International Conference on Current Trends in ELT

The Washback Effects of Task-based Assessment on the Iranian EFL Learners' Grammar Development

Mohammad Ahmadi Safa\textsuperscript{a}, Sa'ideh Goodarzi\textsuperscript{b,*}

\textsuperscript{a}Bu Ali Sina University, Hamedan, Iran
\textsuperscript{b}Islamic Azad University of Hamedan, Science & Research Branch

Abstract

There is some evidence to suggest that tests have washback effects on teaching and learning processes (Alderson & Wall, 1993). Task-based language assessment (TBLA) as an alternative mode of language testing is believed to be of washback effect on language learning (Mislevy, Almond & Steinberg, 2002). However, the nature of this effect on the development of different language elements and skills is yet to be explored. In a partial attempt to address the lacuna, this study investigates the washback effect of TBLA on English as a foreign language learners' (EFL) grammar development. Seventy-four EFL learners were randomly selected from 110 pre-intermediate learners of a language institute and were divided into two groups of control and experiment. To ensure the same level of grammar ability, both groups took a grammar pre-test at the outset of the project. During ten sessions of treatment, the groups received the same kind of grammar instruction, however, the experimental group took a researcher made task-based grammar quiz every three sessions, but the control group took traditional grammar quizzes. After the treatment, the two groups took a grammar post-test. The analyses revealed a positive washback effect of the TBLA on the grammar development of EFL learners. The findings imply that TBLA as a pedagogical measurement tool can well replace the classic assessment procedures as all educational efforts including testing and assessment procedures are planned to maximize the educational gains and developments.

© 2014 The Authors. Published by Elsevier Ltd.
Selection and peer-review under responsibility of Urmia University, Iran.

Keywords: Task-Based Language Assessment; Washback effect; Grammar

1. Introduction

In the fields of education and applied linguistics it is widely believed that testing influences teaching and learning. This influence is referred to as ‘washback’ (Wall & Alderson, 1993), ‘backwash’ (Hughes, 1993, cited in Bailey 1996), or ‘test impact’ (Bachman & Palmer, 1996). McEwen (1995a, p.42) described the effect mechanism

* Corresponding author. Tel.: +98-916-668-2655
E-mail address: Saideh_Goodarzi@yahoo.com
when he stated ‘what is assessed becomes what is valued, which becomes what is taught’. The degree of washback varies over time in accordance with the status of the test, the status of the language being tested, the purpose of the test, the format of the test and skills tested (Shohamy, Donitsa-Schmidt & Ferman, 1996). According to Eckstein and Noah (1993) the majority of washback studies have focused on the positive or negative consequences of standardized tests, however, washback exists in any type of assessment including TBLA in which test results affect test-takers’ future course of development and learning, and thus are regarded as high-stake tests.

Task-based language teaching brings challenges to all areas of the EFL curriculum particularly to the assessment area by recognizing that knowledge of vocabulary and grammar is not sufficient to use a language to achieve ends in social situations. In other words, it embraces a broader conception of communicative competence (Mislevy, Almond & Steinberg, 2002). Task-based tests are defined as any assessment means that ‘require students to engage in some sort of behaviour which stimulates, with as much fidelity as possible, goal-oriented target language use outside the language test situation. Performances on these tasks are then evaluated according to pre-determined, real-world criterion elements and criterion levels’ (Brown, 2004, p. 36).

Washback effect has been one of the greatest concerns for researchers in the field of language testing ample investigation have been done in this area (e.g., Alderson & Wall, 1993; Buck, 1998; Hughes, 2003). However, as Chalhoub-Deville (2001) asserts, ‘while the L2 literature includes numerous investigations of task-based instruction and learning, a cursory examination of testing publications shows that task-based assessment work is scarce’ (p.211). The same status can be found in a local scale in Iran and it appears that task-based assessment has not received the deserving attention in the Iranian ELT programs either. Against this backdrop, the present study is to compare TBLA and traditional assessment in terms of their washback effect on the grammar development of Iranian EFL learners.

2. Literature Review

It is stated that ‘testing is never a neutral process and always has consequences’ (Stobart, 2003, p. 140) as it is a differentiating ritual for students: ‘for everyone who advances there will be some who stay behind (Wall, 2000, p. 500). Language testing as an offshoot of testing is served by the research undertaken in the fields of language acquisition and language teaching (Buck, 1998) within which both testing and teaching are so closely interrelated that it is virtually impossible to work in either field without being constantly concerned with the other (Heaton, 1988).

Alderson (1986) recognized washback as a distinct and emerging area within the field of language testing. Washback is rooted in the notion that tests or examinations can and should drive teaching, and hence learning, and is also referred to as measurement-driven instruction (Popham, 1987). Many linguists have mentioned this term in their works. Hughes (1989, p.1) simply defines washback as ‘the effect of testing on teaching and learning’. Shohamy (1992, p.513) also refers to washback when she describes ‘the utilization of external language tests to affect and drive foreign language learning (in) the school context’ she underlines that ‘this phenomenon is the result of the strong authority of external testing and the major impact it has on the lives of test takers’. Biggs (1995, p. 12) uses the term ‘backwash’ to refer to the fact that testing drives not only the curriculum, but teaching methods and students’ approaches to learning. Messick (1996, p.241) describes washback as ‘the extent to which the test influences language teachers and learners to do what they would not otherwise necessarily do’ and adds an important dimension to the definition of washback when he states ‘evidence of teaching and learning effects should be interpreted as washback… only if that evidence can be linked to the introduction and use of the test.’ Andrews, Fullilove, and Wong, (2002, p.208) state that the term ‘washback’ is used to refer to the effects of tests on teaching and learning, the educational system and the various stakeholders in the education process’. There seems to be at least two major types or areas of washback or backwash studies those relating to traditional, multiple-choice, large-scale test, which are perceived to have mainly negative influences on the quality of teaching and learning (Shepard, 1990) and those studies where a specific test or examination has been modified and improved (e.g., performance-based assessment), in order to exert a positive influence on teaching and learning (Linn& Herman, 1997).

Task based language assessment has been viewed from different perspectives and different groups of language
teachers, researchers and testers share a common interest in addressing task-based assessment questions, but according to their definitions for ‘task-based’ assessment, their orientations vary widely. For some researchers working at the interface between second language acquisition (SLA) research and language pedagogy, the variable influences of task features on examinees’ cognitive processes and resulting performances have drawn special attention, particularly in terms of language production (e.g., Skehan, 1998; Wigglesworth, 2001). For others, ‘task-based’ suggests an idea of assessment alignment with instruction in the form of shared characteristics such as learner-centeredness, contextualization and authenticity (e.g., Chalhoub-Deville, 2001). Still others restrict ‘task-based’ to those assessment examples where interpretations need to be made about abilities of examinees to achieve specific target tasks especially in communication settings (e.g., Long & Norris, 2000). The last group of researchers draw attention to ‘task-based’ tests requiring examinees to engage in the types of activities characteristically encountered in communicative language teaching classrooms (e.g., Paltridge, 1992, cited in Wigglesworth, 2001). In short, the definition of task-based assessment and the incorporation of communication tasks into language testing practice are varied considerably among teachers, researchers and testers, especially depending on the unique aims of assessment within their distinct educational, occupational or research contexts. Growing interest in tasks as means to assess learner ability has resulted from the popularity of performance testing as opposed to multiple-choice and other discrete-point item. The candidate is required to engage in the performance of tasks in task-based assessment which simulate the language demands of real world situations with the aim of eliciting an ‘authentic’ sample of language. McNamara (2001) believes that some scholars are widely researching the properties of such tasks and the influence of these properties on learner performance by focusing on strengthening the links between test tasks and their real world counterparts (e.g., Bachman and Palmer, 1996) and others are focusing on the effects on candidates’ production through manipulation of different task characteristics in the test situation.

Task-based testing is nothing new in the field of language pedagogy. In fact, the integration of task-based language testing into communicative language teaching programs, learner-centered programs (Brindley, 1989; Nunan, 1988) and programs of English for Specific Purposes (ESP) (McNamara, 1989) has long been a goal of curriculum designers (cited in Robinson, 2000). The concern in TBLA extends beyond knowledge of language by itself, to the ability to use language knowledge appropriately and effectively in educationally or professionally important language use setting (Mislevy, Steinberg & Almond, 2002).

Task-based tests need candidates to perform an activity which simulates a performance they will have to engage in outside the test situation’ (Nunan, 2004, p.114). According to Ellis (2003) task-based assessment refers to ‘assessment that utilizes holistic tasks involving either real-world behaviour or the kinds of language processing found in the real-world activities’ (p.285). According to this definition by Ellis, direct performance-referenced tests and direct system-referenced tests employ tasks. Therefore, any tasks whether performance-based or system-based should be direct in essence (Ellis, 2004).

Tavakoli and Skehan (2005) presented a model of task-based performance in relation to language testing. The main purpose of which is to make obvious that the rating assigned to someone on the basis of their performance on a task is the consequence of a whole range of factors, only one of which can possibly be their underlying competence. In addition, the following factors should be considered too:

1. The method by which the rating is done, with the potential this has to introduce error;
2. The context for the performance, including the nature of the interactants involved, and their relationship to one another;
3. The extent to which the testee can engage strategies of performance, and general processing skills, handling rule-based and memory-based language;
4. The task that is involved and conditions under which it is done.
5. These different factors, besides helping us to understand how test scores may be the result of a multiplicity of influences, also provided and agenda for research. This means that we need to advance our understanding of the influence of task characteristics on performance, as well as what impact the conditions under which tasks are done might have on that same performance (Tavakoli & Skehan, 2005).

3. **Research Questions**

The research questions of this study were:
1. Is there any significant difference between the traditional assessment and TBLA concerning their washback effect on EFL learners’ grammar development?
2. Is the task-based assessment's washback effect gender sensitive?
3.

4. Methodology
   4.1. Participants
   The participants of the study were 74 EFL learners of Boroujerd Shokoh English Institute. As it was necessary for all of them to be at the same homogeneous level of language proficiency, first, 110 so-called pre-intermediate EFL learners of the same institute were chosen. Next, the Key English Test (KET) was administered and based on the KET scores, the 74 participants who scored ± 1 SD from the mean were regarded as true pre-intermediate level EFL learning sample of the study. Their age range was 14 to 18 and all were high school students speaking Persian as their first language.

   4.2. Instruments
   The testing instruments used in this study for different purposes were the sample Key English Test (KET), three teacher-made task-based grammar tests, three multiple-choice grammar tests, a 50 item grammar test as pre and post-tests, and a researcher compiled grammar booklet.

   Firstly, based on the proficiency level of the test takers; KET (Key English Test, 2010) was chosen to homogenize the sample. It tests the four skills of reading, writing, listening and speaking and is based on the Waystage specification (1990, Council of Europe). The grammar test (pre-test) consisted of 50 multiple-choice test items and was testing the participants' knowledge of determiners (a/an, the and some), wh-questions, and possessives. Three task-based grammar quizzes (for experimental group) and three traditional grammar quizzes (for control group) that were administered every 3 sessions made the third set of instruments of the research. There were 30 multiple-choice test items in each quiz for control group and 3 grammar task items for the experimental group. The same 50 item multiple choice grammar test (pre-test) was administered to the groups as the post-test after the treatment. The estimated reliability of the tests turned out to be 0.79 (pre-test and post-test), 0.77 (Exp. quiz N 1), 0.71 (Exp. quiz N 2), 0.78 (Exp. quiz N 3), 0.73 (Con. Quiz N 1), 0.76 (Con. Quiz N 2), 0.77 (Con. Quiz N 3). Moreover, the researchers compiled a grammar booklet for the presentation of the intended grammar points to the two groups.

   4.3. Data collection procedure
   This study was implemented on the basis of pre-test-post-test equivalent-group design and the following steps were taken for data collection. In the first stage 110 EFL learners placed at the same level courses were selected from among the EFL learners of Shokoh English Institute in Boroujerd. Then based on the assumed level, a KET sample test (2010) was administered to determine the participants’ true level of language proficiency. Next, 74 subjects who scored within ± 1 SD from the mean were chosen as the participating sample of the study and were randomly put into the experimental and control groups each one including 37 participants.

   In the second stage, both experimental and control groups took the grammar pre-test as a measure of the participants’ grammar knowledge of the selected English grammar points. In the third stage, both experimental and control groups were independently taught the grammar points in the same way using the compiled booklet for 10 ninety minute sessions. The class sessions were held twice a week. The taught grammar points were determiners, wh-questions, and possessives. To nullify the potential effect of methodology, both groups were taught by the same instructor (researcher). After every three sessions, the researchers administered a quiz for both groups. The experimental group was given the task-based quiz and control group was assessed through classic method (multiple-choice, fill-in-the-blanks, true-false and matching test items). So the researchers administered three separate quizzes during the ten sessions and tried to find out the differential effect (if any) of the two assessment types on the following grammar learning and final achievement of the participants in the two groups. Finally, the post-test was administered in the two groups after the treatment sessions.
5. Results

In order to answer the research questions and test the related hypotheses different statistical procedures and tests were run on the data. The pre-test performances of the two groups were compared using an independent samples t test. The results of the analysis are presented in tables 1 and 2.

Table 1. Two Groups' Pre-test Descriptive Statistics

<table>
<thead>
<tr>
<th>Grp</th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expconpre</td>
<td>1</td>
<td>37</td>
<td>30.81</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>37</td>
<td>29.03</td>
</tr>
</tbody>
</table>

Table 1 shows that the mean scores of pre-test results of the experimental group (M= 30.81, SD=5.86) was not strongly higher than that of control group participants (M=29.03, SD=5.00). However, in order to prove the similarity of the results on a statistical basis, a t-test was run on the data (table 2).

Table 2. Independent T-test of the Pre-test Results

<table>
<thead>
<tr>
<th></th>
<th>F</th>
<th>Sig.</th>
<th>T</th>
<th>df</th>
<th>Sig. (2-tailed)</th>
<th>Mean Difference</th>
<th>Std. Error Difference</th>
<th>95% Confidence Interval of the Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equal variances assumed</td>
<td>.80</td>
<td>.37</td>
<td>1.40</td>
<td>72</td>
<td>.16</td>
<td>1.78</td>
<td>1.26</td>
<td>-.74 to 4.31</td>
</tr>
<tr>
<td>Equal variances not assumed</td>
<td>1.40</td>
<td>70.28</td>
<td>.16</td>
<td>1.78</td>
<td>1.26</td>
<td>-.74</td>
<td>4.31</td>
<td></td>
</tr>
</tbody>
</table>

As is shown in table 2, p value is (p= .16) is higher (t=1.40, df =72, two-tailed p=.16) than the critical level of significance (p= .05) (*p>.05). Therefore, as the results indicate there was no statistically significant difference between the two groups concerning their knowledge of the intended grammar points.

Table 3. Descriptive Statistics of Post-test

<table>
<thead>
<tr>
<th>Grp</th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expconpost</td>
<td>1</td>
<td>37</td>
<td>44.03</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>37</td>
<td>40.51</td>
</tr>
</tbody>
</table>

As table 3 reveals the mean score of experimental group in the post-test (M= 44.03, SD=3.71) was apparently higher than that of the control group (M=40.51, SD=3.84).
Table 4. Independent Samples T-test for Post-test Results

<table>
<thead>
<tr>
<th></th>
<th>Levene's Test for Equality of Variances</th>
<th>t-test for Equality of Means</th>
<th>95% Confidence Interval of the Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F</td>
<td>Sig.</td>
<td>T</td>
</tr>
<tr>
<td>expconpost</td>
<td>.45</td>
<td>.50</td>
<td>3.99</td>
</tr>
<tr>
<td>Equal variances</td>
<td>assumed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Equal variances</td>
<td>not assumed</td>
<td></td>
<td>3.99</td>
</tr>
</tbody>
</table>

As is shown in table 4, p value is less (t=3.99, df =72, two-tailed p=.00) than level of significance (p=.05) (*p<.05, two-tailed.). Therefore, the results indicate that there is statistically significant difference between two groups in post-test which confirms the differential impact of TBLA and traditional assessment on the two groups' grammar learning.

The results of pre-test showed that there was no significant difference between the selected experimental and control groups' participants at the outset of the project. The analysis of the post-test and the t-test of group means showed a significant difference between the two groups, suggesting that the students in experimental group improved their grammar skill significantly after they were assessed through the task-based assessment.

The first hypothesis of this study stated that task-based language assessment has an impact on the EFL learners’ grammar development. In order to test the first hypothesis the two groups' performances in the second and third quizzes were also compared to find out the potential differences. The first quizzes' results were not compared as their washback effect should have logically affected the grammar development in a posteriori manner.

Table 5. Descriptive Statistics of Quiz 2

<table>
<thead>
<tr>
<th>Grp</th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>q2expcon</td>
<td>1</td>
<td>37</td>
<td>26.10</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>37</td>
<td>20.29</td>
</tr>
</tbody>
</table>

Table 5 presents the mean score of experimental group participants in the second quiz (M= 26.10, SD=3.09) as apparently higher than that of control group (M=20.29, SD=4.98). Independent samples t-test was applied to see if the apparent difference is statistically the case or not. Table 6 summarizes the t-test results.

Table 6. Independent Samples T-test Quiz 2
As is shown in Table 6, the p value is \( p = .00 \) is less than the critical level of significance \( p = .05 \) (* \( p < .05 \), two-tailed.); therefore, the results indicate that there was statistically significant difference between two groups in their second quiz performance and the task based assessment has had positive washback effect on the follow up grammar learning of the EFL learners. To further test the validity of this finding the third quiz results were compared in a similar fashion.

Table 7. Descriptive Statistics for Quiz 3

<table>
<thead>
<tr>
<th>Grp</th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>q3expcon</td>
<td>1</td>
<td>37</td>
<td>26.59</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>37</td>
<td>17.68</td>
</tr>
</tbody>
</table>

Table 7 presents the mean score of experimental group \((M=26.59, SD=2.84)\) as significantly higher than that of control group participants \((M=17.68, SD=3.86)\).

Table 8. Independent T-test Quiz 3

<table>
<thead>
<tr>
<th>Levene's Test for Equality of Variances</th>
<th>t-test for Equality of Means</th>
<th>95% Confidence Interval of the Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>Sig.</td>
<td>T</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-----------------</td>
<td>----</td>
</tr>
<tr>
<td>q3expcon Equal variances assumed</td>
<td>10.67</td>
<td>.00</td>
</tr>
<tr>
<td>Equal variances not assumed</td>
<td>6.02</td>
<td>60.22</td>
</tr>
</tbody>
</table>

As shown in Table 8, level of significance is \( p = .00 \) which is less than critical \( p \) value \( p = 0.05 \) (* \( p < .05 \), two-tailed.); therefore, the results prove the statistical significance of the difference between two groups in quiz 3. The comparison of the two groups' performances on the second and third quizzed supports the superior washback effect...
of TBLA on the EFL learners' grammar development and hence the first research question is answered and its related null hypothesis is rejected in this way and it is concluded that task-based assessment has positive washback on the EFL learners' grammar development.

The second research question addressed the differential washback effect of traditional assessment and task-based assessment on the grammar development of male and female EFL learners and the null hypothesis for this research question assumes no significant difference between the two testing method's washback effect on the two genders' grammar development. In order to either confirm or reject this hypothesis, the pre-test and post-test mean scores of the male and female participants of the experimental and control groups are compared. The descriptive and inferential statistic information of the comparisons are summarized in single tables to prevent verbosity and save space.

Table 9. Descriptive and Inferential Statistics of Experimental Male and Females' Performances in Pretest

<table>
<thead>
<tr>
<th>Gender</th>
<th>Number</th>
<th>Mean</th>
<th>SD</th>
<th>T</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>21</td>
<td>30.33</td>
<td>6.70</td>
<td>-0.56</td>
<td>0.58</td>
</tr>
<tr>
<td>Male</td>
<td>16</td>
<td>31.44</td>
<td>4.68</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

According to the table 9, p value is (p=0.58) higher than the level of significance (*p>.05), then the results indicate that there was no significant difference between male and female's scores of pre-test in the TBLA group. However the comparison of the same participants' post-test results proved female participants as superior to males.

Table 10. Descriptive and Inferential Statistics of Experimental Male and Females' Performances in Post-test

<table>
<thead>
<tr>
<th>Gender</th>
<th>Number</th>
<th>Mean</th>
<th>SD</th>
<th>T</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>21</td>
<td>45.48</td>
<td>3.59</td>
<td>3.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Male</td>
<td>16</td>
<td>42.12</td>
<td>3.03</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

According to the table10, p value is (p=0.01) less than the level of significance (p=0.05) (*p<.05), hence the results indicate that there was significant difference between male and female's scores and female's scores were higher than male's scores. This finding might indicate that task based language assessment has more positive washback effect on the grammar development of female EFL learners. As a result, the second research question is answered and the related null hypothesis which assumed no differential washback effect of the TBLA testing method on the two genders is rejected.

6. Discussion and Conclusion

The use of assessment as a means of promoting curriculum change has become increasingly common not only in general education (Chapman & Snyder, 2000), but also in language education (for instance, Wall and Alderson, 1993; Cheng, 1997) and its different aspects including the assessment paradigms. This point is best reflected in Elton and Laurillard (1979) as they believe "the quickest way to change student learning is to change the assessment system' (cited in Andrews, et al. 2002, p. 209). The studies of such educational and pedagogical consequences of assessment procedures are mainly carried out through washback studies.

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) while 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. Wall (2000) contends 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. 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. Against this backdrop, in an attempt to consider the differential washback effects of task based language assessment procedure and traditional assessment modes on the follow up grammar learning and development of EFL learners from the students' learning perspective, this study compared the TBLA and traditional assessment procedures and the findings lent a strong support to McNamara (2001) as the assessment modes utilized in this study differentially affected the follow up grammar learning and development of the EFL learners. Furthermore, similar to what McNamara states about the positive washback effect of TBLA, the present study concluded that task based language assessment mode is of a strong positive impact on the EFL learners learning and development. McNamara (2001) holds that task-based assessments that require integrated content and skills have more positive washback than discrete item testing which often stifles communicative teaching approaches.

As for the researcher of the present study, the possible reasons for the superiority of the washback effect of TBLA compared with traditional assessment modes might be best sought in the goal oriented nature of the tasks and their authenticity. McNamara (2000) argues that there are two factors which distinguish second language task-based tests from traditional tests of second language: the fact that there is a performance by the candidate, and that this is judged using an agreed set of criteria. While the researchers of this study admit such differences, they do not believe them to be the most decisive factors as both TBLA and traditional assessment modes entail test taker's performances- though of different natures-and sets of agreed criteria. Norris, Brown, Hudson, and Bonk (2002) add a third criterion arguing that the tasks used in task-based assessments should be as authentic as possible and it is more justifiable to consider this authenticity criterion as the more distinctive feature of TBLA which leads to a more positive washback effect on the learners follow up learning and development as it is believed that only well-designed (Wigglesworth, 2001) and authentic assessment tasks have the potential to provide positive washback into the classroom. Similarly, authenticity of the tasks is what Chalhoub-Deville (2001) raises as the most important and decisive difference though in a different wording. As he puts it, TBLA as a performance assessment differs from traditional paper-and-pencil tests in that the primary focus is to get an accurate picture of students' communicative abilities and to generalize about students' ability beyond the learning and testing situation to real-life communication.

Admitting the positive washback effect of TBLA on instruction and learning, Long and Norris (2000) propose that interest in TBLA can be attributed to such factors as the alignment of task-based assessment with task-based instruction, positive "washback" effects of assessment practices on instruction, and the limitations of discrete-skills assessment (cited in Mislevy, Almond & Steinberg, 2002).

This study compared the washback effect of task-based assessment on grammar development of Iranian EFL learners with that of traditional assessment modes. The results of the study revealed that there is significant difference between TBLA and traditional tests concerning their washback effect on the follow up grammar learning of EFL learners and TBLA is proved to be of a positive washback effect while the traditional assessment modes were not comparatively of the same positive washback effect. Furthermore, the positive washback effect of TBLA on the EFL learners' grammar development is shown to be stronger for female EFL learners than the male participants. The positive washback effect of TBLA underscores the effectiveness of this language assessment mode as an alternative substitute of traditional assessment modes in educational measurement.

A number of implications are conceivable for the results of the study. First, EFL teachers and researchers may need to reflectively think on their assessment practices and beliefs to determine if and how their assessment practices help to improve their learners' language learning processes. Second, if we rightly assume that the main purpose behind relatively all assessment practices is to foster educational development and learning, and if we further believe in the efficiency of the new alternative modes of assessment for the intended development and learning, it is justifiable to replace the traditional assessment procedures with the new alternative modes like performance based assessment procedures. Even if the practicality considerations are preventing such replacements, the testing and assessment processes would yield a more accurate picture of the testees' knowledge and skills level if the traditional assessment modes are integrated with the alternative performance based approaches like task based language assessment procedures.
References