. W. (1978). *Conducting educational research*.New York: Harcourt Brace Jovanovich.

Talebzadeh, Z., & Bagheri, M. S. (2012). Effects of sentence making, composition writing and cloze test assignments on vocabulary learning of pre-intermediate EFL students. *International Journal of English Linguistics, 2*(1), 258-261.

Van den Branden, K. ( 2006): *Task-based language teaching in practice*. Cambridge: Cambridge University Press.

Wall, D. (2000). The impact of high-stakes testing on teaching and learning: Can this be predicted or controlled?. *System, 28*(4), 499-509. doi: 10.1016/s0346-251x(00)00035-x

Watanabe, Y. (2004). Methodology in washback studies.In Cheng L, Y. Watanabe & A. Curtis (Eds.).*Washback in language testing: Research contexts and methods*. (p.19-37). New Jersey: Lawrence Erlbaum Associates.

Watanabe, Y., & Koyama, D. (2008). A meta-analysis of second language cloze testing research. *Second Language Studies, 26*(2), 103–133.

Wang, G. H., & Wang, S. (2013). Roles of metalinguistic awareness and academic extensive reading in the development of EFL/ESL academic writing skills. *Journal of Art and Humanities, 2*(9), 47-55.

Zarei, A., & Neya, S. S. (2014). The effect of vocabulary, syntax, and discourse-oriented activities on short and long-term L2 reading comprehension. *International Journal of Language & Linguistics, 1*(1), 29-39.