Journal of Applied Linguistics and Language Research
Volume 2, Issue 2, 2015, pp. 80-92
Available online at www.jallr.ir
ISSN: 2376-760X

The Effects of the Frequency of TOEFL iBT as Quizzes on Real-life Reading Comprehension Tasks: The Discourse in Focus

Ahmad Reza Beigi Rizi
PhD Candidate, University of Isfahan, Iran

Mansoor Tavakoli
Associate Professor, University of Isfahan, Iran

Abstract
Washback has become an increasingly prevalent and prominent phenomenon in education - what is assessed becomes what is valued, which becomes what is taught. Although many researchers suggest frequent tests as a means of positive washback, others oppose this idea. But what is clear is that many scholars have attempted to provide guidelines in order to achieve positive washback. The present study tried to investigate the effects of the frequency of TOEFL iBT as Quizzes on students' real-life L2 reading comprehension tasks. The participants of this study were 201 intermediate Iranian language students who were randomly selected and divided into three groups. The first group received 10 TOEFL iBT Reading Comprehension tests, i.e. one Reading Comprehension test for every unit of their textbook every session. The second group received five TOEFL iBT Reading Comprehension tests, i.e. one test for every two units of the same textbook every other session. The third group received no tests at all. The performance of the subjects showed that frequent TOEFL iBT Reading Comprehension tests had positive effect on learning. Better performance of the group who received fewer tests revealed that although giving tests was associated with better performance, the amount of improvement (t-observed) was reduced from 14.4 to 11.26 as the number of tests was increased.
Keywords: washback, TOEFL iBT, quiz, reading comprehension tasks, corrective feedback, discourse

BACKGROUND

One of the concerns of TOEFL instructors has been evaluating students' progress during a course and their language achievement at the end of the course (Harris & McCann, 1994, p.26). Researchers believe that measurement provides teachers with the necessary quantitative information about their students' language ability and enables them to make professional judgments within the context of their classes (Bachman & Palmer, 2010).

Correspondence: Ahmad Reza Beigi Rizi, Email: arbeigi@fgn.ui.ac.ir

The most popular use of educational tests in general and language tests in particular, is to identify relative strengths and weaknesses of individual candidates in a given learning context (Sawaki & Sinharay, 2013). Classroom achievement tests are the most common types of language tests from which both learners and teachers could benefit since such tests are helpful not only in providing feedback but also in providing learners with valuable practice and learning opportunities. In fact, such achievement tests provide teachers with valuable information based on which they can assess and evaluate their students' progress toward course objectives and diagnose their areas of difficulty (Spratt, 2005).

Although most of the researchers in the field of TEFL consider testing as an important part of teaching/learning activity (Harris & McCann, 1994), there does not seem to be an agreement on the repeated use of tests or quizzes (Vernon, 1956: 166). Some scholars argue that more frequent testing would increase instructional effectiveness and would encourage students to study and review more often (Morris, 1972). On the other hand, some scholars believe that frequent testing do not help students enough because teachers put their focus only on the tests and teach to the test, providing their students only with the amount of information they need to do well on the tests. Because teachers teach to the test and students read to the test, learning does not last for a long time Marshall (2007).

Therefore the issue with which this study is concerned with is the effects of the frequency of TOEFL iBT as quizzes on real-life Reading Comprehension Tasks. Because this study tries to investigate the influence of testing on learning therefore this study uses the term "washback" according to Fulcher & Davidson (2007, p. 221) to refers to the extent to which the a test influences language learners performance. In this study the real-life reading comprehension tasks are the selected reading sections of a TOEFL test which have been used as a valid test of English language proficiency to measure the subjects' reading ability with the selected skills (one, two, three, twelve & thirteen) of Phillips (2001, pp. 368-442) at the end of the course. This test that served as the post-test consisted of six passages and thirty multiple-choice items. The purpose of this study is to answer that whether administering TOEFL iBT as quizzes results in better performance on real-life Reading Comprehension Tasks.

LITERATURE REVIEW

According to Fulcher and Davidson (2007) "Washback is generally defined as the influence of testing on teaching and learning" (Bailey, 1996, p. 259). The concept of washback is therefore part of what Messick (1989) calls consequential validity. As part of consequential validity, Messick (1996, p. 241) says that: Washback refers to the extent to which the introduction and use of a test influences language teachers and learners to do things that they would not otherwise do that promote or inhibit language learning". (Cited in Fulcher & Davidson, 2007, p.221)

The developers of TOEFL preparation courses around the world have many aims in developing the courses they either tend to supplement regular language classes, or sometimes to replace the regular language classes with TOEFL preparation courses (Alderson & Hamp-Lyons, 1996). The iBT test developers nowadays have many aims in developing the test such as "to maximize the positive consequences of score use" ETS (2008, p.9) and "to create positive washback in TOEFL preparation classrooms through integrated-skills tasks and a speaking test, in the hopes that the emphasis in classroom teaching will shift and [test preparation] courses will more closely resemble communicatively oriented academic English courses" (Reynolds 2010, p. 1).

In this article we have focused on the effect of frequent iBT tests on the improvement of students' reading ability. Testing has many central goals one of which is to provide goals for language teaching and it monitors for both teachers and learners success in reaching the goals (Salim 2007, p.173). Along the same lines, Fulcher & Davidson (2007) believe the focus of washback study has been on those things that we do in the classroom because of the test, but 'would not otherwise do' (p.221). They have also mentioned and focused that washback has different meanings so they have commented that "the concept of washback is to have *any meaning*, it is necessary to identify what changes in learning or teaching can be directly attributed to the use of the test in that context" (p.221). Muñoz & Álvarez (2010) also seem to believe in the different meanings of washback in different contexts. According to them "The majority of washback-intended studies have concentrated on the positive or negative effects of high-stakes examinations on such areas as course content, teachers' methodology, teacher and student attitudes, and learning". (Muñoz & Álvarez 2010, p.35)

Cheng & et al., (2004) also believe that "it is feasible and desirable to bring about beneficial changes in teaching by changing examinations, representing the "positive washback" scenario, which is closely, related to "measurement-driven instruction" in general education." Cheng & et al., (2004, p. 10)

Bachman (1991) demonstrated that the components of language ability included in the test correspond to those covered in the course and that the characteristics of the test tasks correspond to the types of classroom learning activities included in the program (p. 681). What we can infer from Bachman (1991, p. 681) is that if the assessment procedures correspond to the course goals and objectives of an EFL classroom or curriculum, a positive washback effect occurs and if the assessment procedures in an EFL classroom or curriculum do not correspond to its goals and objectives, the tests are likely to create a negative washback effect on those objectives and on the curriculum.

Cheng et al., (2004) have also shown that "Test washback does not always correspond to the effects intended by the inspectorate. As a means for curriculum innovation and implementation, *washback may have some predictable effects* (p.207). Language teaching centers which offer TOEFL preparation courses have often tried to increase learners' achievement. Typical examples are "Wiseman (1961) who believed that paid coaching classes, which were intended for preparing students for exams, were not a good use of the time, because students were practicing exam techniques rather than language learning activities (p. 159), and Davies (1968) believed that testing devices had become teaching devices; that teaching and learning was effectively being directed to past examination papers, making the educational experience narrow and uninteresting (p. 125)." (Cited in Cheng & et al., 2004, p. 9). But it will be very beneficial for TOEFL preparation courses and EFL teachers to improve their language testing skills in order to be able to make appropriate use of TOEFL preparation classroom tests. The importance of the value that TOEFL preparation classroom tests may have in improving teaching and learning will enable EFL teachers and curriculum developers to assign appropriate focus on these tests as practice and learning opportunities.

Many researchers have studied the effect of quizzes on students' performance among some of the recent ones are Ballard and Johnson (2004), Roediger and Karpicke (2006), Marshall (2007), Zarei (2008), Marcell (2008), Johnsom & Kiviniemi (2009), Hashtroudi (2001) and Gholami & Moghaddam (2013). According to Gholami & Moghaddam (2013) who studied the effect of weekly quizzes on students' final achievement score, the performance of the weekly quiz group was significantly better than that of the control group and the reasons behind the success of weekly quizzes may be attributed to class attendance (p.39) and extrinsic motivation (p.40).

In order to become familiar with some implications of the effects of TOEFL quiz frequency on real-life reading comprehension task the method through which the study was carried out and the results obtained will be explained. The study was guided by the following research question:

- What are the effects of TOEFL iBT Quiz frequency on real-life Reading Comprehension Task?

METHOD

Participants

For this study 201 Iranian upper-intermediate students studying English at an English Language institute in Isfahan province of Iran were chosen. They were randomly selected from students of six different classes and were divided into three groups. Each group consisted of two classes to provide sufficient number of subjects for the study. Two classes with a total of 73 students served as the first experimental group, two classes with a total of 76 students served as the second experimental group, and the other two classes with a total of 52 students served as the control group. The first experimental group received 10 quizzes during the 13 sessions of instruction i.e. one quiz for every unit they were taught. The second experimental group received five quizzes i.e., one quiz for every two units of the same textbook and the third group which served as the control group didn't receive any quiz.

Instrumentation

Quick Placement Test

Besides considering these proficiency levels based on the institute' criteria, a proficiency test was administrated to screen the subjects and homogenize them based on their levels of proficiency.

Course book

The course book used in first & second experimental group was "Phillips D. (2001). Longman Complete Course for the TOEFL Test. Preparation for the Computer and Paper Tests. Addison- Wesley Longman, Inc. A Pearson Education Company." This course book was used because it was being taught by the institute and the teachers were familiar with it.

Tests

The treatment was given in the TOEFL preparation courses. Therefore, the instruments for the research procedures focused on reading comprehension and consisted of the following kinds of tests:

Pretest: The original form of the TOEFL Reading Comprehension Test by Phillips (2001) was used to assess and compare the homogeneity of the subjects with regard to their English language proficiency in general, and reading Comprehension ability in particular. This test that served as a pretest contained seven passages with thirty-four, four choice items. The reliability of the test was .94 computed through the KR-21 formula. Special care was taken to choose the most appropriate reading test, the readability, which was geared to the students' level as well as the passages in their textbook. The average readability for the pretest test was 21.9 and its SD was 14.19 and the average readability of the textbooks turned out to be 23.8 that fell within +1SD of the readability indices of the reading comprehension passages in the pretest.

Quizzes: A series of short quizzes served as the independent variable. The quizzes were carefully prepared for the purpose of the present study. In the preparation of the quizzes certain important points were taken into account.

First, each quiz contained two passages of approximately 250-400 words; each followed by 10 items. Five of the 10 items were designed to check the ability of the students on "answering main idea questions correctly"; "recognizing the organization of ideas", "answering stated detail questions correctly" based on TOEFL reading comprehension skills one, two & three of Phillips (2001, pp.368 to 384) and the other five were designed to determine the "tone", "purpose", or "course" of the passages and "where specific information is found" in the passages based on TOEFL reading comprehension skills twelve & thirteen of Phillips (2001, pp.431 to 440). That is, every quiz contained 20 items. Second 20 minutes at the beginning or at the end of each class session was allocated to each quiz. Third the difficulty level of the passages used in the quizzes was adjusted to the difficulty level of the texts used in their textbooks. Forth the items of the quizzes were all of multiple-choice type.

Post-test: The selected reading sections of a TOEFL test was used as a valid test of English language proficiency to measure the subjects' reading ability with the aforementioned skills (one, two, three, twelve & thirteen) of Phillips (2001) at the end of the course. This test that served as the post-test consisted of six passages and thirty multiple-choice items. The reliability of the TOEFL test, computed through the KR-21 formula revealed to be .98. The average readability of the passages turned out to be 24.8 that fell within +1SD of the readability of their textbook.

Procedure

First, at the beginning of the TOEFL preparation course, the Pretest was administered to all three groups, two experimental and one control group to ensure the homogeneity of the participants in the study. Then the homogeneity of the instructional material, course objectives, whole-term syllabus and even the daily lesson plans were strictly controlled in order to increase the precision of the results and to control as many extraneous factors as possible. Later the experimental treatment was conducted (Table 1); because repeated measurements were to serve as the independent variable, the first experimental group received a quiz every session for ten weeks i.e. one quiz for each lesson they were taught during the course. The second experimental group received a combination of the first and second quiz given to the first experimental group every other session i.e. the quiz consisted of one passage from the first quiz and one passage from the second quiz, which contained the same number of items. Each quiz was carefully scored and returned to the students the next session. The control group on the other hand, was not given any type of tests or quizzes. The course consisted of 13 class sessions and they were taught 10 units of their textbook. As a result, the first experimental group received 10 quizzes and the second experimental group received five quizzes. Finally, at the end of the TOEFL preparation course, in order to investigate the impact of the experimental treatment and to determine the relationship between the independent variable (i.e. number of quizzes) and the dependent variable (i.e. students' reading ability) the post test was administered. This test, which served to compare the experimental and control groups' performance, was administered to all three groups during the same week.

Table 1. Sessions, quizzes, pretests and posttests

	Pretest	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	Posttest
G1	✓			Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10		✓
G2	✓			Q1		Q2		Q3		Q4		Q5			✓
Ge	✓														✓

Notes: S = Session, Q = quiz. Second 20 minutes at the beginning or at the end of each class session was allocated to each quiz. All groups had the same number of sessions.

Data Analysis

First the experimental and control groups' performance on the Pretest was compared through one-way ANOVA through (IBM SPSS Statistics version 11) in order to determine the homogeneity of the subjects with respect to their reading ability. The mean scores for the first experimental, second experimental, and control group were 55.78, 55.5, and 54.36, respectively. In this analysis, in order to compare the experimental and control groups achievement at the end of the course, the original form of a TOEFL was administered and the one-way ANOVA was utilized to compare the obtained adjusted means which were 63.63, 66.84, and 54.74 for the first experimental, second experimental and control group, respectively. In this study one-way ANOVA was used because it enables us to compare the means of more than two groups on one dependable variable. Finally, in order to compare the superiority of the two experimental groups the technique of matched t-test was utilized because we had to compare two means obtained from two independent groups of students (the two experimental groups).

RESULTS AND DISCUSSION

Results of the Pretest Stage

In order to show that there was no significance difference in the reading ability of the subjects in the experimental and control groups before providing any treatment, they were pretested through a TOEFL Reading Comprehension. There were a total of 201 subjects in all three groups at this stage. The mean scores of both experimental groups and that of the control group were compared through one-way ANOVA. The mean and the standard deviation of the means are provided in Table 2.

Table 2 one-way ANOVA for comparing the performance of the three groups on the reading test at the pretest stage

Source of Variation	SS	df	MS	F
Between Groups	228.34	2	114.17	.5525
Within Groups	40506.9	196	206.66	
Total	40735.24	198		

Note: SS = Sum of squares, d.f = degrees of freedom, M.S = mean square, F-ratio

The results indicate no significant difference in terms of the reading ability of the subjects in the three groups at the beginning of the study since the F-observed was lower than the F-critical. Thus it could be concluded that the three groups met the condition of homogeneity.

Results of the Post-test Stage

The treatment was completely carried out after 10 weeks and the post-test that was an original form of a TOEFL was administered. The results presented in Table 3 reveal that the treatment has been effective, since the F-observed exceeds the F-critical, i.e. the experimental groups have shown significantly better performance than the control group.

Table 3 one-way ANOVA for comparing the performance of the three groups on the TOEFL reading test at the post-test stage

Source of Variation	SS	df	MS	F
Between Groups	5669.05	2	2834.52	
Within Groups	28859.34	195	147.9966	19.15
Total	3452.39	197		

Note: SS = Sum of squares, d.f = degrees of freedom, M.S = mean square, F-ratio

The mean score of the first experimental group which received 10 quizzes as the treatment changed from 55.78 on the pretest to 63.63 on the post-test, i.e. they showed 7.85 points of improvement. The mean score of the second experimental group which received five quizzes as the treatment changed from 55.5 to 66.84, i.e. 11.34 points of improvement was observed, and that of the control group changed from 54.36 to 54.74 showing 0.38 points of improvement. Consequently, the performance of each group on the pretest with that of the same group on the post-test was compared through matched t-test to investigate the priority of the experimental groups. The results are provided in table 4.

Table 4 Matched t-test for comparing the performance of each group on the pretest and post-test stages

Group	n	X Pre	X Post	df	SD	t-observed
Ex. G1-10Qz	73	55.78	63.63	72	7.2	11.26
Ex. G2-5Qz	76	55.5	66.84	75	7.04	14.04
G c	52	54.36	54.74	51	5.23	.53

The results in this table indicate that the reading ability of both experimental groups improved significantly from the pretest to the post-test, since the observed t-value for both groups were much higher than the t-critical as opposed to that of the control group in which no significant improvement was observed.

However, the improvement in the second experimental group who received five quizzes as the treatment was more than that of the first experimental group who received ten quizzes as a treatment. This can be implied from the t-values and supports the hypothesis that practice, or the frequent use of quizzes is effective to a certain extent, and reveals that better performance has been associated with the use of frequent quizzes, but the amount of improvement has diminished as the number of quizzes increased.

To justify the lack of improvement in the control group, it is worth mentioning that the readability index of the pretest was 22.3 as opposed to that of the post- test which was 25.7, i.e. the post-test showed 3.4 readability index greater than that of the pretest. Since the results of the of the matched t-test showed no significant difference between the control group's performance on the pre and post-tests, equal performance on a test with a higher readability index can show improvement to some extent.

Investigating the Scores on Quizzes

As mentioned before the first experimental group received 10 quizzes, one as for each unit of their textbook and the second experimental group received five quizzes i.e. one quiz for every two units. The question posed here is whether subjects showed improvement in comparison to the previous quiz they received each time. For this reason the means of the quizzes have been compared and no significant improvement is viewed as for each subsequent quiz. This is due to the characteristics of the quizzes. As mentioned the difficulty level of each quiz would increase just as would the difficulty level of the units in their textbook. It seems logical for the subjects to perform equally well on a more difficult quiz after receiving instruction geared to the same level. Tables 5 and 6 in Appendix present the mean scores of individual quizzes of the first and second experimental groups respectively.

CONCLUSION

The analyzed data showed that repeated quizzes led to significantly better performance of the participants. Furthermore, the higher performance of the experimental group who received fewer quizzes revealed that repeated measurements are more effective to a certain frequency. As a result, it can be concluded that quizzes have had positive effect on students' learning and teachers' instruction. This may be due to certain factors, some of which are mentioned below:

In this study, the quizzes were of appropriate difficulty. It was announced in advance and was based on the TOEFL preparation course objectives. So, the subjects had a better chance to become more acquainted with course objectives and areas of emphasis and probably benefited from the constructive role of such quizzes in providing feedback and improving motivation.

The most popular use of educational tests in general and language tests in particular, is to identify relative strengths and weaknesses of individual candidates in a given learning context (Sawaki & Sinharay 2013). In this research each student identifies his/ her relative strengths and weaknesses and each quiz acts as an *activator* for the next quiz. The students also transferred many learnt elements to the next quiz such as their experiences, judgments, strategies and performance feedback, the list is hopefully endless. Each quiz helps the students to make professional judgments for the context of the next quiz. Each quiz also provides students with valuable corrective feedback information based on which they can assess and evaluate their progress toward course objectives and diagnose their areas of difficulty.

This parametric experimental study manipulated the levels of independent variable of TOEFL iBT Quiz frequency on the dependent variable which is real-life reading comprehension task and only selected skills based on Phillips (2001, pp. 368-440) such as “answering main idea questions correctly”, “recognizing the organization of ideas”, “answering stated detail questions correctly”, the “tone”, “purpose”, or “course” of the passages and “where specific information is found”, were in focus. This kind of focus is probably too general in this decade (2010s) and more research is needed to be performed for the backwash effects on each subskill more specifically. One possibility of the students’ success in transferring valuable corrective feedback information based on which they can assess and evaluate their progress toward course objectives and diagnose their areas of difficulty is the familiarity with the needed discourse applied in the *questions* and related to the stated or implied discourse in the reading comprehension texts. What we might follow is to focus on corrective feedback elements or attributes of the discourse which are transferred from one quiz to another so that the phenomenon of washback effects appear in such selected reading comprehension skills. Corrective feedback which seems to be a complex system comes in to being due to certain conditions Beigi Rizi & Ketabi (2015, p. 73). One of such conditions is the repetition of such discourses in different contexts. We also need to have control over the levels or conditions of at least one of the ‘corrective feedback elements or attributes’ to which a subject is exposed to after each quiz by determining what the levels are, how they are implemented, and how and when such washback effects are assigned and exposed to them. The power of discourse Fairclough (1989) found in reading comprehension questions and the texts cannot probably be underestimated when scrutinizing the effects of washback in reading comprehension. This means that separate studies are needed to explore various dimensions of the relations of the power of discourse and language on washback in reading comprehension.

As is the case with most quizzes, since the discourse of the items which were on the quizzes were similar to those of their post-tests, students of the experimental groups had a better chance to become more acquainted with such discourses. This may have reduced the experimental subjects’ anxiety of the new discourses during the post-test and consequently have improved their performance and the risks of experiencing new discourses have subsided.

In an attempt to become prepared for the quizzes, the students in the experimental groups probably had to review the material (discourses) more attentively. Therefore, repeated preparations for the quizzes may have improved the subjects’ familiarity with the discourses in the experimental groups. Furthermore, the students may have also benefited from the reviewed discourses in class discussions, after class discussions with peers or self-mental evaluations of the discourse regarding the corrected papers in subsequent sessions that pinpointed problematic areas of individual students.

The mentioned factors all contribute to better performance of the experimental groups as opposed to that of the control group. The controversial issue in this study lies in the higher performance of the experimental group who received fewer quizzes. Since all groups as mentioned were homogenized based on their levels of proficiency one reason for better performance of the experimental group who received fewer quizzes, according to table 1, can probably be that the second group had more instruction because part of the class time was set for applying the quizzes in the classes so the second group received more instruction based on the quiz results and the course instructions. They probably had more time and more *before test preparation* instructions before each quiz. The time and before test preparation instructions before each quiz was less in the first group from Quiz 2 to Quiz 10 in comparison to group 2 because in group 1 they had one quiz each session from session 3 to session 12 and probably focused more on the discourse of the questions and the texts.

Backwash effects have reached an exciting stage in its development. As the researchers increase our connection with other branches of science such as discourse analysis, psychology, pedagogy, language testing etc. we continue to push the field forward, uncovering new insights and helping both researchers and practitioners reach a better understanding of the dynamic, socially situated, and cognitive processes of the effects of washback.

REFERENCES

Alderson, J. C. (2004). Foreword. In Cheng, L., Watanabe, Y. and Curtis, A. (eds) *Washback in Language Testing: Research Contexts and Methods*. Mahwah, NJ: Erlbaum.

Alderson, J. C. and Wall, D. (1993). Does washback exist? *Applied Linguistics,* 14, 2, 115–129

Alderson, J., & Hamp-Lyons, L. (1996).TOEFL preparation courses: A study of washback. *Language Testing, 14*(2), 280-297.

Bachman L. (1991). What Does Language Testing Have to Offer? *TESOL QUARTERLY*, *25*(4), University of California, Los Angeles

Bachman, L. F. & Palmer, A. S. (2010). *Language Assessment in Practice*. Oxford: Oxford University Press.

Bailey K. (1999). *Washback in Language Testing*. TOEFL Monograph Series. Educational Testing Service. Princeton, New Jersey. RM-99-4

Bailey, K. M. (1996) 'Working for washback: a review of the washback concept in language testing.' *Language Testing, 13(*3), 257–279.

Ballard, C. L. & Johnson, M. F. (2004). Basic Math Skills and Performance in an Introductory Economics Class. *Journal of Economic Education*, 35(1), 3- 24.

Beigi Rizi, A. R. & Ketabi, S. (2015). A Close Look at Sixty Years of Corrective Feedback. *Journal of Applied Linguistics and Language Research*, *2*(1), 2015, pp. 63-77.

Cheng L., Watanabe Y., Curtis A. (2004). *Washback in Language Testing Research Contexts and Methods*. Lawrence Erlbaum Associates, Publishers.

Educational Testing Service. (2008). *Validity evidence supporting the interpretation and use of TOEFL iBT scores*. Princeton, NJ: Author.

Fairclough, N. (1989) *Language and Power*. Harlow: Longman.

Fulcher G. and Davidson F. (2007). *Language Testing and Assessment An advanced resource book*. Routledge Taylor & Francis Group.

Gholami V., Moghaddam M. (2013). The Effect of Weekly Quizzes on Students' Final Achievement Score. *IJMECS*, *5*(1), 36-41.

Harris M. and McCann P. (1994). *Handbooks for the English Classroom Assessment*. Macmillan Heinemann English Language Teaching. ISBN 0 435 28252 2.

Hashtroudi, F. P (2001). *The washback effect of frequent quizzes on EFL learners' grammatical ability* (Master's thesis). Retrieved from http://idochp2.irandoc.ac.ir/fulltextmanager/fulltext15/TH/40/40015.pdf

Johnson, B. C. Kiviniemi, M. T. (2009). *The effect of online chapter Quizzes on Exam performance in an undergraduate Social psychology course*. Teach Psychology, 36(1), 33- 37.

Marcell, M. (2008). Effectiveness of regular online quizzing in increasing class participation and preparation. *International Journal for the Scholarship of Teaching and Learning*, *2*(1), 1- 9.

Marshall, B. (2007). A crisis for efficacy? *Education Review*, 20(1), 29- 35.

McEwen, N. (1995a). Educational accountability in Alberta. *Canadian Journal of Education*, 20, 27–44.

Messick, S. (1989). *Validity*. In Linn, R. L. (ed.) Educational Measurement. New York: Macmillan/American Council on Education, 13–103.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*, 241–256.

Morris, B. (1972). *Objectives and Perspectives in Education: Studies in Educational Theories*. London: Routledge and Kegan Paul.

Muñoz A. and Álvarez M. (2010). Washback of an oral assessment system in the EFL classroom. *Language Testing*. DOI: 10.1177/0265532209347148. Retrieved from http://ltj.sagepub.com/content/27/1/33

Phillips D. (2001). *Longman Complete Course For the TOEFL Test*. Preparation for the Computer and Paper Tests. Addison- Wesley Longman, Inc. A Pearson Education Company.

Reynolds, J (2010). *An exploratory study of TOEFL students as evaluators of washback to the learners*. A thesis Submitted in Partial Fulfilment of the Requirements for the Master of Applied Linguistics, (TESOL Studies), The University of Queensland

Roediger, H. L. & Karpicke, J. D. (2006). Test enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249-255.

Salim (2007). *A Companion to Teaching of English*. Atlantic Publishers & Distributors.

Sawaki Y. & Sinharay S. (2013). Investigating the Value of Section Scores for the TOEFL iBT® Test. TOEFL iBT ® *Research Report TOEFL* iBT–21.

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

Vernon, P. E. (1956) *The Measurement of Abilities*. 2nd ed. London: University of London Press.

Wall D. & Horak T. (2006) *The TOEFL Impact Study*. The Centre for Research in Language Education (CRILE) Seminars 2005 - 2006, the Department of Linguistics and English Language at Lancaster Univeristy. Retrieved from http://www.ling.lancs.ac.uk/groups/crile/seminar_series/sem0506.htm

Zarei, A. A. (2008). On the Learnability of three categories of Idioms by Iranian EFL learners. *Journal of Humanities of the University of Kerman*, *2*(2), 82- 100.

APPENDIX

Table 5. Mean score of the first experimental group on the quizzes

Variable	Mean	Std Dev.
Quiz 01	11.21	2.45
Quiz 02	13.44	2.71
Quiz 03	12.88	2.13
Quiz 04	12.93	1.87
Quiz 05	14.1	1.98
Quiz 06	11.86	2.11
Quiz 07	12.87	2.33
Quiz 08	13.03	2.49
Quiz 09	12.64	2.09
Quiz 10	13.44	2.35

Table 6. Mean scores of the second experimental group on the quizzes

Variable	Mean	Std Dev.
Quiz 01	11.20	2.61
Quiz 02	12.56	2.54
Quiz 03	12.54	2.13
Quiz 04	12.83	1.98
Quiz 05	13.45	2.0