Journal of Effective Teaching in Higher Education

Testing Effect: A Further Examination of Open-book and Closed-book Test Formats

Olesya Senkova, Hajime Otani, Reid L. Skeel, and Renée L. Babcock
Central Michigan University, Mount Pleasant, MI

Abstract. If assessment is the purpose of testing, open-book tests may defeat the purpose. However, a goal of education is to build knowledge, and based on the literature, open-book tests may not be inferior to closed-book tests in promoting long-term retention of information. Participants studied Swahili-English pairs and either re-studied or took an initial quiz, which was cued recall or recognition in an open-book or closed-book format. One week later, the final closed-book recognition test showed higher performance in the quizzed conditions than in the study-twice condition, replicating the testing effect. However, performance was similar across the quizzed conditions, indicating that testing promoted long-term retention regardless of test format (open-book versus closed-book) and test type (cued recall versus recognition). Open-book tests are not inferior to closed-book tests in building knowledge and can be particularly useful in online classes because preventing cheating is difficult when closed-book tests are administered online.

Keywords: testing effect; open-book tests; long-term retention; learning

In educational settings, tests are used as assessment tools because testing students is the primary method of determining how much students have learned. Traditionally, tests are divided into two categories, recall-based and recognition-based, with essay and short-answer questions representing the former and multiple-choice questions representing the latter. Furthermore, tests can be administered with closed-book or open-book formats, with the former requiring students to rely entirely on memory and the latter allowing students to look up what they did not commit to their memory. Choosing the right format for testing students presents a challenge particularly because of the recent popularity of online classes. In most online classes, it is difficult to know whether students are looking up answers to test questions. In fact, there is a debate over whether cheating on the test is more prevalent in online classes than in face-to-face classes, and if so how to prevent it (e.g., Alessio, Malay, Maurer, Bailer, & Rubin, 2017; Christe, 2003; Cluskey, Ehlen, & Raiborn, 2011; Grijalva, Nowell, & Kerkvliet, 2006; Michael & Williams, 2013; Owens, 2016; Rowe, 2004). Although measures can be implemented to minimize cheating such as using a lock-down web browser with a monitor and imposing a time limit, one must realize that no method is completely foolproof. Because a traditional closed-book test is difficult to implement in an online environment, the instructor may adopt an open-book test. However, the consequences of adopting open-book tests, instead of using traditional closed-book tests, are still not clear.

In the present study, the two formats, closed-book and open-book, were compared to examine a commonly held assumption that the latter format is inferior to the former in achieving the goal of promoting long-term retention of studied materials. Additionally, the type of the test (i.e., cued recall and recognition) was manipulated to investigate whether test format would interact with test type. If the purpose of testing is to assess learning, one may argue that an open-book test would defeat the purpose because when one is allowed to look up answers, it would be difficult to assess what one knows and does not know. However, because one of the goals of education is to develop knowledge (see *Bloom's Taxonomy of Educational Objectives*, Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956; also see Anderson et al., 2001, for a revised version), it is important to determine whether using open-book tests, rather than the traditional closed-book tests, would defeat this goal.

Reasons that Open-book Tests May Be Inferior to Closed-book Tests

There are several reasons to assume that open-book tests are inferior to closed-book tests in building knowledge. First, when a test is offered in an open-book format, students may view the impending test as easy because the answers will be available during the test, and thus, they may devote less effort in studying. In fact, Sparrow, Liu, and Wegner (2011) showed that when participants were informed that they would be allowed to look up the answers when they take the test later on, they tended to show on the test lower recall of the target information but enhanced recall of where to find the target information. Accordingly, expecting an open-book test may discourage students from putting enough effort to build their own knowledge.

The second reason that open-book tests might be inferior to closed-book tests in building knowledge is based on the notion that retrieving information from memory enhances learning and facilitates future retrieval. According to Bjork and Bjork (1992), the likelihood of remembering depends on both storage strength and retrieval strength. The storage strength, commonly referred to as memory strength, reflects the amount of learning and determines the availability of information in memory. The storage strength can be increased by repeatedly studying the material. Another component of remembering is retrieval strength, which is how easily memory can be accessed. Bjork and Bjork assumed that building strong memory (storage strength) by repeatedly studying is not sufficient to guarantee successful memory retrieval because in addition to creating strong memory, one needs to practice retrieving memory in order to make it easy to access. Take an old telephone number for example. It may be still available in memory due to repeated use in the past; however, one may experience difficulty retrieving it because it has not been used recently. Another aspect of this theory is that these two types of strength are related such that increasing retrieval strength by repeatedly retrieving memory would also increase storage strength. In effect, retrieving memory acts as another opportunity to learn. What is critical to the issue of test format and building knowledge is that there is an inverse relationship between the ease of retrieval and the amount of increment in storage strength, such that easy retrieval would result in a small increment in storage strength, whereas, difficult retrieval would result in a large increment in storage strength. In

line with this principle, Bjork (1999) argued that training procedures with easy access to correct responses would slow down the growth of storage strength. Based on these notions, it is reasonable to assume that open-book tests, with information readily available during retrieval, would be less efficient in building knowledge because looking up an answer is easier than retrieving an answer from memory, resulting in a smaller increment in storage strength.

Reasons that Open-Book Tests May Not Be Inferior to Closed-Book Tests

There are indications that making a test open-book may not defeat the purpose of building knowledge. First, the format of the test, open-book or closed-book, may not matter as long as the test questions are sufficiently difficult to promote elaborative or deep-levels of processing. Based on the literature on the levels of processing model (Craik, 2002; Craik & Lockhart, 1972; Craik & Tulving, 1975), it is clear that regardless of one's intention to learn, processing information at a deep level (i.e., semantic level) would produce durable memory compared to processing information at a shallow level (i.e., perceptual level). For instance, recognition performance was higher when participants were asked to think whether a given word would fit into a sentence frame (e.g., "He met _____ in the street?") than when participants were asked to judge whether a word was in capital letters (Craik & Tulving, 1975). These orienting questions can be conceptualized as questions on a test; and based on the results of these studies, it is reasonable to assume that if the questions on the open-book test direct one to process the answers at a deep level, durable memories can be formed.

Research on the testing effect also supports the notion that open-book tests may not be inferior to closed-book tests with regard to long-term retention of information. The testing effect is a phenomenon that simply taking tests increases long-term retention of information better than re-studying does. There is substantial evidence showing that the testing effect is a robust phenomenon (see Roediger & Karpicke, 2006a, for an extensive review) that can be observed with a variety of tests (such as free recall, cued recall, and recognition), materials (such as word lists, lists of paired-associates, and prose materials), and settings (such as laboratory and educational settings). There are at least three recently published meta-analyses on the testing effect (Adesope, Trevisan, & Sundararajan, 2017; Pan & Rickard, 2018; Rowland, 2014), and all confirmed that the testing effect is a powerful method of increasing learning. Furthermore, Pan and Rickard (2018) showed that the testing effect is robust even when the format and information being tested on the initial test are different from those on the final test, showing the transfer of learning effect. In addition, some studies have shown that the testing effect is similar between open-book and closed-book tests, even though contrary results have also been reported. In a review of the literature comparing open-book and closed-book test formats, Durning and colleagues (2016) reported that among the five studies that examined the testing effect between these formats, four showed that the testing effect was similar between these test formats (Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008; Agarwal & Roediger, 2011; Gharib, Phillips, & Mathew, 2012; Pauker, 1974), whereas, one showed that the testing effect was lower in the open-book format than in the closed book format

(Moore & Jensen, 2007). However, the results are not as straight forward as it appears. Although Pauker (1974) reported that the testing effect was similar between the open-book and closed book conditions, for the students who started out the semester at the bottom third of the class, open-book tests resulted in lower final test scores than closed-book tests. Agarwal and Roediger (2011) also showed that expecting an open-book test reduced time spent on learning, resulting in lower performance on transfer questions on the final test that probed deep learning. Furthermore, Moore and Jensen (2007) showed that performance on the final test was lower when previous tests were open-book relative to closed-book, and that expecting open-book tests resulted in poor academic behaviors, such as skipping lectures and help-sessions as well as not completing extra assignments. In sum, the literature on testing effect showed mixed results that indicate a complex interaction of various factors (such as motivation) that goes beyond the issue of test format.

Another reason that open-book tests may not be inferior to closed-book tests is based on the issue of information re-exposure. Open-book tests re-expose students to information successfully retrieved as well as information that was not retrieved, providing opportunities for additional learning. Moreover, an open-book test allows students to access correct answers, which would, in turn, minimize commission errors. Butler, Marsh, Goode, and Roediger (2006) showed that when participants made commission errors on the first test and did not receive feedback with the correct answers, they made the same errors on the final test. It is, therefore, possible that open-book tests are superior to closed-book tests by limiting the number of commission errors, preventing long-term retention of incorrect information. Note, however, that a recent meta-analysis by Adesope et al. (2017) showed that providing or not providing feedback was not a significant moderator of the effect size associated with the testing effect. Their explanation was that taking a test itself is cognitively challenging enough to produce the benefit of testing, and therefore, availability of feedback may not produce an additional benefit. Nevertheless, these researchers also cautioned that their finding may be due to unidentified confounding variables given that there are studies that showed the benefit of providing feedback in addition to testing (e.g., Metcalfe, Kornell, & Finn, 2009; Pashler, Cepeda, Wixted, & Rohrer, 2005). With this caveat, Adesope et al. concluded that further research is needed to investigate the effect of providing feedback on the testing effect.

Rationale and Methods of Present Study

It is intriguing to consider a possibility that open-book tests are as effective as closed-book tests in building knowledge, particularly for traditional educators who regard tests only as assessment tools. Because this issue has an important implication for educational practice, it warrants further investigation. As noted above, the results of research are conflicted as to whether open-book and closed-book tests differ in promoting long-term retention of studied material, with some studies showing that these two formats are similar (e.g., Agarwal et al., 2008; Agarwal & Roediger, 2011; Pauker, 1974; Gharib et al., 2012) and the other studies showing that open-book tests are less effective than closed-book tests,

particularly, in promoting deep learning (e.g., Agarwal & Roediger, 2011; Pauker, 1974). However, not all learning situations require deep learning, such as during the early stage of learning. The question, then, is whether the lack of difference between open-book and closed-book tests can be generalized to non-text-based materials. Although text-based materials clearly have educational relevance, building knowledge also involves acquiring facts and building vocabularies. Accordingly, in the present study, we decided to investigate the effect of test format using Swahili-English word pairs in a testing effect paradigm similar to Agarwal et al. (2008). It is possible that the lack of difference between open-book and closed-book tests reported by Agarwal et al. and other researchers in the past was based on the use of text materials. Because text materials are well-organized and meaningful, it is possible that participants could easily engage in elaborative processing regardless of whether the initial test was open-book or closed-book, thereby reducing the difference between the two formats. The question is whether the effect of test format would emerge when the study material is less organized and less meaningful.

The present study consisted of three phases. During Phase I, participants were asked to study a list of Swahili-English word pairs. During Phase II, participants were provided with either additional opportunity to study (re-study) or they received an initial test of the material. During Phase III, which occurred one week later, participants received a final test that assessed their memory for the word pairs. The critical manipulation was that for half of the participants, the initial test was an open-book test, and for the other half of the participants, the initial test was a closed-book test.

Finally, in addition to the format of the test, the type of the test (cued recall versus recognition) was manipulated during Phase II when the initial test was administered. This manipulation was based on the assumption that the difference between open-book and closed-book tests may appear when an initial test is not sufficiently difficult to promote a deep level of processing (i.e., recognition). That is, when a test is sufficiently challenging to promote long-term learning, the test format may not matter, whereas when a test is not challenging enough, a closed-book test may show superiority over an open-book test. In sum, participants in the present study were asked to learn a list of Swahili-English word pairs followed by a cued-recall or recognition initial test, which was administered in an open-book or closed-book format. Also, there was a control condition in which participants were asked to re-study the list instead of taking the initial test. The final recognition test was administered one week later to examine whether the test format (open-book versus closed-book) as well as the test type (cued recall versus recognition) made a difference in long-term retention of the study material.

Methods

Participants

Participants were 39 male and 136 female undergraduate students attending introductory psychology courses at a public university in the Midwestern region of

the United States. They were offered extra course credit for their participation. An equal number (n = 35) of participants were randomly assigned to five between-subjects conditions based on the format (i.e., open-book versus closed-book) and type (i.e., cued recall versus recognition) of the initial quiz, plus a study twice control condition: (1) closed-book recognition, (2) closed-book cued recall, (3) open-book recognition, (4) open-book cued recall, and (5) study 2x (twice). Table 1 summarized the conditions and procedure. Note that for the rest of this paper, the initial test will be referred to as the 'initial quiz' and the final test will be referred to as the 'final test' to make it easy to differentiate these tests. The study was conducted in accordance with approval given by the Institutional Review Board at the university where participants were tested.

Materials

Fifty Swahili-English word pairs were selected from the Nelson and Dunlosky (1994) norms (see Appendix for examples). Based on the normative proportion of correct recall, these pairs were high in difficulty (ranged from .07 to .18). A PowerPoint presentation was used to present these pairs, one at a time in the middle of the computer screen in lowercase letters at the rate of one pair per 5 s. The order of the pairs was randomized once, and the same order was used for all participants.

The initial quiz was either cued recall or recognition. These quizzes were constructed by randomly selecting 35 Swahili words from the study list. Using 35 words rather than 50 words left 15 words for assessing performance on the one-week delayed final test when there was no initial quiz. For the cued recall quiz, these words were presented on a sheet of paper in a random order with a blank space next to each word for a response (e.g., *theluji* - _____). For the recognition quiz, these words were presented with four alternative choices of possible English translation for each Swahili word. The distractor choices were randomly selected from English translation of the other words in the study list, making it associative-recognition rather than item-recognition. Associative recognition was used to increase retrieval effort because unlike item-recognition, associative recognition depends more on retrieval than familiarity (e.g., Hockley & Consoli, 1999; Westerman, 2001). Each Swahili word was presented with four choices next to it, randomly ordered.

Table 1

Conditions and Procedure

Conditions	Session 1		Session 2 (One week later)
	Phase I	Phase II (Initial Quiz or Study)	Phase III (Final Test)
Closed-book Cued Recall	Study	Cued Recall (Closed-book)	Recognition (Closed-Book)
Closed-book Recognition	Study	Recognition (Closed-book)	Recognition (Closed-Book)
Open-book Cued recall	Study	Cued Recall (Open-book)	Recognition (Closed-Book)
Open-book Recognition	Study	Recognition (Open-book)	Recognition (Closed-Book)
Study 2x	Study	Study	Recognition (Closed-Book)

Note: During Phase I, participants studied 50 Swahili-English word pairs. During Phase II, participants were quizzed on 35 word pairs or re-studied 50 word pairs. During Phase III, participants were tested on all 50 word pairs.

The one-week delayed final test consisted of all 50 Swahili words from the study list, each presented with four alternative choices of possible English translation. A recognition test was used as the final test based on the assumption that recognition would provide more sensitive assessment of learning than recall. For the 35 Swahili words that were initially quizzed, the recognition items were the same as those in the initial recognition quiz. For the 15 Swahili words that were not initially quizzed, recognition items were constructed using English translation of other studied words as distractors. These items were randomly ordered once, and the same order was used across participants.

In addition, a sheet of paper with the study list (in the order of presentation) was used for the open-book quiz. A sheet of paper with random two-digit number was used for the filler task (see below), with a stopwatch used to time the duration of the filler task.

Procedure

Small groups up to four individuals were tested in two sessions with one-week delay between the sessions. During Session 1, Phase I and II of the study were administered. During Phase I, participants were instructed to study a list of 50 Swahili-English word pairs. The presentation of the study list was repeated three times to ensure that participants learned the list at a sufficiently high level to avoid a floor effect. They were not informed about the format of the initial quiz nor the final test they took after one-week delay. Following the study phase, participants were asked to perform a filler task for 2 minutes, crossing out the numbers divisible by three. The filler task was administered to eliminate a recency effect. Following the filler task, Phase II commenced, and participants in the initially quizzed condition completed the self-paced initial quiz. Participants in the cued recall condition were asked to write an English equivalent of each Swahili word, and participants in the recognition condition were asked to select the correct English equivalent of each Swahili word among the four alternatives. These quizzes were administered in a closed-book or open-book format; that is, participants in the closed-book condition were asked to take the quiz without looking up the answers whereas participants in the open-book condition were given a sheet of paper with the study list and were allowed to look up the answers. No feedback was given after completing the quiz. In the study 2x condition, instead of taking the initial quiz, participants re-studied all 50 Swahili-English word pairs printed on a sheet of paper one more time. At the end of the first session, participants in all conditions were told that at the second session seven days later, they would be asked about the Swahili words they studied during the first session.

Phase III of the study was administered in Session 2, which was scheduled one week after Session 1. During Phase III, participants took the self-paced final recognition test, with instruction to select the correct English equivalent for each

Swahili word. However, before taking the test, participants were asked to make a global judgment of learning (JOL), predicting how many words they would be able to correctly recognize among 50 Swahili words. Participants were asked to make a JOL rating because it is possible that the format and type of the initial quiz would influence their metacognitive judgments.

Results

The significance level was set at .05, and unless otherwise specified, two-tailed tests were performed. The dependent measures were the proportion of correct responses on the initial quiz, the proportion of correct responses on the final recognition test, and JOL, which was converted to a proportion. Note that on the initial quiz, participants were quizzed on 35 word pairs out of 50 word pairs they studied whereas on the final recognition test and JOL, participants were tested on all 50 word pairs. Because the goal of the study was to investigate the effect of the initial quiz format (i.e., open-book versus closed-book) as well as the initial quiz type (i.e., cued recall versus recognition) on final test performance, the proportion of correct responses on the final recognition test was compared across the conditions.

A one-way analysis of variance (ANOVA) on the final test performance for all 50 word pairs indicated that the difference among the conditions was significant, $F(4, 170) = 2.95$, $MSE = 0.03$, $p = .02$, $\eta_p^2 = .07$. As shown in Table 2, least significant difference (LSD) tests revealed that the study 2x condition ($M = .45$, $SD = .13$) showed significantly lower performance than the open-book cued recall ($M = .57$, $SD = .18$), open-book recognition ($M = .54$, $SD = .19$), and closed-book cued recall ($M = .57$, $SD = .17$) conditions. The difference between the study 2x condition ($M = .45$, $SD = .13$) and the closed-book recognition condition ($M = .52$, $SD = .14$) did not reach statistical significance with a two-tailed test ($p = .09$); however, based on *a priori* hypothesis that the initial testing would produce a testing effect, the difference was significant with a one-tailed test ($p = .04$). No other difference was significant, indicating that the testing effect was similar across the quizzed conditions.[1]

In order to gain insight as to how the testing effect had occurred, different groups of word pairs were analyzed. Because 35 word pairs out of 50 studied word pairs were quizzed on the initial quiz, these quizzed word pairs should show a testing effect on the final test, and that the effect should be similar across the quizzed conditions. This expectation was confirmed. A one-way ANOVA on the final test performance for 35 words that were quizzed on the initial quiz indicated that the difference among the conditions was significant, $F(4, 170) = 4.36$, $MSE = 0.03$, $p = .002$, $\eta_p^2 = .09$. As shown in Table 2, LSD tests indicated that the study 2x condition ($M = .44$, $SD = .14$) showed significantly lower performance than the open-book cued recall ($M = .59$, $SD = .18$), open-book recognition ($M = .55$, $SD = .19$), closed-book cued recall ($M = .57$, $SD = .18$), and closed-book recognition ($M = .53$, $SD = .15$) conditions. No other comparison was significant, indicating that the testing effect was similar across the quizzed conditions.

Next, 15 word pairs that were not quizzed on the initial quiz were analyzed to test a possibility that these non-quizzed words also showed a testing effect. This expectation was not confirmed. A one-way ANOVA on final recognition performance for 15 word pairs that were not quizzed on the initial quiz indicated that the difference among the conditions was not significant, $F(4, 170) = 0.83$, $MSE = 0.04$, $p = .51$, indicating that the testing effect was not observed with these words (see Table 2).

Table 2

Mean Proportion of Correct Responses on the Initial Quiz and on the Final Recognition Test as a Function of Initially Quizzed Conditions and Different Groups of Words on Final Test

Conditions		Initial Quiz (35 word pairs)	Final Test					
			All 50 Word Pairs	35 Quizzed Word Pairs	15 Not Quizzed Word Pairs	Correct on Initial Quiz	Incorrect on Initial Quiz	JOL
Closed-book Cued Recall	*M*	.36[a]	.58[a]	.57[a]	.58[a]	.82[a]	.48[a]	.29[a]
	SD	.19	.18	.18	.19	.16	.18	.19
Closed-book Recognition	*M*	.73[b]	.52[a]	.53[a]	.51[a]	.61[b]	.28[b]	.33[a]
	SD	.18	.14	.15	.14	.14	.21	.20
Open-book Cued Recall	*M*	.99[c]	.57[a]	.59[a]	.53[a]	.59[b]	-	.28[a]
	SD	.01	.18	.18	.24	.18	-	.14
Open-book Recognition	*M*	.98[c]	.54[a]	.55[a]	.52[a]	.55[b]	-	.30[a]
	SD	.04	.19	.19	.24	.19	-	.17
Study 2x	*M*	-	.45[b]	.44[b]	.50[a]	-	-	.29[a]
	SD	-	.14	.14	.18	-	-	.19

Note: For each column, significant differences ($p < .05$) are indicated by different superscript letters. All comparisons were based on two-tailed tests except for the comparison between the closed-book recognition and study 2x conditions for the final test with all 50 word pairs, which was based on a one-tailed test.

To investigate whether correctly responding on the initial quiz influenced the final test performance, the next analysis examined what proportion of the correct responses on the initial quiz was also correct on the final test. Because the initial quiz was not administered in the study 2x condition, this condition was excluded from the analysis. A one-way ANOVA indicated that the difference among the conditions was significant, $F(3, 136) = 19.23$, $MSE = 0.03$, $p < .001$, $\eta_p^2 = .30$. As shown in Table 2, LSD tests revealed that the closed-book cued condition ($M = .82$, $SD = .16$) showed significantly higher performance than the open-book cued recall ($M = .59$, $SD = .18$), open-book recognition ($M = .55$, $SD = .18$), and closed-book recognition ($M = .61$, $SD = .13$) conditions, indicating that successfully recalling without looking up answers on the initial quiz (i.e., closed-book cued recall) led to a higher success on the final test. No other comparison was significant.

An analysis was also performed to investigate whether the initial quiz had a positive

effect on the final test performance for those word pairs for which incorrect responses (i.e., omission and commission errors) were given on the initial quiz. However, in the open-book cued recall and recognition conditions, the number of such errors was close to zero, and therefore, a *t*-test was conducted to compare the closed-book cued recall and recognition conditions. Final recognition performance on the words participants failed to remember on the initial quiz indicated that the closed-book cued recall condition (M = .48, SD = .18) produced higher performance than the closed-book recognition condition (M = .28, SD = .21), t (66) = 4.23, p < .001, d = 1.02, on the final test. This finding shows that the participants were able to recognize items on the final tests that they did not remember on the initial quiz, and that cued recall on the initial quiz showed higher likelihood of such success than recognition.

Next, JOL was analyzed to investigate whether metacognition was influenced by the format and type of the initial quiz. A one-way ANOVA on JOL indicated that the difference among the conditions was not significant, F (4, 170) = 0.45, MSE = 0.03, p = .77. In all conditions, participants predicted that they would be able to correctly recognize 30 percent from 50 Swahili words they studied one week earlier (M = .30, SD = .18).

Lastly, a one-way ANOVA on the initial quiz performance showed that the difference among the conditions was significant, F (3, 136) = 175.01, MSE = 0.02, p < .001, η_p^2 = .79. As mentioned, the number of errors was closed to zero for the open-book cured recall (M = .99, SD = .01) and recognition tests (M = .98, SD = .04), and LSD tests showed that the difference was non-significant between these two conditions. All the other comparisons were significant, indicating that recognition was easier than cued recall when the initial quiz was closed-book.

Discussion

The present study examined whether open-book and closed-book formats of an initial quiz would influence performance on a delayed final recognition test when Swahili-English word pairs, as opposed to text materials, are used as study material. As mentioned in the introduction section, in the online learning environment, testing is challenging because it is difficult, if not impossible, to make a test cheat proof.[2] However, such a concern may only arise when a test is considered only as an assessment tool, consistent with the traditional view of education. In contrast, if the focus is shifted toward a long-term goal of education (i.e., building knowledge), it may not matter whether a test is open-book or closed-book because what matters most is whether students will develop knowledge of whatever they are learning. In fact, the past studies investigating the effect of test format using the testing effect paradigm showed that both open-book and closed-book formats produced similar performance on the final closed-book test, indicating that both formats would promote long-term memory (e.g., Agarwal et al., 2008; Agarwal & Roediger, 2011; Pauker, 1974; Gharib et al., 2012). However, a concern with these studies is that they used text-based materials, and therefore, it is possible that the lack of difference between open-book and closed-book formats was simply reflecting the fact that the materials were well-organized and

meaningful, thereby conducive to elaborative processing regardless of the format of the initial quiz. The present study, therefore, used a simple material (i.e., Swahili-English word pairs) in order to reduce the possible influence of other factors (such as organization, meaningfulness, and familiarity). Furthermore, the present study investigated the type of initial quiz (cued recall versus recognition) because the difference between the test formats might emerge when the initial quiz is not sufficiently difficult to induce a deep level of processing (i.e., recognition).

There were several major findings. First, the final test performance was similar among the four quizzed conditions, regardless of whether all 50 words from the study list (i.e., the whole list) or 35 words that were quizzed on the initial quiz were examined. The results, therefore, indicated that neither the initial quiz format (i.e., open-book versus closed-book) nor the initial quiz type (i.e., cued recall versus recognition) influenced the final test performance.

Second, the study 2x control condition produced significantly lower final recognition performance than the four quizzed conditions for both the whole list and the quizzed words. This finding is consistent with a phenomenon referred to as the testing effect (e.g., Roediger & Karpicke, 2006), indicating that taking a quiz increases long-term retention of information relative to re-studying. This notion is further supported by the finding that the testing effect was only found with the quizzed words, as opposed to the non-quizzed words (i.e., 15 control words that were not quizzed on the initial quiz).

Third, when the final test performance was conditionalized on correct responses on the initial quiz, the closed-book cued recall condition showed higher performance than the other quizzed conditions. This finding indicates that there is an advantage in making the initial quiz closed-book cued recall in line with the notion of desirable difficulty by Bjork (1994, 1999), which contends that difficult processing, whether it is at encoding or retrieval, benefits long-term retention.

Finally, JOLs of the final test was similar across conditions: In all conditions, including the study 2x condition, participants predicted that they would be able to correctly recognize about 30% of 50 Swahili words on the final test. Because actual performance was higher than 30% in all conditions, participants underestimated their performance. Although JOL ratings were not accurate in all conditions, it is important to note that JOLs were similar between the study 2x and quizzed conditions, indicating that JOLs did not show a testing effect. This finding is consistent with the result of other studies (e.g., Agarwal et al., 2008; Roediger & Karpicke, 2006b) that participants are not sensitive to the beneficial effect of testing. Furthermore, note that in many studies, participants show a tendency to overestimate rather than underestimate their performance (e.g., Agarwal et al., 2008; Ayton & McClelland, 1997). It is not clear the reason that participants underestimated their performance in the present study. A possibility is that the material they learned in this study was unfamiliar foreign vocabularies, and therefore, participants did not have high confidence.

Overall, the results of the present study showed that although performance on the

initial quiz was predictably higher in the open-book condition than in the closed-book condition, this advantage vanished after one-week delay, resulting in similar final test performance between the open-book and closed-book conditions. Furthermore, the final test performance did not show the effect of initial quiz type (i.e., cued recall and recognition). Taken together, an open-book test format is as effective as a closed-book test format in promoting long-term retention, even when the study material is not text materials.

In addition, the initial quiz could be either cued recall or recognition, indicating that in building knowledge, the act of being quizzed is the critical factor across quiz formats and quiz types. Why is it that the initial quiz format did not make a difference in the amount of testing effect? It is possible that even when the quiz was open-book, participants took it as if it was a closed-book quiz. If so, there was no functional difference between the two quiz formats. Although this is plausible, it is unlikely based on the initial quiz performance, showing that performance was almost 100% on the open-book quiz whereas performance was much lower on the closed-book quiz. Alternatively, it is possible that what is critical is the opportunity to process the words deeply regardless of whether it is done by open-book or closed-book quizzes. It appears that both open-book and closed-book quizzes acted as orienting questions in an incidental learning study (e.g., Craik & Tulving, 1975) inducing a deep level of processing. Further studies are needed to examine the processes that are induced by open-book and closed-book formats, which led to higher final test performance for the quizzed conditions relative to the study 2x condition.

Although the final test performance was similar across the quizzed conditions, there was an indication that a closed-book format with cued recall on the initial quiz may yield some advantage. As mentioned, consistent with the notion of desirable difficulty (Bjork, 1994, 1999), the final test performance was higher in the closed-book cued recall condition than in other quizzed conditions when the words that were correctly responded on the initial quiz were examined. This finding indicated that difficult retrieval would produce greater increment in storage strength than easy retrieval, in line with the notion that there is a negative correlation between storage and retrieval strength (Bjork & Bjork, 1992). It is possible that the difficulty of retrieval may ultimately prevail when memory is tested after a retention interval longer than one week. In fact, meta-analyses by Pan and Rickard (2018) and Rowland (2014) showed that retrieval effort or elaborative retrieval was a moderating variable that increased the testing effect. Note however that a meta-analysis by Adesope et al. (2017) showed that the testing effect was greater with less effort such that the testing effect was greater with recognition tests than with cued-recall tests. Accordingly, the role of retrieval difficulty in the testing effect is not clear, and therefore, further investigation is needed.

Another interesting finding was that when the words that were not correctly responded on the initial quiz (i.e., omission and commission errors) were examined, the closed-book cued recall condition showed higher final recognition performance than the closed-book recognition condition. However, this finding may simply reflect the fact that recognition is easier than recall, such that the increase in performance

from the initial quiz to the final test was as a result of comparing a difficult initial quiz (i.e., cued recall) and an easy final test (i.e., recognition).

Another issue that needs to be investigated in the future is the nature of the final test. In the present study, a recognition test was used to test final performance. However, it is possible that the results might be different when a cued recall test is used as the final test. The type of final test might be important because in the present study, the advantage of initial cued recall over recognition may have been masked due to a mismatch between the initial and final tests in the cued recall conditions. That is, a transfer appropriate processing (Morris, Bransford, & Franks, 1977) is confounded between the initial cued recall and recognition quiz conditions. However, ultimately how knowledge will be tested would be critically dependent on the type of knowledge and how it is used. For instance, learning a foreign language would require more than just recognizing vocabulary words and their English equivalents. In this sense, a decision to adopt a particular test type may require a domain specific approach. Nevertheless, the present study showed that it is premature to assume that a test is inferior just because it is administered using an open-book test format.

In conclusion, the results of the present study showed that the testing effect is similar between open-book and closed-book quizzes, even when the study material is an unrelated set of Swahili-English pairs, as opposed to well-organized and meaningful text materials. Furthermore, initial quiz type, cued recall or recognition, did not make a difference. These results, therefore, supported the notion that an open-book test is not necessarily inferior to a closed-book test in promoting long-term retention. However, there was an indication that making the initial quiz difficult, as in the closed-book cued recall condition, has an added advantage, which needs to be investigated in future research.

Practical Implications for Classroom Practice

Given that testing promotes long-term retention of studied material, coupled with the present result that there is no difference between open-book and closed-book formats, how can these research findings be translated to classroom practice? On the one hand, it seems to be safe to replace traditional closed-book tests with open-book tests if the purpose of education is to build knowledge. On the other hand, such practice would represent a radical departure from the traditional method, and as such, it may be difficult to convince teachers in traditional face-to-face classrooms to adopt such new practice. However, the situation may be different for teachers in online classes because these teachers may be more experienced with non-traditional methods. Nevertheless, doing away with closed-book tests entirely may not be practical because it would be difficult if not impossible to document the outcome of education unless there is an assessment.[3] Based on these considerations, a preferable approach would be to mix open-book tests and closed-book tests within a particular course (or a curriculum) with the former being used for building knowledge and the latter being used for assessment. In line with this recommendation, the second author of this paper began using this hybrid approach in a 300-level course at his university. In this face-to-face class,

he implemented open-book quizzes (multiple-choice questions), which he allowed students to take multiple times. He used these open-book quizzes for building knowledge, but for assessment, he used closed-book exams. The results were encouraging. For the semester when he implemented the quizzes, there was a modest increase in the average score of the closed-book exams (9% improvement overall) compared to prior semesters. Although the increase was not dramatic, the results were encouraging, given that the quizzes were not mandatory and did not contribute to the course grade. With this modest success, this hybrid approach should be explored further, particularly in online classes. We argue that this hybrid approach is especially relevant for online classes for at least two reasons. First, in these classes, in-person test proctoring can be expensive and time-consuming, and second, as mentioned earlier, other methods of minimizing the potential for cheating on online closed-book tests have limitations. Accordingly, by incorporating open-book quizzes, the number of closed-book tests can be reduced without jeopardizing the development of long-term knowledge. In conclusion, any laboratory finding requires extensive translational research before it becomes useful in practice. However, the results of the present study showed that open-book tests are a viable method of building knowledge, which we regard as one of the important goals of education.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

References

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, *87*, 659-701.

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J., & Wittrock, M. C. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of Educational Objectives* (Complete edition). New York: Longman.

Alessio, H. M., Malay, N., Maurer, K., Bailer, A. J., & Rubin, B. (2017). Examining the effect of proctoring on online test scores. *Online Learning*, *21*, 146-161.

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L. III, & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology, 22*, 861-876.

Agarwal, P. K., & Roediger III, H. L. (2011). Expectancy of an open-book test decreases performance on a delayed closed-book test. *Memory*, *19*, 836-852.

Ayton, P., & McClelland, A. G. R. (1997). How real is overconfidence? *Journal of Behavioral Decision Making, 10,* 279–285.

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.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. P Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185-205). Cambridge, MA: MIT Press.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (pp. 35-67). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives. The classification of educational goals. Handbook 1 Cognitive domain.* New York: Logmans, Green and Co.

Butler, A. C., Marsh, E. J., Goode, M. K., & Roediger, H. L., III. (2006). When additional multiple-choice lures aid versus hinder later memory. *Applied Cognitive Psychology, 20*, 941-956.

Christe, B. (2003). Designing online courses to discourage dishonesty. *Educause Quarterly, 26*, 54-58.

Cluskey Jr, G. R., Ehlen, C. R., & Raiborn, M. H. (2011). Thwarting online exam cheating without proctor supervision. *Journal of Academic and Business Ethics*, *4*, 1-7.

Craik, F. I. (2002). Levels of processing: Past, present... and future? *Memory*, *10*, 305-318.

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning & Verbal Behavior, 11*, 671-684.

Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*, 268-294.

Durning, S. J., Dong, T., Ratcliffe, T., Schuwirth, L., Artino, A. R., Boulet, J. R., & Eva, K. (2016). Comparing open-book and closed-book examinations: A systematic review. *Academic Medicine*, *91*, 583-599.

Gharib, A., Phillips, W., & Mathew, N. (2012). Cheat sheet or open-book? A comparison of the effects of exam types on performance, retention, and anxiety. *Psychology Research*, *2*, 469-478.

Grijalva, T. C., Nowell, C., & Kerkvliet, J. (2006). Academic honesty and online courses. *College Student Journal, 40*, 180-185.

Hockley, W. E., & Consoli, A. (1999). Familiarity and recollection in item and associative recognition. *Memory & Cognition, 27*, 657-664.

Metcalfe, J., Kornell, N., & Finn, B. (2009). Delayed versus immediate feedback in children's and adults' vocabulary learning. *Memory & Cognition*, *37*, 1077-1087.

Michael, T. B., & Williams, M. A. (2013). Student equity: Discouraging cheating in online courses. *Administrative Issues Journal*, *3(2).* Retrieved from https://www.swosu.edu/academics/aij/2013/v3i2/michael-williams.pdf

Moore, R., & Jensen, P. A. (2007). Do open-book exams impede long-term learning in introductory biology courses? *Journal of College Science Teaching, 36*, 46-49.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*, 519-533.

Nelson, T. O., & Dunlosky, J. (1994). Norms of paired-associate recall during multitrial learning of Swahili-English translation equivalents. *Memory*, *2*, 325-335.

Owens, H. S. (2016). *Cheating within online assessments: A comparison of cheating behaviors in proctored and unproctored environments* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3737143).

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin*, *144*, 710-756.

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*, 3-8.

Pauker, J. D. (1974). Effect of open book examinations on test performance in an undergraduate child psychology course. *Teaching of Psychology*, *1*, 71-73.

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

Rowe, N. C. (2004). Cheating in online student assessment: Beyond plagiarism. *Online Journal of Distance Learning Administration*, *7*. Retrieved from http://www.westga.edu/~distance/ojdla/summer72/rowe72.html

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*, 1432-1463.

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science, 333*, 476-478.

Westerman, D. L. (2001). The role of familiarity in item recognition, associative recognition, and plurality recognition on self-paced and speeded tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 723-732.

Footnotes

[1] We also conducted a 2 (quiz type: cued recall and recognition) x 2 (quiz format: open- book and closed-book) ANOVA without the study 2x condition. The results showed that no effect was significant.

[2] We acknowledge that the problem of cheating is not limited to online tests.

[3] We also acknowledge that assessments can be conducted in a variety of ways including (and not limited to) using closed-book tests.

Appendix

Examples of Swahili-English Word Pairs

Swahili Word	English Equivalent
Jani	Leaf
Chura	Frog
Lozi	Almond
Nira	Yoke
Wakili	Agent
Yatima	Orphan
Bahasha	Envelope
Chaza	Oyster
Fumbo	Mystery