

The Effect of Various Testing Conditions on Long-Term Retention of Reading Materials: The Case of Initial and Delayed Test Types, and Feedback on Test

Mahmoud Qaracholloo *

PhD candidate of TEFL at Allameh Tabataba'i University of Tehran

Mohammad Reza Ghorbani

Assistant professor of TESL at University of Bojnord

Farhad Ghiasvand

MA student of TEFL at Allameh Tabataba'i University of Tehran

Abstract

This study investigated various testing conditions for their influence on long-term retention of reading materials. To do so, 84 English as Foreign Language (EFL) learners were randomly selected from a total of 746 and were randomly divided into two equal groups to participate in two experiments. In each experiment, the participants studied some texts and participated in some initial testing conditions before taking a 10-day delayed final exam. The testing conditions of the first experiment were 1) study + simultaneous open-book test, 2) study + open-book test, 3) study + closed-book test + feedback, 4) study + closed-book test, 5) no study no test, and 6) study with no test. The second phase was a replication of the first 5 testing conditions of the first phase accompanied by 3 more conditions, namely, 6) study + study, 7) study + study + study, 8) study + study + study + study. Analysis of variance results showed that different test types, feedback on test, and restudying could differently influence long-term retention. It was found that feedback on test had the highest effect on retention. Similarly, taking a test after study was more influential than restudying. Finally, open-book testing worked better than closed-book testing.

Keywords: closed-book test, open-book test, closed-book test plus feedback, initial study period, and retention

* - PhD candidate of TEFL at Allameh Tabataba'i University of Tehran

-Received on: 14/12/2015

Accepted on: 27/02/2016

Email: *m_qaracholloo@yahoo.com*

1. Introduction

A large body of research has revealed that test-taking not only is used to assess learning, but also it can be used to enhance learning and improve long-term retention (Carpenter, Pashler, & Vul, 2006; Karpicke & Roediger, 2007a; McDaniel, Roediger, & McDermott, 2007; Roediger & Karpicke, 2006a). When students study and then take a test over the materials, they recall them more effectively and at a greater ease. According to the well-established psychological phenomenon known as ‘the testing effect’, it is believed that retrieval processes being used when taking a test have powerful effects on learning in general and long-term retention in particular (Roediger & Butler, 2011). Put more clearly, having taken a test, students get involved in some cognitive processes which in turn can lead to an elevation of learning.

2. Literature Review

Closed-book tests (CBTs) and Open-book tests (OBTs) are two main types of testing commonly used in educational settings both internationally and in the Iranian educational system. CBT or the traditional method of testing students, as the name speaks, is a kind of test which is taken without any concurrent consultation with notes, textbooks or supplementary materials. On the contrary, OBT allows students to make use of their notes and textbooks while taking the test (Mohanani, 1997). These test types have been of growing interest among experts and educators in measurement and psychology of learning and many have theorized on the probable advantages attributed to each type.

The supporters of CBTs have their own rationales and reasons why these kinds of tests might enhance learning more than their counterparts. As stated by Hoffman (1996), CBTs give students the encouragement to engage in rote memorization which necessitates more challenging retrieval processes. Many studies have proved this in favor of CBTs, shedding light on the fact that those tests which require greater challenging retrieval processes lead to greater long-term retention (Bjork, 1999; Karpicke & Roediger, 2007b; Roediger & Karpicke, 2006b). Even the comparison of various CBTs has showed that those which are more challenging in terms of retrieval processes produce greater benefits for retention in the long run. Two of such tests are recall and recognition tests which are found to include different levels of testing effect when compared (Butler & Roediger, 2007; Glover, 1989; Kang, McDermott, & Roediger, 2007; McDaniel, Anderson, Derbish, & Morrisette, 2007).

On the contrary, some educators proclaim that OBTs improve and measure learning more efficiently than CBTs (Cnop & Grandsard, 1994; Eilertsen & Valdermo, 2000; Theophilides & Koutselini, 2000). Another superiority of OBTs over CBTs is that they are more realistic (Feller, 1994), in

that, students have access to their complementary materials and commit fewer errors. More importantly, OBTs motivate students to apply higher-order thinking skills such as reasoning and problem-solving, which are not attributed to traditional CBTs (Feller, 1994; Jacobs & Chase, 1992). Additionally, students may feel less anxiety while getting prepared for OBTs than CBTs (Theophilides & Dionysiou, 1996). Another important feature of OBTs is that they allow a greater degree of involvement in and understanding of course materials (Eilertsen & Valdermo, 2000).

Another line of research strand has focused on the difficulty involved in delayed test and the way delaying a test can affect the subsequent retention of the materials. It has generally been reported that delaying an initial test positively affects the subsequent retention (Jacoby, 1978; Karpicke & Roediger, 2007a; Modigliani, 1976; Pashler, Zarow, & Tripplett, 2003; Whitten & Bjork, 1977). The repetitive retrieval processes involved in delayed tests increment the cognitive engagement of the students and correspondingly increase their test outcomes.

In addition to CBT, OBT, the effect of feedback on the retention of studied material has also been highly reflected in the literature (Mendenhall, Beaver, & Beaver, 2002). Immediate versus delayed feedback have been of central attention in prior research. In contrast with OBT, which can be followed by immediate feedback, CBT has the possibility of feedback provision only after the test is completely done. More clearly, it is possible for students to receive immediate feedback about their performance during OBT. However, in CBTs, instructors do not provide feedback during the test, and feedback occurs after students have finished the test. Both types of immediate and delayed feedback have been found to promote long-term retention, though delayed feedback has been more widely reported in this regard. Schmidt, Young, Swinnen, and Shapiro (1989) investigated the phenomenon of motor learning and found that delayed feedback often promotes better long-term retention than immediate feedback. As for multiple-choice tests, it has also been reported that delayed feedback leads to better retention than immediate feedback (Epstein, Epstein & Brosvic, 2001; Epstein, Lazarus, Calvano, Matthews, Hendel, Epstein & Brosvic, 2002). Bangert-Drowns, Kulik, Kulik, and Morgan (1991) have also come up with the same findings. Similarly, the feedback given after initial tests is found to influence long-term retention (Jacob & Lefgren, 2004; Pashler, Cepeda, Wixted & Rohrer, 2005).

A review of the literature show that each one of the test types (CBT, OBT, and delayed), and feedback types (immediate and delayed) has its own influence on long-term retention. Although many studies can be tracked in the literature with a focus on each test type or feedback type for its influence on long-term retention, to the best of our knowledge, there is only one scholarly paper comparing the probable effects of these test types and feedback types on long-term retention to see which one is more influential in this regard.

Agarwal, Karpicke, Kang, Roediger and McDermott (2007) found that both CBT and OBT produced a testing effect on long-term retention. It was also found that both types of feedback could affect long-term retention similarly, but more than test types. Moreover, OBT accompanied by feedback turned out to be more effective than the CBT without feedback. Although the study by Agarwal, Karpicke, Kang, Roediger and McDermott (2007) has come up with invaluable findings about the effects of different test types and feedback types on long-term retention, the field is still in need of more well-documented studies to better clarify the issue. In response to this shortcoming, a two-phase experimental study was run to answer the following research question:

RQ: Are there significant differences between *CBT*, *OBT*, *Simultaneous OBT*, *CBT plus Feedback*, *Initial Study Period* and *Restudying* regarding their effect on long-term retention?

3. Method

3.1. Participants

The participants of the study were selected from among the EFL learners (nearly 746) of Iranmehr Language Institute in Tehran. In order to have a homogeneous group of participants in terms of English proficiency, only those EFL learners who were studying *American English File 3* in their sixth semester (123 persons) were included in the study. In order to have a more discreet sampling, the participants were administered Oxford Placement Test (Version 1.1), which includes three parts. To have a more reliable sampling, the writing part was put aside and only Part 1(40 items) and Part 2 (20 items) of the test include a total of 60 multiple-choice items were administered to the participants. Then, those learners whose scores fell within one standard deviation below and above the mean (84 learners) were included in the study. The selected 84 participants (31 males and 53 females, whose age ranged from 16 to 23), were randomly divided into two groups of 42. Finally, each group was also randomly assigned to each phase of the study.

3.2. Materials

The texts were selected from *Reading and Vocabulary Development 4: Concepts and Comments* (Achert & Lee, 2005). Each text of the book is nearly 950 words long. The book includes a total of 20 texts divided into 5 units based on their topics. Each text is followed by a variety of question types, namely, vocabulary, true-false comprehension, short-answer comprehension, reading strategy, word forms, grammar review, multiple-choice comprehension, and matching questions. 12 texts were randomly selected to be used in the study, 6 texts for the first and 6 texts for the second phase. Only the multiple-choice comprehension and short-answer comprehension questions provided after each text were selected to be used in the initial and final tests.

The same questions were used in the initial and final tests. However, all of the questions used in the first phase of the study were multiple-choice comprehension ones and the questions selected for the second phase were short-answer comprehension ones. The rationale for this policy was to control for the probable effect of question types. Table 1 shows the title of texts selected for each phase of the study along with the number of questions provided after each text.

Table 1

Texts and Questions for each Phase of the Study

<i>Texts for the first phase</i>	<i>Number and type of questions</i>
1. Deserts	8 multiple-choice comprehension questions
2. Cave paintings	8 multiple-choice comprehension questions
3. National parks	8 multiple-choice comprehension questions
4. Comets	10 multiple-choice comprehension questions
5. New plants	10 multiple-choice comprehension questions
6. Memory	10 multiple-choice comprehension questions
<i>Texts for the second phase</i>	
1. Navajo sand painting	7 short-answer comprehension questions
2. The United Nations	10 short-answer comprehension questions
3. Satellites	10 short-answer comprehension questions
4. Motor vehicles: The pros and cons	10 short-answer comprehension questions
5. Obesity: The new epidemic	10 short-answer comprehension questions
6. Can fashion be hazardous to your health	10 short-answer comprehension questions

3.3. Data Collection Procedure

In the first phase of the study, which included six testing conditions, the participants were told to study six texts. They were not told whether their study would be followed by a test or feedback. The order of the presentation of all the six texts and the testing conditions were counterbalanced to control the probable effect of texts and testing condition order. The participants were tested in six groups of seven members. By this, the order of presentation of testing conditions for the members of each group was the same, but different from one group to another. Five out of six testing conditions started with a study period. The group members were asked to read the text and give it back to the tester. Additionally, four out of five study periods were followed by an initial test (Table 2).

As shown in Table 2, one of the groups had no study to be followed by an initial test. This group functioned as the control group to see if the lack or presence of study period would influence the score of participants on a final delayed test. Another group had a study period, but it was not followed by an initial test. This was done to see to what extent studying without taking an

initial test would influence the delayed test. The other four testing conditions were followed by one type of test. The ‘simultaneous OBT’ was actually taking a test without previous study. More specifically, it was an over-material test without previous studying. The ‘OBT’ was an over-material test with previous study period. There were two additional CBTs with the same logistics of traditional testing; however, one of them was followed by feedback. Actually, the ‘CBT plus feedback’ was self-graded by the participants and then was followed by the feedback provided by the teacher on the wrong answers. The second session was a 10-day delayed test on the same material using the same questions used in the initial test. In this session, all of the groups took the test simultaneously because there were no differences in the testing procedures from one group to another.

Table 2
Testing Conditions of the First Phase of the Study

Study	Conditions in the first session		Second session	
	Test or study	Feedback	Final test	
---	-----	---	✓	
---	✓ Simultaneous OBT	---	✓	
✓	✓ OBT	---	✓	
✓	✓ CBT	✓	✓	
✓	✓ CBT	---	✓	
✓	✓ -----	---	✓	

Note: Check marks indicate the presence of study, test, or feedback in the first or second session

The second phase was conducted like the first one. The participants were asked to read six texts and were not told whether studying the texts would be followed by a test, another study, or feedback. Just like the first phase, the texts and the testing conditions were presented in a counterbalanced manner to control for the probable effect of texts and testing conditions. The participants were tested in six groups of five members and two groups of six members. For all the members of each group, the order of presentation of texts was counterbalanced, but the testing conditions were presented similarly. Seven out of eight conditions started with a study period. The participants were asked to read the texts and hand them back to the tester. Four out of seven study periods were followed by an initial test.

One of the groups had no study and no initial test. This was actually a control group to see whether, and to what extent, a concurrent lack of study and test would influence the final test result. Three groups had restudy periods without a test to see if restudying would lead to the same effects of initial tests on the final test. The other four testing conditions were exactly similar to those of the first phase. The second session, which was for a 10-day delayed test on the same materials using the same questions, was also conducted like that of the first phase.

Since the questions of this phase were in short-answer format and not multiple-choice, a 5-point scale was used by two raters for rating the questions. Incorrect answers and blanks were scored zero by the raters. The answers with some hints or indication of the correct answer were scored as one. The correct but less detailed answers were scored as two. The correct answers with moderate detailed answers were scored as three. And finally, the correct answers with enough detailed explanation were given the score four. The average of the scores given by the two raters was computed to be used as the final scores for each participant.

The raters were two PhD students of TEFL who had the required theoretical and practical qualifications to rate the questions. They had nearly ten years of teaching experience. More importantly, they had passed several courses of language testing and had designed many language tests as term projects for the course they had taken as students or the courses they had taught as teacher.

4. Results and Discussion

4.1. Phase 1

4.1.1. Initial test

Two out of the six testing conditions had no initial test before the final test; therefore, only the results for four testing conditions are presented here. The participants answered a total of 54 multiple-choice questions; three of the texts had 8 questions and the other three had 10 questions. The results of descriptive statistics indicated that the lowest and the highest means belonged to the groups that took the CBT ($M = 38.42$) and CBT plus feedback ($M = 51.85$), respectively. The two kinds of OBT had more moderate effects on scores with a mean of 43 for simultaneous OBT and a mean of 48 for OBT. It was found that OBT and CBT plus feedback had the closest scores (mean difference = 3.57). The difference between simultaneous OBT and OBT (mean difference = 5.28) was not considerable either. The mean difference for simultaneous OBT and CBT was more or less the same (4.57). The biggest differences were found between the mean scores of OBT and CBT (9.85), on the one hand, and CBT plus feedback and CBT (13.42), on the other (Tables 3 and 4).

Table 3

Descriptive Statistics of Phase 1

<i>Groups</i>	<i>Mean</i>	<i>Std. Deviation</i>
Simultaneous OBT	43.00	3.08
CBT	48.28	4.12
CBT plus feedback	51.85	2.30
CBT	38.42	5.11

The results of one-way ANOVA indicated a significant difference between the four kinds of testing conditions ($F = 62.18, p < 0.01$). Post hoc results also indicated that there were significant differences between all testing conditions (see Table 4).

Table 4
Post Hoc Test Results

<i>Groups</i>	<i>Groups</i>	<i>Mean difference</i>	<i>Std. error</i>	<i>Sig.</i>
Simultaneous OBT	OBT	-5.28	1.05	.000**
	CBT plus feedback	-8.85	1.05	.000**
	CBT	4.57	1.05	.001**
OBT	CBT plus feedback	-3.57	1.05	.014*
	CBT	9.85	1.05	.000**
CBT plus feedback	CBT	13.42	1.05	.000**

Note: ** = $p < 0.01$; * = $p < 0.05$

4.1.2. Final Test

Although only four groups of participants took the initial tests, all of the groups took the final test. The questions in the final test were the same multiple-choice comprehension questions of the first session initial test. It was found that the lowest mean was for the control group with no study and no test ($M = 20.57$). Just like that of the initial test, the highest mean belonged to CBT plus feedback ($M = 50.14$). In a similar manner, simultaneous OBT ($M = 37.57$) and OBT ($M = 41.85$) had a higher mean than CBT ($M = 32.71, SD = 2.05$). It was also found that the mean score for ‘Study without Test’ ($M = 29.71$) was lower than the four testing conditions which had a kind of test, but was higher than the control group with no study no test. The highest difference in mean scores was found to be between ‘No Study No Test’ and ‘CBT plus feedback’ (mean difference = 29.57) and the lowest was between CBT and ‘Study without Test’ (mean difference = 3) (Tables 5 and 6).

Table 5
Descriptive Statistics

<i>Groups</i>	<i>Mean</i>	<i>Std. Deviation</i>
No Study No Test	20.57	3.40
Simultaneous OBT	37.57	2.50
OBT	41.85	3.13
CBT plus feedback	50.14	1.77
CBT	32.71	2.05
Study Without Test	29.71	2.87

A look at the results of the testing conditions CBT, simultaneous OBT, OBT and CBT plus feedback shows that although the scores for each testing

condition have decreased on the final delayed test, the ascending order of mean scores is the same; CBT (initial mean = 38.42, delayed mean = 32.71), simultaneous OBT (initial mean = 43, delayed mean = 37.57), OBT (initial mean = 48.28, delayed mean = 41.85) and CBT plus feedback (initial mean = 51.85, delayed mean = 50.14). Although the mean difference from the initial to the delayed test was nearly 6 for the three test conditions of CBT, simultaneous OBT and OBT, the mean for CBT plus feedback only minimally decreased on the final delayed test (1.61) (Tables 3 and 5).

The results of one-way ANOVA indicated a significant difference between the six testing conditions on their final test ($F = 101.41, p < .01$). Post hoc results showed that except for simultaneous OBT vs. OBT (mean difference = 4.28, $p > .05$) and CBT vs. Study without Test (mean difference = 3, $p > 0.05$), there were significant differences between all the other testing conditions on the final tests.

4.1.3. Discussion of Phase 1

It was found that the ascending order of mean scores for the four testing conditions of CBT, simultaneous OBT, OBT and CBT plus feedback was the same on the initial and the final tests. The results of ANOVA indicated significant differences between all the four testing conditions on the initial and the final tests, except for the difference between simultaneous OBT and OBT on the final test. Not surprisingly, it was found that OBT, simultaneous OBT and CBT plus feedback outperformed CBT. This finding is in tune with the results of the study by Agarwal, Karpicke, Kang, Roediger and McDermott (2007). Having repeated exposure to supplementary material while taking the tests and receiving feedback on the test are the reasons why CBT, which only allows one exposure to the questions, was less influential than the other three testing conditions. The same results were obtained for the final test. This indicates the deeper engagement in the testing materials attributed to OBTs and CBT plus feedback.

There was a significant difference between simultaneous OBT and OBT. This is an innovative finding in that no previous study has compared the two. Although it has been stated that generally OBTs raise higher-level thinking (Feller, 1994; Jacobs & Chase, 1992) and alleviate stress, which consequently result in more efficient learning (Theophilides & Dionysiou, 1996), no study has already compared simultaneous OBT with OBT for their probable difference. It can be claimed that though, in both kinds of OBTs, the participants have access to supplementary materials, previous study period can function as a warm-up activity which provides them with deeper cognitive engagement to take a more successful test. Of course, there was no significant difference between the two on the final test. This could be attributed to the 10-day time interval between the initial and final test. Maybe a 10-day time

interval was enough to moderate the effect of previous cognitive engagement with the text under test.

Table 6
Post Hoc Test Results

<i>Groups</i>	<i>Groups</i>	<i>Mean difference</i>	<i>Std. error</i>	<i>Sig.</i>
No Study No Test	Simultaneous OBT	-17.00	1.43	.000**
	OBT	-21.28	1.43	.000**
	CBT plus feedback	-29.57	1.43	.000**
	CBT	-12.14	1.43	.000**
	Study Without Test	-9.14	1.43	.000**
Simultaneous OBT	OBT	-4.28	1.43	.075
	CBT feedback test	-12.57	1.43	.000**
	CBT	4.85	1.43	.026*
	Study Without Test	7.85	1.43	.000**
OBT	CBT plus feedback	-8.28	1.43	.000**
	CBT	9.14	1.43	.000**
	Study Without Test	12.14	1.43	.000**
CBT plus feedback	CBT	17.42	1.43	.000**
	Study Without Test	20.42	1.43	.000**
CBT	Study Without Test	3.00	1.43	.65

Note: ** = $p < 0.01$; * = $p < 0.05$ *

There were also significant differences between CBT plus feedback on the one hand, and simultaneous OBT and OBT, on the other. Although in OBTs, the participants had access to the material and could get higher scores, the CBT plus feedback group had a higher mean. This could be explained in two ways. First, the number of exposures to the materials was two for OBT and one for simultaneous OBT; however, CBT plus feedback had four exposures to the materials tested. More specifically, the participants first studied the materials as the first exposure. For the second exposure, they took the test. When they self-graded, they had the chance for the third exposure to the materials. And finally, the participants received feedback on their test answers, which allowed for a deeper focus on testing points. The second reason why the CBT plus feedback group outperformed simultaneous OBT and OBT groups can be attributed to the feedback accompanied with CBT. This finding is in harmony with the previous assertions about the positive effects of feedback on test results and retention (Schmidt, Young, Swinnen & Shapiro, 1989), but the fresh finding is that the effect of feedback accompanied with CBT was much higher than the effect of simultaneous OBT and OBT. It can be asserted that feedback on test can result in deeper cognitive engagement with the testing materials.

It was also found that the control group, which had no previous study or test, had the lowest score on the final test. Additionally, the mean score for

the group which had only previous study of the materials was lower than all the other groups which had initial tests. This indicates that an initial test necessitates more cognitive engagement than studying without taking a test. Moreover, although the mean scores for CBT, simultaneous OBT and OBT noticeably decreased on the final test, this was really minimal for CBT plus feedback. It can be claimed that the cognitive processes needed for feedback application is more lasting than that of testing alone.

Generally speaking, it can be asserted that initial study period, closed-book test, closed-book test plus feedback, open-book test, and simultaneous open-book test could influence the long-term retention of materials, but to different degrees. Taking an initial test was more influential than studying. However, a test plus feedback was better than other kinds of tests. Additionally, open-book tests influenced long-retention more than closed-book tests.

4.2. Phase 2

There are several rationales for conducting a second phase. First, some criticism can be leveled against the superiority of test types and feedback over studying without taking a test for their effects on long-term retention. It could be claimed that this superiority can be the consequence of re-exposure to the same material and not the test effect. Hence, a second phase was run to compare the test types with restudying conditions. More clearly, it was assumed that maybe second, third, and fourth exposures to the studied materials could similarly improve test results.

The second rationale for conducting a second phase was to control for the possible effects of different texts and test types. On the one hand, different texts were studied in the second phase. Of course, the length of texts and their reading difficulty were more or less equal to those of the first phase. On the other hand, the test types used in the second phase were short-answer comprehension questions and not multiple-choice ones. This was done to control for the possible effect of test types on long-term retention.

The second phase was similar to the first one in the five testing conditions 1) No Study No Test, 2) Simultaneous OBT, 3) OBT, 4) CBT plus feedback, and 5) CBT. The point of difference was that there was no 'study without test' condition, and three restudying conditions were added (Table 7).

4.2.1. Initial Test

Just like that of the first phase, in the initial test of the second phase only four testing conditions had a test. Similarly, the highest and the lowest means belonged to CBT plus feedback ($M = 224$) and CBT ($M = 160.60$), respectively. Likewise, simultaneous OBT ($M = 180$) and OBT ($M = 202$) had more moderate scores. In the same manner, the highest difference was between CBT plus feedback and CBT (Table 8).

Table 7
Testing Conditions of the Second Phase of the Study

<i>Study</i>	<i>Conditions in the first session</i>		<i>Second session</i>
	<i>Test or study</i>	<i>Feedback</i>	<i>Final test</i>
---	-----	---	✓
---	Simultaneous open-book test	---	✓
✓	Open-book test	---	✓
✓	Closed-book test	✓	✓
✓	Closed-book test	---	✓
✓	Study	---	✓
✓	Study+Study	---	✓
✓	Study+Study+Study	---	✓

Note: Check marks indicate the presence of study, test, or feedback in the first or second session

Table 8
Descriptive Statistics for the Initial Test of Phase Two

<i>Groups</i>	<i>Mean</i>	<i>Std. Deviation</i>
Simultaneous OBT	180	14.17
OBT	202.40	4.15
CBT plus feedback	224	3.70
CBT	160.60	5.77

The results of one-way ANOVA were the same as those of the initial test of the first phase. The results of one-way ANOVA indicated a significant difference between the testing conditions ($F = 57.28, p < 0.01$). Post hoc comparisons indicated significant differences between all the testing conditions as well (Table 9).

4.2.2. Final Test

All the eight groups took the final test. The questions of the final test were the same short-answer comprehension questions used on the initial test. The lowest and highest means belonged to the ‘No Study No Test Condition’ ($M = 100$) and ‘CBT plus feedback’ ($M = 215.40$), respectively. Simultaneous OBT ($M = 178$) and OBT ($M = 190.60$) had higher means than CBT ($M = 154$). Although the three study conditions had higher mean scores than the ‘No Study No Test’ condition, their mean scores were noticeably lower than the four testing conditions which had an initial test in the first session (Tables 10).

The results of one-way ANOVA indicated a significant difference between the testing conditions ($F = 24.22, p < 0.01$). Post hoc comparisons indicated that out of a total of 28 possible comparisons, there were no significant differences between ‘No Study No Test’ vs. ‘Study + Study’ ($MD = 35, p > 0.05$), ‘Simultaneous OBT’ vs. OBT ($MD = 12.80, p > 0.05$), ‘Simultaneous OBT’ vs. CBT ($MD = 23, p > 0.05$), OBT vs. ‘CBT plus feedback’ ($MD = 25, p > 0.05$), CBT vs. ‘Study + Study’ ($MD = 19.40, p >$

0.05), CBT vs. ‘Study + Study + Study’ (MD = 12.56, $p > 0.05$), CBT vs. ‘Study + Study + Study + Study’ (MD = 13.56, $p > 0.05$), ‘Study + Study’ vs. ‘Study + Study + Study’ (MD = 6.83, $p > 0.05$), ‘Study + Study’ vs. ‘Study + Study + Study + Study’ (MD = 5.83, $p > 0.05$), and ‘Study + Study + Study’ vs. ‘Study + Study + Study + Study’ (MD = 1.00, $p > 0.05$). There were significant differences between all the other testing conditions on the final test (Table 11).

Table 9

Post Hoc Comparisons for the Initial Test of Phase Two

<i>Groups</i>	<i>Groups</i>	<i>Mean difference</i>	<i>Std. error</i>	<i>Sig.</i>
Simultaneous OBT	OBT	-22.00	5.14	.003**
	CBTplus feedback	-43.00	5.14	.000**
	CBT	20.00	5.14	.008**
OBT	CBTplus feedback	-21.80	5.14	.004**
	CBT	42.00	5.14	.000**
CBTplus feedback	CBT	63.80	5.14	.000**

Note: ** = $p < 0.01$; * = $p < 0.05$ *

Table 10

Descriptive Statistics of Scores on the Final Test of Phase Two

<i>Groups</i>	<i>Mean</i>	<i>Std. Deviation</i>	<i>N</i>
No study no test	100.00	8.09	5
Simultaneous open-book test	178	7.38	5
Open-book test	190.60	7.15	5
Closed-book test plus feedback	215.40	13.21	5
Closed-book test	154	6.43	5
Study + study	135	4.52	5
Study + study + study	141.83	6.55	6
Study + study + study + study	140	9.95	6

4.2.3. Discussion of Phase 2

The second phase of the study was a replication of the first one. Those testing conditions which were the exact duplicate of the same conditions of the first phase led to similar findings. CBT plus feedback was found to be better than the other three test types both in the initial and final tests. By the same token, OBT was found to be more effective than simultaneous OBT. Additionally, CBT was the least influential test type in terms of improving long-term retention. Moreover, the effect of No Study No Test control condition on delayed retention was lower than all the four testing conditions. These findings were yielded using a different kind of question (short-answer comprehension

questions). It can be stated that question types could not influence the way feedback, OBT, and CBT improve long-term retention.

It was also found that repeated study of three or four times had a higher effect on long-term retention than No Study No Test, which was a control condition. In the first phase of the study, it was found that there was no difference between studying once and no study at all for their influence on retention; however, in the second phase, it was found that if the number of studies increased to three or four, it could improve retention. More clearly, though the 10-day time interval from the initial to the final test could alleviate the effect of studying, this alleviation of effect could be refunded through restudying or repeated studying. An increase in the number of exposures to the materials increased the effect on long-term retention. Of course, the effect of repeated exposure was found only when the number of exposures increased to three. Since three and four exposures affected the long-term retention equally, it can be claimed that through three or four exposures to the same materials, test-takers gain as much cognitive engagement as needed to take a successful test.

The comparison of restudying conditions with the test conditions showed that although restudying by itself could improve retention; this impact was much lower than that of the four test types. Clearly speaking, taking a test, whether OBT, CBT, and CBT plus feedback, could promote long-term retention in a way that no restudying condition could match. This could be attributed to the exposure to different kinds of materials. In the initial tests, the participants had access to the same questions tested on the final exam. This exposure to the specific testing points allowed the participants to be free from preoccupation with all the information presented in the texts. On the contrary, in the restudying condition, the participants were re-exposed to the total materials without any specific reference to what particular point is supposed to be tested finally.

5. Conclusion and Implications

The present study about various test types, initial study period, feedback, and restudying showed that they could improve long-term retention of materials in different degrees. Due to the meta-materials benefits associated with feedback provided by teachers and testers, it was found that feedback on test could improve long-term retention more than any test types, initial study period and restudying of the materials. Additionally, initial tests could result in a deeper cognitive engagement and correspondingly greater long-term retention than initial study period and restudying. Of course, OBTs, due to their over-material nature, outperformed CBTs in this regard. Moreover, it was found that restudying the material could catch up with CBTs in terms of their effect on long-term retention.

Based on the findings of this study, it may be concluded that teachers need to encourage their students to study the learning materials repeatedly. They can do so by giving learners frequent tests and providing them with

Table 11

Post Hoc Results for the Final Test of Phase Two

<i>Groups</i>	<i>Groups</i>	<i>Mean difference</i>	<i>Std. error</i>	<i>Sig.</i>
No study no test	Simultaneous OBT	-78.00	10.55	.000**
	OBT	-90.80	10.55	.000**
	CBT plus feedback	-116	10.55	.000**
	CBT	-54	10.55	.000**
	Study + study	-35	10.55	.061
	Study + study + study	-41.83	10.10	.006**
	Study + study + study + study	-40.83	10.10	.008**
Simultaneous OBT	OBT	-12.80	10.55	1.00
	CBT plus feedback	-38.40	10.55	.025*
	CBT	23	10.55	.89
	Study + study	43	10.55	.007**
	Study + study + study	36.15	10.10	.030*
	Study + study + study + study	37.15	10.10	.022*
OBT	CBT plus feedback	-25.00	10.55	.58
	CBT	36.40	10.55	.043*
	Study + study	55.80	10.55	.000**
	Study + study + study	48.96	10.10	.001**
	Study + study + study + study	49.96	10.10	.001**
CBT plus feedback	CBT	62.00	10.55	.000**
	Study + study	81.40	10.55	.000**
	Study + study + study	74.56	10.10	.000**
	Study + study + study + study	75.56	10.10	.000**
CBT	Study + study	19.40	10.55	1.00
	Study + study + study	12.56	10.10	1.00
	Study + study + study + study	13.56	10.10	1.00
Study + study	Study + study + study	-6.83	10.10	1.00
	Study + study + study + study	-5.83	10.10	1.00
Study + study + study	Study + study + study + study	1.00	9.63	1.00

Note: ** = $p < 0.01$; * = $p < 0.05$ *

frequent feedback. It may also be concluded that test format does not influence the learners' retention of materials. This implies that teachers need not worry about the type of test they give. In other words, as long as learners take tests and subsequent feedback, retention happens regardless of the type of test and feedback.

These findings can have theoretical and practical implications. Apart from the theoretical value of these findings in shedding light on some of the theoretical issues surrounding this topic, these findings can also have implications for teachers and syllabus designers. These findings may help teachers spend more time on testing and providing feedback. Instead of spending all the class time on teaching, teachers can spend a portion of the time on testing to ensure a higher level of learner engagement and learning. Syllabus designers can also incorporate these findings in preparing materials. They can design materials in such a way to allow for repeated exposure as well as more teacher feedback.

References

- Achert, P., & Lee, L. (2005). *Reading and vocabulary development 4: Concepts and comments (3rd ed)*. Boston: Thomson Heinle.
- Agarwal, P. K., Karpicke, J. D., Kang, S. H., Roediger, H. L., & McDermott, K. B. (2008). Examining the testing effect with open-and closed-book tests. *Applied Cognitive Psychology*, 22(7), 861-876.
- Bangert-Drowns, R. L., Kulik, C. L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of educational research*, 61(2), 213-238.
- Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher, & A. Koriat (Eds.), *Attention and performance XVII. Cognitive regulation of performance: Interaction of theory and application* (pp. 435-459). Cambridge, MA: MIT Press.
- Butler, A. C., & Roediger III, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology*, 19(4-5), 514-527.
- Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review*, 13(5), 826-830.
- Cnop, I., & Grandsard, F. (1994). An open-book exam for non-mathematics majors. *International Journal of Mathematical Education in Science and Technology*, 25(1), 125-130.
- Eilertsen, T. V., & Valdermo, O. (2000). Open-book assessment: A contribution to improved learning? *Studies in Educational Evaluation*, 26(2), 91-103.

- Ellis, R., Sheen, Y., Murakami, M., & Takashima, H. (2008). The effects of focused and unfocused written corrective feedback in an English as a foreign language context. *System*, 36(3), 353-371.
- Epstein, M. L., Epstein, B. B., & Brosvic, G. M. (2001). Immediate feedback during academic testing. *Psychological reports*, 88(3), 889-894.
- Epstein, M. L., Lazarus, A. D., Calvano, T. B., Matthews, K. A., Hendel, R. A., Epstein, B. B., & Brosvic, G. M. (2010). Immediate feedback assessment technique promotes learning and corrects inaccurate first responses. *The Psychological Record*, 52(2), 187-201.
- Farrokhi, F., & Sattarpour, S. (2012). The effects of direct written corrective feedback on improvement of grammatical accuracy of high-proficient L2 learners. *World Journal of Education*, 2(2), p49.
- Feller, M. (1994). Open-book testing and education for the future. *Studies in Educational Evaluation*, 20, 235-238.
- Ferris, D., & Roberts, B. (2001). Error feedback in L2 writing classes: How explicit does it need to be? *Journal of second language writing*, 10(3), 161-184.
- Glover, J. A. (1989). The “testing” phenomenon: Not gone, but nearly forgotten. *Journal of Educational Psychology*, 81, 392-399.
- Jacob, B. A., & Lefgren, L. (2004). Remedial education and student achievement: A regression-discontinuity analysis. *Review of economics and statistics*, 86(1), 226-244.
- Jacobs, L. C., & Chase, C. I. (1992). *Developing and using tests effectively. A guide for faculty*. Jossey-Bass Inc: San Francisco, CA.
- Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning & Verbal Behavior*, 17, 649-667.
- Kang, S. H., McDermott, K. B., & Roediger III, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, 19(4-5), 528-558.
- Karpicke, J. D., & Roediger, H. L. (2007a). Expanding retrieval promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33, 704-719.
- Karpicke, J. D., & Roediger, H. L. (2007b). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, 57, 151-162.
- McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, 19(4-5), 494-513.
- McDaniel, M. A., Roediger, H. L., & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic bulletin & review*, 14(2), 200-206.

- Mendenhall, W., Beaver, R. J., & Beaver, B. M. (2002). *A brief introduction to probability and statistics*. Cengage Learning Editors.
- Modigliani, V. (1976). Effects on a later recall by delaying initial recall. *Journal of Experimental Psychology: Human Learning & Memory*, 2, 609-622.
- Mohanan, K. P. (1997). *Open book examination*. A report and a response to some recurrent concerns, Seminar on conducting open book examinations organized by Centre for Development of Teaching and Learning at National University of Singapore, January.
- Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(1), 3-8
- Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in cognitive sciences*, 15(1), 20-27.
- Roediger, H. L., & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, 1, 181-210.
- Roediger, H. L., & Karpicke, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17, 249-255.
- Schmidt, R. A., Young, D. E., Swinnen, S., & Shapiro, D. C. (1989). Summary knowledge of results for skill acquisition: support for the guidance hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(2), 352.
- Sheen, Y. (2007). The effect of focused written corrective feedback and language aptitude on ESL learners' acquisition of articles. *TESOL Quarterly*, 41(2), 255-283.
- Theophilides, C., & Koutselini, M. (2000). Study behavior in the closed-book and the open-book examination: A comparative analysis. *Educational Research and Evaluation*, 6(4), 379-393.
- Whitten, W. B., & Bjork, R. A. (1977). Learning from tests: Effects of spacing. *Journal of Verbal Learning and Verbal Behavior*, 16(4), 465-478.