Journal of Educational Psychology
2014, Vol. 106, No. 2, 435–447

0022-0663/14/$12.00 DOI: 10.1037/a0035715

Manipulations of Choice Familiarity in Multiple-Choice Testing Support a Retrieval Practice Account of the Testing Effect

Yoonhee Jang
University of Montana

Hal Pashler
University of California, San Diego

David E. Huber
University of Massachusetts, Amherst

We performed 4 experiments assessing the learning that occurs when taking a test. Our experiments used multiple-choice tests because the processes deployed during testing can be manipulated by varying the nature of the choice alternatives. Previous research revealed that a multiple-choice test that includes "none of the above" (NOTA) produces better performance on a subsequent test only when the correct answer is something other than NOTA (Odegard & Koen, 2007). However, when NOTA was an incorrect choice alternative, the correct answer was the only familiar alternative. Thus, familiarity may have allowed participants to identify the answer, which was then restudied. In other words, the testing benefit might have reflected a familiarity-guided restudy process rather than retrieval practice. In the current study, we examined the role of familiarity in the multiple-choice testing effect, manipulating the familiarity of alternatives. If NOTA was the correct answer, there was no testing benefit when the alternatives were all novel (Experiment 1) or all familiar (Experiment 3). Familiarity-guided restudy predicts memory impairment when there is a single familiar alternative for a NOTA-correct question. In contradiction to this hypothesis, there was a testing benefit for this condition (Experiments 2 and 4). Experiment 4 further collected metacognitive confidence ratings during the multiple-choice test, providing evidence of a recall-to-reject strategy for this condition. These results suggest that learning from multiple-choice tests is mainly due to retrieval practice rather than the use of familiarity.

Keywords: testing effect, multiple-choice test, none of the above, choice familiarity, retrieval practice

A test not only measures knowledge or skill but can also serve to strengthen that knowledge. The benefit of testing upon subsequent testing is called the *testing effect*—an item that is retrieved on the initial test is found to be strengthened as seen in performance on subsequent tests (e.g., Gates, 1917; Glover, 1989; Spitzer, 1939; for reviews, see Bjork, 1975; Dempster, 1996; Roediger & Karpicke, 2006). The testing effect has been found in various situations, with a variety of test formats as well as different types of material (for a recent review, see Roediger & Karpicke, 2006; see also Jang, Wixted, Pecher, Zeelenberg, & Huber, 2012). The current study used multiple-choice testing as the form of test practice, with learning assessed in a final cued-recall test. By manipulating the nature of the choice alternatives during multiple-choice testing, such as with the inclusion of "none of the above" (NOTA) as an option and with the inclusion of familiar versus novel alternatives, we investigated the processes deployed during testing to determine which ones support learning from testing.

This article was published Online First February 3, 2014.

Yoonhee Jang, Department of Psychology, University of Montana; Hal Pashler, Department of Psychology, University of California, San Diego; David E. Huber, Department of Psychology, University of Massachusetts, Amherst.

Correspondence concerning this article should be addressed to Yoonhee Jang, Department of Psychology, University of Montana, 32 Campus Drive, Missoula, MT 59812-1584. E-mail: yoonhee.jang@umontana.edu

Effects of multiple-choice tests are potentially complex, and there are different strategies that a test taker might deploy in a multiple-choice test. Thus, multiple-choice testing provides a unique opportunity for studying the processes that support learning from testing. For instance, one strategy is to choose the most familiar alternative, rather than engaging in effortful recall. A more effortful strategy for taking a multiple-choice test is to use the question as a probe and to attempt recall of the answer and then to search the alternatives to see if any of them match the retrieved answer (e.g., Park, 2005). These different strategies could potentially have different effects upon the learning that is revealed in a final cued-recall test.

Recently, Roediger and his colleagues reported a benefit of multiple-choice testing using nonfiction passages of general knowledge topics (e.g., Butler, Marsh, Goode, & Roediger, 2006; Butler & Roediger, 2008; Roediger & Marsh, 2005). They found that participants correctly answered more questions that were previously tested during the multiple-choice testing phase compared with questions that were not previously tested. Odegard and Koen (2007) extended the results of Roediger and Marsh (2005) by including NOTA as a choice alternative. After reading nonfiction passages, participants took an initial multiple-choice test that contained three different question formats: (a) the correct answer was offered without a NOTA option (target present); (b) NOTA was included as the correct choice (target absent); and (c) NOTA was included as one of the incorrect alternatives (target and NOTA present). They

found that the benefit of multiple-choice testing occurred only when the target was presented on the multiple-choice test (i.e., the first and last cases). These findings are analogous to the eyewitness testimony experiments of Schooler, Foster, and Loftus (1988) in which questions that did not contain the correct answer during an initial test produced impairment for a later test, compared with those that contained the correct answer.

The inclusion of NOTA is useful for several reasons. For questions that include NOTA, the test taker can adopt one of several different strategies, and these strategies may inform understanding of the testing effect. For instance, when NOTA appears as a choice alternative, the test taker may be more inclined to adopt a "recall-to-reject" strategy: They might attempt to recall any information associated with each alternative and use a mismatch between the recalled information and the question to reject that alternative (e.g., Gronlund & Ratcliff, 1989; Hintzman & Curran, 1994; for a review, see Clark & Gronlund, 1996), selecting NOTA after rejecting all other alternatives. More generally, the inclusion of NOTA introduces new processes into the intervening test and new ways in which learning during testing may involve different processes compared with a final test that does not include NOTA. Finally, the inclusion of NOTA is necessary for a properly controlled investigation into the role of familiarity in the multiple-choice testing effect—an investigation of familiarity requires trials for which a familiar alternative is the correct choice but also trials for which the familiar alternative is an incorrect choice. Because there needs to be a correct choice on all trials (otherwise, test takers could develop strategies that have nothing to do with specific knowledge for the items), trials that have only incorrect familiar alternatives must also present NOTA as the correct choice.

There are obvious applications of this research to education. For classes with many students, multiple-choice tests have become the norm in many educational settings. Therefore, the optimal method for constructing an effective multiple-choice test is of substantial interest. It was originally believed that the inclusion of NOTA as a choice alternative is a better method for assessing knowledge because it requires more thought (e.g., Boynton, 1950). However, other educators have argued that NOTA items are misleading (e.g., Haladyna & Downing, 1989a, 1989b; Knowles & Welch, 1992; Oosterhof & Coats, 1984) and should only be used when the information being tested is purely factual, such as with mathematics and historical dates (e.g., Frary, 1991). In light of this debate, there is a need for a better empirical understanding of the actual effects of taking NOTA tests and, more generally, of the role of familiarity in multiple-choice testing.

The Goal of the Present Study: The Role of Familiarity

The learning process that underlies the multiple-choice testing effect and the role of NOTA are still poorly understood. This is largely because it is unclear when and how test takers use familiarity to guide their choices, rather than engaging in effortful retrieval attempts (i.e., a testing benefit due to retrieval practice). Because no feedback is provided while taking a multiple-choice test, some form of selection is needed to guide the choice and the learning that underlies the testing benefit. This might involve choosing an alternative based solely on familiarity—the participant might assume that a highly familiar choice is correct and then learn that choice for future recall. In other words, the multiple-choice testing effect may be due to familiarity-guided restudy of the information rather than retrieval practice (for further discussion, see, e.g., Reder & Ritter, 1992). To test this hypothesis, we performed a series of experiments that manipulated the familiarity of the incorrect choice alternatives (henceforth called lures).

It is important to note that our use of familiarity in developing this familiarity-guided restudy hypothesis is not necessarily the same as the use of familiarity as it is commonly used in dual-process accounts of recognition (e.g., Atkinson & Juola, 1973; Mandler, 1980). Instead, we make a distinction between knowledge that a word was previously studied (i.e., familiar) versus knowledge for the tested association. For instance, the test taker may realize that one of the choice words was previously studied because they recall what they were thinking at the time that they initially viewed that word. However, if they fail to recall the tested association, then this memory retrieval only serves as an indication that the word is familiar (i.e., certainly a word that was previously studied). Although such a retrieval process would be considered recollection, it might nonetheless serve as the basis of familiarity-guided restudy because the selection of that word on the multiple-choice test was not based on retrieved knowledge for the tested association.

Figure 1 illustrates our basic paradigm in which participants studied a series of Swahili–English word pairs followed by a four-alternative multiple-choice test and then the final cued-recall test. Among the examples shown in this figure, Question A in-

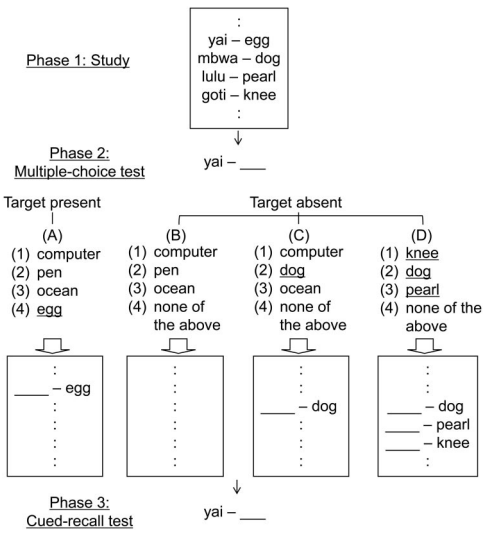

Figure 1. An example of four-alternative multiple-choice tests in the study of Swahili–English word pairs followed by a cued-recall test used in Experiments 1–4. Question A includes the target and three lures that were not seen during the study phase (target present). The other three questions include "none of the above" as the correct choice (target absent). Question B includes no familiar-choice alternatives: The three lures were not seen during the study phase (Experiment 1). Question C includes a single familiar-choice alternative: One of the three lures was seen in another word pair during the study phase (Experiments 2 & 4). Question D includes all familiar-choice alternatives: Each of the three lures was seen in another word pair during the study phase (Experiment 3). The familiar-choice alternatives are underlined.

cludes the target and three lures that were not seen during the study (target present), and therefore, there is only one familiar alternative (which is underlined in the figure). Not shown in the figure is a different target-present question for which the target is the sole familiar choice alternative, created by including both the target and NOTA as a lure (target and NOTA present). Similar to Question A, such a question also reveals testing benefits (Odegard & Koen, 2007). Question B shows a question where NOTA is the correct answer (target absent), and there are no familiar choice alternatives. Suggesting that target presence/absence is the key variable behind the multiple-choice testing effect, the target-absent condition of Odegard and Koen (2007), such as Question B, did not produce testing benefits. However, it is unknown whether there is a testing benefit for a target-absent question with familiar lures.

In the target-present questions with unfamiliar lures (such as Question A of Figure 1), there are at least two equally possible explanations for why a testing benefit might occur: (a) familiarity-guided restudy: the familiar choice alternative might be assumed during the multiple-choice phase to be the correct answer and used for additional study; and (b) retrieval practice: effortful retrieval using the most familiar choice alternative as a retrieval cue (i.e., rather than assuming that the most familiar alternative is correct, the test taker attempts to retrieve the associated Swahili word to see if it matches). These two explanations assume fundamentally different processes for the multiple-choice testing effect, considering that one is essentially the beneficial effect of restudying (because the target is given as a choice alternative) but that the other is the process thought to underlie most testing effects reported in the literature (for details, see Roediger & Karpicke, 2006). Nevertheless, the results of Odegard and Koen (2007) are compatible with both explanations. Thus, to differentiate between familiarity-guided restudy versus retrieval practice, we manipulated the familiarity of the lures, as explained next.

Consider Question C of Figure 1, which has not been examined in the literature. This is a target-absent question with a single familiar lure as well as the correct answer of NOTA. If participants use retrieval to check a familiar choice alternative (i.e., recall-to-reject), then this condition would produce a testing benefit even though it is a target-absent condition. This would occur if participants engage in backward retrieval based on the familiar lure, which could result in retrieving a different Swahili word than is currently tested (i.e., a mismatch). However, consider what would happen if familiarity is used to guide restudy (i.e., the familiar alternative is assumed to be correct, without retrieval-based checking). In this case, the familiar lure would be incorrectly learned, which would harm performance rather than help in the absence of feedback (e.g., McConnell & Hunt, 2007). Thus, familiarity would reduce memory accuracy for the final test compared with the control condition, and hereafter the reduced performance is called a testing cost (in contrast to a testing benefit). It is important to realize that these processes are not mutually exclusive—the test taker might use recall for rejection of some of the choice alternatives but also use familiarity to guide their final selection. Because a pure reliance on one process or the other predicts testing benefits versus testing costs for this condition, it can be used to determine which process contributes more substantially to learning from testing. Thus, while this condition cannot falsify the existence of either process, it can determine which one provides a better explanation.

The last question of Figure 1, Question D, is another type of target-absent question that has not been examined in the literature, and it tests a situation in which all the lures are familiar while NOTA is the correct answer. In this case, retrieval is expected to again produce a testing benefit for the same reason as Question C. However, because the lures are all equally familiar, familiarity-guided restudy might no longer apply (i.e., no choice is more familiar than the other choice alternatives).

Finally, we acknowledge that there may be other hypotheses beyond familiarity-guided restudy and retrieval practice. Nevertheless, these two hypotheses are the most obvious alternatives, and it will help identify the nature of the testing effect in general to determine whether the multiple-choice testing effect is better explained by one or the other of these two hypotheses.

Methodological Overview

There are two key findings in studies of the testing effect. First, the testing effect is often found even in the absence of feedback (e.g., Allen, Mahler, & Estes, 1969; Runquist, 1983; Wheeler & Roediger, 1992). Accordingly, all four experiments of this study did not provide feedback. Second, in the absence of feedback, test practice typically results in better performance on a final test than study practice when there is a delay between practice and the final test, whereas the opposite pattern is found for an immediate final test (e.g., Jang et al., 2012; Roediger & Karpicke, 2006; Thompson, Wenger, & Bartling, 1978). However, the goal of this study was to identify the processes that support learning from testing rather than to assess the effectiveness of different kinds of practice prior to a final test. Therefore, the current study did not include different interventions (test vs. study practice) and did not manipulate delay. Instead, all experiments used a control condition that did not include any form of practice prior to the final cued-recall test. This condition served as a baseline against which learning from multiple-choice testing was measured.

Table 1 presents an overview for all four experiments in this study. Experiment 1 sought to replicate the basic conditions of Odegard and Koen's (2007) study with a direct application to second language learning by using Swahili–English word pairs. For the items in the control condition of all experiments, neither the Swahili cue word nor the associated English word appeared during multiple-choice testing. All experiments used three experimental conditions: The "target" condition that included the target as the correct answer, the "NOTA" condition that included NOTA as the correct answer, and the "target + NOTA" condition that included the target as the correct answer and NOTA as a lure. Going beyond what has been previously examined, Experiments 2 and 4 used a single lure that was familiar in the NOTA condition (mixed familiarity), and Experiment 3 used all familiar lures in all three experimental conditions (equally familiar). Specifically, the NOTA condition is illustrated as Question B (Experiment 1), C (Experiments 2 and 4), and D (Experiment 3) of Figure 1. It is important to note that familiar lures were created by rearranging studied items (i.e., a rearrangement between a tested Swahili word and an English lure that belonged to another previously studied Swahili word) in such a way that all items were kept within the same experimental condition. Thus, any retrieval practice specific to the rejection of lures would apply to the experimental condition of interest.

Table 1
Number of Presentations of the English Words as a Choice Alternative (Target or Lure) During the Four-Alternative Multiple-Choice Test for Each Condition

Experiment	Target	NOTA	Target + NOTA	Control
1	× 1: target	Not presented	× 1: target	Not presented
2 & 4	× 1: target	× 1: lure in the NOTA condition	× 1: target	Not presented
3	× 1: target × 3: lure in the target condition	× 3: lure in the NOTA condition	× 1: target × 2: lure in the target + NOTA condition	Not presented

Note. NOTA = none of the above.

Experiment 1

Experiment 1 was designed to investigate the multiple-choice testing effect, using Swahili words paired with their English translations. Odegard and Koen (2007) found a benefit of multiple-choice testing when the correct answer was a previously studied item. However, because Odegard and Koen (2007) tested multiple facts from the same passage, it is possible that the testing benefit they observed reflected a progressive untangling of the associations contained in that passage over the course of the multiple-choice questions. In that case, it might not generalize to the current situation in which the multiple-choice questions tested separately learned associations rather than a common reading passage. To further reduce any relations between questions, in Experiment 1, we used Swahili words as cues that were initially devoid of meaning.

Method

Participants. Forty-one undergraduate students at the University of California, San Diego, were recruited and received credit for psychology courses in return for their participation.

Materials. Study stimuli for experimental trials consisted of 96 Swahili–English translation equivalents drawn from the norms of Nelson and Dunlosky (1994). According to the normative likelihood of the English word being recalled when the Swahili word was presented, each pair falls into one of the three difficulty categories: 32 easy (M = .26), 32 medium (M = .12), and 32 hard (M = .06) pairs. One-fourth of the 32 pairs per category were randomly assigned to each of the four conditions of the multiple-choice test: control, target, NOTA, and target + NOTA conditions. The 192 lures and the translation targets were moderately high frequency (an average of 80 times per million; norms from Kucera & Francis, 1967), singular nouns from four to eight letters in length. The practice trials used four different Swahili–English study pairs and 11 English lures. The same practice words were used for all participants.

Procedure. The experiment consisted of two blocks that each had three phases: learning Swahili–English word pairs, multiple-choice testing, and final cued-recall.

During Phase 1 of each block, participants studied 48 Swahili–English word pairs (16 randomly chosen from each of the three item-difficulty categories). The pairs were randomly assigned to conditions in such a way that they were used equally in all conditions across participants. The study phase had three presentation cycles (i.e., a total of three presentations per pair). Each pair was presented for 4 s once per presentation cycle, and list order was randomized anew for each presentation cycle. The study phase was followed by a simple math distractor task that took approximately 30 s.

During Phase 2 of each block, participants gave a response to 36 multiple-choice questions, with a question consisting of a single Swahili word and four choice alternatives. Responses were self-paced, and no feedback was given. The four conditions (target, NOTA, target + NOTA, and control) contained 12 study pairs (four from each of the three item-difficulty categories) each. The 36 trials, representing 12 occurrences of each experimental condition, appeared in a randomly determined order. As seen in Table 1, in the target condition, one of the four alternatives was the English translation associated with the Swahili word, which was the correct answer. In the NOTA condition, NOTA was the correct answer (the English translation associated with the Swahili word did not appear). In the target + NOTA condition, one of the alternatives was the English translation associated with the Swahili word, and NOTA appeared as a lure. The presentation order of the target in the target-present conditions was counterbalanced, and the NOTA alternative was always presented at the fourth position. Word pairs assigned to the control condition did not appear during this phase. All lures in all experimental conditions were new English words. After testing, there was again a math distractor task for approximately 30 s.

During Phase 3 of each block, participants were given self-paced cued-recall tests (i.e., given a Swahili cue and asked to recall the English translation) of all 48 word pairs, appearing in a randomly determined order. If they did not know the answer, then they were allowed to skip the test trial and proceed to the next trial. As with the multiple-choice test, no feedback was given.

Results

For each experiment, we first report correct cued-recall performance to investigate the multiple-choice testing effect. Then, we report different types of incorrect cued recall (error responses) as a function of condition. Finally, we report multiple-choice test performance. Throughout, statistical significance was determined with an alpha of .05, and estimates of effect size are reported as partial eta squared (η_p^2) for statistically significant effects. There were no main effects of block order and interactions relevant to block order, and so the data combined across blocks are reported hereafter.

Correct cued recall. Cued-recall accuracy of all experiments is reported in Table 2. To obtain a measure of the multiple-choice

Table 2
Mean Cued-Recall and Multiple-Choice Accuracy (Proportion Correct) for Each Condition

Test	Target	NOTA	Target + NOTA	Control
Cued recall				
Exp 1	.51 (.039)	.34 (.040)	.51 (.042)	.33 (.037)
Exp 2	.51 (.041)	.36 (.031)	.48 (.039)	.28 (.033)
Exp 3	.49 (.041)	.38 (.031)	.49 (.040)	.34 (.036)
Exp 4	.48 (.040)	.39 (.035)	.48 (.038)	.32 (.038)
Multiple choice				
Exp 1	.88 (.022)	.85 (.029)	.80 (.025)	
Exp 2	.89 (.023)	.75 (.031)	.81 (.026)	
Exp 3	.77 (.037)	.65 (.033)	.72 (.036)	
Exp 4	.91 (.017)	.74 (.035)	.85 (.024)	

Note. Standard errors of the mean are in parentheses. NOTA = none of the above; Exp = experiment.

testing effect, for each experiment, we calculated the difference between the proportion correctly recalled for the experimental condition and the proportion correctly recalled for the control condition (i.e., accuracy change from baseline).

Figure 2 (first panel) shows that the magnitude of the testing effect was significantly different across the three experimental conditions, $F(2, 80) = 46.08$, mean square error (MSE) = 0.01, $\eta_p^2 = .54$. The magnitude of the testing effect was greater for the target condition than for the NOTA condition, $t(40) = 8.54$, and greater for the target + NOTA condition than for the NOTA condition, $t(40) = 8.05$, but it did not differ between the two target-present conditions, $t(40) = 0.28$, $p = .78$. Furthermore, for the NOTA condition, the magnitude of the testing effect was not significantly different from zero, $t(40) = 0.50$, $p = .62$, which suggests no benefit from multiple-choice testing when NOTA is the correct option. These findings replicate Odegard and Koen's (2007) results, revealing testing benefits in the target-present conditions (i.e., the target and target + NOTA conditions), but not in the target-absent condition (i.e., the NOTA condition).

Incorrect cued recall. We examined the nature of the errors during final cued-recall. There were four types of error response on final cued recall: (a) previously studied English translations to different Swahili words, regardless of which condition that word belonged to (intrusions of other study words); (b) lures from any of the multiple-choice tests, regardless of which condition that lure belonged to (intrusions of multiple-choice lures); (c) new words that were not presented during the study and multiple-choice test phases; and (d) no response (omissions). For completeness, the descriptive statistics of all error response probabilities of all experiments are reported in Table 3; the first two types of errors (i.e., intrusions of previously seen words during study or multiple-choice testing) are of primary interest and are further discussed.

One of the critical findings from the previous studies in which nonfiction passages were used (e.g., Roediger & Marsh, 2005) was that prior testing also increased recall of multiple-choice lures as responses on the cued-recall test, compared with the control condition. More specifically, participants often produced lures that were incorrectly selected during multiple-choice testing: negative testing effects. In Experiment 1, however, the probability of recalling any lures from the initial multiple-choice test (regardless of whether they were selected or even from the same question or condition) was very low (see Table 3) and not significantly different across the conditions, $F(3, 120) < 1$. Moreover, the probability of recalling a lure that was mistakenly selected during multiple-choice testing was too low for additional analyses (collapsed across the three experimental conditions, $M = .005$, $SE = .002$). These results are consistent with other multiple-choice testing experiments that used word pairs (e.g., Whitten & Leonard, 1980).

Interestingly, the probability of recalling a word during the final cued-recall test that was one of the previously studied translations to a different Swahili word was substantially higher than the lure intrusion rate. This suggests that participants suffered more interference (i.e., more learning) during study of the word pairs than during multiple-choice testing. Furthermore, the intrusion rate of previously studied words differed across the conditions, $F(3, 120) = 6.65$, $MSE = 0.004$, $\eta_p^2 = .14$, being greater for the NOTA condition than for the target condition, $t(40) = 2.92$, or the target + NOTA condition, $t(40) = 4.75$. Additionally, there were more intrusion errors of other study words for the control condition than for the target + NOTA condition, $t(40) = 3.02$. The remaining comparisons were not significantly different, $ts < 1.73$, $ps > .09$. These differences between conditions likely reflect the different accuracy levels during final cued recall—in conditions with lower accuracy, there is a higher proportion of error trials and thus greater opportunity for producing this type of error.

Multiple-choice test performance. Multiple-choice accuracy of all experiments is found in Table 2. The proportion of correct responses to multiple-choice questions was significantly different across the three experimental conditions, $F(2, 80) = 6.40$, $MSE = 0.01$, $\eta_p^2 = .14$. Proportion correct was greater for the target condition than for the target + NOTA condition, $t(40) = 4.40$, and the remaining comparisons were not significantly different, $ts < 1.75$, $ps > .09$.

Discussion

Experiment 1 successfully replicated Odegard and Koen (2007) while generalizing the multiple-choice testing effect to a second-language learning situation. Both the target and the target + NOTA conditions produced testing benefits, but the NOTA condition did not, which suggests that the benefit of testing requires the presence of the target. However, it is unclear whether the presence of the target is necessary because the target is the correct answer or because the target is the only familiar alternative. According to the familiarity-guided restudy hypothesis, test takers might assume that the familiar alternative is the correct answer and then use this knowledge, without retrieval attempts, to restudy the association between the Swahili word and the English target. In Experiment 2, we tested this hypothesis by including a familiar alternative in the NOTA condition.

Experiment 2

Experiment 2 investigated whether a single familiar alternative in a multiple-choice test produces a testing benefit even if that alternative is incorrect. Both the familiarity-guided restudy hypothesis and retrieval practice are compatible with the results of Experiment 1 because the only familiar alternative in Experiment 1 was the correct answer—both restudying the association be-

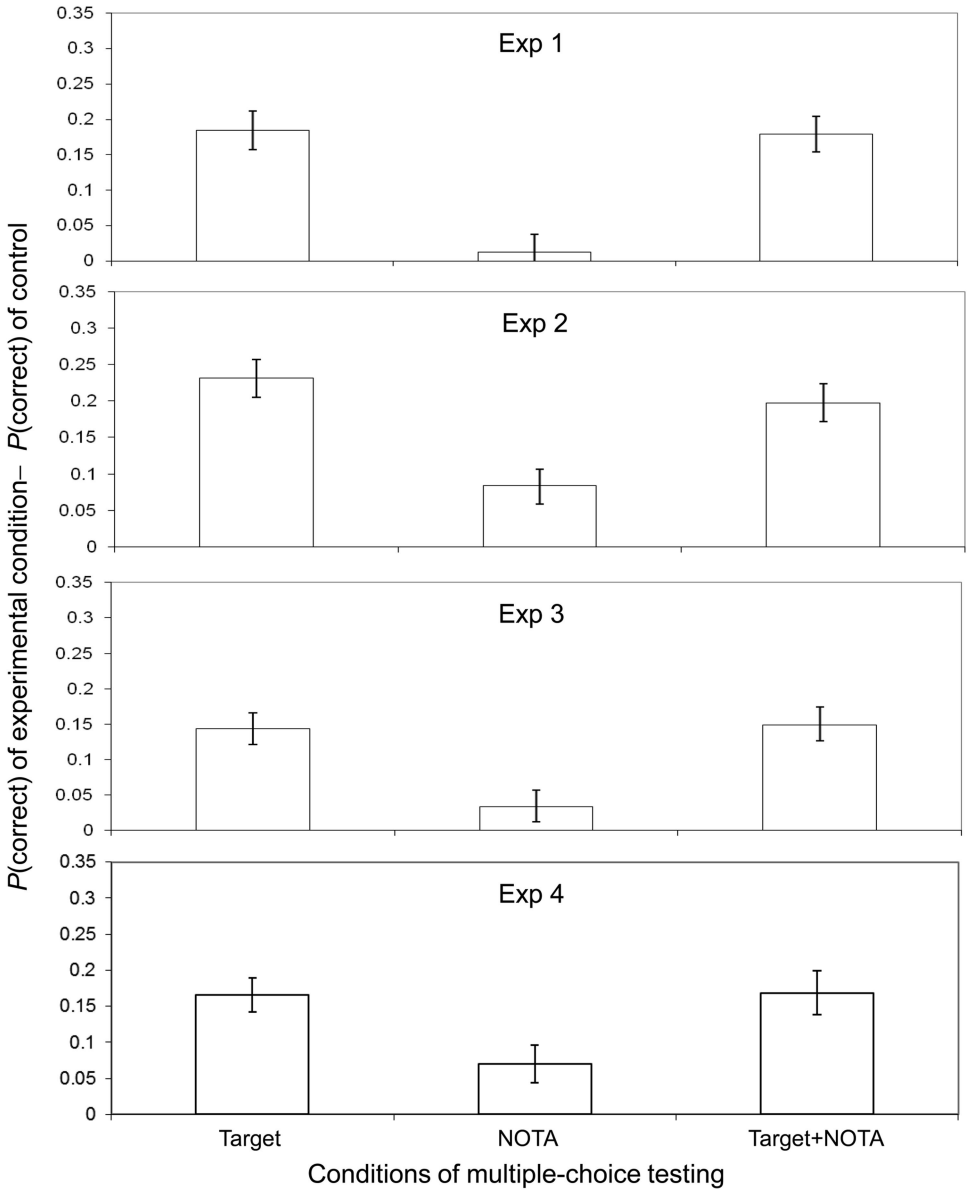

Figure 2. Cued-recall accuracy change from baseline. The *y* axis represents a measure of the multiple-choice testing effect, which is the difference in the proportion (*P*) correct between an experimental condition (target, none of the above [NOTA], and target + NOTA) and the control condition, which served as a baseline: A zero means that there is no benefit from multiple-choice testing. Error bars depict ±1 standard error of the mean. Exp = experiment.

tween the Swahili word and the familiar alternative and using the familiar alternative to retrieve the associated Swahili word would improve performance. In contrast to Experiment 1, all the conditions of Experiment 2 included exactly one familiar alternative. In other words, unlike Experiment 1, the NOTA condition of Experiment 2 included a familiar lure. If familiarity guides the choice during multiple-choice testing, this condition should produce a testing cost rather than a testing benefit. However, if test takers attempt to retrieve the Swahili word associated with the familiar alternative, this condition should produce a testing benefit even though the target is not presented as a choice alternative. This is predicted to be the case because the familiar lure in the NOTA condition is the translation to a different Swahili word from the same condition, and retrieval practice of the associated Swahili word should improve performance for that condition when the Swahili word is tested during the final cued-recall test.

Method

Participants. Forty-one undergraduate students at the University of California, San Diego, were recruited and received credit for psychology courses in return for their participation.

Table 3
Mean Incorrect Cued-Recall Rate as a Function of Error Type for Each Condition

Error type	Target	NOTA	Target + NOTA	Control
Other study words				
Exp 1	.11 (.017)	.15 (.020)	.09 (.017)	.13 (.018)
Exp 2	.12 (.019)	.17 (.023)	.13 (.019)	.19 (.024)
Exp 3	.18 (.030)	.22 (.028)	.18 (.025)	.23 (.031)
Exp 4	.19 (.025)	.24 (.028)	.20 (.026)	.26 (.033)
Lures				
Exp 1	.02 (.005)	.03 (.006)	.02 (.005)	.03 (.009)
Exp 2	.02 (.005)	.03 (.008)	.03 (.007)	.02 (.007)
Exp 3[a]				
Exp 4	.03 (.005)	.03 (.007)	.02 (.004)	.02 (.005)
New words				
Exp 1	.11 (.028)	.15 (.036)	.12 (.032)	.15 (.034)
Exp 2	.14 (.030)	.14 (.031)	.13 (.031)	.17 (.033)
Exp 3	.09 (.019)	.14 (.026)	.12 (.025)	.16 (.027)
Exp 4	.30 (.041)	.34 (.036)	.30 (.039)	.40 (.039)
No response				
Exp 1	.25 (.039)	.33 (.048)	.26 (.043)	.36 (.049)
Exp 2	.22 (.042)	.30 (.044)	.24 (.038)	.34 (.050)
Exp 3	.24 (.041)	.26 (.040)	.21 (.041)	.27 (.047)

Note. Standard errors of the mean are in parentheses. NOTA = none of the above; Exp = experiment.
[a] The lures used in Experiment 3 were selected from other study words (i.e., English translations associated with different Swahili words) in each experimental condition.

Materials. Experiment 2 used the same Swahili–English word pairs as Experiment 1. The lures on the multiple-choice test consisted of 168 English words randomly selected from the 192 lures used in Experiment 1.

Procedure. The procedure was identical to that of Experiment 1 except as noted. Each English translation of the Swahili words in the NOTA condition was randomly reassigned to a different Swahili word in the NOTA condition. This random reassignment was done without replacement such that each translation appeared exactly once as a lure to a different Swahili word in the NOTA condition. As a consequence, the NOTA condition consisted of one familiar and two new lures (mixed familiarity), as seen in Table 1. In the NOTA condition, the alternative presentation order of the familiar lure was counterbalanced, and the question order used as a lure was randomized.

Results

Correct cued recall. Figure 2 (second panel) shows that the magnitude of the testing effect differed across the three experimental conditions, $F(2, 80) = 25.85$, $MSE = 0.01$, $\eta_p^2 = .39$. The magnitude of the testing effect was greater for the target condition than for the NOTA condition, $t(40) = 6.23$, and greater for the target + NOTA condition than for the NOTA condition, $t(40) = 6.32$, but it did not differ between the two target-present conditions, $t(40) = 1.51$, $p = .14$. The relative ordering positions of these three conditions are the same as in Experiment 1. However, in contrast to Experiment 1, the magnitude of the testing effect for the NOTA condition was significantly above zero, $t(40) = 3.77$. Bearing in mind that the familiarity-guided restudy hypothesis predicted this condition to be below zero, whereas retrieval practice predicted this condition to be above zero, this result demonstrates that retrieval practice played a greater role in determining learning from testing.

Incorrect cued recall. As in Experiment 1, the probability of responding with a multiple-choice lure during final cued recall was very small (see Table 3) and did not differ across the conditions, $F(3, 120) = 2.02$, $MSE = 0.001$, $p = .12$. Additionally, the probability of recalling a lure that was incorrectly selected during multiple-choice testing was too small to analyze as a function of condition (collapsed across the three experimental conditions, $M = .01$, $SE = .003$). Especially, for the NOTA condition, the probability of recalling the familiar lure that was incorrectly selected during multiple-choice testing was also very small ($M = .02$, $SE = .004$).

The probability of responding with a previously studied English translation of a different Swahili word was significantly different across the conditions, $F(3, 120) = 6.96$, $MSE = 0.006$, $\eta_p^2 = .15$. This type of intrusion was greater for the NOTA condition than for the target condition, $t(40) = 3.08$, or the target + NOTA condition, $t(40) = 3.13$. It was greater for the control condition than for the target condition, $t(40) = 3.27$, or the target + NOTA condition, $t(40) = 2.92$. The rest of the comparisons were not significantly different, ts < 1.16, ps $> .26$. These findings were consistent with those of Experiment 1.

Multiple-choice test performance. As shown in Table 2, the proportion of correct responses to multiple-choice questions was significantly different across the three experimental conditions, $F(2, 80) = 19.04$, $MSE = 0.01$, $\eta_p^2 = .32$. Proportion correct was greater for the target condition than for the NOTA condition, $t(40) = 5.32$, or the target + NOTA condition, $t(40) = 5.14$;, and it was greater for the target + NOTA condition than for the NOTA condition, $t(40) = 2.37$.

Multiple-choice errors in the NOTA condition (.25) mainly reflected selection of the familiar lure ($M = .20$, $SE = .02$). However, as discussed earlier, participants rarely recalled these familiar lures during the final cued-recall test.

Discussion

Like Experiment 1, the target and target + NOTA conditions of Experiment 2 produced a substantial testing benefit. A benefit in these conditions was expected according to both the familiarity-guided restudy hypothesis and retrieval practice. However, unlike Experiment 1, the NOTA condition of Experiment 2 contained a single familiar lure. This manipulation allowed a critical test of the two hypotheses, which made opposite predictions. There was a significant testing benefit in this condition, which supports the retrieval practice account. Furthermore, as in Experiment 1, test takers rarely recalled one of the lures from the multiple-choice questions in the final cued-recall test regardless of whether the lure was previously seen or not, which provides additional evidence against the restudy hypothesis.

Experiment 3

The NOTA condition of Experiment 2 provided support for the claim that a familiar alternative on a multiple-choice test prompts test takers to engage in effortful retrieval, which can enhance later performance. In other words, by retrieving the Swahili word associated with the familiar English lure on a multiple-choice test,

not only do the test takers reject that choice during multiple-choice testing, but this retrieval makes it more likely for them to correctly recall that same English word when they are tested with the associated Swahili word during final cued recall. A remaining question is whether this retrieval attempt occurs only when there is a single familiar alternative (i.e., a highly likely candidate) or whether a retrieval attempt is initiated by any familiar alternative, even if all of the alternatives are familiar. To differentiate between these possibilities, in Experiment 3, we used familiar lures for all conditions. This experimental design is important not only for theoretical reasons but also because it mimics a real-world application of multiple-choice testing: Multiple-choice questions in education applications often consist of a choice between a correct answer versus lures that were also covered in class. Thus, Experiment 3 directly assesses the magnitude of the multiple-choice testing effect as might occur in a typical education setting.

Method

Participants. Forty-one undergraduate students at the University of California, San Diego, were recruited and received credit for psychology courses in return for their participation.

Materials. Experiment 3 used the same Swahili–English word pairs as Experiment 1.

Procedure. The procedure was identical to that of Experiment 1 except as noted. All lures on the multiple-choice test were selected from the other English translations of the Swahili words assigned to that condition. Thus, as seen in Table 1, during multiple-choice testing, an English word in the target condition appeared once as a target and three times as a lure for three other questions randomly selected in the target condition (because each question needed three lures). An English word in the NOTA condition was presented three times as a lure for three other questions randomly selected in the NOTA condition (because each question needed three lures). An English word in the target + NOTA condition appeared once as a target and twice as a lure for two other questions randomly selected in the target + NOTA condition (because each question needed two lures).

Results

Correct cued recall. As illustrated in Figure 2 (third panel), the magnitude of the testing effect was significantly different across the three experimental conditions, $F(2, 80) = 18.21$, $MSE = 0.01$, $\eta_p^2 = .31$. It was greater for the target condition than for the NOTA condition, $t(40) = 5.14$, and greater for the target + NOTA condition than for the NOTA condition, $t(40) = 5.07$; but it did not differ between the two target-present conditions, $t(40) = 0.25$, $p = .80$. These results are consistent with those of Experiments 1 and 2. In contrast to Experiment 2, but in accordance with Experiment 1, the magnitude of testing effect in the NOTA condition was not reliably different from zero, $t(40) = 1.50$, $p = .14$. Thus, when NOTA is the correct answer, there is no benefit of testing if the lures are equally familiar (Experiment 3) or equally unfamiliar (Experiment 1).

Incorrect cued recall. Unlike Experiments 1 and 2, all of the lures in Experiment 3 were English translations that were initially studied with different Swahili words (see Table 3). The intrusion rate of other study/lure words was significantly different across the conditions, $F(3, 120) = 4.84$, $MSE = 0.006$, $\eta_p^2 = .11$, being greater for the NOTA condition than for the target + NOTA condition, $t(40) = 2.86$, and being greater for the control condition than for the target condition, $t(40) = 2.60$, or the target + NOTA condition, $t(40) = 3.72$. The rest of the comparisons were not significantly different, $ts < 1.92$, $ps > .06$. As in Experiments 1 and 2, the probability of recalling a selected lure during final cued recall was too low to analyze as a function of condition (collapsed across the three experimental conditions, $M = .05$, $SE = .001$).

Multiple-choice test performance. As reported in Table 2, the proportion of correct responses to multiple-choice questions was significantly different across the three experimental conditions, $F(2, 80) = 15.73$, $MSE = 0.01$, $\eta_p^2 = .28$, being greater for the target condition than for the NOTA condition, $t(40) = 5.24$, or the target + NOTA condition, $t(40) = 2.32$, and being greater for the target + NOTA condition than for the NOTA condition, $t(40) = 3.47$.

Discussion

Experiment 3 produced the same pattern of results as Experiment 1. The testing effect was observed in the target and target + NOTA conditions, but not in the NOTA condition. Across the three experiments, there was no benefit of testing in the NOTA condition when the lures were equally unfamiliar or equally familiar.

Experiment 4

Experiment 2 was the only experiment that produced a testing benefit in the NOTA condition, and it was also the only experiment that included a single familiar alternative in multiple-choice test for this condition (compared with no familiar alternatives or all familiar alternatives). The implications of Experiment 2 should be considered cautiously as the Type I error rate for mistakenly finding a testing benefit in the NOTA condition among any of three different experiments is larger than the chosen alpha level of .05. Thus, it is prudent to replicate the results of Experiment 2. Therefore, Experiment 4 was designed to replicate the results of Experiment 2 while making two important changes to rule out some less interesting alternative explanations.[1]

First, unlike the previous experiments, participants in Experiment 4 were required to give a response during final cued recall (i.e., there were no errors of omission during cued recall). As seen in Table 3, omission rates were between 21% and 36% for the first three experiments (30% for the NOTA condition of Experiment 2), and the option of not responding might have interacted with the familiarity of the lures during multiple-choice testing.

Second, Experiment 4 collected metacognitive judgments on multiple-choice responses to directly assess the types of strategies that were deployed during multiple-choice testing. Many researchers in the memory literature assume that recollection and familiarity are separate processes that can contribute to recognition decisions (e.g., Atkinson & Juola, 1973; Mandler, 1980; for reviews, see Dunn, 2004; Wixted, 2007; Yonelinas, 2002). Recollection is the retrieval of specific details associated with the test

[1] We are grateful to Andrew Butler and an anonymous reviewer for bringing these points to our attention.

item, whereas familiarity reflects memory strength for the test item itself, and remember–know (R–K) judgments are often used to assess the relative contribution of these processes. However, R–K judgments would not be appropriate for multiple-choice questions, considering that test takers are confronted with more than one test item. Instead, we used metacognitive judgments that captured overall confidence (i.e., certainty about the selected item compared to the alternatives. Therefore, in Experiment 4, participants were asked to choose one of the three confidence judgment categories—*pure guess*, *educated guess*, and *certainly remember*—for each multiple-choice question. More specifically, test takers were instructed that certainly remember judgments should be used only when they recalled the English translation of the tested Swahili word but not when their selection was made only by rejecting the alternative choices. Instead, the rejection of choice alternatives should result in an educated guess judgment. If the testing effect in the NOTA condition with a single familiar lure reflected recall to reject for the familiar lure (with this recall boosting the cue–target association for the familiar lure), we expected to see a relatively higher proportion of educated guess judgments for correct multiple-choice questions in the NOTA condition compared with in the other conditions. This type of educated guess through recall to reject would not be possible in the target and target + NOTA conditions. In this analysis, we focused on the relative proportions of these judgments, rather than absolute proportions because we know from the results of Experiment 2 that multiple-choice accuracy was likely to be lower in the NOTA condition.

Method

Participants. Forty-one undergraduate students at the University of Montana were recruited and received $15 or credit for psychology courses in return for their participation.

Materials. Experiment 4 used the same Swahili–English word pairs as Experiment 1.

Procedure. The procedure was identical to that of Experiment 2 except as noted. Upon choosing an option for the multiple-choice question, participants were told to press the *1* key when their choice was a pure guess, the *2* key when their choice was an educated guess, or the *3* key when they certainly remembered the English translation associated with the Swahili word, with this resulting in a selection of the correct translation in the target and target + NOTA conditions or the selection of NOTA if they failed to see the translation among the choices. During the final cued-recall test, participants were told that they had to type in a word before continuing to the next trial.

Results

Correct cued recall. Figure 2 (fourth panel) shows that the magnitude of the testing effect differed across the three experimental conditions, $F(2, 80) = 14.60$, $MSE = 0.01$, $\eta_p^2 = .27$. The magnitude of the testing effect was greater for the target condition than for the NOTA condition, $t(40) = 4.32$, and greater for the target + NOTA condition than for the NOTA condition, $t(40) = 5.15$, but it did not differ between the two target-present conditions, $t(40) = 0.15$, $p = .88$. Importantly, the magnitude of the testing effect for the NOTA condition was significantly above zero, $t(40) = 2.71$, which is a successful replication of Experiment 2.

Incorrect cued recall. As seen in Table 3, the pattern of intrusions across the conditions was the same as with prior experiments, demonstrating that the prior results were not an artifact of test takers withholding guesses during cued recall. The forced response instructions greatly increased the intrusion rate of new words (more than double that of Experiment 2) and also increased the intrusion rate of other study words compared with Experiment 2 (an increase of about 7%). However, the intrusion rate of lures remained almost the same as found in the previous experiments. Specifically, the probability of recalling a multiple-choice lure during final cued recall was very small (see Table 3) and did not differ across the conditions, $F(3, 120) = 1.68$, $MSE = 0.001$, $p = .18$. Also, the probability of recalling a lure that was incorrectly selected during multiple-choice was too small to analyze as a function of condition (collapsed across the three experimental conditions, $M = .01$, $SE = .002$). For the NOTA condition, the probability of recalling the familiar lure that was incorrectly selected during multiple-choice testing was extremely small ($M = .01$, $SE = .003$).

The probability of recalling a previously studied English translation of a different Swahili word was significantly different across the conditions, $F(3, 120) = 7.96$, $MSE = 0.007$, $\eta_p^2 = .17$. This intrusion rate was greater for the NOTA condition than for the target condition, $t(40) = 3.33$, or the target + NOTA condition, $t(40) = 3.00$. It was also greater for the control condition than for the target condition, $t(40) = 3.60$, or the target + NOTA condition, $t(40) = 3.15$. The rest of the comparisons were not significantly different, $ts < 0.85$, $ps > .40$.

Multiple-choice test performance. As shown in Table 2, the proportion of correct responses to multiple-choice questions was significantly different across the three experimental conditions, $F(2, 80) = 26.88$, $MSE = 0.01$, $\eta_p^2 = .40$. Proportion correct was greater for the target condition than for the NOTA condition, $t(40) = 7.23$, or the target + NOTA condition, $t(40) = 3.90$;, and it was greater for the target + NOTA condition than for the NOTA condition, $t(40) = 3.72$.

Multiple-choice confidence judgments. Figure 3 shows the probability of giving each type of confidence judgment for correct

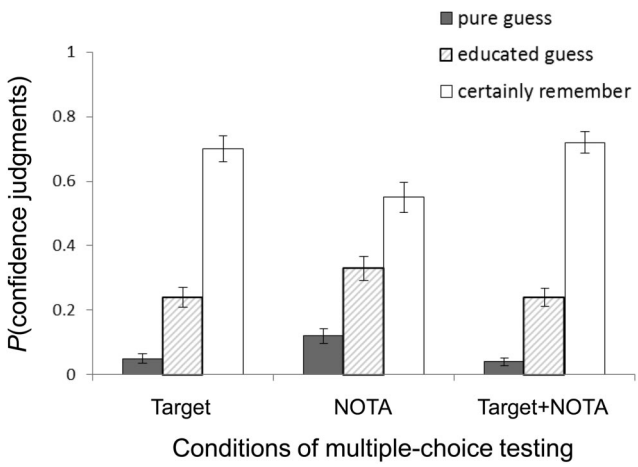

Figure 3. The relative proportion (*P*) of confidence judgments for correct multiple-choice responses in Experiment 4. The sum of the three bars for each condition equals 1.0. Error bars depict ±1 standard error of the mean. NOTA = none of the above.

multiple-choice responses. This is a conditional probability (conditioned on correct multiple-choice responses). It is important to use the conditional probability because the multiple-choice accuracy differed across the experimental conditions (thus confounding analysis of the joint probability).[2] By calculating the conditional probability, we ascertain the relative proportions of the different types of confidence judgments that occur for correct multiple-choice responses (i.e., the three types for each condition in the figure add up to 1.0). As seen in the figure, certainly remember judgments were made to more than half of correct multiple-choice responses.

As predicted, the proportion of educated guess judgments differed across the conditions, $F(2, 80) = 6.65$, $MSE = 0.02$, $\eta_p^2 = .14$. Particularly, it was greater for the NOTA condition than for the target condition, $t(40) = 2.89$, and greater for the NOTA condition than for the target + NOTA condition, $t(40) = 2.83$. There was no difference between the two target-present conditions, $t(40) = 0.42$, $p = .67$. These findings provide evidence that test takers used recall to reject the familiar lure, which led them to choose NOTA not because they had recalled the English translation but rather because they had rejected the only familiar alternative. The proportion of pure guess judgments also differed across the conditions, $F(2, 80) = 11.99$, $MSE = 0.005$, $\eta_p^2 = .23$. It was greater for the NOTA condition than for the target condition, $t(40) = 3.44$, and greater for the NOTA condition than for the target + NOTA condition, $t(40) = 3.91$. There was no difference between the target and target + NOTA conditions, $t(40) = 1.30$, $p = .20$. This is not surprising, considering that multiple-choice accuracy was lowest in the NOTA condition; even though only correct multiple-choice trials were considered in these analyses, it is more likely that test takers were correct by reason of pure guess in a condition with low accuracy. Finally, the proportion of certainly remember judgments also differed across the conditions, $F(2, 80) = 19.10$, $MSE = 0.02$, $\eta_p^2 = .32$. It was greater for the target condition than for the NOTA condition, $t(40) = 4.56$, and greater for the target + NOTA condition than for the NOTA condition, $t(40) = 5.16$. There was no difference between the two target-present conditions, $t(40) = 0.98$, $p = .34$. These findings are not surprising because in the NOTA condition, the target did not appear among the alternatives, and so even if test takers correctly recalled the English translation, they may have been uncertain about their retrieval (i.e., the lack of a target may have caused them to doubt their retrieval).

Discussion

The findings of Experiment 4 are consistent with those of Experiment 2, replicating the testing benefit in the NOTA condition, which contained a single familiar lure. This is notable not only for being a replication but also because this replication occurred despite forcing test takers to give a response on all cued-recall trials (i.e., the pattern of results is not an artifact of withholding guesses during cued recall). In addition, this experiment examined metacognitive confidence judgments in regard to the multiple-choice test, providing additional evidence that test takers used recall to reject the familiar lure in the NOTA condition. As predicted by a recall-to-reject strategy, the proportion of educated guess judgments for correct multiple-choice responses increased in the NOTA condition compared with the two target-present conditions.

General Discussion

Multiple-choice testing provides a unique opportunity for examining the processes that support learning during testing because the strategies deployed during testing can be manipulated by varying the nature of the choice alternatives on a multiple-choice test. In the current study, we manipulated the presence/absence of NOTA as a choice alternative, and we also manipulated, across experiments, the familiarity of the choice alternatives. In addition to shedding new light on the testing effect, this study is useful from an applied perspective. Multiple-choice testing has become ubiquitous in education, but little is known about the internal processes that test takers deploy when answering multiple-choice questions and the consequences for learning. In the current study, these processes were investigated. We examined a realistic learning situation in four experiments involving learning of foreign vocabulary (e.g., Swahili–English word pairs), reaching the conclusion that learning during testing primarily occurs through retrieval practice. Furthermore, retrieval practice appears to be occurring when there is a single standout familiar choice alternative regardless of whether that choice is correct.

Previous research (e.g., Odegard & Koen, 2007) found enhanced performance on a later cued-recall test following multiple-choice questions for which the correct answer was a previously studied item (target present) but not when the correct answer was NOTA (target absent). However, in this situation, the incorrect choice alternatives were all novel, whereas only the correct choice alternative was familiar in the target-present conditions. Thus, test takers could potentially have used familiarity to guide their choice and then engaged in restudy of the information, leading to enhanced performance on a later test. On this account, learning from a multiple-choice test would reflect a benefit from restudy rather than retrieval practice. For questions where NOTA was the correct answer, we tested this hypothesis by replicating the results when the incorrect choice alternatives were all novel (Experiment 1) and by additionally examining situations where the incorrect choice alternatives were all familiar (Experiment 3) or a mixture, with just one familiar alternative (Experiments 2 and 4). We consistently found a testing benefit in the NOTA condition only when one of the incorrect choice alternatives was familiar. If the only type of learning from testing was familiarity-guided restudy, then there should have been a testing cost rather than a testing benefit when the only familiar item was an incorrect choice.

Instead of the familiarity-guided restudy hypothesis, we suggest that the multiple-choice testing benefit reflects some form of retrieval practice. Specifically, we propose that test takers generally use the stated question to attempt to retrieve the answer. This enhances later performance on cued recall but only if the correct answer is among the choice alternatives. It may be that this form of retrieval practice operates even when the target does not appear among the alternatives. However, the failure to find a match may

[2] For completeness, the joint probabilities are reported in Appendix Table A1 where the data are further broken down by final cued-recall accuracy, such that the sum of the six values for each column equals the multiple-choice accuracy of each condition.

make the test taker uncertain as to the accuracy of the retrieved information. In other words, confirmation is needed (e.g., error corrective feedback; see Carrier & Pashler, 1992). This explains the testing benefit for questions that included the correct item and also the lack of a testing benefit when NOTA was the correct answer, provided that the choice alternatives were all novel or all familiar. In addition to this form of retrieval, we suggest that when there is a single familiar choice alternative (i.e., a clear standout), test takers focus on that alternative and use it to attempt retrieval in the backward associative direction (e.g., using a familiar English word to attempt to retrieve the Swahili word), and this retrieval practice underlies the testing benefit for the NOTA condition of Experiments 2 and 4. The multiple-choice confidence judgments of Experiment 4 supported this claim, revealing that the proportion of educated guess judgments for correct multiple-choice responses was greater for the NOTA condition than for the target-present conditions; if test takers focused on the familiar lure and recalled that the associated Swahili word was something other than the Swahili word on that multiple-choice question, they could rule out that choice alternative and thus make an educated guess judgment that NOTA was the correct answer. Because this rejection requires retrieval of the associated Swahili word, there would be retrieval practice for the rejected choice alternative, which would enhance retrieval of that word during the final cued-recall test when prompted with its Swahili equivalent. This explains the testing benefit in the mixed familiarity NOTA condition of Experiments 2 and 4. In theory, this same recall-to-reject process might have been deployed in Experiment 3, which presented multiple familiar lures on every multiple-choice test trial. That this did not occur may reflect some sort of optimal decision-making strategy based on cost/benefit calculation by the participants when confronted with multiple familiar alternatives (and if the costs they sought to minimize included effort, then this hypothesis might be summarized as saying that participants may have been showing laziness).

While the specific assumptions of this retrieval practice account await additional experimentation, it is clear from our results that a strong form of the familiarity-guided restudy hypothesis is insufficient—if test takers always assumed that a single familiar alternative was correct and then used this knowledge to restudy, we would have observed a testing cost rather than a testing benefit for the NOTA condition of Experiments 2 and 4. However, there is good reason to assume that some form of restudy occurs during multiple-choice testing. For instance, Butler and Roediger (2007), and Kang, McDermott, and Roediger (2007) reported that restudying versus taking a multiple-choice test improved final recall to the same degree. These findings may explain why there was a larger testing benefit in the target-present conditions compared with in the NOTA condition of Experiments 2 and 4. However, the failure to see testing costs in the NOTA condition of Experiments 2 and 4 suggests that test takers are unwilling to engage in restudy simply on the basis of familiarity. Instead, the application of familiarity-guided restudy in multiple-choice testing may be limited to situations where the familiar choice alternative provides confirmation regarding the answer recalled in response to the question.

In brief, our results have both theoretical and practical implications. Theoretically, our results are consistent with the claim that learning from multiple-choice testing primarily occurs from retrieval practice. Practically, our results suggest that including a single familiar alternative on a multiple-choice test, whether that alternative is correct or incorrect, promotes additional learning even when the correct answer is NOTA. If one goal of a test is to promote additional learning, then it is best to avoid easy multiple-choice tests where a choice alternative can be correctly chosen simply by virtue of being familiar but also to avoid overly difficult multiple-choice tests where all of the alternatives are familiar.

References

Allen, G. A., Mahler, W. A., & Estes, W. K. (1969). Effects of recall tests on long-term retention of paired associates. *Journal of Verbal Learning and Verbal Behavior, 8,* 463–470. doi:10.1016/S0022-5371(69)80090-3

Atkinson, R. C., & Juola, J. F. (1973). Factors influencing the speed and accuracy of word recognition. In S. Kornblum (Ed.), *Attention and performance IV* (pp. 583–612). New York, NY: Academic Press.

Bjork, R. A. (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123–144). Hillsdale, NJ: Erlbaum.

Boynton, M. (1950). Inclusion of "none of the above" makes spelling items more difficult. *Educational and Psychological Measurement, 10,* 431–432.

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. doi:10.1002/acp.1239

Butler, A. C., & Roediger, H. L., III. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19,* 514–527. doi:10.1080/09541440701326097

Butler, A. C., & Roediger, H. L., III. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36,* 604–616. doi:10.3758/MC.36.3.604

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20,* 633–642. doi:10.3758/BF03202713

Clark, S. E., & Gronlund, S. D. (1996). Global matching models of memory: How the models match the data. *Psychonomic Bulletin & Review, 3,* 37–60. doi:10.3758/BF03210740

Dempster, F. N. (1996). Distributing and managing the conditions of encoding and practice. In E. L. Bjork & R. A. Bjork (Eds.), *Memory* (pp. 317–344). San Diego, CA: Academic Press. doi:10.1016/B978-012102570-0/50011-2

Dunn, J. C. (2004). Remember–know: A matter of confidence. *Psychological Review, 111,* 524–542. doi:10.1037/0033-295X.111.2.524

Frary, R. B. (1991). The none-of-the-above option: An empirical study. *Applied Measurement in Education, 4,* 115–124. doi:10.1207/s15324818ame0402_2

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology, 6*(40), 1–104.

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81,* 392–399. doi:10.1037/0022-0663.81.3.392

Gronlund, S. D., & Ratcliff, R. (1989). Time course of item and associative information: Implications for global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 846–858. doi:10.1037/0278-7393.15.5.846

Haladyna, T. M., & Downing, S. M. (1989a). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2,* 37–50. doi:10.1207/s15324818ame0201_3

Haladyna, T. M., & Downing, S. M. (1989b). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2,* 51–78. doi:10.1207/s15324818ame0201_4

Hintzman, D. L., & Curran, T. (1994). Retrieval dynamics of recognition and frequency judgments: Evidence for separate processes of familiarity

and recall. *Journal of Memory and Language, 33,* 1–18. doi:10.1006/jmla.1994.1001

Jang, Y., Wixted, J. T., Pecher, D., Zeelenberg, R., & Huber, D. E. (2012). Decomposing the interaction between retention interval and study/test practice: The role of retrievability. *Quarterly Journal of Experimental Psychology, 65,* 962–975. doi:10.1080/17470218.2011.638079

Kang, S. H. K., McDermott, K. B., & & Roediger, H. L., III. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology, 19,* 528–558. doi:10.1080/09541440601056620

Knowles, S. L., & Welch, C. A. (1992). A meta-analytic review of item discrimination and difficulty in multiple-choice items using "none-of-the-above". *Educational and Psychological Measurement, 52,* 571–577. doi:10.1177/0013164492052003006

Kucera, H., & Francis, W. H. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University.

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review, 87,* 252–271. doi:10.1037/0033-295X.87.3.252

McConnell, M. D., & Hunt, R. R. (2007). Can false memories be corrected by feedback in the DRM paradigm? *Memory & Cognition, 35,* 999–1006. doi:10.3758/BF03193472

Nelson, T. O., & Dunlosky, J. (1994). Norms of paired-associate recall during multitrial learning of Swahili–English translation equivalents. *Memory, 2,* 325–335. doi:10.1080/09658219408258951

Odegard, T. N., & Koen, J. D. (2007). "None of the above" as a correct and incorrect alternative on a multiple-choice test: Implications for the testing effect. *Memory, 15,* 873–885. doi:10.1080/09658210701746621

Oosterhof, A. C., & Coats, P. K. (1984). Comparison of difficulties and reliabilities of quantitative word problems in completion and multiple-choice item formats. *Applied Psychological Measurement, 8,* 287–294. doi:10.1177/014662168400800305

Park, J. (2005). Learning in a new computerized testing system. *Journal of Educational Psychology, 97,* 436–443. doi:10.1037/0022-0663.97.3.436

Reder, L. M., & Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 435–451. doi:10.1037/0278-7393.18.3.435

Roediger, H. L., III, & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1,* 181–210. doi:10.1111/j.1745-6916.2006.00012.x

Roediger, H. L., III, & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 1155–1159. doi:10.1037/0278-7393.31.5.1155

Runquist, W. N. (1983). Some effects of remembering on forgetting. *Memory & Cognition, 11,* 641–650. doi:10.3758/BF03198289

Schooler, J. W., Foster, R. A., & Loftus, E. F. (1988). Some deleterious consequences of the act of recollection. *Memory & Cognition, 16,* 243–251. doi:10.3758/BF03197757

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology, 30,* 641–656. doi:10.1037/h0063404

Thompson, C. P., Wenger, S. K., & Bartling, C. A. (1978). How recall facilitates subsequent recall: A reappraisal. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 210–221. doi:10.1037/0278-7393.4.3.210

Wheeler, M. A., & Roediger, H. L., III (1992). Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science, 3,* 240–245. doi:10.1111/j.1467-9280.1992.tb00036.x

Whitten, W. B., II, & Leonard, J. M. (1980). Learning from tests: Facilitation of delayed recall by initial recognition alternatives. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 127–134. doi:10.1037/0278-7393.6.2.127

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114,* 152–176. doi:10.1037/0033-295X.114.1.152

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46,* 441–517. doi:10.1006/jmla.2002.2864

Appendix

Experiment 4 Findings

Table A1
Proportion of Trials With Correct Multiple-Choice Responses With the Indicated Confidence Judgment and Indicated Final Cued-Recall Accuracy in Experiment 4

Response	Target	NOTA	Target + NOTA
Correct			
Pure guess	.003 (.002)	.01 (.003)	.004 (.002)
Educated guess	.04 (.013)	.04 (.009)	.03 (.010)
Certainly remember	.44 (.040)	.30 (.036)	.44 (.038)
Incorrect			
Pure guess	.04 (.012)	.06 (.012)	.03 (.008)
Educated guess	.18 (.022)	.17 (.019)	.15 (.018)
Certainly remember	.21 (.020)	.15 (.020)	.19 (.015)

Note. Standard errors of the mean are in parentheses. NOTA = none of the above.

Table A1 reports the metacognitive judgments from Experiment 4 as the proportion of trials in each condition that resulted in correct multiple-choice response that also had the indicated level of metacognitive judgment and the indicated level of cued-recall accuracy. In other words, summing up a column of Table A1 produces the multiple-choice accuracy rate for each of the experimental conditions. The results from the 3 × 3 (multiple-choice testing condition and confidence judgment category) analyses of variance for final cued-recall accuracy are reported in Table A2.

Table A2
Results From the 3 × 3 (Multiple-Choice Testing Condition and Confidence Judgment Category) Analyses of Variance for Final Cued-Recall Accuracy in Experiment 4

		Correct cued-recall				Incorrect cued-recall			
Variable	*df*	*F*	*MSE*	*p*	η_p^2	*F*	*MSE*	*p*	η_p^2
Condition (C)	2, 80	23.85	.003		.37	2.58	.006	.08	
Judgment (J)	2, 80	100.28	.06		.72	32.08	.02		.44
C × J	4, 160	24.53	.004		.38	3.52	.007		.08

Note. Effect size (η_p^2) is reported only when the *F* value was significant. Multiple-choice testing conditions = target, NOTA, and target + NOTA; confidence judgment categories = pure guess, educated guess, and certainly remember. df = degrees of freedom; MSE = mean square error.

Received January 26, 2013
Revision received December 2, 2013
Accepted December 18, 2013 ■