International Journal of Research in English Education (IJREE)

Original Article **Published online: 20 September 2021.**

The Washback Impacts of Task-based Assessment on Iranian EFL Learners' Vocabulary Learning and Grammatical Ability

Maryam Abdollahi[1] & Siros Izadpanah[1]*

* Correspondence:
cyrosIzadpanah@yahoo.com
1. Department of English Language Teaching, Zanjan Branch, Islamic Azad University, Zanjan, Iran

Received: 21 March 2021
Revision: 13 May 2021
Accepted: 13 July 2021
Published online: 20 September 2021

Abstract

This study was an investigation of the washback impacts of task-based assessment on Iranian EFL learners' vocabulary learning and grammatical ability at the intermediate level. For this purpose, 184 male and female students who were in their third year of school were chosen from 12 schools of (X) (X). This election was done via a multistage cluster random sampling technique using the Cambridge Placement Test (2010). All of the sessions were divided into two parts, 20 minutes were considered for teaching grammar and 20 minutes were for teaching vocabulary. The components of this study were Oxford Practice Grammar Intermediate Diagnostic Test and a standard vocabulary test extracted from TOEFL exams between 2017-2020. The researchers prepared a test for every two groups at the end of every three sessions. It was found that by removing the pre-test factor, the covariance, the task-based evaluation's washback impact makes the grammatical and vocabulary learning of students better. Considering the reality that every academic endeavor contains planned testing and evaluation techniques to maximize academic achievement and progress, the findings hint that TBLA as an academic measurement device can nicely replace the traditional evaluation techniques.

Keywords: grammatical ability, task-based assessment, vocabulary learning, washback impact

1. Introduction

Assessment is one of the most considerable prevalent discussions in the era of education and learning that plays an essential role in learning the second or foreign language (Lam, 2018; Liu & Brantmeier, 2018; Knight, 2020). As Bachman (1995) and Levi and Inbar-Lourie (2020) stated, assessment and testing are important parts of every teaching and learning process. Two fundamental roles of assessment in education are that firstly, it is used for evaluation and proper decisions, and secondly it would help teachers as a teaching device (Kermad & Kang, 2018). The assessment incorporates teacher observation, classroom discussion, and it helps to analyze student's work, including homework and tests. Therefore, in the fields of second or foreign language education and applied linguistics, it is widely believed that testing and assessments influence teaching and learning. This influence is referred to as 'washback' (McKinley & Thompson, 2018; Pitoyo, 2020), or 'test impact' (Bachman & Palmer, 1996). Washback's degree differs by the time corresponding to the test's situation, the language which is being examined, the aim of the test, the format of the test, and examined capabilities (Gebril, 2018; Kafipour, Jafari, & Khojasteh, 2018; Salehi & Yunus, 2012; Sumera, Barua, & Navamoney, 2015).

One of the enormous concerns for researchers in the field of language testing is a washback effect that there are some explorations about this (Damankesh & Babaii, 2015). Thus the L2 literature contains several investigations of task-based instruction and learning a hasty exploration of testing, publications clarify that task-based assessment work is few (Kafipour, Jafari, & Khojasteh, 2018). Like the other elements of this study, vocabulary is one of five important elements in learning a foreign or second language. Krashen (1989) defined vocabulary as words in a specific language or freestanding items of language that have meaning. Vocabulary knowledge is very important because it consists of all the important words we have to know to access our background information, show our thoughts and speaking well, and learn about new views. It is commonly believed that an important part of learning a foreign language is learning its terminology (Cate, 2018). One can't read, compose, communicate in, or handle a language without knowing its wording (Çakmak & Erçetin, 2018; Goldstein, Ziolkowski, Bojczyk, Marty, Schneider, Harpring, & Haring, 2017). It shows up that from approximately 1945s to the late 1970s and early 1980s, essentially all techniques and usages of language instruction gave lexicon learning little or no importance.

Another element of this study is grammar and one of the important goals of learning a second or foreign language is helping learners to face the complex rules of grammar and it is because grammar plays a very important role in all four skills and also it is the teachers' duty to clarify any hard points in learning for learners (Alsied, Ibrahim, & Pathan, 2018; Spierings & Cate, 2016). The rules of language structure are around how words alter and how they are put together into sentences (Jean & Simard, 2011). The language structure of dialect is what happens to words when they got to be plural or negative, or what word arrange is utilized when we make address or connect two clauses to create on sentence (Cate, 2018; Jean & Simard, 2011; Milne, Petkov, & Wilson, 2018). Grammar instruction has been relatively unaltered by research findings.

1.1 Statement of the Problem

One branch of foreign language education, which seems to receive not much attention, is the innovation of differentiated instruction and creative assessment in the foreign language classroom and as mentioned earlier, the problem is to demonstrate the degree of efficiency of washback effect of two different types of quizzes, the task based and traditional quizzes, on vocabulary learning and grammatical ability of the English as a Foreign Language (EFL) learners. More recently, the best way of teaching and testing in general and in foreign language teaching and learning in particular, have obsessed the mind of many scholars and teachers. To date there has been little agreement on these issues and finding a perfect methodology suitable for both teaching and testing of foreign languages has been an unaccomplished wish for which there is a long way to be paved. Consequently, in the recent decades, the attitudes towards assessing students have been changed dramatically. The use of assessment as a means of promoting curriculum change has become increasingly common not only in general education (Sattar, Abdullah, & Mirzaei, 2018), but also in language education and its different aspects including the assessment paradigms. It is also believed that the quickest way to change student learning is to change the assessment system (Andrews, Fullilove, & Wong, 2002). The studies of such educational and pedagogical consequences of assessment procedures are mainly carried out through washback studies.

In addition, considering the important role of teaching and learning process in the field of Teaching English as a Foreign Language (TEFL), there are controversial issues regarding the repeated use of tests or quizzes in EFL classes. Some educators believe that frequent tests will increase instructional effectiveness and encourage students to study

and review more (Kafipour, Jafari, & Khojasteh, 2018), while some others believe that frequent use of quizzes may not be useful for learners and may led to their frustration and maybe the teachers will rely on testing more than it is needed to motivate students and plan for their own teaching (Safa & Jafari, 2017). Washback studies have been primarily concerned with the teachers' perspectives and have barely addresses this effect from students' points of views (Alderson & Hamp-Lyons, 1996). However, for a better understanding of how washback occurs as a result of different assessment procedures within the classroom, researchers need to investigate changes in students' motivation, learning styles, learning strategies, and educational outcomes and achievements.

So far, however, it was contended that many washback studies do not investigate learning outcomes, so it is necessary to investigate whether washback of exams affects learning, and if so, how (Taylor, 2005). The same point is raised by McNamara (2001) who supports the possibility of the type of assessment to be an important factor for the follow-up or preceding learning. In this attempt we'll consider the differential washback effects of task based language assessment procedure and traditional assessment modes on vocabulary learning and development of Iranian EFL learners. In addition to what was mentioned, according to Morris (2016), there are a number of problems associated with the traditional testing and the student may also be facing extenuating circumstances (e.g., personal problems and illness) at the time she is being tested, thus also hampering her performance on the test. The problems associated with traditional testing often mask what the student really knows, or in the case of English as a second language (ESL), what the student can do in her second language.

Considering these problems and difficulties and since there are a little empirical studies about the effects of repeated quizzes on students' structural ability, this study was set to use task-based assessment and to compare the washback effects of frequent task based and traditional quizzes for enhancing and developing vocabulary learning and grammatical ability of learners. There are some studies examining the use of task-based assessment and its effects on four essential skills including listening, speaking, and reading. However, there are few research studies on the use of task-based assessments and its washback effects in teaching a specific skill, such as vocabulary and grammar which this current research aimed at. The purpose of this study was to investigate the washback impacts of task-based assessment on Iranian EFL learners' vocabulary learning and grammatical ability.

1.2 Research Questions

The purpose of this study was to demonstrate the washback impacts of task-based assessment on Iranian intermediate EFL learners' vocabulary learning and grammatical ability. The goal of this investigation was to answer the following research questions:

1. Does the washback impact of task-based assessment significantly improve Iranian intermediate EFL learners' grammatical ability?

2. Does the washback impact of task-based assessment significantly improve Iranian intermediate EFL learners' vocabulary learning?

1.3 Null Hypotheses

The hypotheses of this research are as follows:

1. The washback impact of task-based assessment doesn't significantly improve Iranian intermediate EFL learners' grammatical ability.

2. The washback impact of task-based assessment doesn't significantly improve Iranian intermediate EFL learners' vocabulary learning.

2. Review of the Literature

2.1 Testing vs. Assessment

The importance of testing and assessment in language teaching is well known to all. We often use tests to make decisions about individuals' abilities and our decisions might influence their academic as well as personal lives. Information about people's language ability is often very useful and sometimes necessary within teaching systems and as long as it is thought appropriate for individuals to be given a statement of what they have achieved in a second or foreign language, the tests of some kind or other will be needed in order to provide information about the achievement of learners (Tollefsen et al., 2014). This has made testing and assessment an important component of

teaching and instruction. However, care should be taken about using the two terms testing and assessment. Some applied linguists use the term "testing" to apply to the construction and administration of formal or standardized tests such as Test of English as a Foreign Language (TOEFL) and "assessment" to refer to more informal methods such as group and peer assessment. As cited in Clapham (2000), For example, Valette (1977) states that "tests" are large-scale proficiency tests and that "assessments" are school-based tests. This, however, is a rough illustration of the dichotomy between testing and assessment. Bachman (1995) gives a comprehensive definition of testing: "A test is a measurement instrument designed to elicit a specific sample of an individual's performance. As one type of measurement, a test necessarily quantifies characteristics of individuals according to explicit procedures and rules" (p.20).

Moreover, for Farhady, Jafarpur, and Birjandi (1995), testing often connotes the presentation of a set of questions to be answered. Assessment, nevertheless, requires a different definition. According to Shohamy (1992), assessment is a super-ordinate term which includes all forms of assessment. It not only assigns scores to students, but also diagnoses their problems and remedies them through employing specific methods and techniques. Gipps (1994) also defines assessment as "a wide range of methods for evaluating pupils' performance and attainment" including formal testing and examinations, practical and oral assessment and classroom-based assessment carried out by teachers (p.10). Regarding the importance of assessment in contrast to testing, Inger (1993) argues that testing is designed to be administered during a normal school period and it presents a series of discrete tasks that force students to move repeatedly from one unconnected item into the next. Inger concludes that this shortcoming of language testing can be overcome by assessment techniques and procedures.

A test is one form of assessment and refers to procedures used to measure a learners' learning at a specific point in time and often involves collecting information in numerical form. Common forms of tests are multiple choice questions and gap-fill or cloze tests. In English classes, teachers also need to assess their students' learning to determine the effectiveness of their teaching and of the materials they use. Assessment refers to any of the procedures teachers use to do this, which may include interviews, observations, administering questionnaires and reviewing students' work. Assessment covers a broader range of procedures than testing and includes both formal and informal measures (Morrow, 2018).

Assessment is the systematic process of documenting and using empirical data on the knowledge, skills, attitudes, and beliefs. By taking the assessment, teachers try to improve student learning. Almost everybody has experienced testing during his or her life. A test is used to examine someone's knowledge of something to determine what that person knows or has learned. It measures the level of skill or knowledge that has been reached. An evaluative device or procedure in which a sample of an examinee's behavior in a specified domain is obtained and subsequently evaluated and scored using a standardized process . Test and assessment are used interchangeably, but they do mean something different. A test is a "product" that measures a particular behavior or set of objectives. Meanwhile assessment is seen as a procedure instead of a product. Assessment is used during and after the instruction has taken place. After you've received the results of your assessment, you can interpret the results and in case needed alter the instruction. Tests are done after the instruction has taken place, it's a way to complete the instruction and get the results. The results of the tests don't have to be interpreted, unlike assessment (Duque & Cuesta, 2017).

2.2 Task-Based Language Teaching (TBLT)

TBLT is an educational framework and an approach for the theory and practice of second/foreign language (L2) learning and teaching, and a teaching methodology in which classroom tasks constitute the main focus of instruction. A classroom task is defined as an activity that is (1) goal-oriented, (2) content-focused, (3) has a real outcome, and (4) reflects real-life language use and language need. The syllabus in TBLT is organized around activities and tasks rather than in terms of grammar or vocabulary (Kessler, Solheim, & Zhao, 2021; Sert & Amri, 2021).The interest in TBLT is based on the strong belief that it facilitates second language acquisition (SLA) and makes L2 learning and teaching more principled and more effective.

2.3 Task

In commencing to talk about TBI, it seems crucial to specify what is meant by task. Various definitions are presented for task. Lyle, Rowland, Ostrovski, and Dabney (2021) suggest that a communication task must stimulate real communicative exchange, provide incentive for the L2 speaker/learner to convey information, provide control for the information items required for investigation, and fulfill the needs to be used for the goals of the experiment.

2.4 Utilization and Classroom Application of Task-Based Language Assessment

TBLA has been utilized by language teachers in the L2 classroom in a broad range of formal and informal educational settings that serve a wide range of language learners who come from different age groups, different proficiency levels, and different educational and cultural backgrounds (Eicher & Joyner, 2021).

2.5 The Washback Effect

Dong, Fan, and Xu (2021) define washback as the direct and indirect effect of examination on teaching methods. They maintain that according to the effect of examination on what to do in the classroom, we may refer to 'positive' and 'negative' washback. Ma (2021) maintains that the term washback is itself a neutral one, can be related to influence. If the test is poor, then the washback may be felt to be negative. But if the washback hypothesis holds, the good tests should have good effects (as yet undefined) rather than negative effective effects.

2.6. Iranian and International Studies on Washback Impacts of Task-based Assessment and Vocabulary Learning
Nazari et al. (2021) attempted to explore the 'washback effect' of task-based assessment (TBLA) on EFL Iranian learners' pragmatic development. To this end, through conducting KET (Key English Test), 60 out of 120 EFL Iranian learners studying in an English language school, were randomly selected. The treatment group was assessed through TBLA and the control group was assessed via non-task based assessment for 20 ninety-minute sessions. The class sessions were held twice a week. The obtained data were examined through independent sample *t*-test. The findings implied that TBLA as a pedagogical measurement tool can well replace the classic assessment procedures, since all educational efforts including testing and assessment procedures are planned to maximize the educational gains and developments.

Bokiev and Samad (2021) investigated the influence of language assessment on teaching and learning. In contrast to the wealth of studies involving external large-scale language examinations, scant research has been conducted to explore the influence of internal language assessment on instruction, particularly in the context of a university foundation programme. This qualitative study investigated the washback effects of an English language assessment system (ELAS) on the teaching and learning of English in a Malaysian university foundation programme. Apart from an in-depth analysis of official documents on the ELAS, they conducted individual semi-structured interviews with three curriculum and assessment developers, three English language instructors, four students and four alumni of the foundation programme and analysed the collected data using Miles and Huberman's (1994) framework for qualitative data analysis. Findings indicated that the ELAS, with its different assessment forms, exerted an overall positive washback on various aspects of English teaching and learning. Yet, a number of factors related to the assessment, teachers, students as well as context mediated the extent of washback experienced. Based on the findings of the study, we put forward a few recommendations on how to encourage positive washback.

Salma and Prastikawati (2021) studied the authentic assessments to be applied in the English classroom as the completion of the newest curriculum demands. Thus, teachers are suggested to implement this assessment in the English teaching and learning process. Concerning that fact, this current study aims at investigating the washback of the use of performance-based assessment in the English class and the barriers faced by the teachers in implementing performance-based assessment in the English learning process. This study applied a qualitative research design. Using purposive sampling, there were two teachers and 72 students who participated in this study. An open-ended questionnaire and semi-structured interview were carried out in collecting the data. The result showed that performance-based assessment affected positively to the teachers. They could identify the students' real competence, improve their teaching quality, and assess students objectively. Further, it gives positive effects for the EFL learners in matters such as learning enthusiasm and self-confidence, boredom reduction, and skill improvement. In conclusion, the use of performance-based assessment is good at improving some teaching aspects both on teachers and students. This finding proves the importance of performance-based assessment to boost the teaching and learning quality.

Tayeb, Aziz, and Ismail (2018) in their research about predominant washback of the general secondary English examination on teachers, investigated the dominant wash back effect of the GSEE on Yemeni teachers who are highly engrossed by the test. The study focuses on four pedagogical dimensions namely teachers' teaching methods, content assessment, attitudes and motivation. Mixed mode approach (qualitative and quantitative methods) was applied using classroom observations, semi-structured interviews, group discussions and a questionnaire of 72 items administered to 46 English teachers. NVivo10 was used to analyse the qualitative data. SPSS/V22 was used for analysing the quantitative data in which the Cronbach's Alpha reached (.88). The results revealed that the test had a great influence

on teachers (P<.001) mainly on their teaching methods. Triangulation with the qualitative analysis confirmed equivalent implications. The study contributes a clear evidence of the powerful exam washback on the factors of the language learning practices and its influence on how and what teachers teach.

The literature reviewed here provided evidence-based interventions designed to facilitate vocabulary and grammar knowledge by the help of task based assessments. The body of research on improving vocabulary and grammar skills and strategies found to successfully assist students in accessing the content curriculum. Instruction that is highly structured, explicit, provides scaffolding, and is intense was recommended within a balanced literacy program. The literature identified a balanced literacy program as one containing word study, grammar and vocabulary development. Using varying strategies within this framework have been found to be efficacious for struggling students.

Vocabulary learning and grammar ability were two important components of balanced literacy instruction. It plays a critical role in the development of students' knowledge and a relationship between the washback effects of task based assessments has been well documented in this study. The use of task based and traditional assessments within the literacy classroom has proven to improve word and grammar knowledge within the content areas. While, based on a number of interventions were presented in this review, it was found that task based assessments were the most beneficial for all types of learners. In the research described in this study, the efficacy of the washback effects of task based assessments was explored in social studies classrooms to determine if equally positive results for this intervention could be obtained with heterogeneous groups of learners. Effect sizes showed significant differences in favor of the treatment groups. In addition, positive results for second language learners within the social studies indicate that task based assessments can positively benefit all learners. Therefore, this investigation replicated and extended the research using task based assessments within the heterogeneous, inclusive setting in order to enhance the vocabulary learning and grammar ability of the students**.**

3. Methodology

3.1 Design of Study

The design of this study was quasi-experimental. A quasi-experiment is an empirical interventional study used to estimate the causal impact of an intervention on target population without random_assignment. Quasi-experimental research shares similarities with the traditional experimental design or randomized_controlled_trial, but it specifically lacks the element of random assignment to treatment or control. Instead, quasi-experimental designs typically allow the researcher to control the assignment to the treatment condition, but using some criterion other than random assignment. In this quantitative study, an experimental design was used to determine the washback impacts of task-based assessment on the Iranian intermediate EFL learners' vocabulary learning and grammatical ability.

3.2 Participants

The participants of this research were 184 first-grade senior high school learners in X. X province has 8 cities, among these cities, X city was chosen. This city includes two districts that the district two were randomly chosen. In 2020 in district 2 there were 433 schools that among them senior high schools were randomly chosen. In order to guarantee the homogeneity of the individuals of this examination and to satisfy the goals of the examination; first, a Cambridge placement test (2010) by Cambridge University Press shared among all the learners to determine the stage of their skill. Based on the rules of CPT (2010), out of 359 learners, 184 learners were proved to be intermediate that out of them 92 learners were female and 92 learners were male. After that, the student's consent form was distributed among learners in order to make them familiar with the processes of the current study. Finally, in order to motivate the learners for participating in this study, an English storybook with a pen was provided for them as a gift. There were 92 female learners in 2 different groups, one experimental task-based group (A1) (n=46) and one experimental traditional group (A2) (n=46) in the first level. Also, there were 92 male learners in 2 different groups, one experimental task-based group (B1) (n=46) and one experimental traditional group (B2) (n=46) in the first level. The type of sampling in this study was multistage cluster random sampling.

3.3 Instruments

The following instruments were employed in order to collect the required data for the present study.

a) Cambridge Placement Test (2010) by Cambridge University Press

To have a homogeneous group of participants, to neutralize any effect of proficiency level on participants' performance and to fulfill the objectives of the study, first, a Cambridge placement test (2010) by Cambridge University Press was distributed among all the student participants of this study in order to determine their level of proficiency. The aim was to select those students with the intermediate level of proficiency. Out of 359 students, 184 proved to be intermediate.

b) Oxford Practice Grammar Intermediate Diagnostic Test

The oxford practice grammar intermediate diagnostic test extracted from TOEFL exams between 2017-2020 were administered. The learners seemed to be frustrated by the high level of the test. However, the researchers tried to reassure them by explaining that the sessions they will take part in will hopefully lead them to find it less difficult. The teaching material covered during this intervention was Oxford Practice Grammar (Intermediate level) by John Eastwood.

c) A Standard Vocabulary Test Extracted from TOEFL Exams between 2017-2020

All experimental and control groups took the vocabulary pre-test as a measure of the participants' vocabulary knowledge of the selected English vocabulary points by using a standard pre-test which was administered to the students at the beginning of the training course.

d) Three Teacher Made Task-based Vocabulary Quizzes

In commencing to talk about TBI, it seems crucial to specify what is meant by task. Various definitions are presented for task. Beglar and Hunt (2002) suggest that a communication task must stimulate real communicative exchange, provide incentive for the L2 speaker/learner to convey information, and provide control for the information items required for investigation.

e) Three Teacher Made Traditional Vocabulary Quizzes

When the term traditional assessment is used, what is usually meant is summative assessment. Summative assessments seek to determine what students know at the end of a chapter, unit, or series of lectures on a topic. Tests are just one of the tools used in traditional assessment, although they often receive the most attention and are the most pervasive. This is due to factors such as the ease with which they can be administered and scored and the fact that they enable teachers to judge students' progress quickly and easily (Mede & Atay, 2017).

f) Pre and Post Tests

Students were given pre-tests on the first day of the investigation, and post-tests on the last day of the study. Oxford practice grammar intermediate diagnostic test (OPGIDT) and a standard vocabulary test (VT) extracted from TOEFL exams, as the pre and post-tests, were provided for participants to determine prior knowledge, to inform and guide instruction for use in the subsequent data analysis, and to compare results of the experimental groups.

3.4 Data Collection

The main data collection stage took place for 2 months, based on the works of previous researchers in the field of washback effects and mostly due to the rules of the education department of X, because they don't issue the required licenses for the researchers to have more time in order to perform their research in a longer period, in different schools, and they believe these kinds of research are a time-consuming process for learners and make them stand back from their original curriculum. The study, as mentioned previously, involved two main groups (N=184), group A consisted of (A1, A2) and group B consisted of (B1, B2); one class of every twelve senior high schools for Male and Female learners were chosen and out of 184 learners 92 learners were female and 92 learners were male. There were 92 female learners in 2 experimental groups, one experimental task-based group (A1) (n=46) and one experimental traditional group (A2) (n=46) in the first level. Also, there were 92 male learners in 2 experimental groups, one experimental task-based group (B1) (n=46) and one experimental traditional group (B2) (n=46) in the first level. The intervention had two main objectives; one was to raise the participants' grammar and vocabulary awareness, and then measure the washback impacts of task-based assessment on the X intermediate EFL learners' vocabulary learning and grammatical ability. The chief principle behind this instruction was to encourage learners to take a more active role in developing their vocabulary learning and grammatical ability.

At first, the informed consent letter was distributed among learners to read and sign. They were also asked to write their emails to arrange for future sessions. During that session, the oxford practice grammar intermediate diagnostic test and a standard vocabulary test extracted from TOEFL exams between 2017-2020 were administered for the first time. The learners seemed to be frustrated by the high level of the test. However, the researchers tried to reassure them by explaining that the sessions they will take part in will hopefully lead them to find it less difficult. The teaching material covered during this intervention was Oxford Practice Grammar (Intermediate level) by John Eastwood. These books are specially designed for teaching and training the grammar points at the intermediate level. Also, fifty words were selected from TOEFL exams between 2017-2020, and the Academic Word Lists textbook and the meaning of these words were taught traditionally to the learners during these 10 sessions.

3.5 Data Analysis Sessions

SPSS Statistics 24 program was used to evaluate the data obtained from these participants. To analyze the results, the following statistical tests were used:

3.5.1 The Analysis of Covariance (ANCOVA)

The covariance was used to compare the procedures of one or more groups and to approximate one or more independent variables and derive the effect from the equation of one or more interfering variables, covariance, or covariate.

3.5.2 Kolmogorov Smirnov

This test (K–S test or KS test) was a nonparametric test of the equality of continuous, one-dimensional probability distributions that could be used to compare a sample with a reference probability distribution (one-sample KS test), or to compare two samples (two-sample K–S test) to decide between parametric or non-parametric tests.

3.5.3 Levene's Test

This test was used to check the homogeneity of the variances.

3.5.4 Independent T-Test

The independent-samples t-test (or independent t-test, for short) compares the means between two unrelated groups on the same continuous, the dependent variable, in order to determine whether there is statistical evidence that the associated population means are significantly different.

4. Results

The distribution of variables depends on the most significant central indexes of mean, dispersion, and the standard deviation. Descriptive data on grammatical abilities and vocabulary learning scores in experimental classes, pre-test, and post-test, are seen in Table 1(a) and Table 1 (b).

Table 1(a). Descriptive statistics of the POST-OPGIDT and POST-VT test scores in experimental (A1, B1) and (A2, B2) groups

	OPGIDT TEST		VT TEST	
Groups	Post-test (A1, B1)	Post-test (A2, B2)	Post-test (A1, B1)	Post-test (A2, B2)
Mean	66.7391	66.1196	63.9348	62.1957
Median	67.0000	66.0000	65.0000	62.0000
Mode	66.00a	66.00	66.00a	61.00
Std. Deviation	7.45370	7.17573	6.62298	7.11173
Skewness	-.149	-.042	-.374	-.089
Kurtosis	-.939	-.898	-.622	-.901

Considering Table 1(a), the mean scores of the (A1, B1) experimental groups in the post- OPGIDT test is 66/739, the mean scores of the (A2, B2) experimental groups in the post- OPGIDT test is 66/119, and the mean scores of the (A1, B1) experimental groups in the post-VT test is 66/934 and the mean scores of the (A2, B2) experimental groups in the post-VT test is 66/195. The results demonstrated that the mean scores of the (A1, B1) experimental groups in both post- OPGIDT and post-VT tests are more than the mean scores of the (A2, B2) groups in both post- OPGIDT and post-VT tests. Since the kurtosis and Skewness values were between (2, -2) so the data were normally distributed.

Table 1(b). Descriptive statistics of the pre-OPGIDT and pre-VT test scores in experimental (A1, B1) and (A2, B2) groups

	OPGIDT TEST		VT TEST	
Groups	Pre-test (A1, B1)	Pre-test (A2, B2)	Pre-test (A1, B1)	Pre-test (A2, B2)
Mean	55.8370	56.0652	55.9348	55.9674
Median	56.0000	56.0000	56.0000	56.0000
Mode	55.00a	50.00a	60.00	60.00
Std. Deviation	7.30539	7.14651	7.24425	7.21027
Skewness	-.093	-.039	-.081	-.055
Kurtosis	-.986	-.903	-.915	-.972

Considering Table 1(b), the mean scores of the (A1, B1) experimental groups in the pre- OPGIDT test is 55/837, the mean scores of the (A2, B2) experimental groups in the pre- OPGIDT test is 56/065 and the mean scores of the (A1, B1) experimental groups in the pre-VT test is 55/934 and the mean scores of the (A2, B2) experimental groups in the pre-VT test is 55/967. The results demonstrated that the mean scores of the (A1, B1) experimental groups in both pre- OPGIDT and pre-VT tests are almost equal to the mean scores of the (A2, B2) groups in both pre- OPGIDT and pre-VT tests. Since the kurtosis and Skewness values were between (2, -2) so the data were normally distributed.

4.1 The Pre-assumptions of the Covariance Analysis

4.1.1 Normality of the Scores

In order to check the normal distribution of the data, Kolmogorov-Smirnov test were conducted. In other words, the following assumptions were tested. The null distribution of this statistics is calculated under the null hypothesis that the sample is drawn from the reference distribution (in the one-sample case) or that the samples are drawn from the same distribution (in the two-sample case). If the probability value or Sig is smaller than the value of 0.05, the null hypothesis or the assumption of the normal distribution of the sample is rejected at the error level of 5%, otherwise, the null hypothesis will be confirmed, which means that the data are normally distributed. It should be noted that if the assumption of the normal distribution of data is accepted, then in order to test the research hypotheses, the parametric method and otherwise non-parametric methods will be used. The results of the Kolmogorov-Smirnov tests were demonstrated in Tables 2 and 3.

Table 2. The normality of the post-test scores in experimental groups

One-Sample Kolmogorov-Smirnov Test

Groups		Post-test (A1, B1)	Post-test (A2, B2)	Result
OPGIDT test	Kolmogorov-Smirnov Z	.726	.625	Distribution of normal data
	Asymp. Sig. (2-tailed)	.667	.830	Distribution of normal data
VT test	Control	.627	.650	Distribution of normal data
	Asymp. Sig. (2-tailed)	.826	.792	Distribution of normal data

a. Test distribution is Normal

Table 3. The normality of the pre-tests scores in experimental groups

One-Sample Kolmogorov-Smirnov Test

Groups		Pre-test (A1, B1)	Pre-test (A2, B2)	Result
OPGIDT test	Kolmogorov-Smirnov Z	.825	.638	Distribution of normal data
	Asymp. Sig. (2-tailed)	.504	.810	Distribution of normal data
VT test	Control	1.030	.577	Distribution of normal data
	Asymp. Sig. (2-tailed)	.239	.893	Distribution of normal data

a. Test distribution is Normal.

Considering the Sig values obtained within the two and three tables, all of them were over 0.05, H0 that was the normality of the variables in the pre and post-test scores at the significance level of 0.05 was accepted.

4.2 Homogeneity of the Variances

In this study, Levee's test was an inferential statistical test that was used to evaluate the equality of variance for a variable that is computed for two or more groups. Some common statistical methods assume that the variance of the population from different samples was taken as equal. In this method, a test was used to evaluate the homogeneity of variance and results were presented in Tables 4 and 5.

Table 4. Homogeneity of variance between experimental groups in pre and post-OPDGTD test

Test of Homogeneity of Variances

	Levene's Statistic	df1	df2	Sig.	Result
Pre-test (A1, B1)	.047	1	182	.829	The assumption of the equality of variances is accepted
Post-test (A2, B2)	.087	1	182	.768	The assumption of the equality of variances is accepted

Table 5. Homogeneity of variance between experimental groups in pre and post-VT test

Test of Homogeneity of Variances

	Levene's Statistic	df1	df2	Sig.	Result
Pre-test (A1, B1)	.000	1	182	.983	The assumption of the equality of variances is accepted
Post-test (A2, B2)	.971	1	182	.326	The assumption of the equality of variances is accepted

Considering the Sig values obtained in Tables 4 and 5, all of that were over 0.05, the H0 that was concerning homogeneity of the variances at the significance level of 0.05 was accepted and so the belief of the homogeneity of the variances of the participants within the pre and post-tests scores were accepted with the 5% level of error.

4.3 Covariance Running before Beginning the Study

This presupposition was followed and a pre-test has been performed for LEARNERS at intermediate level, before the implementation of the independent variable.

4.4 Homogeneity of Regression Slope

To analyze the homogeneity of regression slope, the F value was calculated between covariance and independent variables the results which were presented in Tables 6 showed that this index was significant (Sig> 0.05).

Table 6. Regression Slope homogeneity test between covariance and independent variable

Tests of Between-Subjects Effects

Dependent Variable: Posttest

Source	Type III Sum of Squares	df	Mean Square	F	Sig	
group * pretest (OPGIDT Test)	159.473	2	79.736	1.503	.123	Reception of homogeneous regression slope
group * pretest (VT Test)	60.468	2	30.234	0.631	.341	Reception of homogeneous regression slope

According to Table 6 and the obtained sig values, which were all more than 0.05. H0 means that the homogeneity of the regression line slope between covariance and the independent variable was accepted at an important level of 0.05. The F value of the covariance variable was calculated in order to analyze the linearity of the correlation between covariance and an independent variable. Also, it was calculated between them and results were presented in Table 7 that express an index which was suitable (Sig >0/05).

Table 7. The test of linearity of the correlation of covariance and the independent variable

Source	Type III Sum of Squares	df	Mean Square	F	Sig
group * pretest (OPGIDT Test)	8648.742	1	8648.742	1.433E3	.000
group * pretest (VT Test)	4980.887	1	4980.887	249.513	.000

a. R Squared = .143 (Adjusted R Squared = .138)

In addition to using analysis of covariance, the necessary assumptions for analysis of covariance were examined. These assumptions were available and the results of this analysis of covariance are shown in Table 8.

4.5 Data Analysis of the First Research Question

First research question was: Does the washback impact of task-based assessment significantly improve Iranian intermediate EFL learners' grammatical ability?

For data analysis of the null hypothesis, which was: "The washback impact of task-based assessment doesn't significantly improve Iranian intermediate EFL learners' grammatical ability", and the alternative hypothesis "The washback impacts of task-based assessment significantly improve Iranian intermediate EFL learners' grammatical ability", as it is mentioned before, covariance analysis was used. Also, the necessary pre assumptions for analysis of covariance were investigated and these assumptions were existing. The results of covariance analysis were demonstrated in Table 8.

Table 8. The results of covariance analysis

Tests of Between-Subjects Effects

Dependent Variable: Post. OPGIDT

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	8666.400a	2	4333.200	717.784	.000
Intercept	509.073	1	509.073	84.327	.000
Pre.OPGIDT.1	8648.742	1	8648.742	1.433E3	.000
group1	32.242	1	32.242	5.341	.022
Error	1092.682	181	6.037		
Total	821725.000	184			
Corrected Total	9759.082	183			

a. R Squared = .888 (Adjusted R Squared = .887)

As was shown in Table 8, given the main output of covariance analysis in the fourth column, the value of (F= 32.242) in covariance analysis for grammatical ability scores was significant and the H0 was rejected and the H1 was accepted. It can also be inferred that, after the pre-test scores correction, there was a clear contrast between the two study classes (A1, B1) and (A2, B2) mean in the post-test.

4.6 Data Analysis of the Second Research Question

The second research question was: Does the washback impact of task-based assessment significantly improve Iranian intermediate EFL learners' vocabulary learning? For data analysis of this research question, covariance analyses were used. The study of covariance findings has been shown in Table 9.

Table 9. The results of covariance analysis

Tests of Between-Subjects Effects

Dependent Variable: Post OPGIDT

Source	Type III Sum of Squares	df	Mean Square	F	Sig
Corrected Model	5120.017a	2	2560.009	128.241	.000
Intercept	1521.204	1	1521.204	76.203	.000
pre	4980.887	1	4980.887	249.513	.000
group1	142.932	1	142.932	7.160	.008
Error	3613.200	181	19.962		
Total	740542.000	184			
Corrected Total	8733.217	183			

a. R Squared = .586 (Adjusted R Squared = .582)

As was shown in Table 9, given the main output of covariance analysis in the fourth column, the value of (F= 142.932) in covariance analysis for vocabulary test scores was significant, so the H0 was rejected and the H1 was accepted. Therefore, comparing the mean pre-test scores of the two experimental groups (A1, B1) and (A2, B2), it can be concluded that there is a significant difference. It can also be inferred that, after the pre-test scores correction, there was a clear contrast between the two study classes (A1, B1) and (A2, B2) mean in the post-test. In the pre-test, (A1, B1) elementary experimental group mean was 55.935 and it was 63.934 in the post-test, since the (A2, B2) experimental group mean was 56.967 in the pre-test and 62.195 in the post-test. The significant difference in the experimental groups' post-test scores gives us the conclusion which says the pre-test covariance factor reduction causes that X intermediate EFL learners' vocabulary learning substantially increased by the washback impacts of task-based evaluation.

5. 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. 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.

5.1 Research Question One

The primary research question for this investigation was: Does the washback impact of task-based assessment significantly improve X intermediate EFL learners' grammatical ability? In both classes, the researchers conducted a questionnaire after every three sessions. Task-based quizzes were offered to the experimental task-based groups and standard experimental groups were evaluated by standard quizzes (multiple-choice, fill-in-the-blanks, true-false, and matching test items) to understand the washback effects of task-based evaluation on the grammatical ability of EFL learners. Learners made progress on all factors from pre-test to post-test in the (A1, B1) categories and all learners showed positive changes. The final results showed that there are considerable variations in student success depending on pre-and post-test factors and the results showed the positive washback impacts of task-based assessment that significantly improved X intermediate EFL learners' grammatical ability.

In addition, in their research on the influence of dynamic assessment on elementary EFL learners' L2 grammar learning, Sharafi and Sardareh (2016) explored the effect of dynamic assessment on the grammar learning of elementary EFL learners. To this end, forty-six male adult elementary EFL learners took part in the study in two groups, namely the experimental and control groups. Based on intact group sampling, the participants were picked. Their homogeneity was also verified by the Michigan Test in Cambridge. The initial knowledge of the target grammatical items (time and location prepositions) by the participants was tested by the grammar pre-test. Then while the experimental group underwent its unique treatment for ten sessions in the form of dynamic evaluation, the control group witnessed the activation of their routine classroom. Both groups took post-test grammar at the conclusion of the treatment sessions. The findings of an independent sample t-test indicated that dynamic evaluation has a major effect on the learning of time and place prepositions by elementary EFL learners and the impact was strong.

And at last, Azarian, Nourdad, and Nouri (2016) in their research about the influence of dynamic assessment on elementary EFL learners' overall language achievement, the standard level test of Top Notch was conducted on 75 male learners to provide the homogeneity of the members in both classes, and 40 learners were selected for this study. Each including 20 participants was randomly placed in two control and experimental groups. In the usual class protocol, the participants in the control group were then taught Top NOTCH-FUNDAMENTALS A before the post-examination test. But, via dynamic evaluation methodology, participants of the experimental group was taught. In both groups, contributors were given a post-test to see if the overall language achievement of the participants was enhanced by dynamic evaluation. According to the test results, it was found that there was a fundamental difference between the overall language development in the two groups, i.e. the dynamic evaluation of the group is better than the control group. The results may have educational consequences for curriculum designers, testers, language tutors, and teacher educators.

The mentioned finding is in line with Brame and Biel (2015), Chehrazad and Ajideh (2012), Talebzadeh and Bagheri (2012), and 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 effect 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. The importance of this research question was using of task-based conclusion, as a method to improve grammatical knowledge of learners in the first level of high school and also can be used for different levels in different academic places.

5.2 Research Question Two

The second research question for this investigation was: Does the washback impact of task-based assessment significantly improve X intermediate EFL learners' vocabulary learning? The main objective was to raise the participants' vocabulary awareness, and then measure the washback impacts of task-based assessment on the X intermediate EFL learners' vocabulary learning. During two months and over period of 10 twenty-minute sessions learners in experimental task-based and traditional groups received direct vocabulary instruction. After every three sessions, a quiz was administered for both groups by researchers in order to find out the washback effect of task-based assessment on EFL learners' vocabulary learning. The experimental task based groups were given the task based quizzes and the experimental traditional group were assessed through classic quizzes, such as multiple choices, fill in blanks, true false, and matching test items.

Learners in (A1, B1) groups demonstrated gains on all measures from pretest to posttest and all learners demonstrated improvements. The overall findings determine that there are significant differences in student performance for condition on pre and post-tests and the results showed the positive washback effect of task based assessment that significantly improves EFL learners' vocabulary learning. Previous research about around the washback effect of task-based evaluation on X intermediate EFL learners' lexicon learning has given a sound contention for utilization of task-based appraisal to progress the lexicon learning of the learners. The current investigation was in line in a few points of interest with the past investigation on utilization of the task-based assessment within the classroom area. This finding supports the result of research by Alderson and Hamp-Lyons (1996) that the TOEFL affected both what and how teachers teach but contradicts the conclusion drawn by Alderson and Wall (1993), Watanabe (1996), and Green (2003), whose observations of International English Language Testing System (IELTS) preparation and English for academic purposes (EAP) classrooms indicated that course content was very clearly influenced by the test, but any influence on teaching and learning method was less obvious.

Interestingly, Reynolds, Shih, and Wu (2018) in their investigation about modeling Taiwanese adolescent learners' English vocabulary acquisition and retention: the washback impact of the college placement test center's reference word list examined the influence of the RWL and word property factors (Polysemy, Part of Speech, Word Length, & Word Family Size) on non-English majors' (n= 566) vocabulary acquisition (VA). Results demonstrated medium to large relationships between RWL Level/Inclusion and VA as well as small to medium correlations between Polysemy, Frequency, Word Length, and VA. An eight illustrative variable consecutive relapse represented half of the change in VA, with RWL Level and RWL Inclusion adding the most explanatory capacity to the model. Next, washback impacts of the RWL on Taiwanese high school English learners' VA was considered and proposals on compiling an experimentally new RWL were demonstrated. They showed that the assessments and suggestions for revision of the RWL are comparable in relation to the creation and revision of word records for other learning.

6. Conclusion

Based on the results achieved from the first research question, there are meaningful differences between the washback impacts of task-based assessment in (A1, B1) groups and the washback impacts of traditional assessment in (A2, B2) groups and the washback impacts of task-based assessment significantly improved the grammatical ability of the male and female LEARNERS in (A1, B1) groups more than traditional assessment in (A2, B2) groups. Based on the results achieved from the second research question, there are meaningful differences between the washback impacts of task-based assessment in (A1, B1) groups and the washback impacts of traditional assessment in (A2, B2) groups and the washback impacts of task-based assessment significantly improved the vocabulary learning of the male and female learners in (A1, B1) groups more than traditional assessment in (A2, B2) groups.

6.1 Implications for Instructors, Curriculum Designers, and Students

Generally, the significance of the TBLA and its washback effect are the pedagogical implications arising from this study. In conventional teacher-centered English classrooms, vocabulary is not mostly regarded a main and important skill and grammar is regarded as a very important skill for students. The teacher teaches the vocabulary only explaining or translating the lexicon in a very easy way and also teaches grammar in a traditional way that in the most of the cases are boring for students. Also the students are expected to memorize according to the routines they are learning through the textbooks and teachers. As it is suggested by researchers and communicative language educators, there could be a number of pedagogical tasks for the students to develop their vocabularies and grammatical ability in learning EFL. Thus, teachers should increase the students' motivation and confidence in using the appropriate and authentic tasks and enhance their effort to provide appropriate washback effect themselves. Pedagogically, this study suggests that TBLA and its washback effect should be used and will help EFL students improve their knowledge of vocabularies and grammar. The results of the study demonstrated that using task based assessments and quizzes in meaningful contexts expands the learners' mental representation in a way to proliferate their effects to utilize words and grammatical points in the future properly.

Therefore, the roles of language teachers in classroom are so important because they should help students become skilled problem (task) solvers so that they can continue to make practice outside of the classroom. Language teachers must not simply teach the vocabulary and grammar. Rather, they should train learners how to use them as a tool to reach their own particular goals. So, teachers may take TBLA so seriously in vocabulary learning and grammatical ability and other language skill areas and consider it as a part of task-based language teaching and assessment. Another implication is for curriculum designers or syllabus writers that should consider step by step to guide tasks and processes and design and improve textbooks based on this important subject. Students also can use TBLA in their learning process as it helps them to know their difficulty in using language in authentic and real tasks in real worlds. Finally, the results demonstrated new aspects of the TBLA effectiveness in promoting students' vocabulary and grammar. The study may raise teachers' awareness and encourage them to utilize TBLA in their classroom. The results also provided both teachers and students with valuable perspectives into how TBLA plays an important role in the process of teaching and learning vocabulary and grammar.

6.2 Limitations and Delimitations of the Study

Some difficulties and limitations imposed upon the research process. Firstly, there are some learners' characteristics such as motivation, attitudes toward English language learning, personality and so on, which couldn't be controlled hence and may have affected the internal validity of this research. The second one was due to the laws of Education Department of Zanjan Province, in this study the researchers were in the shortage of time, during two months and over period of 10 forty-minute sessions, students in the experimental task based and traditional groups received the vocabulary grammar instruction and only six quizzes were used to investigate the washback effects. However, previous researches have garnered statistically significant results with shorter time periods for the intervention during a course in one month or for longer time periods for the intervention during three or four months and also the researchers used more or less than six quizzes during the intervention.

The third limitation addresses the number of the participants of this study. In addition, the researchers had no opportunity to have access more participants due to the shortage of time. The study had two delimitations as follows: the participants were deliberately only 184 Iranian intermediate EFL students in Zanjan city. Thus, the research was done on this proficiency level only and not on other proficiency level and not in other cities of Zanjan province for example Abhar and Khoramdare. In addition, this study focused on washback impacts of task-based assessment on Iranian intermediate EFL learners' vocabulary learning grammatical ability, and other aspects like writing, listening, and reading were not investigated.

References

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

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

Alsied, S. M., Ibrahim, N. W., & Pathan, M. M. (2018). The use of grammar learning strategies by Libyan EFL learners at Sebha University. *ASIAN TEFL, 3*(1), 37-51. doi: http://dx.doi.org/10.21462/asiantefl.v1i1.40

Andrews, S., Fullilove, J., & Wong, Y. (2002). Targeting washback-A case study. *System, 30*(2), 207-223. https://doi.org/10.1016/S0346-251X(02)00005-2

Azarian, F., Nourdad, N., & Nouri, N. (2016). The effect of dynamic assessment on elementary EFL learners' overall language attainment. *Theory and Practice in Language Studies, 6*(1), 203-208. doi: http://dx.doi.org/10.17507/tpls.0601.27

Bachman, L. F. (1995). *Fundamental considerations in language testing.* Hong Kong: Oxford University Press.

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

Beglar, D., & Hunt, A. (2002). Implementing task-based language teaching. In J. Richards & W. A. Renandya (Eds.), *Methodology in language teaching: An anthology of current practice* (pp. 96-106). Cambridge: Cambridge University Press.

Bokiev, U., & Abd Samad, A. (2021). Washback of an English language assessment system in a Malaysian university foundation programme. *The Qualitative Report, 26*(2), 555-587. https://doi.org/10.46743/2160-3715/2021.4349

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

Çakmak, F., & Erçetin, G. (2018). Effects of gloss type on text recall and incidental vocabulary learning in mobile-assisted L2 listening. *ReCALL, 30*(1), 24-47. doi: https://doi.org/10.1017/S0958344017000155

Cate, C. (2018). The comparative study of grammar learning mechanisms: Birds as models. *Current Opinion in Behavioral Sciences, 21*, 13-18. doi: https://doi.org/10.1016/j.cobeha.2017.11.008

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

Clapham, C. (2000). Assessment and testing. *Annual Review of Applied Linguistics, 20*(5), 147-161. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.455.4548&rep=rep1&type=pdf

Damankesh, M., & Babaii, E. (2015). The washback effect of Iranian high school final examinations on LEARNERS' test-taking and test-preparation strategies. *Studies in Educational Evaluation, 45*, 62-69. doi: https://doi.org/10.1016/j.stueduc.2015.03.009

Dong, M., Fan, J., & Xu, J. (2021). Differential washback effects of a high-stakes test on students' English learning process: evidence from a large-scale stratified survey in China. *Asia Pacific Journal of Education, 3*(1), 1-18. https://doi.org/10.1080/02188791.2021.1918057

Duque Micán, A., & Cuesta Medina, L. (2017). Boosting vocabulary learning through self-assessment in an English language teaching context. *Assessment & Evaluation in Higher Education, 42*(3), 398-414. doi: https://doi.org/10.1080/02602938.2015.1118433

Eicher, B. L., & Joyner, D. (2021). Components of assessments and grading at scale. In *Proceedings of the Eighth ACM Conference on Learning@ Scale* (pp. 303-306). https://doi.org/10.1145/3430895.3460165

Farhady, H., Jafarpur, A., & Birjandi, F. (1995). *Testing language skills from theory to practice.* Tehran: SAMT.

Gebril, A. (2018). Integrated-skills assessment. *The TESOL Encyclopedia of English Language Teaching*, 1-7. doi: https://doi.org/10.1002/9781118784235.eelt0544

Gipps, C. V. (1994). *Beyond testing: Towards a theory of educational assessment*. London: Falmer Press.

Goldstein, H., Ziolkowski, R. A., Bojczyk, K. E., Marty, A., Schneider, N., Harpring, J., & Haring, C. D. (2017). Academic vocabulary learning in first through third grade in low-income schools: Effects of automated supplemental instruction. *Journal of Speech, Language, and Hearing Research, 60*(11), 3237-3258. doi: https://doi.org/10.1044/2017_JSLHR-L-17-0100

Green, A. (2003). *Test impact and English for academic purposes: A comparative study in backwash between IELTS preparation and university pre-sessional courses.* Unpublished PhD thesis, Centre for Research in Testing, Evaluation and Curriculum in ELT, University of Surrey, Roehampton.

Inger, M. (1993). Authentic assessment in secondary education, *IEE BRIEF*. https://eric.ed.gov/?id=ED365711

Jean, G., & Simard, D. (2011). Grammar learning in English and French L2: LEARNERS' and teachers' beliefs and perceptions. *Foreign Language Annals, 44*(4), 465–492.

Kafipour, R., Jafari, S., & Khojasteh, L. (2018). The effect of task-based instruction on L2 grammar learning and motivation of Iranian EFL learners' junior high school LEARNERS. *Revista Publicando, 5*(16), 769-795.

Kermad, A., & Kang, O. (2018). Effect of classroom assessment stakes on English language learners' oral performance. *TESOL Journal*, *10*(2). doi: https://doi.org/10.1002/tesj.392

Kessler, M., Solheim, I., & Zhao, M. (2021). Can task-based language teaching be "authentic" in foreign language contexts? Exploring the case of China. *TESOL Journal*, *12*(1), 1-16. https://doi.org/10.1002/tesj.534

Knight, S. (2020). Augmenting assessment with learning analytics. In: Bearman M., Dawson P., Ajjawi R., Tai J., Boud D. (eds) *Re-imagining university assessment in a digital world. The enabling power of assessment*, vol 7. Springer, Cham. https://doi.org/10.1007/978-3-030-41956-1_10

Krashen, S. D. (1989). We acquire vocabulary and spelling by reading: Additional evidence for the input hypothesis. *The Modern Language Journal*, *73*(4), 440-463. https://doi.org/10.2307/326879

Lam, R. (2018). Processes in portfolio development journey. In *Portfolio Assessment for the Teaching and Learning of Writing* (pp. 29-42). Springer, Singapore.

Levi, T., & Inbar-Lourie, O. (2020). Assessment literacy or language assessment literacy: Learning from the teachers. *Language Assessment Quarterly*, *17*(2), 168-182. https://doi.org/10.1080/15434303.2019.1692347

Liu, H., & Brantmeier, C. (2019). I know English: Self-assessment (SA) of foreign language (FL) reading and writing abilities among young Chinese learners of English. *System, 80*, 60-72. doi: https://doi.org/10.1016/j.system.2018.10.013

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).

Lyle, C., Rowland, M., Ostrovski, G., & Dabney, W. (2021). On The effect of auxiliary tasks on representation dynamics. In *International Conference on Artificial Intelligence and Statistics* (pp. 1-9). PMLR.

Ma, H. (2021). *Washback effects of the IELTS test: views and experiences of Chinese students in the context of a Sino-UK joint programme with English as the medium of instruction*. Doctoral dissertation, Queen's University Belfast.

McKinley, J., & Thompson, G. (2018). Washback effect in teaching English as an international language. *The TESOL Encyclopedia of English Language Teaching*, 1-12. doi: https://doi.org/10.1002/9781118784235.eelt0656

McNamara, T. (2001). Language assessment as social practice: Challenges for research. *Language Testing, 18*(4), 333–350. https://doi.org/10.1177/026553220101800402

Miles, M., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Thousand Oaks, CA: Sage.

Milne, A. E., Petkov, C. I., & Wilson, B. (2018). Auditory and visual sequence learning in humans and monkeys using an artificial grammar learning paradigm. *Neuroscience, 389*, 104-117. https://doi.org/10.1016/j.neuroscience.2017.06.059

Morris, S. B. (2016). *Using non-traditional testing methods to identify talent and potential in early childhood* (Doctoral dissertation, University of Southern California).

Morrow, C. K. (2018). Communicative language testing. *The TESOL Encyclopedia of English Language Teaching*, *12*, 1-7. doi: https://doi.org/10.1002/9781118784235.eelt0383

Nazari, M., Bayati, A., & Rajabi, P. (2021). The washback effect of task-based assessment on the Iranian EFL learners' development of pragmatic competence. *International Journal of Foreign Language Teaching and Research*, *9*(34), 177-189. http://jfl.iaun.ac.ir/article_677957.html

Pitoyo, M. D. (2020). Gamification-based assessment: The washback effect of quizizz on students' learning in higher education. *International Journal of Language Education*, *4*(1), 1-10. https://doi.org/10.26858/ijole.v4i2.8188

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. http://www.psychologie-aktuell.com/fileadmin/download/ptam/4-2010_20101218/02_Rauch.pdf

Reynolds, B. L., Shih, Y. C., & Wu, W. H. (2018). Modeling Taiwanese adolescent learners' English vocabulary acquisition and retention: The washback effect of the College entrance examination center's reference word list. *English for Specific Purposes*, *52*, 47-59. https://daneshyari.com/article/preview/9951973.pdf

Safa, M. A., & Jafari, F. (2017). The washback effect of dynamic assessment on grammar learning of Iranian EFL learners. *Journal of Language Teaching and Learning*, *7*(1), 55-68. doi: 10.1016/j.sbspro.2014.03.393

Salehi, H., & Yunus, M. M. (2012). The washback effect of the Iranian universities entrance exam: Teachers' insights. *GEMA Online® Journal of Language Studies*, *12*(2), 609-628. https://ejournal.ukm.my/gema/article/view/630

Salma, N., & Prastikawati, E. F. (2021). Performance-based assessment in the English learning process: washback and barriers. *Getsempena English Education Journal*, *8*(1), 164-176. https://doi.org/10.46244/geej.v8i1.1305

Sattar, W., Abdullah, M. R. T. L. B., & Mirzaei, F. (2018). A FAHP approach to select students' performance assessment criteria in task-based English language teaching. In *SHS Web of Conferences*, *53*, 1-6. doi: https://doi.org/10.1051/shsconf/20185303005

Sert, O., & Amri, M. (2021). Learning potentials afforded by a film in task-based language classroom interactions. *The Modern Language Journal*, *105*(1), 126-141. https://doi.org/10.1111/modl.12684

Sharafi, M., & Sardareh, S. A. (2016). The effect of dynamic assessment on elementary EFL LEARNERS' L2 grammar learning. *Journal of Applied Linguistics and Language Research*, *3*(3), 102-120. http://www.jallr.com/index.php/JALLR/article/view/291

Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: Washback effect over time. *Language Testing*, *13*(3), 298-317. doi: https://doi.org/10.1177/026553229601300305

Spierings, M. J., & ten Cate, C. (2016). Budgerigars and zebra finches differ in how they generalize in an artificial grammar learning experiment. *Proceedings of the National Academy of Sciences*, *113*(27), 3977-3984. doi: https://doi.org/10.1073/pnas.1600483113

Sumera, A., Barua, A., & Navamoney, A. (2015). Exploring the effect of backwash in first year medical LEARNERS and comparison with their academic performances. *Procedia-Social and Behavioral Sciences*, *174*, 491-495. doi: https://doi.org/10.1016/j.sbspro.2015.01.693

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. doi**:**10.5539/ijel.v2n1p257

Tayeb, Y. A., Aziz, M. S. A., & Ismail, K. (2018). Predominant washback of the general secondary English examination on teachers. *International Journal of Engineering & Technology*, *7*(3.21), 448-456. https://www.sciencepubco.com/index.php/ijet/article/view/17211

Taylor, L. (2005). Washback and impact. *ELT Journal*, *59*(2), 154-155. doi:10.1093/eltj/cci030

Tollefsen, K. E., Scholz, S., Cronin, M. T., Edwards, S. W., de Knecht, J., Crofton, K., & Patlewicz, G. (2014). Applying adverse outcome pathways (AOPs) to support integrated approaches to testing and assessment (IATA). *Regulatory Toxicology and Pharmacology*, *70*(3), 629-640. doi: https://doi.org/10.1016/j.yrtph.2014.09.009

Valette, R. (1977). *Modern language testing* (2nd ed.). New York: Harcourt Brace Jovanovich.

Watanabe, Y. (1996). Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research. *Language Testing*, *13*(1), 318-333. http://dx.doi.org/10.1177/026553229601300306

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. https://ijllnet.com/journals/Vol_1_No_1_June_2014/4.pdf