The Washback Effect of Oral Interview Test on the Speaking Ability of Iranian High School Students

Badiee Moghaddam, Nazanin

Master of Teaching English as a Foreign Language, Tabriz Branch of Islamic Azad University
Nazanin_babiee@yahoo.com

Abstract

There is a consensus that tests which create positive washback should be used in educational settings. The tests which are used in schools in Iran mostly concentrate on grammar points, thereby, overlooking oral production. Therefore, the high school graduates would be have good knowledge of grammar, but they would be unable to use English in real life. The present study was carried out in order to observe the effect of an interview as a testing tool on the speaking ability of high school students. To complete this study, the researcher formulated one hypothesis for the speaking ability. Two classes from Absal High School in Tehran were selected. One class was considered as the experimental class and the other one as the control class in the present study. The experimental group received a sample of the interview test, and they were told that they would need to take a mandatory interview test which would account for five points of the final exam. The results indicated that the interview test did not have significant effects on the speaking ability of the students.

Keywords: Interview, Speaking ability, Washback.
1. Literature review

There have been scholars who consider language tests as powerful tools to bring about change in the curriculum and the students’ learning. Alderson (1986) pointed out an additional area to language testing research should be added and the researchers’ attention should be directed to the washback effect. He strongly believes that innovation in the language curriculum should be made through innovations in language testing. Washback is often introduced in language testing courses for teachers as a powerful concept that all test designers need to strongly consider, and of which most classroom teachers are all well aware of. Davies (1985) asks whether tests should necessarily be in line with the curriculum, and suggests that perhaps tests should lead and influence curriculum. Morrow (1986) further used the term ‘washback validity’ to establish a relationship between testing and teaching and learning. He claimed that one advantage of washback validity would be that testing researchers will need to come into the classroom in order to observe the effect of their tests in action.

Research in language testing has centered on questions about how to assess the specific characteristics of a given group of test takers and how to incorporate these characteristics into the design of language tests. Cheng (2005, p. 25) believes that “the most important theoretical development in language testing, since the 1980s, has perhaps been the realization that a language test score represents a complexity of multiple influences”. According to Bachman (2004), one should not interpret language test scores simplistically as being an indicator of the particular language ability to be measured because test scores are also affected by the characteristics and content of the test tasks, the characteristics of the test taker, the strategies the test taker uses in completing the test task, and the inferences we wish to draw from them (Bachman & Palmer, 2010). The interpretation of test scores is made particularly difficult, according to Bachman and Palmer (1996), by these factors which undoubtedly have interactions with each other.

Traditionally, tests were administered at the end of the teaching and learning process for the purpose of evaluation. However, due to the existence of high-stakes public examination systems today, the direction has completely changed because testing precedes the teaching and learning process so that it can influence the process (Cheng & Curtis, 2004). They also argue that when tests are used as levers for change, new textbooks will be designed to fit the purposes of a new test, and school administrative staff, teachers, and students will all try to achieve good scores on the tests. Shohamy (1993, p. 2) pointed out that “the need to include aspects of test use in construct validation originates in the fact that testing is not an isolated event; rather, it is connected to a whole set of variables that interact in the educational process”.

Messick (1989) maintains that washback effect should be placed under the broader category of construct validity (consequential validity). Messick believes that aspects of test use, the effect that tests have on test takers and teachers, the interpretation which are made of scores by decision makers, and the misuses, abuses, and unintended uses of tests are all
included in construct validity. Washback, according to Messick, is an integral quality of any kind of test, especially when people's futures are influenced by the examination results, irrespective of the quality of the examination. However, whether the washback effect is negative or positive could depend on the examination measures.

Positive and negative washback effects have been associated with whether an examination is congruent with the sentiment and the purposes of the course objectives. Beneficial washback can be achieved if such a relationship is established, and it is negative if the test does not hold any congruence with course objectives.

1.1. The Origin of Washback

Washback has been used differently by scholars in the field. For instance, it was the term first used by Alderson and Wall (1993); Another illustration of this is that it has been applied as test impact by Bachman and Palmer (1996), systemic validity by Fredericksen and Collins (1989), consequential validity by Messick (1989), measurement-driven instruction by Popham (1983), and curriculum alignment by Shepard (1990). Cheng (2005) maintains that all these terms refer to the same phenomenon. He further believes that a lot of development in language testing has contributed to the fact that testing method and test taker characteristics influence test performance.

1.2. What is Washback?

It is important to keep in mind that washback is a neutral term, which may refer to positive or intended (Bachman & Palmer, 1996) or beneficial effects or harmful effects (Buck, 1988). In short, washback is an educational concept which comes from research studies into (a) the relationship between teaching, learning, and testing and into the relationship between different curriculum components, and (b) into the relationship of curriculum change and innovation. However, the word backwash can be found in certain dictionaries and has been defined as “the unwelcome repercussions of some social action” (Cheng, 2005, p. 27).

Washback or backwash refers to the effects of language testing on teaching and learning (Alderson & Wall, 1993). The influence of testing on teaching and learning washback or backwash can be sought in the notion that tests should push forward teaching and hence learning (measurement-driven instruction). This push is mainly looked for in high stake tests. If a match is created between the content and format of the test and the content and format of the curriculum (e.g. curriculum surrogate such as the textbook), this goal is achieved. The closer the fit or match between test content and curriculum, the greater the potential improvement on the test. Such an agreement can be found in the Hong Kong context in terms of (especially in the textbooks), where new or revised examinations are introduced into the education system with the purpose of improving teaching and learning (systemic validity).
Tests influence educational process in various ways. One commonly held perspective is that teachers will be influenced by the fact that their students are planning to take a certain test and will adapt their teaching methodology and lesson content accordingly. Popham (1987) laid out the structure of the traditional notion of measurement driven instruction to give an account of the relationship between instruction and assessment. Popham argued that “assessment directs teachers’ attention toward the content of test items, acting as powerful curricular magnets” (p. 680). In addition, measurement-driven instruction will happen when a high-stakes test of educational achievement has some influences on the instructional program that aims to prepare students for the test (Popham, 1987). Popham also stated that few educators would contend the idea that these kinds of high stakes tests have dramatic influences on the nature of educational programs. Accordingly, teachers desire to concentrate on a significant part of their instructional activities on the knowledge and skills assessed by such tests.

It seems that measurement-driven instruction can be a powerful force for educational improvement, but they should be properly conceived and implemented. Popham (1987, p. 682) believes that “creative teachers can ‘efficiently’ promote mastery of content-to-be tested and then get on with other classroom pursuits”. He further believes that “in a high-stakes environment in which the results of tests leads to rewards, sanctions, or public scrutiny and loss of professional status, teachers will be urged to pursue the objectives that the test embodies” (p. 682).

For decades, examinations have been long used to control educational systems. Arnove, Altbach, and Kelly believe (1992) that tests have been with us for a long time, at least a thousand years or more, if their use in Imperial China to select the highest officials of the land is taken into consideration. Those examinations were probably the first Civil Service Examinations ever constructed by the human beings. Although the examination aimed to choose civil servants, its washback effect involved creating and controlling an educational program (Spolsky, 1994). Latham (1877) characterized examinations as a trespassing power “to exert control over the internal educational systems that were becoming increasingly complex” (Spolsky, 1994, p. 58).

Tests definitely leave their effects on learning and teaching. Tests are not made and utilized in a test-tube without psychometric value; they are devised to meet the demands of an educational system or of society at large (Bachman, 1990). Cronbach (1984) has stated this fact as follows: “testing of abilities has always been intended as an impartial way to perform a political function that of determining who gets what” (p. 5). Cronbach believes that most are concerned about the consequences of using test results for the educational system or society. One consequence that is associated with language testing, for example, is that of ‘washback’, or ‘the effect of testing on instruction’ (Bachman, 1990, p. 283). Cheng (2005, p. 36) argues that “beliefs about testing reflect beliefs about teaching and learning”. Referring to the traditional teaching and learning views which have been grounded on behaviorist psychology.
and pedagogy, Cheng (2005, p. 36) argues that “the desired performance of language learners is brought about by reinforcing successive approximations of correct performance”.

There have been a lot of criticisms against examinations. One criticism is made by Madaus (1988, p. 85), who pointed out that “tests can become the ferocious masters of the educational process, not the compliant servants they should be”. Madaus also believes that measurement-driven instruction without a doubt leads to cramming, narrows the curriculum, concentrates attention on “those skills most amenable to testing” (p. 85). “It also constrains the creativity and spontaneity of teachers and students, and finally demeans the professional judgment of teachers” (Madaus, 1988, p. 85).

However, examinations have taken a prominent role in the educational policies of many countries (Eckstein & Noah, 1992). Such use of tests for power and control, as pointed out by Shohamy (1993), is an especially common practice in countries that have “centrally controlled educational agencies” (Shohamy, Donista-Schmidt & Ferman, 1996, p. 299). Shohamy (1993), who is against the use of tests to instigate change in education, argues that policymakers in central agencies, aware of the power of tests, use them to manipulate educational systems, to control curricula, and to impose new textbooks and new teaching methods. She also states that “tests represent a social technology deeply embedded in education, government and business, and they provide a mechanism for enforcing power and control” (p. 299). Furthermore, Shohamy et al. (1996, p. 299) state ,“ the power and authority of tests enable policy-makers to use them as effective tools for controlling educational systems and prescribing the behavior of those who are affected by their results, namely administrators, teachers and students”. Under those centrally controlled educational systems, “tests are viewed as the primary tools through which changes in the educational system can be introduced without having to change other educational components such as teacher training or curricula” (Cheng, 2005, p. 41). Petrie (1987, p.175) emphasizes that “it would not be too much of an exaggeration to say that evaluation and testing have become the engine for implementing educational policy”. Moreover, Linn (1992) commented that educational assessments are expected not only to serve as a monitor of educational achievement, but also to be powerful tools of educational reform.

Exams are used differently by people in charge of education. In other words, they have different attitudes to tests. Exams are used extensively in schools by principals and administrators to induce learning whereas in classrooms, tests and quizzes are employed by teachers to set discipline and to encourage learning (Stiggins & Conklin, 1992). Foucault (1979) views the examination as the most efficient tool through which society imposes discipline as it contains all features needed for power and control. Shohamy (2001) indicates that decisions -makers use tests for exercising power and control. The degree of a test impact, according to Madaus (1990), is often influenced by several factors, i.e. the subject-matter which is tested, the nature of the test, and the use to which the test scores are put.
Various aspects are affected by tests when they are administered for the first time. A number of studies have been done to investigate verbal, behavioral, attitudinal, and educational aspects immediately after a new test has been introduced into the educational system. According to this view, “academic tasks are broken down into discrete units and presented to the learners” (Noble & Smith, 1994, p. 2). However, according to the more recent psychological and pedagogical views of learning, three interrelated dimensions are emphasized:

1. Learning is regarded as a process which involves interpretation and construction of meaning (Glaser & Bassok, 1989)
2. “Learning is knowledge-dependent, and it is not merely an act of receiving information but one of interpreting information through earlier learning.” (Noble & Smith, 1994, p. 2)
3. Learning is situated in a social context (Resnick & Resnick, 1992). Teaching that reflects the cognitive-constructivist view of learning is likely to be “holistic, integrated, project oriented, long-term, discovery based, and social” (Cheng, 2005, p. 37).

Thus, performance assessment formed on the basis of constructivist model of learning is defined by Gipps (1994, p. 99) as “a systematic attempt to measure a learner’s ability to use previously acquired knowledge in solving novel problems or completing specific tasks”. Resnick and Resnick (1992, p. 59) stated that, “1) you get what you assess. 2) You do not get what you do not assess. 3) You should build assessments towards how you want educators to teach”. If tests have effects on curriculum and instruction, the argument goes; performance assessment could trigger thinking-oriented curriculum directed towards developing higher order abilities and problem-solving skills (Resnick & Resnick, 1992).

Referring to a ‘unified validity’ concept, Messick (1994) indicates that when an assessment model is designed to make inferences about a certain construct, the inferences drawn from that assessment model should not only derive from test score interpretation but also from other variables in the social context. In this context, Linn (1992, p. 29) stated that “it is incumbent upon the measurement research community to make the case that the introduction of any new high-stakes examination system should include more provisions for paying greater attention to investigations of both the intended and unintended consequences of the system”. “The issues involved in the ethics of test use are numerous and complex and their impact on test use will vary from culture to culture and from one testing context to another” (Bachman & Palmer, 2010, p. 280).

1.3. Theoretical Background of Washback

Washback, which is sometimes confused with backwash (Biggs, 1995, 1996, cited in Cheng, 2000), can be generally considered as the influence of an examination on teaching and learning (Chen, 2002b; Hughes, 2003; Cheng, 2003). However, not all scholars, have agreed to its definition. Alderson and Wall (1993) limited the application of the term ‘washback’ to “classroom behaviors of teachers and learners rather than the nature of printed and other
pedagogic material” (p.118). They would also regard washback to be as something teachers and learners do that “they would not necessarily otherwise do” (p. 117). Messick (1996) believes that for washback to take place, good or bad teaching has to be “evidentially linked to the introduction and use of the test” (p. 16). In addition, Wall (1997) has made a distinction between washback and test impact. He believes that these two are different, the latter referring to the effect of a test on “individuals, policies or practices, within the classroom, the school, the educational system or society as a whole” (cited in Cheng and Curtis, 2004, p. 4). Other researchers (Andrews, Fullilove & Wong, 2002) do not make such a distinction and they deem that narrow and wider effects can be categorized under the term washback. For the purposes of this study, washback will be understood in the wider sense that is, including what some scholars call ‘impact’.

The issue of washback has been linked to validity. Morrow (1986) refers to “washback validity” to depict “the quality of the relationship between testing and teaching and learning” (cited in Cheng, 2000, p. 4). For Messick (1996) washback is included within construct validity, and it is an integral feature of any kind of assessment, especially when the results are applied for significant judgments. For him, washback includes the consequential part of construct validity, but information about the operative level of test validity should help one distinguish test washback itself from the effects of good or bad educational practices irrespective of the quality of the test. Hence, washback in itself cannot be considered as a reliable criterion to establish test validity. It should be regarded as other test properties, such as authenticity and directness that may probably generate washback. Messick (1996) claims that in case a test is seen as defective, that is because it involves construct underrepresentation or construct-irrelevant features, then good teaching cannot be considered an influence of the test, and conversely, if a test is construct-validated, but there is poor teaching, then negative washback cannot be related to the test. Only valid tests (which minimize construct Washback of underrepresentation and construct irrelevancies), can increase the likelihood of positive washback.

1.4. Types of Washback

Alderson and Wall (1993) keep themselves away from a simplistic idea about the way a test can influence behaviors. Therefore, they constructed 15 washback hypotheses according to what is influenced: teaching, learning, content, rate, sequence, degree, depth, attitudes and also the number of teachers or learners affected by a test. Which hypotheses will come into play hinges upon the nature of the test, the educational context, and the nature of the judgments that are adopted on the basis of the test results. Actually, there appears to be a myriad of variables in society, education, and schools that play an important role in how washback will emerge.

Studying washback, there should probably some focus on participants (teachers, students, material developers, publishers), process (actions by participants towards learning), and products (what is learned and the quality of learning), as suggested in Hughes’s trichotomy.
model (Hughes, 1993). Watanabe (2004) proposes removing the complexity of washback by regarding it in terms of: Dimension (specificity, intensity, length, intentionality and value of the washback), aspects of learning and teaching that may be impacted by the examination, and the factors mediating the process of washback being produced (test factors, prestige factors, personal factors, macro-context-factors). Researchers normally put focus on one aspect or type of washback.

In Alderson’s and Wall’s study in Sri Lanka (Alderson & Wall, 1993; Wall, 1996), the administration of a test of English as a foreign language demonstrated to generate faster changes in the content of teaching than changes in teaching methodology. Cheng (1997), in the preliminary results of a study of the washback effect of the Hong Kong Certificate of Education Examination in English in Hong Kong secondary schools, believes that washback effect “works quickly and efficiently in bringing about changes in teaching materials and slowly and reluctantly and with difficulties in the methodology teachers employ” (p. 1).

Cheng puts forward the term ‘washback intensity’ to consider the “degree of washback effect in an area or a number of areas that an examination affects most” (p.7). Madaus (1990, cited in Shohami, Donitza-Schmidt & Ferman, 1996) recognizes as ‘high’ such situations when admission, promotion, placement or graduation are contingent on the test. Another aspect that has been investigated is whether the test has been applied as a leverage for change (Pearson, 1988 in Cheng, 1997), so everything, from textbooks to staff, works to obtain better scores. Cheng (2000) reports on how tests are often included into the education system to enhance teaching and learning, especially in centralized countries where tests are considered an effective device for making changes into an educational system without having to change other educational elements. In some countries these tests can be considered “the engine for implementing educational policy” (Petrie, 1987, p. 175 in Cheng, 2000, p. 6).

It has been proven that it is simplistic to say that a test can lead to all desired changes in teaching and learning. Education must be regarded as a complex element and there are many factors which can induce changes, like the school environment, messages from administration, expectations of teachers and students. Hence, Saif (2000) believes that an investigation of the needs and objectives of learners and educational systems should be implemented as a starting point for the research in washback. Wesche (1983, as cited in Bailey, 1996), maintains that when tests are the reflections of the situations, content and purpose where learners will use the language, they are likely to upgrade motivation.

Shohamy et al. (1996) regard factors like the state of the subject-matter tested, the nature of the test, and the application to which the test scores are applied. Wall (1996) offers a list of factors which might have hindered the examination in Sri Lanka from building an effective ‘lever for change’. These include micro factors such as teachers’ ones (lack of understanding of exam, the nature of the change desired, resistance to change, unfamiliarity with the test format and content) as well as more macro factors like gap between designers of test and teachers, lack of well trained teachers, overload of teachers, etc. Moreover, according to
Andrews et al. (2002), the new influence of a testing innovation is mediated by the teachers and how they interpret the innovation, which is different from what the conceivers of the test had in mind. Another factor can be the published materials which are being used (Andrews et al., 2002). Obviously, washback is a very complex notion. It includes the effect of an examination in the classroom, but also in the school, in the educational system and also in the society. Besides, this effect does not always happen directly but it entails a number of factors, like the teachers’ understanding of the test, the status of the test as well as that of the subject – matter tested, the macro – context where the examination is administered, the objective of learning the language in the context, among others. In addition, to study the washback effect, it is essential to examine the people that take part in the educational process, the actual classroom events and activities, and the outcomes of these processes.

2. Methodology

One of the most important findings is that washback is a complex phenomenon. It is really challenging to control all of the variables. According to Cheng (2004), aspects such as specificity, intensity, length, intentionality, and value have to be taken into consideration. Furthermore, the research to date, suggests various factors seem to mediate the process of washback. These factors, according to Cheng, include test factors, prestige factors, personal factors, micro-context factors, and macro-context factors. Given the complexity of washback, the purpose of the study is to investigate the impact of interview tests on high school students’ speaking ability and study habits in the Iranian context. The following research question was developed:

Does the interview test make any statistically significant difference between the experimental group and the control group in their speaking ability?

To answer the above research question, the researcher selected 60 students from a high school in Tehran, employed an interview test and a KET test, and adopted a quasi-experimental design.

2.1. Participants

Sixty students from Absal High School participated in this study. The researcher was allowed to carry out this study in this high school after she obtained permission from the Ministry of Education, District No. 4. The interview speaking test was administered to the participants who had already been divided into four classes by the school at the beginning of the educational year. The interview test showed that students in two classes enjoyed homogeneity in their speaking ability. The students in those two classes were chosen as the sample of this study.

To ensure their homogeneity in their language proficiency, the researcher administered a KET test to the students in the two classes. The results indicated that the participants of the two classes were homogeneous.
Generally, the participants in the current study had some characteristics. They had learned English for three years. Therefore, they had some linguistic background. They were about 16 years old. Their proficiency in English was very low, except the ones who had taken English classes. They were female students.

2.2. Instruments

To answer the research questions in the present study, the researcher used a KET test, and interview test questionnaire.

1. A KET test was administered to homogenize the participants in terms of their English language proficiency. This test was administered to the control and the experimental classes. Furthermore, the reliability of this test was calculated. The K-R21 reliability for the proficiency test was 0.91 as indicated in Table 1.

<table>
<thead>
<tr>
<th>K-R21 Reliability Indices of Proficiency Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homogeneity</td>
</tr>
<tr>
<td>-------------</td>
</tr>
<tr>
<td>Homogeneity</td>
</tr>
</tbody>
</table>

2. To measure the students’ speaking ability, a constructed validated interview test was designed on the basis of the students’ textbook in the first grade of high school. This test was based on the grammar points in the first five units of the book. This test was used as both the pre-test and the post-test in this study. The following procedures were taken in the construction of the interview test:
- A table of specification of the grammar points of the first five units of the first grade English textbook was prepared.
- Interview questions were constructed in such a way that they would elicit the grammar points in the first five units of the book.
- The questions were reviewed by some high school teachers. Their advice was taken into account. Necessary revisions were made.
- It was pretested with some similar high school students.
- Reliability was calculated. To ensure reliability, the researcher calculated the inter-rater reliability. The inter-rater reliability index for the two raters who rated the students on the pretest of speaking (r (58) = .88, P = .000 < .05). Table 2 indicates significant agreement between the two raters.
Table 2
**Inter-Rater Reliability**

<table>
<thead>
<tr>
<th></th>
<th>Pearson Correlation</th>
<th>PreR2</th>
<th>PostR2</th>
</tr>
</thead>
<tbody>
<tr>
<td>PreR1</td>
<td>Sig. (2-tailed)</td>
<td>.886</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>PostR1</td>
<td>Pearson Correlation</td>
<td></td>
<td>.878</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td></td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>60</td>
<td></td>
</tr>
</tbody>
</table>

**Correlation is significant at the 0.01 level (2-tailed).**

A factor analysis through varimax rotation was carried out to probe the underlying constructs of the questionnaire. The SPSS extracted three factors which account for 63.45 percent of the variance.

3. The teachers’ performances were videotaped and further analyzed. The researcher with prior arrangements attended their classes and videotaped their classes while they were teaching a grammar point.

4. An interview was conducted with the two teachers and several students participating in this study. The researcher aimed at reaching a more accurate finding in this study.

2.3. Procedures

This study, as already mentioned, aimed to investigate the effect of interview tests on the Iranian EFL learners’ speaking skill as well as their study habits. This study was completed through six stages, and it took four months to be completed.

1. Two experienced speaking raters rated the speaking ability of 120 freshman students (four classes) in Absal High School. The ratings were based on the analytic scale of speaking (Council of Europe, 2001, as cited in Luoma, 2004). As mentioned before, only accuracy was considered in rating the speaking ability.

2. Then, students in the two classes which had similar speaking ability were chosen as the participants of this study. To ensure their homogeneity in language proficiency, the researcher administered a KET test to both classes. Through statistical analysis, it was observed that the two classes were homogenous with regard to their overall English language proficiency.

3. For the teachers as well as the students in the two classes not to be aware that a study was being done, the principal of the high school convinced the teachers and students that this project was run by the Ministry of Education. Also, the teacher and the students in the experimental class were informed that 5 point from the final exam score would be allocated to their speaking test. This was done in order to inculcate the idea that speaking would account for the final exam score. All in all, the researcher did everything to eliminate the image of an experimental study in the teachers’ and the students’ mind.

4. As the students were passing through the first semester of their freshman year, the researcher did two things. First, the researcher provided the experimental class with a sample of speaking questions which were similar to the questions in the interview test. She
intended to give these students more awareness on the form of the speaking test. Of course she herself did not do so directly, she handed in the sample to the principal who also made sure that the copies were distributed to the students of the experimental class by the teacher. Secondly, she filmed the control and the experimental classes to catch any possible differences in the way they taught grammar.

5. Through coordination with the principal and the teachers, the post-test interview was conducted. The same raters rated the speaking ability of the two classes. The post-test interview involved the same questions as those of the pretest.

6. Interviews were conducted with the teachers and some students participating in this study to further investigate the effects of the interview test on teachers’ teaching and students’ learning.

2.4. Design

The quantitative nature of this study with some hypotheses formulated at the outset demanded that it be an experimental study (Mackey & Gass, 2005). In high schools in Iran, the students had been divided into classes by the school authorities, so the researcher could not randomly divide the students into classes. Therefore, the design of this study was quasi experimental with pretest and posttest design and intact groups.

Regarding the statistical analyses, an analysis of covariance (ANCOVA) was run to compare the experimental and comparison groups’ mean scores on posttest of speaking in order to probe the effect of final interview on the improvement of the speaking of the Iranian EFL learners while controlling for possible effects of their entry speaking ability as measured through pretest. ANCOVA has two main assumptions, i.e. homogeneity of regression slope and linear relationship between the covariate (pretest of speaking) and the dependent variable (posttest of speaking) (Mackey & Gass, 2005).

3. Results of the data analysis and discussion

In the following section, the results of this study will presented. First of all, the results of the proficiency test which was used to homogenize students with regard to their proficiency level will be presented. Second, the results concerning the individual hypothesis will be represented.

The researcher administered KET to assure the homogeneity of the two classes with regard to over language proficiency. An independent t-test was run to compare the experimental and control groups on the proficiency test in order to prove that they enjoyed the same level of general language proficiency prior to the main study. As displayed in Table 3 the mean scores for experimental and control groups on proficiency test were 35.70 and 36.81 respectively.
As indicated in Table 4, the results of the independent t-test (t (58) = .30, P = .763 > .05, r = .04 it represents a weak effect size) indicated that there was not any significant difference between experimental and control groups on proficiency test. Thus it could be concluded that the two groups enjoyed the same level of general language proficiency prior to the main study.

**Table 3**

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>experimental</td>
<td>33</td>
<td>35.70</td>
<td>12.778</td>
<td>2.224</td>
</tr>
<tr>
<td>Control</td>
<td>27</td>
<td>36.81</td>
<td>15.860</td>
<td>3.052</td>
</tr>
</tbody>
</table>

**Table 4**

<table>
<thead>
<tr>
<th>Levene’s Test for Equality of Variances</th>
<th>t-test for Equality of Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>Sig.</td>
</tr>
<tr>
<td>Equal variances assumed</td>
<td>2.184</td>
</tr>
<tr>
<td>Equal variances not assumed</td>
<td>.296</td>
</tr>
</tbody>
</table>

It should be noted that the assumption of homogeneity of variances is met (Levene’s F = 2.18, P = .145 > .05). That is why the first row of Table 4, i.e. “Equal variances assumed” is reported because t-test has an extra assumption called homogeneity of variances. That is to say the groups should enjoy homogeneous variances. The scores of a group should not be too spread while the scores of the other group are condensed. Levene has devised a test to probe this assumption.

As mentioned already, the main focus of this study was upon this hypothesis, which was put forward in the following way:

H0: The interview test does not make any statistically significant difference between the experimental group and the control group in their speaking ability.

The researcher administered an interview test to come up with data on the speaking ability of experimental and the control classes. For the data analysis, an analysis of covariance (ANCOVA) was run to compare the experimental and comparison groups’ mean scores on posttest of speaking in order to probe the effect of final interview on the improvement of the speaking of the Iranian EFL learners while controlling for possible effects of their entry speaking ability as measured through the pretest. ANCOVA has two main assumptions, i.e. homogeneity of the regression slope and linear relationship between the covariate (pretest of speaking) and the dependent variable (posttest of speaking). As displayed in graph 1, the
assumption of homogeneity of regression slope is met. Both the experimental (upper line) and the control (lower line) show the same regression slopes.

**Graph 1** Homogeneity of Regression Slope Posttest of CT by Groups with Pretest

The second assumption, i.e. linear relationship between the dependent variable and covariate is examined within the main table of ANCOVA results. The F-observed value for the effect of covariate (pretest of speaking) is significant (F (1, 57) = 66.91, P = .000 < .05; Partial η² = .54 it represents a large effect size) (Table 5). Based on these results, it could be concluded that there was a linear relationship between the dependent variable and covariate. Thus, the second assumption was also met.

**Table 5**

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Sum of Squares</th>
<th>Df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
<th>Partial Eta Squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pretest</td>
<td>1545.578</td>
<td>1</td>
<td>1545.578</td>
<td>66.915</td>
<td>.000</td>
<td>.540</td>
</tr>
<tr>
<td>Group</td>
<td>65.583</td>
<td>1</td>
<td>65.583</td>
<td>2.839</td>
<td>.097</td>
<td>.047</td>
</tr>
<tr>
<td>Error</td>
<td>1316.557</td>
<td>57</td>
<td>23.097</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>10327.000</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The assumption of homogeneity of variances was also met. As displayed in Table 6 the Levene’s F-value (F (1, 58) = 1.05, P = .309 > .05) was not significant.
Table 6  
*Homogeneity of Variances*

<table>
<thead>
<tr>
<th></th>
<th>F</th>
<th>df1</th>
<th>df2</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.052</td>
<td>1</td>
<td>58</td>
<td>.309</td>
<td></td>
</tr>
</tbody>
</table>

The F-observed value for the effect of the independent variable (experimental vs. control) was not significant (F (1, 57) = 2.83, P = .097 > .05; Partial $\eta^2 = .047$ it represents a weak to moderate effect size). Based on these results it could be concluded that there was not any significant difference between the mean scores of the experimental and control groups on the posttest of speaking after controlling for possible effect of their entry speaking ability as measured through the pretest of speaking. Thus, the null-hypothesis that there is no difference between the experimental group and the control group in their speaking ability as a result of the final interview test was supported.

As displayed in Table 7 the mean scores for experimental and control groups on posttest of speaking are 12.11 and 9.98 respectively.

Table 7  
*Descriptive Statistics Posttest of Speaking by Groups*

<table>
<thead>
<tr>
<th>Group</th>
<th>Mean</th>
<th>Std. Error</th>
<th>95% Confidence Interval</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>9.981</td>
<td>.931</td>
<td></td>
<td>8.118</td>
<td>11.845</td>
</tr>
<tr>
<td>Experimental</td>
<td>12.106</td>
<td>.841</td>
<td></td>
<td>10.422</td>
<td>13.790</td>
</tr>
</tbody>
</table>

As indicated in Graph 2, the speaking ability of the experimental group has improved, but this improvement was not significant enough to reject the null hypothesis.
The main purpose of the present study was to observe the effect of incorporating an interview test on the learners’ speaking ability. The results indicated that the interview test did not lead to significant progress in the students’ speaking ability. Although a communicative test was employed and five points was allocated, the students in the experimental class did not take considerable steps forward in their speaking ability. This can be sought in what Alderson and Wall (1993) argue. They believe that if a test does not have the desired washback, this is not necessarily due to a lack of validity of the test.

Despite the communicative quality of a test, this can then hardly be attributed to a problem with the test, whereas validity is a property of a test. Alderson and Wall (1993, p.118) argue that washback, if it exists, is probably a complex feature which cannot be associated directly to test validity. According to Alderson and Wall, the issue of washback or systematic validity, “is at best premature and at worst ill-conceived” (p.118). According to the washback hypothesis, any test, whether it is good or bad, can have beneficial washback if it increases motivation, facilitative anxiety, and makes teachers and learners do good things they would not otherwise do, that is, the term ‘washback’ is itself ‘a neutral one’ (Alderson, 2004, p. xi). Thus the quality of washback might not be related to the quality of the test (Alderson, 2004). Therefore, although the test was communicative, and it was of a good quality, it failed to leave a positive washback effect.

The finding of this study that the incorporation of an interview test does not have any effects on the students’ speaking ability is conjunction with Kohn’s idea (2002, as cited in Cheng, 2004) that a good test does not necessarily guarantee to leave positive washback effects. Kohn believes that teachers do not resist change, but they resist being changed. According to Kohn, a teacher is charge of any reform, irrespective of any reform. He can shut the door and do whatever he desires to do. Cheng (2004) believes that the relationship between testing and teaching and learning tends to be very sophisticated. He maintains that there are many intervening influences in an educational context that play an important role in the effect that a test might leave. Washback is a complex phenomenon, so although a test might have validity, it might not a positive washback effect because it has not made teachers and students do good things.

The interview with the teachers revealed good things in the present study. The researcher has come up with a good finding in the current research. If a good test is designed to leave a positive effect, teachers should have the necessary skills demanded by the test. In other words, the teacher’s training would be an obstacle in the way of positive washback effects. Teachers’ professionalism affects the influence of tests on learning and teaching. This finding is in accordance with Madaus (1988) assertion that teachers will focus on the skills which appear in the test when their professional value is judged based on their students’ success in the exam. Teachers’ education and training are of great importance. Factors such as the teacher’s own education and educational experience (Watanabe, 2000), the amount of general methodological training teachers have received (Andrews, 1994 and), their training in teaching towards specific exams and in how to use exam related textbooks (Wall & Alderson, 1993),
their access to and familiarity with exam support materials such as exam specifications, and their understanding of the exam’s rationale or philosophy are regarded as significant. Watanabe (1996) believes that teachers’ educational background, to some extent affects their perception of tests. He believes that teachers’ factors are more important than the influence of examinations.

Moreover, Khaniya (1990) states that “a large number of teachers help students cope with examinations in order to preserve their reputation as good teachers” (p.51). Furthermore, Fullan and Stiegelbauer (1991) believe that the subjective reality of the teacher concerning a new test is always different from the objective reality that the proponents of change originally imagined. He believes that teachers work on their own, without considering what experts and other colleagues say. They are forced to accomplish a lot, but they are given little time to achieve their goals. The interview with the teacher of the experimental group in this study also helped to reveal that the fact that washback required much more time to emerge. She pointed out that the time was not sufficient, so they needed more time to work on the students’ speaking ability. She also said that administering a speaking test only once does not have enough force to make improvement.

The finding of this study indicates that if one intends to design a test which aims at positive washback effects, he should embark upon eliminating obstacles such as teachers’ experience and skills in teaching. The teacher of the experimental class, despite graduated in language teaching from an accredited university in Iran, was not prepared to teach to the test, which was found by the videotaped shot as well as the interview with her. In other words, there would be a positive washback effect for a good test if the teacher is well-trained.

The researcher, through interview with teachers and students participating in this study, has come up with the conclusion that a test cannot be definitely used as a leverage to instigate change or improvement. A test can be designed to have positive washback effects in case everything is well-prepared. For instance, teachers, students, the material, and authorities should be ready to accommodate a new test.

The videotaping of the class indicated that there was a mismatch between the activities done in the class and what was demanded of the test. The activities done by the students and the teacher’s performance bore little resemblance, if ever, to the features of the test. The researcher believes that if washback is going to be positive in a washback study, the teacher should be not only guided and trained, but also provided with activities which meet the objectives of the course. In other words, if the syllabus is not in conjunction with the content of the test, it should be modified by the teacher. Therefore, this research can further put a stamp of approval on the idea that the test should be in line with the syllabus if it is going to have positive effects.

Generally, the researcher believes that the interview test could not produce positive results in speaking ability of the high school students because of the following reasons:
1. The teacher was not well-trained for the test. This was shown by the interview with the teacher of the experimental class.
2. The students did not take the test seriously because they observed that the other classes of the same grade did not have such a test at the end.
3. Furthermore, the students pointed out the teachers did not work on the sample test quite well and asked them to work on it at home.
4. For a new test to be quite and seriously attended by teachers and students, more time is needed. Also, more forces are needed. Just the principal of the school cannot implement a change.

4. Conclusion

This study had been concluded with this finding: the interview achievement test might not positively affect the speaking ability of high school students in Iran. This finding of this study had both theoretical and practical implications in the area of applied linguistics. The researchers in the field of language testing and assessment should know that an array of factors needs to be taken into consideration in the implementation of a new test, the most important of which is the teacher. Teachers are the axis on which many developments rotate. Because of this the teacher suggests the following triangle to better display the relationship between test, learning and teaching, and teacher.

Teacher

Test                   Learning and Teaching

In other words, it should be born in mind that the relationship between test and learning is mediated by the teacher in the classroom.
References


Assessing content area explanations, Los Angeles: University of California, CRESST.


Baker, E. L. (1989). *Can we fairly measure the quality of education*? Center for Study of


C.K., Johnson, K. (Eds.), *The communicative approach to language teaching* (pp. 143- 159). Oxford: Oxford University Press.
Cambridge, Deighton: Bell and Company.
Cambridge: CUP.

Chapman, D. & Snyder, C. (2000). Can high stakes national testing improve instruction:

Chen, L. (2002a). *Taiwanese junior high school English teachers’ perceptions of the


Cheng, L. (2000). *Washback or backwash: A review of the impact of testing on


Entrance examinations on instruction in pre-college level EFL. Language Testing Update, 27, 42–47.

ERIC Number: ED 472167.

Erlbaum.


for Study of Evaluation.


Herman, J. L. (1989). *Priorities of educational testing and evaluation: The testimony of the 

Herman, J. L. (1992). *Accountability and alternative assessment: Research and development 

Herman, J. L., & Golan, S. (1991). *Effects of standardized testing on teachers and 

Hong Kong: University of Hong Kong.


Hughes, A. (1993). *Backwash and TOEFL 2000*. Unpublished manuscript, University of


issues, CSE Technical Report 384, Los Angeles: University of California, CRESST.


Berry (Eds.), *Bringing about change in language education* (pp. 83-99), Hong Kong: 

University

Language and Education, 11, 38-54.

language testing: Research contexts and methods. (pp. 3-17). Mahwah, N. J.: Lawrence Erlbaum.


Messick, S. (1994). The interplay of evidence and consequences in the validation of


don teaching and learning. In L.

Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in over time.* Language Testing, 13, 298-317.


Reading, UK.


Researcher, 20 (2), 21–27.

Resources Information Center on Assessment and Evaluation (ERIC/AE)


Shepard, L. A. (1990). Inflated test score gains: is the problem old norms or teaching the test?


University of California. Center for Research on Evaluation, Standards and
University Press.
education and from innovation theory. Language Testing, 13(3), 334-354.

(Eds.), Encyclopedia of language and education: Vol. 7. Language testing and
assessment (pp.

education

Wall, D., (2000). The impact of high-stakes testing on teaching and learning: can this be
washback effect of the basic competence test in English. Unpublished Doctoral
washback effect of the Nepalese English examinations. Unpublished doctoral dissertation,
University of Edinburg, UK.

Watanabe, Y. (1996). Does grammar translation come from the entrance examination?


Press, Manchester.