American Psychological Association

Journal of Experimental Psychology: Applied

ISSN: 1076-898X

https://doi.org/10.1037/xap0000363

True–False Tests Enhance Retention Relative to Rereading

Oyku Uner, Eylul Tekin, and Henry L. Roediger III
Department of Psychological and Brain Sciences, Washington University in St. Louis

Testing with various formats enhances long-term retention of studied information; however, little is known whether true–false tests produce this benefit despite their frequent use in the classroom. We conducted four experiments to explore the retention benefits of true–false tests. College students read passages and reviewed them by answering true–false questions or by restudying correct information from the passages. They then took a criterial test 2 days later that consisted of short-answer questions (Experiments 1 and 2) or short-answer and true–false questions (Experiments 3 and 4). True–false tests enhanced retention compared to rereading correct statements and compared to typing those statements while rereading (the latter in a mini meta-analysis). Evaluating both true and false statements yielded a testing effect on short-answer criterial tests, whereas evaluating only true statements produced a testing effect on true–false criterial tests. Finally, a simple modification that asked students to correct statements they marked as false on true–false tests improved retention of those items when feedback was provided. True–false tests can be an effective and practical learning tool to improve students' retention of text material.

Public Significance Statement

This study shows that true–false quizzes help students retain more information on a later test compared to passive restudy, when students get feedback on their quizzes. Importantly, these quizzes do not only improve memory on later true–false tests, but also on short-answer tests. This study also suggests that a possible method to increase the effectiveness of true–false quizzes is asking students to try correcting true–false questions they consider to be "false."

Keywords: retrieval practice, testing effect, true–false questions

Retrieval of previously encountered information improves long-term retention of that information relative to restudying—a finding known as the "testing effect" (Roediger & Karpicke, 2006, see Rowland, 2014, for a meta-analytic review)—and retrieval practice is now widely recommended as an effective strategy for students and instructors (Dunlosky et al., 2013; Miyatsu et al., 2018; Pashler et al., 2007; Putnam et al., 2016). Given its advantage over restudying, however, numerous questions arise regarding how to implement testing, such as when practice tests should occur, how many tests are optimal, and whether the format of the test matters, to name a few. Test format, in particular, is pertinent to the effectiveness of retrieval practice, as it varies widely depending on the learning context. A large body of research has found testing effects with different practice test formats—namely, free recall, cued recall, short-answer, multiple-choice, and recognition tests (e.g., Butler & Roediger, 2007; Carpenter & DeLosh, 2006; Glover, 1989; Hogan & Kintsch, 1971; Kang et al., 2007; Roediger & Karpicke, 2006)—yet, the sizes of observed testing effects differ across different formats (Rowland, 2014). Therefore, one might simply conclude that test formats producing larger testing effects would be more effective tools for learning.

Nonetheless, when adopting testing as a study strategy, practicality is undoubtedly important for both students and instructors. For example, although short-answer tests with feedback frequently promote greater retention of targeted information than multiple-choice tests do (Kang et al., 2007; Little et al., 2012; Rowland, 2014), administering short-answer tests is harder and scoring them is time-consuming. Multiple-choice tests, on the other hand, are easier to administer and score, preferred by students (Zeidner, 1987), and frequently used by instructors. Therefore, understanding the potential benefits of test formats that students and instructors are more willing to use is imperative. Additionally, exploring whether slight modifications of these test formats can increase their retention benefit is an important applied issue. As an example, several studies using multiple-choice tests have shown that presenting competitive lures to increase question difficulty (Little et al., 2012) or providing correct-answer feedback to prevent the acquisition of misinformation (Butler & Roediger, 2008) can increase the effectiveness of these practice tests (see Butler, 2018, for multiple-choice tests best practices).

Oyku Uner https://orcid.org/0000-0002-1689-9095
Henry L. Roediger III https://orcid.org/0000-0002-3314-2895

The data reported in this manuscript are available on the Open Science Framework repository (https://osf.io/vtswx/). We thank Kassandra Diaz, Cameron Perrin, Elise Grever, Kirby Knapp, and Angela Qiu for their help with data collection and scoring. We also thank Andrew Butler for providing us with materials and for his helpful feedback on the project. Some of the results of this study were presented in the Annual Meetings of the Psychonomic Society (2018 and 2019) and the Midwestern Psychological Association Conference (2019).

Correspondence concerning this article should be addressed to Oyku Uner, Department of Psychological and Brain Sciences, Washington University in St. Louis, One Brookings Drive, St. Louis, MO 63130-4899, United States. Email: uner@wustl.edu

In the present study, we investigated whether answering true–false questions on previously learned material improves long-term retention relative to restudying the material. True–false tests may be particularly advantageous when instructors cannot easily create more than one competitive alternative to the correct answer on a multiple-choice question (Burton, 2001). Similar to multiple-choice tests, true–false tests are objective measures of performance, and they are easy to administer and score. Thus, if true–false tests enhance long-term retention as do other test formats, adopting them as learning tools would be both effective and practical.

Unfortunately, researchers have been mostly concerned with the assessment value of true–false tests rather than their effect on learning. Specifically, they have discussed the advantages and disadvantages of the true–false test, its reliability and validity, and how guessing can be corrected and discouraged (e.g., Burton, 2002; Ebel, 1970; Frisbie & Becker, 1991; Kellogg & Payne, 1938). Some researchers argued that students could easily blindly guess on a true–false test and that this guessing makes the test unreliable; however, others have argued that students can make informed guesses with partial knowledge and that guessing can be corrected (Burton, 2002, 2005). In addition, some researchers stated that true–false tests can only assess factual knowledge, limiting its validity as an assessment tool, yet others claimed that these tests could be constructed to assess different types of knowledge (Ebel, 1970). Despite these contrasting perspectives, however, empirical evidence on this matter is scarce (Burton, 2005; Frisbie & Becker, 1991) and true–false tests are still commonly used in the classroom.

Among the few studies that examined the effects of true–false tests on learning, most focused on their potential negative consequences. In one experiment, for example, college students first took true–false tests on passages they read (Toppino & Brochin, 1989). One week later, they were given a list of statements and were asked to rate them on a scale that ranged from *definitely false* to *definitely true*. Critically, some of these statements were repeated from the true–false tests, and some were new statements. Regardless of the statements' objective validity (i.e., whether the statements were actually true or false), students rated repeated statements as more true than new statements. In other words, prior exposure to false statements made students more likely to think those statements were true (Toppino & Brochin, 1989, for similar findings, see Brown & Nix, 1996; Hasher et al., 1977; Roberts & Ruch, 1928; Sproule, 1934; Toppino & Luipersbeck, 1993). However, students in this experiment were not given a criterial test that assessed memory, so whether they would have produced incorrect information presented in the false statements on a later memory test is unclear. Findings from studies employing multiple-choice tests, however, suggest that exposure to incorrect information on practice tests may in fact increase the likelihood of producing it on later tests (Roediger & Marsh, 2005). In addition, one classroom study demonstrated that, although multiple-choice or essay preliminary tests led to better exam performance than no tests, true–false preliminary tests did not yield any benefits (Jersild, 1929). In fact, Jersild concluded that "the true–false test is of dubious value as a pedagogical instrument" (Jersild, 1929, p. 608).

To our knowledge, only a few studies have examined the potential positive effects of answering true–false questions on later retention. In a classroom study by Standlee and Popham (1960), two of four sections of an educational psychology class took weekly true–false quizzes. The instructor graded quizzes in the first section, and students graded their own quizzes in the second section. In the third section, the instructor read the quiz questions and immediately provided the answer, whereas in the fourth section, they only presented the quiz content as a normal part of the class material. The instructor-graded weekly quizzes improved midterm grades compared to having no quizzes, but no other differences in the study were reliable. However, the sections that took instructor- and student-graded quizzes consistently performed 7%–9% better on both the midterm and final exams than the other two sections, suggesting that true–false tests may benefit long-term retention (for similar findings, also see Kellogg & Payne, 1938; Keys, 1934; Remmers & Remmers, 1926). Unfortunately, whether true–false tests enhance long-term retention relative to restudying remains unclear, given that a majority of prior studies occurred in a classroom setting across multiple sessions, and lacked methodological details necessary to determine whether they isolated the effects of true–false questions on retention.

A recently published study conducted about the same time as our research asked if true–false tests can improve retention of tested and related information (Brabec et al., 2020). In a series of experiments, answering "true" questions did improve retention of tested information (but not related information), whereas answering "false" questions improved retention of related information (but not tested information), both relative to reading a passage once. When each true–false question was structured to target both tested and related information (e.g., *True or false? Castle Geyser (not Steamboat Geyser) is the tallest geyser*), retention on tested and related information improved after answering both "true" and "false" questions. However, the criterial test occurred only 5 min after the initial study and exposure to the material was not controlled (i.e., a read-only rather than a restudy control condition was used). In a fourth experiment, Brabec et al. included a restudy control group and manipulated whether participants took the criterial test after 5 min or 2 days. Although participants retained more from the passages after restudying or being tested relative to reading a passage once, participants who were tested did not outperform those who restudied on the immediate or the delayed test. However, Brabec et al. did not provide feedback on the true–false tests, and this may have masked a potential testing effect. Thus, as it stands, there is no conclusive evidence regarding the possible retention benefits of true–false tests relative to restudying.

Although direct evidence regarding testing effects with true–false questions is mostly lacking, previous research with different practice test formats is informative. For example, results from a meta-analysis comparing the effects of testing and restudying on later retention show that the initial test format influences the magnitude of the testing effect (Rowland, 2014). Specifically, test formats that rely on generating an answer (i.e., free recall, cued recall, short answer) tend to produce larger testing effects than those that rely on selecting an answer (i.e., recognition, multiple choice) (e.g., Carpenter & DeLosh, 2006; Glover, 1989; Rawson & Zamary, 2019). This general superiority of retrieval practice with formats like cued recall over those like recognition probably stems from an elaborative or effortful process associated with generating a response. Given that true–false tests do not require response generation and resemble a recognition test, they could produce small yet reliable testing effects as with other formats relying on answer selection (e.g., multiple-choice tests).

Considering the robustness of test-enhanced learning (Rowland, 2014) and some previous findings on true–false tests (e.g., Keys,

1934; Schaap et al., 2014; Standlee & Popham, 1960), answering true–false questions may enhance retention relative to restudying. However, prior testing effect studies examining other test formats suggest that the effect may be small with true–false questions, and some studies even point out to possible negative consequences of these questions (e.g., Toppino & Brochin, 1989). Given the ubiquity of true–false tests in the classroom, but the lack of a complete understanding of their value as a learning tool in comparison to restudying (a technique commonly used by students, e.g., Karpicke et al., 2009), research examining the effectiveness of true–false tests on learning is imperative. Therefore, in the present study, we asked whether these tests enhance retention by comparing groups that took true–false tests to groups that restudied only correct information.

A secondary goal of our study was to explore whether a modification to the true–false test would increase its possible retention benefits. Specifically, in addition to asking some students to indicate whether presented statements are true or false, we asked some students to correct the statements that they thought were false. We predicted that this correction procedure would promote the kinds of generative or elaborative processes that typically enhance learning (Dunlosky et al., 2013). Elaborating on to-be-learned material typically improves memory, and it is critical that the elaborations are generated by the learner rather than provided within the learning material (Dunlosky et al., 2013; Pressley et al., 1987). Similar to elaboration, providing explanations for to-be-learned material also has robust mnemonic benefits (Chi et al., 1994; Rittle-Johnson & Loehr, 2017). By asking participants to correct true–false questions they considered as false, our goal was to encourage elaborative retrieval practice within a recognition-based retrieval practice format. In addition, given that generation-based retrieval practice typically produces a larger testing effect than recognition-based retrieval practice (Rowland, 2014), we expected the correction procedure to increase the potential benefits of true–false tests.

Bayles and Bedell (1931) described a similar modification to the true–false test in which students had to correct false statements, though the authors' purpose was not to examine if this modification would boost retention. They showed that scores on various tests (e.g., multiple-choice, completion) had higher correlations with scores on the modified true–false test than scores on the unmodified version. Based on this finding, Bayles and Bedell concluded that the modified true–false test was more valid than the unmodified true–false test.

Recently, Schaap et al. (2014) used a modified true–false test, where participants were asked to justify their answers by writing down why a statement is true or false. The authors' primary interest was not the mnemonic benefits of testing; therefore, their methodology differed from the paradigm used in most testing effect studies. Participants did not go through a controlled initial learning phase; they were administered a pretest and posttest (same questions on both) assessing knowledge from prior courses. Between the pre- and posttests, half of the participants took standard true–false tests, and the other half took modified true–false tests that required explaining why presented statements were true or false. Although about three-fourths of the justifications written by the latter group were elaborate, their posttest performance was similar to participants who took the standard true–false test—a counterintuitive finding from an elaboration or explanation standpoint. However, whether participants received feedback on the intermediate true–false tests and whether the same knowledge was assessed on these true–false tests and the posttest were both unclear, and these may have influenced the null findings. Insufficient power could also explain these findings, as there were only 13 participants in each group, though average posttest performance did not display a pattern that would be expected if the modification had any retention benefits.

As mentioned above, in our study, we asked participants to provide corrections to statements they marked as false, rather than asking them to provide explanations for all statements. We thought that asking participants to justify all their answers on a true–false quiz would defeat the purpose of adopting these tests for their practicality. That is, if students and instructors prefer true–false tests for ease of administering and grading (similar to multiple-choice tests), requiring an explanation on all questions turns the test into a short-answer test and diminishes its practicality. Furthermore, when the study material is factual (e.g., learning names, dates), asking students to explain why a statement is correct may not be a productive learning method. Instead, asking students to explain why a statement is false and ask for the corrected version of the statement, to the extent that students can successfully do so, may be more useful.

In our study, we were also interested in the different types of information that could be assessed with true–false questions. Though true–false tests have been critiqued as only assessing factual information, they can be constructed to assess more complex knowledge (Ebel, 1970). In our experiments, participants studied passages about historical figures, places, or important events. Some of our questions required memory for facts that appeared in only one part of the passage (factual questions), whereas other questions required integrating information that appeared on different parts of the passage (relational questions). By including both question types, we aimed to represent the variety of information students may be tested on, as well as have different levels of question difficulty. Since relational questions required more than just remembering facts presented in the passages, we expected these questions to be more difficult than factual questions.

Overview of the Experiments

We conducted four experiments to examine the possible benefits of answering true–false questions on a later test and if so, to determine under what conditions true-false tests are most effective. In all the experiments, college students read eight brief passages about important figures, events, or places (e.g., the Taj Mahal, the KGB) in a laboratory setting. Students then reviewed the passages according to their assigned condition, either by answering true–false questions or by restudying correct information from the passages (the true version of the true–false statements). In all experiments, students returned to the laboratory 2 days later to take a criterial test. The test consisted of either short-answer questions (Experiments 1 and 2) or short-answer and true–false questions (Experiments 3 and 4).

The four experiments primarily differed in what students did to review the information from the passages. In the first two experiments, one group took true–false quizzes (True–false), a second group took the same quizzes and tried to correct the statements they identified as false (Correction), and a third group simply read correct statements while typing them into a text box presented underneath (Typing Control). We used the Typing Control condition to ensure that participants were rereading the statements, which was

admittedly more atypical and stricter than the standard rereading control condition in which participants are just instructed to restudy the material without additional activities. In Experiment 1, we did not provide feedback on the true–false quizzes, whereas we provided correct-answer feedback on these quizzes in other experiments. In Experiments 3 and 4, we added a second control condition in which participants silently reread correct statements instead of typing them into a textbox (Rereading Control). In addition, we manipulated the format of the criterial test to have some true–false and some short-answer questions.

Experiments 1 and 2

In Experiments 1 and 2, we asked if true–false quizzes enhance long-term retention relative to restudying and whether trying to correct statements identified as false increases the potential retention benefits of true–false quizzes. Students read eight brief passages and reviewed them by taking true–false quizzes (True–false condition), taking true–false quizzes while also providing corrections to statements marked as false (Correction condition), or reading and copying only correct statements (Typing Control condition). Two days later, students came back for a short-answer criterial test. Experiments 1 and 2 differed only on the provision of feedback. Quizzed participants in Experiment 1 did not receive feedback, whereas those in Experiment 2 received correct-answer feedback after each quiz question. If true–false quizzes do enhance long-term retention, the True–false and Correction groups should outperform the Typing Control group. Furthermore, if providing corrections on the true–false quiz boosts the retention benefit of true–false quizzes, the Correction group should perform even better than the True–false group.

Experiment 1

Method

Participants. To determine sample size, we conducted an a priori power analysis to be able to detect a testing effect, targeting the between-groups variable of review condition in a mixed-factor design, assuming a medium-sized testing effect (Cohen's $f = 0.30$), an α of 0.05, and power of 0.80. The power analysis called for 84 participants. One hundred and five Washington University undergraduates participated in the experiment for payment ($15) or course credit (1.5), and they were randomly assigned to one of three review conditions. Seven participants started the experiment but did not come back for the second session, and one other participant could not complete the first session due to a programming error. Ninety-seven participants completed both sessions of the experiment. We excluded data from five participants in the Typing Control condition because they did not type statements and data from one participant in the Correction group as they scored 0 on the final test. After these exclusions, we had a sample of 91 participants ($n = 32$ in the True–false and Correction conditions, and $n = 27$ in the Typing Control condition). The study was approved by Washington University's Institutional Review Board.

Design. A 2 (question type) × 3 (review condition) mixed-factorial design was used, where the question type on the quiz and criterial test (factual, relational) was manipulated within subjects and review condition (True–false, Correction, Typing Control) was manipulated between subjects. Participants read passages and reviewed the information from the passages based on their assigned condition. One group took true–false quizzes (True–false), a second group took true–false quizzes and corrected the statements they indicated as false (Correction), and a third group read a series of correct statements while typing them in a textbox (Typing Control). All participants returned 2 days later for a short-answer criterial test on the passages.

Materials. Students read eight passages about important figures, events, or places. The passages were developed from Wikipedia pages on each passage topic, and they were used in prior research conducted in our lab (Butler et al., 2007). The length of the passages ranged from 533 to 605 words ($M = 572$ words) and the passages had similar difficulty levels.[1] We split the passages into two sets for counterbalancing purposes. Set A consisted of passages on Galileo Galilei, the Arab-Israeli War, Venice, and the KGB, whereas Set B consisted of passages on Salvador Dali, the Taj Mahal, Chernobyl, and the First Crusade. Half of the participants studied Set A first and Set B second, and the other half studied Set B first and Set A second. The presentation order of passages within each set was randomized.

For the quizzed groups, eight true–false questions were created per passage. Half of these questions assessed factual information, whereas the other half assessed relational information. Factual questions corresponded to a specific sentence in the passage, whereas relational ones required integrating information from various parts of the passage. On each quiz, the answer to half of the questions was "true" and the answer to the other half was "false." This assignment was counterbalanced, such that the answer to a specific question was "true" for half of the participants and "false" for the other half of the participants.

For the control group, 64 statements corresponding to true–false quiz questions were created. Half of these statements were the "true" statements from the quizzes, whereas the remaining half was corrected versions of "false" statements from these quizzes (i.e., the control group only read correct statements). Similar to the true–false questions, half the statements targeted factual and the other half targeted relational information from each passage.

The criterial test consisted of 64 short-answer questions. Of the eight questions from each passage, four corresponded to factual and the other four corresponded to relational statements. Questions on the criterial test were blocked by passage, but the order of the blocks and the order of questions within each block were randomized. All of our materials can be found in the OSF repository (https://osf.io/vtswx/).

Procedure. The study consisted of two sessions. Participants read and reviewed the passages according to their assigned condition in the first session, and took a criterial test during the second session that occurred 2 days after the first.

Session 1. In the first session, participants read four passages (Set A or Set B), each for 5 min. Each passage was presented on a computer screen and participants were allowed to scroll through it. Presentation order of the four passages was randomized and participants took 30-s unfilled breaks in between each passage. After reading the passages, participants completed a filler task for 5-min

[1] As a proxy of passage difficulty, we examined the True–false group's quiz performance on each passage across the four experiments ($n = 189$). Quiz performance ranged from 0.71 to 0.81 ($M = 0.75$), suggesting similar difficulty levels.

that consisted of basic arithmetic problems. Then, participants either took true–false quizzes on each passage, took these true–false quizzes and were prompted to correct the statements they indicated as false, or read and typed true statements from each passage, totaling to 32 questions or statements. Participants completed the review activity after a set of four passages, rather than after each passage, in order to provide interference and to avoid possible ceiling effects on the true–false quizzes.

Participants in both the True–false and Correction conditions made a response on each question by selecting "True" or "False" on the computer screen. The statements were presented one at a time, participants had to spend at least 3-s on each statement, and they were required to make a response before they could move on. Feedback was not provided on the true–false quizzes for either group. Statements corresponding to a passage were presented one after another; however, the order of the four quizzes and of questions within a quiz were randomized. After indicating whether a statement was true or false, participants in the Correction condition were asked to type the correct version of each statement they considered to be false. Providing a correction was not forced, so participants could leave the space blank. Participants in the Typing Control condition read true statements corresponding to the questions on the true–false quizzes. These statements were presented one at a time, and statements corresponding to a passage were blocked. The order of the four blocks and of the eight statements within a block were randomized. To ensure that participants in the control condition were paying attention to the statements they were presented, we asked the Typing Control group to copy each statement in a textbox provided below. All the review conditions were self-paced.

After participants reviewed the first set of passages according to their assigned condition, they went through the same procedure for the second passage set. The review condition was consistent across the two sets of passages (e.g., participants who were in the True–false group for Set A were also in the True–false group to review Set B). At the end of the first session, participants were asked to rate their prior knowledge on each passage topic on a 7-point scale.

Session 2. This session took place 2 days after the first session and consisted of a criterial test on the passages. The test was the same for all participants and contained 64 short-answer questions that assessed the same information targeted on the true–false quizzes. The test was self-paced, but participants had to spend at least 5-s on each question.

Results

Scoring. Whether participants in the Typing Control group were copying the statements was scored by the computer by calculating the percentage of overlap between participant responses and presented statements. Participants who showed less than 80% overlap were excluded from our analyses, which corresponded to five participants that had completed both sessions. A closer examination of data from these participants revealed that participants were judging whether the statements were true or false even though they were not given any such instructions (all statements presented in this condition were true). This was the primary reason for poor typing performance in the following experiments as well.

Short-answer responses on the criterial test were scored by the experimenters, where correct answers were given two points, partially correct answers were given one point, and incorrect answers were given zero point. Raters who were blind to conditions scored approximately 13% of all criterial test responses. Interrater reliability was calculated by correlating scores of each rater, and the raters showed good agreement (all rs > .91). After disagreements were resolved, the remaining responses were equally split among the raters for scoring. The same exclusion criteria and scoring procedure were used across all experiments.

All omnibus tests of statistical significance used an α level of 0.05 and all pairwise comparisons are reported with a Bonferroni correction. We report effect sizes using partial eta-squared for main effects and interactions, and Cohen's d for pairwise comparisons. In all our analyses, we first included the counterbalancing variables. The results reported below are without these variables, unless they interacted with the review condition—the critical manipulation in our experiments.

Performance on the True–False Quizzes. Table 1 shows quiz performance based on review condition (True–false or Correction) and correct answer (true or false) for Experiment 1. We examined quiz performance by conducting a 2 (review condition) × 2 (correct answer) × 2 (question type) mixed-factorial ANOVA. Because the question type variable did not lead to a main effect and did not interact with the other variables, we collapsed the data across this variable and conducted a 2 × 2 mixed ANOVA. On average, the True–false group scored 0.78, which was significantly better than the performance of the Correction group that scored 0.72, $F(1, 62) = 5.79$, $p = .02$, $\eta_p^2 = 0.09$. In addition, both groups answered true questions correctly more often ($M = 0.83$, $SD = 0.09$) than they did false questions ($M = 0.67, SD = 0.09$), $F(1, 62) = 65.19, p < .001, \eta_p^2 = 0.51$. Critically, there was an interaction between review condition and correct answer of quiz questions, $F(1, 62) = 16.20$, $p < .001$, $\eta_p^2 = 0.21$. Even though the True–false ($M = 0.82$, $SD = 0.09$) and Correction ($M = 0.84$, $SD = 0.09$) groups did not differ on how well they answered true questions ($p = .38$, $d = 0.22$), the True–false group ($M = 0.74$, $SD = 0.13$) performed significantly better on the false

Table 1
Quiz and Criterial Test Performances on True and False Questions in Experiments 1 and 2

	Test type	
Conditions	Quiz	Criterial test
Experiment 1		
True questions		
True–false	.82 (.09)	.57 (.18)
Correction	.84 (.09)	.52 (.18)
False questions		
True–false	.74 (.13)	.39 (.16)
Correction	.60 (.18)	.37 (.16)
Experiment 2		
True questions		
True–false	.84 (.08)	.54 (.18)
Correction	.83 (.11)	.49 (.18)
False questions		
True–false	.68 (.13)	.51 (.18)
Correction	.61 (.16)	.53 (.19)

Note. Standard deviations are reported in parentheses. True questions correspond to statements that are true on the true–false quizzes, and False questions correspond to statements that are false on the true–false quizzes. The table describes performance of the True–false and Correction groups on the two types of questions (true or false) on the quiz and the criterial test. The criterial test in Experiments 1 and 2 consists of short-answer questions.

questions than the Correction group ($M = 0.60$, $SD = 0.18$) ($p = .001$, $d = 0.89$).

Although we did not expect these differences in quiz performance, they may have been caused by the correction procedure. Participants in the Correction group were prompted to correct a sentence only if they selected "false" on a true–false question, whereas they could move to the next question if they selected "true." If selecting "false" was aversive to participants in the Correction group because of the desire to avoid writing a correction, they would select "false" less often than participants in the True–false group, which would lower their accuracy. In fact, when we examined the rate of selecting "false" for both groups, we found that the Correction group picked "false" 38% of the time, which was less than the True–false group, who picked this option 46% of the time, $t(62) = 4.03$, $p < .001$, $d = 1.01$. This might explain why the two groups' accuracy differed only on the false questions.

Prior Knowledge. At the end of the first session, participants rated their prior knowledge on each passage on a 7-point scale. Averaged across all passages, the True–false ($M = 2.49$, $SD = 0.95$), Correction ($M = 2.16$, $SD = 0.73$), and Typing Control ($M = 2.45$, $SD = 0.79$) groups had similar prior knowledge ratings, $F(2, 88) = 1.51$, $p = .23$. Furthermore, across all participants, average prior knowledge ratings for each passage ranged between 1.80 and 3.08 ($M = 2.36$). That is, our groups did not differ in terms of their prior knowledge, and participants' ratings, on average, did not exceed the halfway point of the scale.

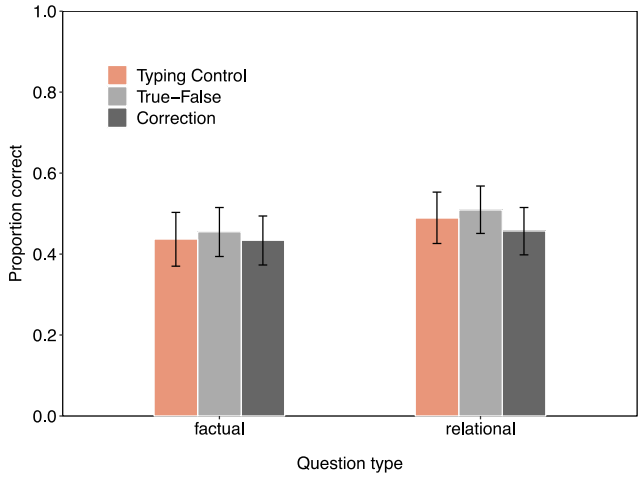
We collected prior knowledge ratings after participants had read and reviewed the passages, and these may have influenced prior knowledge ratings. Therefore, differences between groups may not accurately reflect participants' actual prior knowledge on the passages. Given this consideration, and because we found similar results across experiments, we do not report analyses on this measure in the remaining experiments. These analyses can be found in the OSF repository (https://osf.io/vtswx/).

Performance on the Criterial Test. Figure 1 shows performance on the criterial test based on review condition (True–false, Correction, Typing Control) and question type (factual, relational). If taking true–false quizzes boosts retention relative to typing correct statements, we would expect both the True–false and Correction groups to outperform the Typing Control group on the criterial test. Furthermore, if trying to correct statements considered to be false increases the testing benefit, we would expect the Correction group to outperform the True–false group. Surprisingly, a 2 (question type) × 3 (review condition) mixed ANOVA showed that performance on the criterial test did not differ among review conditions, $F(2, 88) = 0.42$, $p = .66$, $\eta_p^2 = 0.01$. The True–false ($M = 0.48$, $SD = 0.16$), Correction ($M = 0.45$, $SD = 0.16$), and Typing Control ($M = 0.46$, $SD = 0.16$) groups performed similarly on the criterial test.

We also examined performance on factual and relational questions on the final test, and whether the review condition interacted with this variable. Overall, participants did better on relational ($M = 0.49$, $SD = 0.16$) compared to factual ($M = 0.44$, $SD = 0.17$) questions, $F(1, 88) = 16.41$, $p < .001$, $\eta_p^2 = 0.16$, and there was no interaction between question type and review condition, $F(2, 88) = 0.96$, $p = .39$, $\eta_p^2 = 0.02$. Although these results may be in opposition to our prediction, they could be an artifact of our scoring procedure. That is, because relational questions require integrating two pieces of information from a passage, answers on these questions can get more partial credit than factual questions, allowing participants to demonstrate partial knowledge. In favor of this hypothesis, when we examined performance with strict scoring, where partially correct answers did not earn any points, we found that performance on factual and relational questions was identical ($M = 0.40$), $F(1, 88) < 1$.

Figure 1

Performance of Review Groups on the Criterial Test Across Question Types in Experiment 1



Note. Error bars represent 95% confidence intervals. See the online article for the color version of this figure.

Finally, we explored whether the quizzed groups performed differently on the criterial test questions corresponding to true or false quiz questions by conducting a 2 (true or false) × 2 (review condition) mixed-factorial ANOVA. As mentioned above, the two groups did not differ in their criterial test performance, $F(1, 62) = 0.86$, $p = .36$, $\eta_p^2 = 0.01$. Overall, quizzed participants answered criterial test questions corresponding to true quiz questions more correctly ($M = 0.55$, $SD = 0.18$) than those corresponding to false quiz questions ($M = 0.38$, $SD = 0.16$), $F(1, 62) = 131.55$, $p < .001$, $\eta_p^2 = 0.68$. Critically, this was similar for both groups (i.e., no interaction, $F(1, 62) = 1.17$, $p = .28$, $\eta_p^2 = 0.02$, see Table 1).

Performance of the Correction Group. The Correction group selected "false" on the initial quizzes 38% of the time, and thus could generate a correction for those items. Participants in this group provided a correction on 70% of the statements they marked as false, and 82% of these corrections corresponded to statements that were actually false on the quiz. Furthermore, 54% of all provided corrections were accurate.

We examined the Correction group's criterial test performance specifically on the items that they attempted to correct on the quizzes, regardless of the accuracy of selecting "false" and the accuracy of the provided correction. Although the Correction group, when only considering the performance on the corrected items, performed numerically better ($M = 0.56$, $SD = 0.17$) than the Typing Control ($M = 0.46$) and the True–false ($M = 0.46$) groups on the criterial test, a one-way ANOVA revealed no differences among groups, $F(2, 87) = 2.77$, $p = .07$, $\eta_p^2 = 0.06$.

Time on Task. We calculated how much time the three groups spent in the first session, given that typing, answering true–false

questions, and providing corrections were all self-paced. Unsurprisingly, groups differed in how much time they spent in this session, $F(2, 88) = 103.05$, $p < .001$. The Typing Control group spent the most time in the first session ($M = 87.72$ min, $SD = 7.60$) as they were writing 64 sentences. This was followed by the Correction group ($M = 73.71$ min, $SD = 6.19$), who were asked to correct the statements they considered to be false. The True–false group spent the least amount of time in the first session ($M = 66.02$ min, $SD = 3.11$). All pairwise comparisons were significant ($ps < .001$ and $ds > 1.57$). Because the Typing Control group spent the most time with the learning material, it could be possible that this contributed to the lack of a testing effect. However, time spent during the first session was not correlated with criterial test performance, $r(91) = .06$, $p = .58$, and this was true when we examined this correlation separately for the True–false and Typing Control groups.

We do not report time on task analyses for the remaining experiments, because they do not undermine our conclusions regarding group differences. These analyses can be found in the OSF repository (https://osf.io/vtswx/).

Experiment 2

Method

Participants. Ninety-seven Washington University undergraduates participated in the experiment. We excluded data from two participants due to a programming error, data from four participants in the Typing Control condition as they were not typing the statements, and data from one participant in the Correction condition as they did not follow instructions. Thus, we had a sample of 90 participants ($n = 31$ for the True–false condition, $n = 30$ for the Typing Control condition, and $n = 29$ for the Correction condition).

Materials, Design, and Procedure. The materials, design, and procedure were identical to Experiment 1 except for the provision of correct-answer feedback on the true–false quizzes. The feedback statement was presented regardless of response accuracy and remained on the screen for at least 4-s before participants could move on to the next quiz question.

Results

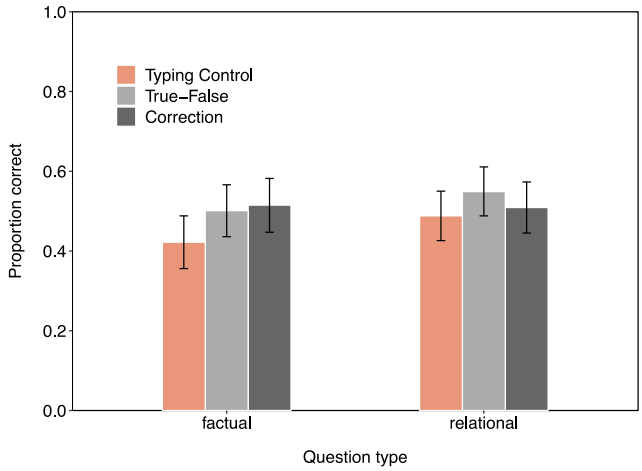
Performance on True–False Quizzes. As in the previous experiment, we examined quiz performance based on review condition and correct answers on these quizzes in a 2 × 2 mixed-factorial ANOVA (see Table 1). The data were collapsed across the question type variable (factual or relational), as this variable did not produce a main effect and did not interact with the two variables. The True–false group ($M = 0.76$) performed similarly to the Correction group ($M = 0.72$), $F(1, 58) = 3.12$, $p = .08$, $\eta_p^2 = 0.05$. Similar to Experiment 1, quizzed participants performed better on true questions ($M = 0.84$, $SD = 0.10$) than they did on false questions ($M = 0.64$, $SD = 0.15$), $F(1, 58) = 87.06$, $p < .001$, $\eta_p^2 = 0.60$. Although the quizzed groups performed similarly on the true statements, the Correction group numerically scored lower than the True–false group on the false statements; however, the interaction of review condition and item type was not significant, $F(1, 58) = 2.38$, $p = .13$, $\eta_p^2 = 0.04$. Once again, even though we did not expect the two groups to perform differently on the true–false quiz, the True–false group still performed numerically better. Similar to the first experiment, we examined both groups' rate of choosing "false" on the quizzes. Unlike the previous experiment, the Correction group ($M = 39\%$) did not select "false" significantly less than the True–false group ($M = 42\%$), $t(58) = 1.54$, $p = .13$, $d = 0.37$. Similar rates of selecting "false" may explain why the quiz accuracy of the two groups did not differ in this experiment.

Performance on the Criterial Test. Figure 2 shows performance on the criterial test based on review condition (True–false, Correction, Typing Control) and question type (factual, relational). If there is a testing effect with true–false questions when feedback is given, the two quizzed groups should perform better than the control condition. Additionally, if providing corrections to statements perceived as false benefits retention even more, the Correction group should outperform the True–false group. We conducted a 2 (question type) × 3 (review condition) mixed-factorial ANOVA to test these hypotheses. Although the True–false ($M = 0.53$, $SD = 0.17$) and Correction ($M = 0.51$, $SD = 0.18$) groups performed numerically better than the Typing Control group ($M = 0.46$, $SD = 0.16$), we found that all three groups performed similarly on the criterial test, $F(2, 87) = 1.49$, $p = .23$, $\eta_p^2 = 0.03$, replicating the previous experiment.

When we examined criterial test performance split by factual and relational questions, we again found that participants did better on the relational ($M = 0.52$, $SD = 0.17$) than factual questions ($M = 0.48$, $SD = 0.19$), $F(1, 87) = 9.96$, $p = .002$, $\eta_p^2 = 0.10$, even though we had predicted the opposite pattern. When we compared performance on these questions using strict scoring (no partial credit given), similar to the previous experiment, we found that there was no longer a meaningful difference between relational ($M = 0.42$, $SD = 0.17$) and factual ($M = 0.44$, $SD = 0.19$) questions, $F(1, 87) = 3.60$, $p = .06$, $\eta_p^2 = 0.04$. Furthermore, although there was an interaction between review condition and question type when partial credit was given, $F(2, 87) = 3.59$, $p = .03$, $\eta_p^2 = 0.08$, this interaction was no longer significant when strict scoring was used, $F(2, 87) = 2.93$, $p = .06$, $\eta_p^2 = 0.06$.

Figure 2

Performance of Review Groups on the Criterial Test Across Question Types in Experiment 2



Note. Error bars represent 95% confidence intervals. See the online article for the color version of this figure.

Similar to Experiment 1, we examined how the two quizzed groups performed on the criterial test questions that corresponded to true or false quiz questions in a 2 (true or false) × 2 (review condition) mixed-factorial ANOVA (see Table 1). The two groups performed similarly on the criterial test, $F(1, 58) = 0.09$, $p = .77$, $\eta_p^2 = 002$. Unlike Experiment 1, the criterial test performance for questions corresponding to true and false items on the initial quizzes were not different, $F(1, 58) = 0.45$, $p = .50$, $\eta_p^2 = 0.01$, likely due to the correct-answer feedback in Experiment 2. However, there was a significant interaction between the two variables, $F(1, 58) = 6.93$, $p = .01$, $\eta_p^2 = 0.11$. Although the True–false group performed slightly better on the questions corresponding to true statements than the Correction group, and the Correction group performed slightly better on the questions corresponding to false statements than the True–false group (see Table 1), none of the pairwise comparisons were reliable (all ps $> .05$, ds < 0.27)

Performance of the Correction Group. Again, we examined what the Correction group did during the quizzes and how that reflected on the criterial test. This group selected "false" 39% of the time on the initial quizzes and could provide a correction for those items. They provided a correction on 62% of the statements marked as false, and 84% of these corresponded to statements that were in fact false on the quiz. 54% of all provided corrections were accurate. These percentages are similar to Experiment 1.

We examined the Correction group's criterial test performance for the items they tried to correct on the quizzes, regardless of the accuracy of selecting "false" and of the provided correction. Unlike Experiment 1, a one-way ANOVA showed differences among groups, $F(2, 87) = 18.31$, $p < .001$, $\eta_p^2 = 0.30$. Pairwise comparisons revealed that the Correction group ($M = 0.72$, $SD = 0.18$) performed better than the Typing Control ($M = 0.46$) and the True–false ($M = 0.53$) groups, ps $< .001$, ds > 1.09. These results suggest that the correction attempt and the provision of feedback boosted the Correction group's performance for items they elaborated on.

Discussion

In Experiment 1, criterial test performance of the two quizzed groups (True–false and Correction) was not greater than that of the control group (Typing Control). However, we did not provide feedback to the quizzed groups, and this might have prevented us from observing a testing effect. On the true–false quizzes, students were presented with incorrect statements half of the time, whereas students in the control group only saw correct statements. In the absence of correct-answer feedback, quizzed students may have accepted erroneous information and carried it through to the criterial test (Roediger & Marsh, 2005). In fact, when we examined incorrect responses on the criterial test of participants who were quizzed during the first session—only considering the items presented as false statements—we found that misinformation was carried to the criterial test about 14% of the time. This finding may also explain why both quizzed groups performed worse on criterial test questions corresponding to false quiz questions than those corresponding to true quiz questions.

In Experiment 2, we investigated whether the quizzed groups would outperform the control group when we provided them with correct-answer feedback after each true–false question. Performance of the quizzed groups was about 6% higher compared to Experiment 1, and this difference across experiments likely stemmed from the increased performance on criterial questions that corresponded to false quiz questions. The provision of feedback was likely beneficial, as quizzed participants produced incorrect information from false statements only about 5% of the time on the criterial test relative to 14% in Experiment 1. Furthermore, in Experiment 2, there were no performance differences between true and false statements on the short-answer criterial test, likely due to correct-answer feedback. Yet, neither quizzed group in Experiment 2 performed reliably better compared to the control group on the criterial test. These results are surprising given robust testing effects reported with other test formats in the literature (Rowland, 2014).

Our results suggest that the correction procedure may be useful, particularly when students are provided with feedback on a quiz. In both experiments, the Correction group outperformed the other groups on the criterial test (numerically in Experiment 1 and reliably in Experiment 2) when their accuracy was calculated based on the items they attempted to correct on the true–false quizzes. With the provision of feedback in Experiment 2, the Correction group showed a 26% improvement relative to the Typing Control group and a 19% improvement relative to the True–false group. This improvement, however, was not reflected in the Correction group's overall performance as we did not require participants to correct all the statements they deemed as "false" (i.e., averaging across both experiments, only 66% of statements were corrected). Nonetheless, these results suggest that asking, and potentially requiring, students to generate corrections to false statements and then providing them with feedback may be an effective way to use true–false tests in the classroom.

In both Experiments 1 and 2, we examined the role of question type (factual or relational) on the quiz and criterial test accuracy. We included this distinction to better represent the variability in the kinds of knowledge students are assessed on, and to introduce different levels of question difficulty. However, contrary to our prediction that performance on factual questions would be higher than performance on relational questions, participants performed similarly on the two types of quiz questions. Performance on the short-answer criterial test did not confirm our predictions either; participants performed better on relational than factual questions when partial credit was awarded. Because relational questions required integrating multiple parts of the passages, whereas factual questions assessed one part of the passage only, more partial credit was available on the relational questions. As such, lenient scoring allowed participants to demonstrate partial knowledge. However, this made the comparison between factual and relational questions unfair, as the two were no longer assessed on the same scale. Therefore, we also examined criterial performance using strict scoring and found no differences between performance on the factual and relational questions—replicating our quiz findings. Given that this manipulation did not show the expected differences in performance, and that question type was not of primary interest in our study, we dropped this variable from future experiments, and we do not discuss it further.

Experiments 3 and 4

In the first two experiments, we did not observe a testing effect with true–false questions. In Experiment 3, we explored this surprising finding by making three major changes to our

methodology: increasing sample size, adding a second control condition, and matching the format of the criterial test to the quizzes.

The testing effect is a robust finding, but the magnitude of the effect depends on many variables, such as the initial test format (Rowland, 2014). In particular, initial tests that ask students to generate a response (e.g., short-answer) tend to produce a larger testing effect than tests that require students to select a response (e.g., recognition). Given that a true–false test is more similar to a yes/no recognition test than it is to a short-answer or essay test, we hypothesized that it may produce a testing effect smaller than we had initially predicted. The results of the second experiment showed that answering true–false questions and receiving correct-answer feedback boosted retention 6% relative to the control condition. Although this difference was not reliable, the performance of the quizzed and nonquizzed groups resembled what we would expect if there was a testing effect. Therefore, we considered insufficient sample size to be a possible reason for the lack of a testing effect in Experiment 2. In Experiment 3, we assumed a smaller effect size in our power analysis and increased our sample size to detect a testing effect with true–false questions, if it exists. If insufficient power prevented us from observing testing effects in Experiment 2, we expected that with an increased sample size in Experiment 3 students who answer true–false questions should do reliably better on the 2-day delayed criterial test than the control group.

In addition to insufficient power, another possible reason why we did not observe a testing effect may be the control group used in our experiments. Participants in the control condition read statements that were all correct, and typed each statement into a textbox presented underneath. We used this typing requirement to make sure participants in the control condition read the statements. However, typing the statements in a textbox may have recruited additional processes that simply rereading those statements does not. As an example, participants might have engaged in retrieval practice if they read statements and then typed them from memory, rather than directly typing them while looking at them. In addition, typing the statements might have yielded a production effect (Forrin et al., 2012; Ozubko & Macleod, 2010). Given these concerns, in Experiment 3, we added a second control condition in which participants only read correct statements without having to type them (Rereading Control). If our previous control condition prevented us from observing a testing effect, then, in Experiment 3, students who take true–false quizzes should perform better on the criterial test than those who simply reread correct statements, but not better than those who reread and type the same statements.

A final issue regards the format of the criterial test. Although the quizzes consisted of true–false questions in the first two experiments, the criterial test consisted of short-answer questions. It is possible that the mismatch between initial and final test formats was the reason for the lack of a testing effect (cf., Pan & Rickard, 2018). In particular, because true–false questions likely require less generative retrieval processes than short-answer questions, the criterial test in the first two experiments might have demanded more from the participants than they were prepared for. In turn, this mismatch of processes required by the two tests might have prevented the quizzed participants to do as well as they would have had the criterial test contained true–false questions (see Morris et al., 1977, for a transfer-appropriate processing view). To address this possibility, in Experiment 3, we manipulated the format of the criterial test, so that some questions were short-answer and others were true–false. If the mismatch between quiz and criterial test formats precluded a testing effect, in Experiment 3, quizzed participants should do better on the true–false criterial test than participants in the control condition, but there should not be any reliable performance differences on the short-answer criterial test.

By making these three changes in Experiment 3, we explored why true–false tests did not improve retention in the first two experiments. We also dropped the Correction condition to first examine a testing effect with standard true–false tests. Experiment 4 was a replication of Experiment 3 to provide more support for our findings.

Experiment 3

Method

Participants. To determine sample size, we conducted an a priori power analysis for the between-groups variable of review condition assuming a small-sized testing effect (Cohen's $f = 0.21$ based on the effect size comparing True–false and Typing Control groups in Experiment 2), an α of 0.05, and power of 0.80. The power analysis called for 168 participants. Of the 198 Washington University undergraduates who started the experiment, 18 did not come back for the second session and 1 could not complete the first session due to a programming error. Of the participants who completed both sessions, 1 participant's data were excluded because they did not follow instructions, and 12 participants' data were excluded as they were not typing the statements. Thus, our sample consisted of 166 participants ($n = 58$ for True–false, $n = 48$ for Typing Control, and $n = 60$ for Rereading Control conditions).

Design. A 3 (review condition) × 2 (test format) mixed-factorial design was used, where review condition (True–false, Typing Control, Rereading Control) was manipulated between-subjects and criterial test format (short-answer, true–false) was manipulated within subjects. The True–false and the Typing Control groups had the same procedures as in Experiment 2, and the Rereading Control group simply read a series of correct statements. The criterial test consisted of short-answer and true–false questions.

Materials. The materials used in Experiment 3 only differed from Experiments 1 and 2 on the criterial test format. Although the test still consisted of 64 questions, half of the questions were in short-answer format (similar to the first two experiments), but the other half was in true–false format. This was manipulated within each passage and counterbalanced so that the same question appeared in either format an equal number of times across participants. In addition, the true–false questions on the criterial test matched those on the quiz for the participants who took quizzes in the first session. That is, if the True–false group saw a true statement on the initial quiz, and if they were tested on this information on the true–false criterial test, then they would see the same correct statement on the criterial test rather than the incorrect version (or vice versa).

Procedure. The procedure of Experiment 3 was mostly identical to that of Experiment 2. Below we only list the differences between the two experiments.

Session 1. In Experiment 3, all participants read each passage for at least 3-min instead of 5-min. In addition, the Rereading Control group simply read eight true statements per passage. The procedures

for the True–false and Typing Control groups were identical to Experiment 2.

Session 2. All participants completed a criterial test that consisted of 32 short-answer and 32 true–false questions.

Results

Performance on True–False Quizzes. In Experiment 3, only the True–false group took quizzes during the first session. This group scored 0.74 overall on the quizzes, selected "false" 44% of the time, and scored better on the true statements ($M = 0.80$, $SD = 0.10$) than they did on the false statements ($M = 0.68$, $SD = 0.16$), $t(57) = 5.67$, $p < .01$, $d = .74$ (see Table 2).

Performance on the Criterial Test. Figure 3 shows criterial test performance based on review condition (True–false, Rereading Control, Typing Control) and test format (short answer, true–false). We examined criterial test performance based on how participants reviewed the passages during the first session and what format the questions were on the criterial test by conducting a 3 × 2 mixed ANOVA. Unlike the first two experiments, criterial test performance differed among groups, $F(2, 163) = 8.48$, $p < .001$, $\eta_p^2 = 0.10$. Pairwise comparisons revealed that the True–false group ($M = 0.64$, $SD = 0.14$) performed significantly better than the Reread Control group ($M = 0.54$, $SD = 0.15$), $p = .001$, $d = .68$. However, the True–false group did not outperform the Typing Control group ($M = 0.59$, $SD = 0.14$), $p = .23$, $d = .36$. Performance of the two control groups did not differ either, $p = .25$, $d = .33$. In other words, we replicated the first two experiments where testing with true–false questions did not improve performance relative to copying correct statements. However, testing with true–false questions did improve performance relative to rereading correct statements, the standard control condition in testing effect research.

Unsurprisingly, participants performed better on true–false ($M = 0.80$, $SD = 0.12$) than short-answer ($M = 0.48$, $SD = 0.18$) criterial test questions, $F(1, 163) = 870.62$, $p < .001$, $\eta_p^2 = 0.84$. Critically, however, there was no interaction between review condition and test format, $F(2, 163) = 1.18$, $p = .31$, $\eta_p^2 = 0.01$, suggesting that the magnitude of the testing effect is similar regardless of the criterial test format.

Table 2

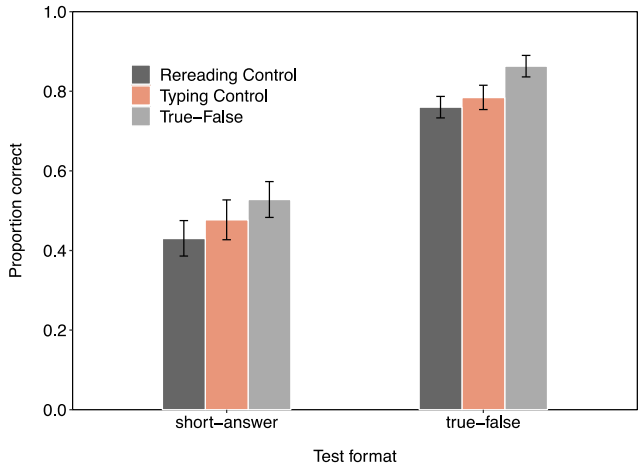
Quiz and Criterial Test Performances on True and False Questions in Experiments 3 and 4

Conditions	Test type		
	Quiz	Short-answer criterial test	True–false criterial test
Experiment 3			
True questions	.80 (.10)	.53 (.20)	.96 (.07)
False questions	.68 (.16)	.53 (.19)	.77 (.18)
Experiment 4			
True questions	.80 (.11)	.52 (.22)	.93 (.07)
False questions	.70 (.15)	.53 (.21)	.73 (.19)

Note. Standard deviations are reported in parentheses. True questions correspond to statements that are true on the true–false quizzes, and False questions correspond to statements that are false on the true–false quizzes. The table describes performance of the True–false group on the two types of questions (true or false) on the quiz, short-answer criterial test, and true–false criterial test.

Figure 3

Performance of Review Groups on the Criterial Test Across Test Formats in Experiment 3



Note. Error bars represent 95% confidence intervals. See the online article for the color version of this figure.

We also examined how the True–false group performed on each type of criterial test questions based on if they corresponded to true or false questions from the quizzes using a 2 × 2 repeated measures ANOVA (see Table 2). As mentioned above, the True–false group was more accurate on true–false than short-answer criterial test questions, $F(1, 57) = 277.02$, $p < .001$, $\eta_p^2 = 0.83$. Furthermore, these participants were more accurate on criterial test questions that corresponded to true quiz questions ($M = 0.67$, $SD = 0.14$) than those corresponding to false quiz questions ($M = 0.61$, $SD = 0.16$), $F(1, 57) = 42.59$, $p < .001$, $\eta_p^2 = 0.43$. A significant interaction, $F(1, 57) = 34.85$, $p < .001$, $\eta_p^2 = 0.38$, revealed that short-answer criterial test accuracy for questions corresponding to true or false quiz questions were not different, $p = .77$, $d = .04$, but true–false criterial test accuracy for questions corresponding to true quiz questions was significantly better than those corresponding to false quiz questions, $p < .001$, $d = 1.12$. These results were similar to the performance difference between true and false quiz questions obtained on the true–false quiz. The lack of a performance difference between true and false statements on the short-answer criterial test also replicates findings reported in Experiment 2.

A closer examination of our results reveals that, relative to the two control groups, the testing effect on the true–false criterial test arose from true, but not false quiz questions. We conducted a 3 × 2 mixed ANOVA to see how each review group performed on the true–false criterial test questions based on if they corresponded to true or false questions from the quizzes. For this analysis, we used the overall true–false criterial performance of each control group, as they did not take initial quizzes to split between true or false quiz questions. Critically, there was a significant interaction between review condition and true–false criterial test answer, $F(1, 163) = 66.77$, $p < .001$, $\eta_p^2 = 0.45$. Pairwise comparisons showed that the True–false group ($M = 0.96$, $SD = 0.07$) outperformed the Rereading Control ($M = 0.76$, $SD = 0.12$) and the Typing Control ($M = 0.79$, $SD = 0.10$) groups when the true–false criterial test questions corresponded to true quiz

questions ($ps < .001$, $ds > 1.97$). However, this group ($M = 0.77$, $SD = 0.18$) did not perform differently than the control groups (0.76 for the Rereading Control and 0.79 for the Typing Control groups) on questions corresponding to false quiz questions ($ps > .05$, $ds < 0.14$).

Experiment 4

In Experiment 4, we wanted to replicate our findings regarding testing effects with true–false questions. Because we consistently found that testing did not improve retention relative to copying correct sentences, we also dropped the Typing Control condition in Experiment 4.

Method

Participants. One hundred and forty-eight Washington University undergraduates started the experiment, but 16 did not come back for the second session, resulting in 132 participants who completed both sessions. We excluded data from two participants due to experimenter error. As such, our sample consisted of 130 participants ($n = 68$ in the True–false condition, and $n = 62$ in the Rereading Control condition).

Materials, Design, and Procedure. We used the same methodology as the previous experiment with the exception of dropping the Typing Control condition.

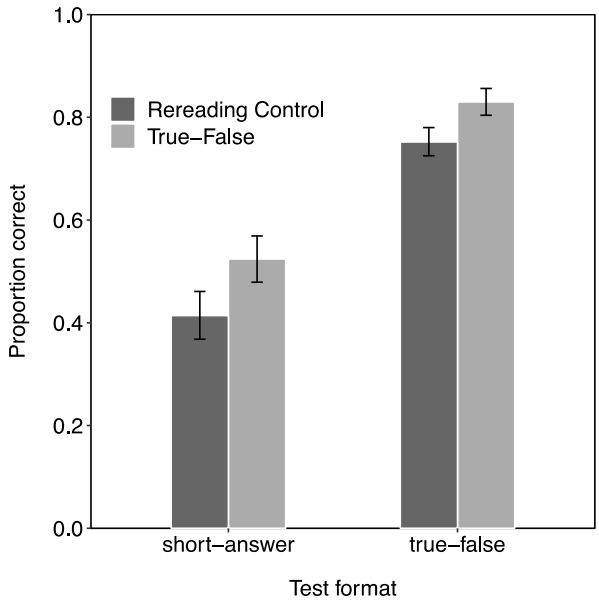
Results

Performance on True–False Quizzes. The True–false group scored 0.75 on the quizzes, and they selected "false" 45% of the time. Furthermore, they performed better on the true statements ($M = 0.80$, $SD = 0.11$) than they did on the false statements ($M = 0.70$, $SD = 0.15$), replicating previous experiments, $t(67) = 5.96$, $p < .001$, $d = .72$ (see Table 2).

Performance on the Criterial Test. Figure 4 shows criterial test performance in Experiment 4. We again examined performance on this test based on review condition (True–false, Rereading Control) and test format (short-answer, true–false) through a 2 × 2 mixed ANOVA. The True–false group ($M = 0.63$, $SD = 0.16$) outperformed the Rereading Control group ($M = 0.53$, $SD = 0.15$) on the criterial test ($F(1, 128) = 15.17$, $p < .001$, $\eta_p^2 = 0.11$), replicating the testing effect we observed in Experiment 3. Similarly, participants performed better on the true–false criterial test questions ($M = 0.79$, $SD = 0.12$) than they did on the short-answer criterial test questions ($M = 0.47$, $SD = 0.2$), $F(1, 128) = 649.06$, $p < .001$, $\eta_p^2 = 0.84$. The format of the criterial test did not interact with the review condition, $F(1, 128) = 1.79$, $p = .18$, $\eta_p^2 = 0.01$, suggesting once again that the testing effect is similar in size regardless of the criterial test format. These results replicate our findings from Experiment 3.

Similar to Experiment 3, we examined how the True–false group performed on the true–false and short-answer criterial test questions based on what the correct answers to those questions were on the initial quizzes by conducting a 2 × 2 repeated measures ANOVA (see Table 2). This group was more accurate on true–false than on short-answer criterial test questions, $F(1, 67) = 293.20$, $p < .001$, $\eta_p^2 = 0.81$. This group was also more accurate on criterial test questions corresponding to true quiz questions ($M = 0.66$, $SD = 0.16$) than those corresponding to false quiz questions ($M = 0.60$, $SD = 0.18$), $F(1, 67) = 61.46$, $p < .001$, $\eta_p^2 = 0.48$. There was a significant interaction, $F(1, 67) = 51.40$, $p < .001$, $\eta_p^2 = 0.43$. Short-answer criterial test accuracy for questions corresponding to true or false quiz questions were not significantly different, $p = .55$, $d = .07$, but true–false criterial test accuracy for questions corresponding to true quiz questions was significantly better than those corresponding to false quiz questions, $p < .001$, $d = 1.14$. These results are similar to the results of Experiment 3.

Figure 4
Performance of Review Groups on the Criterial Test Across Test Formats in Experiment 4



Note. Error bars represent 95% confidence intervals.

We again compared how review groups performed on the true–false criterial test split by true and false quiz questions, conducting a 2 × 2 mixed ANOVA. Similar to Experiment 3, we used the overall true–false criterial performance of the Rereading Control group as the split between true and false questions is not meaningful for this group. We found a significant interaction between review condition and true–false criterial test answer, $F(1, 128) = 80.92$, $p < .001$, $\eta_p^2 = 0.39$. Replicating Experiment 3, the True–false group ($M = 0.93$, $SD = 0.07$) outperformed the Rereading Control group ($M = 0.75$, $SD = 0.11$) only on the true questions ($p < .001$, $d = 1.95$), but the two groups performed similarly on the false questions ($M = 0.73$, $SD = 0.19$ for the True–false group, $p = .39$, $d = 0.13$). These results suggest that the testing effect observed on the true–false criterial test was entirely driven by the true questions.

Discussion

In Experiment 3, our goal was to test three hypotheses regarding why we did not observe testing effects in Experiments 1 and 2. These possibilities were insufficient sample size to detect a small testing effect, response requirements in the control condition used in previous experiments, and the mismatch between quiz and criterial

test format. The results of Experiment 3 suggest that the Typing Control condition that we used in Experiments 1 and 2 possibly masked a testing effect. In Experiment 3, we found that testing with true–false questions improved long-term retention relative to simply rereading the statements (Rereading Control), but not relative to rereading and typing these statements (Typing Control). That is, observing a testing effect depended on the control condition that we used.

Our results also suggest that the mismatch between quiz and criterial test format was not the reason that a testing effect was not observed in Experiments 1 and 2. In Experiment 3, the True–false group performed 10% better than the Rereading Control group on both the short-answer (mismatch) and true–false (match) criterial test questions. That is, the magnitude of the testing effect was similar regardless of the format of the criterial test (for a meta-analytic review of similar findings, see Pan & Rickard, 2018).

In Experiment 3, we found a similar performance difference between the True–false and Typing Control groups as in Experiment 2. This difference, however, was still not reliable in Experiment 3 even after we increased our power to detect a smaller testing effect. Although it is possible that there is an even smaller effect between these two groups that can be detected with a larger sample size, and that insufficient power in both Experiments 2 and 3 precluded a testing effect, such a small effect may not be practically important in the classroom.

Finally, we found that true questions in Experiment 3 produced better performance when the test format was true–false (for both the quiz and criterial test). In this experiment, this difference meant that a testing effect on the true–false criterial test was only observed on true questions, but not false questions. However, the advantage of true questions disappeared when the test format was short-answer. These results replicated the results of Experiment 2, where we obtained a difference between true and false questions on the true–false quiz, but not on the short-answer criterial test. These findings suggest that, when the following test requires response generation, true and false statements lead to a similar performance, whereas when the following test requires recognition of correct information, true statements have an advantage over false statements even after feedback. The difference on the true–false tests might stem from higher familiarity with the true statements, as they had been presented before during the initial study.

We conducted Experiment 4 to replicate our findings from Experiment 3. The True–false group outperformed the Rereading Control group on both the short-answer and true–false criterial test questions. Furthermore, the testing effect on the short-answer criterial test emerged both from true and false quiz questions, whereas the testing effect on the true–false criterial test stemmed only from true quiz questions. These findings exactly replicate Experiment 3.

General Discussion

The impetus of this study was to examine whether true–false tests enhance long-term retention relative to rereading. We asked students to read and review passages by answering true–false questions (True–false and Correction groups) or reading correct statements corresponding to those true–false questions (Typing Control and Rereading Control groups). Students then took a 2-day delayed criterial test, which consisted of short-answer questions only (Experiments 1 and 2) or a mixture of short-answer and true–false questions (Experiments 3 and 4). Across four experiments, we found that (a) true–false tests enhance retention compared to simply rereading correct statements, but not compared to rereading and copying correct statements, (b) the retention benefit of true–false tests transfers to a short-answer test, (c) how much participants retain from false quiz questions may depend on the criterial test, and (d) asking students to correct statements they consider as false can be beneficial. These findings suggest that true–false tests, given their widespread use in the classroom to *assess* learning, can also be used to *enhance* learning. Critically, this study begins to identify the conditions under which true–false tests can support learning and the findings that are discussed below have the potential to guide the use of these tests in classroom settings.

True–False Tests Enhance Retention Relative to Rereading but not Relative to Typing

In our study, we investigated if students who answered true–false questions retained more from passages compared to students who reread correct statements corresponding to the true–false questions. We considered restudying as the appropriate control condition relative to a condition in which students study the material once (e.g., Wheeler & Roediger, 1992). Compared to a restudying condition, the study-once condition is less competitive and increases the likelihood to observe a testing effect (see Brabec et al., 2020, for a testing effect with true–false questions relative to a study-once control). Among test formats used in the classroom, the true–false test likely resembles rereading the most, as students are presented with multiple statements and are asked to judge which ones are true and which ones are false. Therefore, to determine that testing is what improves retention, rather than the exposure to correct information (i.e., about half of the statements in a true–false test are typically true), we used a restudy control condition.

To ensure that participants were rereading the statements, we also asked them to type the presented statements into a textbox in Experiments 1 and 2. Despite robust testing effects reported in previous research using other test formats, we failed to obtain a testing effect using true–false quizzes. Findings of Experiment 3 revealed that testing effects do extend to true–false tests; however, whether or not a testing effect emerged depended on the control condition. When we compared the performance of a tested group to a group that reread all-correct statements and copied them to a textbox, we did not find a reliable testing effect in three experiments. However, when we compared the criterial test performance of a group that took true–false tests to a group that simply reread all-correct statements, we did find a testing effect.

Why did typing correct statements while reading them preclude a testing effect? It is possible that typing statements, rather than reading them, may have resulted in additional processing of the material. For example, participants could have been holding a statement in mind as they were typing it into a textbox. As such, participants may have engaged in incidental covert retrieval practice, slightly improving their retention of those statements (Smith et al., 2013). Additionally, typing the statements as opposed to only reading them may have resulted in a production effect (Ozubko & Macleod, 2010). Although the production effect originally compared reading to-be-learned information out loud and reading it silently, other studies have compared silent reading to typing

to-be-learned material and found similar retention benefits after production (e.g., Forrin et al., 2012). However, both of these explanations are post hoc, and our experiments were not designed to test these hypotheses.

In two of the three experiments that compared testing to typing, testing numerically but not reliably improved performance. To better understand this effect, we conducted a mini meta-analysis using data from Experiments 2 and 3 and comparing the difference between the True–false and Typing Control groups on the criterial test when the True–false group received feedback. Figure 5 shows the average weighted effect size (Cohen's *d*) comparing testing to typing using a fixed-effects model. The small but reliable effect size, $d = 0.38$, 95% CI [0.07, 0.69], suggests that testing effects with true–false questions can be observed relative to typing. We also conducted a similar mini meta-analysis using data from Experiments 3 and 4, comparing the True–false and Rereading Control groups on the criterial test. Figure 6 shows the average weighted effect size (Cohen's *d*) comparing testing to rereading using a fixed-effects model. The weighted effect size obtained from these experiments is larger, $d = 0.67$, 95% CI [0.41, 0.92], suggesting that true–false tests have a sizable retention benefit relative to rereading. In consideration of these results, it appears that testing enhances retention relative to both control conditions. However, the testing effect comparing testing to typing is much smaller compared to the testing effect observed comparing testing to rereading, and might not be of practical significance for students and instructors.

Our results show that a true–false test can be an effective tool to improve memory for passages, especially in comparison to restudying those passages. These findings fit well within a large body of research demonstrating retention benefits of testing (Rowland, 2014). Although Brabec et al. (2020, Experiment 4) did not find a testing effect comparing true–false testing to a restudy control condition, the discrepancy between the two studies may stem from the fact that participants in our study received correct-answer feedback after true–false questions with the exception of Experiment 1. Given that participants who take a true–false test are exposed to incorrect information on roughly half of the questions, providing feedback is critical.

Although our findings suggest that true–false tests enhance retention relative to rereading, note that participants in our Rereading Control condition were not restudying all of the passages. Instead, participants in this condition were presented with several correct statements from each passage. Asking participants to read to-be-tested information rather than the entire passage allowed us to isolate the benefit of testing, as the only difference between the True–false and Rereading Control groups was the act of retrieval. However, in the classroom, students will not know what questions will be asked on an exam, and such a targeted restudy may not be realistic outside the laboratory. Even though prior testing effect studies have employed an isolated restudy condition similar to our Rereading Control condition (e.g., LaPorte & Voss, 1975; Szpunar et al., 2013), and isolated restudy and restudy of the full learning material may not lead to different performance on a later test (Butler, 2010), mnemonic benefits of testing with true–false questions may differ when compared to a condition where participants reread the complete learning material. However, if anything, our targeted rereading condition would seem to represent a stricter control condition than reading the whole passage.

Figure 5

Forest Plot of Effect Sizes (Cohen's d) With 95% Confidence Intervals for Testing Effects Observed Comparing the True–False and Typing Control Groups in Experiments 2 and 3

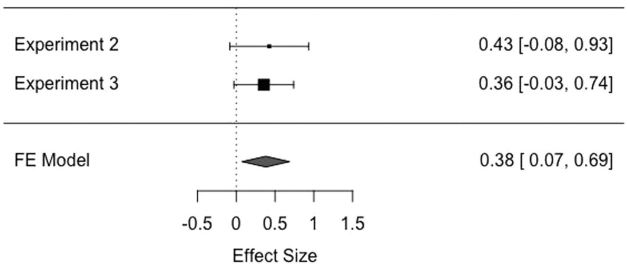
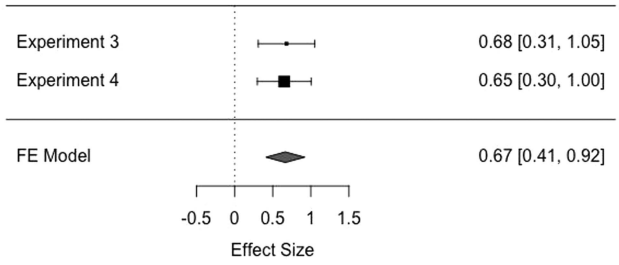


Figure 6

Forest Plot of Effect Sizes (Cohen's d) With 95% Confidence Intervals for Testing Effects Observed Comparing the True–False and Rereading Control Groups in Experiments 3 and 4



The Benefit of True–False Tests Transfers to a Short-Answer Test

An important outcome in our study is the finding that retention benefits arising from answering true–false questions can transfer to a criterial test that includes short-answer questions. In Experiments 3 and 4, we obtained a reliable testing effect comparing testing to rereading. Although the initial test consisted of true–false questions (for the True–false group), the criterial test consisted of some true–false and some short-answer questions. In both experiments, the True–false group outperformed the Rereading Control group on the short-answer criterial test. In other words, true–false tests not only enhanced retention relative to rereading; their benefits transferred to a different and presumably more demanding short-answer criterial test. That the match between initial and criterial test formats does not affect whether testing effects are observed replicates the extant literature (for a meta-analytic review, see Pan & Rickard, 2018). Furthermore, considering that true–false tests can be easily constructed, administered, and scored, the finding that these tests promote better learning than restudying even on tests with open-ended questions may have important implications for education.

Retention After Evaluating False Statements Depends on the Criterial Test

In all experiments, groups that initially took quizzes saw some true and some false questions. We examined participants' criterial

test performance when criterial test questions corresponded to a true or a false question from the initial quiz. With the exception of the first experiment, we found that participants in the True–false and Correction groups performed similarly on the true and false questions when the criterial test had short-answer questions. In other words, the testing benefit stemmed from both the true and false questions from the initial true–false quizzes. In Experiment 1, however, we found that participants performed better on the criterial test questions corresponding to true questions from the quizzes. It is likely that, in the absence of feedback, participants could not correct the misinformation presented in false questions and carried it to the criterial test. Nevertheless, when feedback was provided on the quizzes in Experiments 2–4, participants benefitted equally on the short-answer criterial test from answering true and false questions on the quizzes.

A different pattern emerged, however, when the criterial test also contained true–false questions (Experiments 3 and 4). Specifically, true questions from the quizzes yielded much better retention than the false questions from the quizzes. In our experiments, this meant that evaluating the accuracy of false statements did not improve true–false criterial test performance, whereas evaluating the accuracy of true statements yielded a testing effect relative to rereading. That is, the testing effect on the true–false criterial test was entirely driven by the true questions (see Brabec et al., 2020, for a similar finding on a short-answer criterial test). The differential processes required by true–false and short-answer test formats can provide a possible explanation for this finding. For short-answer questions, participants had to generate a response, whereas for true–false questions, they had to recognize whether the statement was presented (true statements) or not presented (false statement) during the initial study. Given that true statements are previously studied by participants, they might lead to higher familiarity than false statements, enhancing their recognition on a true–false test relative to false statements (Yonelinas, 2002). Answering short-answer questions, on the other hand, involves recalling studied material (i.e., recollection) rather than familiarity, which might have eliminated the benefit of true statements over false ones on a later test. This explanation is post hoc and needs further investigation.

Our results warrant a more nuanced conclusion regarding the benefit of true–false testing. When the criterial test has short-answer questions, evaluating the accuracy of true *and* false statements on an initial true–false quiz improves criterial test performance relative to rereading correct statements. However, when the criterial test has true–false questions, only evaluating the accuracy of true statements seems to improve subsequent test performance. Therefore, true–false quizzes enhance retention relative to rereading on a short-answer test, but only true questions on true–false quizzes do so on a true–false test.

Correcting Statements Identified as False Can Improve Retention

In Experiments 1 and 2, we examined the role of self-generated corrections in testing effects with true–false questions. Specifically, whenever participants in the Correction group selected "false" on the true–false quiz, we asked them to write the correct version of the statement as well as they could. We predicted that this elaborative component would enhance memory for the corrected items on the criterial test, which would then increase the testing effect observed with this modification. In other words, we predicted the Correction group to outperform both the True–false and the Typing Control groups on the criterial test.

Contrary to our predictions, there were no reliable differences among groups in both experiments when we looked at overall criterial test performance. However, because providing corrections were not required, not all statements identified as false received a correction. Furthermore, even though participants in the Correction group should have selected "false" on the quizzes half of the time in both experiments (i.e., half of the statements on the quiz were false), they selected false less than 40% of the time. The finding that the Correction group selected "false" suboptimally on the quizzes could be due to participants' desire to avoid providing a correction. This would decrease their accuracy on the quizzes, and thus on the criterial test, possibly resulting in the lack of a benefit of the correction procedure at the group level. In classroom settings, however, students will likely be more motivated to perform well on quizzes and exams, and therefore will be less likely to try to avoid providing a correction.

To better understand the effects of our correction procedure, we examined the Correction group's criterial test performance only for the statements that they corrected, regardless of the need for a correction (i.e., participants could have tried to correct an objectively true statement on the quiz) and the accuracy of the provided correction (i.e., participants may have generated an incorrect statement). In Experiment 1, the correction procedure improved the performance of the corrected items on the criterial test around 9% relative to the Typing Control and True–false groups, but this difference was not statistically significant. In Experiment 2, when correct-answer feedback was provided on the quizzes, the Correction group outperformed both groups on the criterial test by 23%. These findings suggest that, when participants elaborate on the true–false questions they think are false and receive feedback, their retention of those items may improve.

One issue to consider, however, is that when the Correction group's performance is calculated by considering accuracy only on the corrected items, the reliability of the measure is limited. That is, by asking participants to correct false statements only, we are reducing the number of possible trials by half, as half the statements are incorrect in reality. If we had informed participants that half of the quiz questions were true, the Correction group might have selected "false" on the quizzes more frequently, and therefore might have provided even more corrections than they did in Experiments 1 and 2. The number of correction trials is also reduced by not requiring participants to correct the statement every time they picked false.

Nonetheless, this simple modification to the true–false test seems to be a promising way of implementing true–false tests in the classroom. Criterial test performance of the Correction group is especially striking given that we considered all items that are corrected, regardless of the accuracy of selecting "false" in the first place and the accuracy of the correction. If we consider criterial test performance on items that only correspond to correctly corrected false statements, the Correction group's performance increases even more in both experiments (0.85 in Experiment 1, 0.89 in Experiment 2). However, this outcome is based on an even smaller subset of our data, and given the aforementioned consideration on the number of trials, we chose to report on performance on all of the corrected items regardless of their accuracy.

It is also important to note that the lack of an overall benefit of the correction procedure in Experiments 1 and 2 may have been caused by insufficient power. Since Experiments 3 and 4 were conducted to explore testing effects with standard true–false tests, the Correction group was dropped. If we had included this group in Experiment 3, with an increased sample size, we may have observed differences between the overall criterial test accuracy of the Correction and True–false groups. Future studies that include a larger sample could obtain such differences, especially if participants are also forced to provide a correction for questions they selected "false," if there are more "false" than "true" questions, or if participants are informed of the breakdown of answers on the true–false tests.

Educational Implications

An important outcome in our study is that taking true–false tests on previously read passages improved memory for those passages on a 2-day delayed short-answer criterial test, relative to simply rereading targeted information from those passages. Although tests are ubiquitous in the classroom, they have been strongly recommended as a learning rather than an assessment tool only in the last two decades (Dunlosky et al., 2013). Despite mixed evidence regarding the utility of true–false tests as an assessment tool (Burton, 2002, 2005; Ebel, 1970; Storey, 1966), our results suggest that true–false tests can be employed as a learning tool, particularly when the criterial test consists of short-answer questions. Importantly, the ease of creating, administering, and scoring true–false tests can make instructors and students more likely to use them. Especially in situations where instructors have difficulty creating good multiple-choice questions with multiple plausible lures, a true–false test may be particularly useful. However, we do not yet know whether true–false and multiple-choice tests lead to testing effects of similar magnitude, so this recommendation is tentative. Furthermore, examining the benefits of true–false tests in relation to a condition that studies the full learning material (rather than isolated statements) will extend the practical implications of our findings. Nonetheless, bearing in mind the frequent use of true–false tests in educational practice, our findings establish conditions under which these tests are beneficial as a learning tool and thus help guide their effective implementation in the classroom.

Another important outcome in our study is the utility of the correction procedure on true–false tests. We found that a simple modification to the true–false test where participants are prompted to correct statements they think are false, combined with correct-answer feedback, boosted retention of the corrected items on a delayed short-answer test, compared to a control group that copied all-correct statements as well as compared to taking a standard true–false test. Specifically, the modified true–false test with feedback, relative to the standard true–false test, improved accuracy from 53% to 72% when students learned from text passages—a change from a failing grade to a C- in most classrooms. However, it is important to note that this procedure did not lead to an overall retention benefit, but a benefit only for the true–false questions that received corrections. Therefore, when implementing this procedure, it is important to encourage students to attempt to correct statements whenever they mark a question as "false." Furthermore, providing students with correct-answer feedback is vital, as we did not find the benefits of the correction procedure when feedback was not provided. Through the guidelines outlined by our findings, true–false questions, particularly with the correction procedure, could offer a good balance between practicality and effectiveness.

References

Bayles, E. E., & Bedell, R. C. (1931). A study of comparative validity as shown by a group of objective tests. *The Journal of Educational Research*, *23*(1), 8–16. https://doi.org/10.1080/00220671.1931.10880121

Brabec, J. A., Pan, S. C., Bjork, E. L., & Bjork, R. A. (2020). True-false testing on trial: Guilty as charged or falsely accused? *Educational Psychology Review*, 1–26. https://doi.org/10.1007/s10648-020-09546-w

Brown, A. S., & Nix, L. A. (1996). Turning lies into truths: Referential validation of falsehoods. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(5), 1088–1100. https://doi.org/10.1037/0278-7393.22.5.1088

Burton, R. F. (2001). Quantifying the effects of chance in multiple choice and true/false tests: Question selection and guessing of answers. *Assessment & Evaluation in Higher Education*, *26*(1), 41–50. https://doi.org/10.1080/02602930020022273

Burton, R. F. (2002). Misinformation, partial knowledge and guessing in true/false tests. *Medical Education*, *36*(9), 805–811. https://doi.org/10.1046/j.1365-2923.2002.01299.x

Burton, R. F. (2005). Multiple-choice and true/false tests: Myths and misapprehensions. *Assessment & Evaluation in Higher Education*, *30*(1), 65–72. https://doi.org/10.1080/0260293042003243904

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(5), 1118–1133. https://doi.org/10.1037/a0019902

Butler, A. C. (2018). Multiple-choice testing in education: Are the best practices for assessment also good for learning? *Journal of Applied Research in Memory and Cognition*, *7*(3), 323–331. https://doi.org/10.1016/j.jarmac.2018.07.002

Butler, A. C., Flanagan, P., Roediger, H. L., III, & McDaniel, M. A. (2007, November). *The benefit of generative study activities depends on the nature of the criterial test* [Poster presentation]. The annual meeting of the Psychonomic Society, Long Beach, CA.

Butler, A. C., & Roediger, H. L. , III. (2007). Testing improves long-term retention in a simulated classroom setting. *The European Journal of Cognitive Psychology*, *19*(4–5), 514–527. https://doi.org/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*(3), 604–616. https://doi.org/10.3758/MC.36.3.604

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition*, *34*(2), 268–276. https://doi.org/10.3758/BF03193405

Chi, M. T., De Leeuw, N., Chiu, M. H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*(3), 439–477.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4–58. https://doi.org/10.1177/1529100612453266

Ebel, R. L. (1970). The case for true-false test items. *The School Review*, *78*(3), 373–389. https://doi.org/10.1086/442915

Forrin, N. D., Macleod, C. M., & Ozubko, J. D. (2012). Widening the boundaries of the production effect. *Memory & Cognition*, *40*(7), 1046–1055. https://doi.org/10.3758/s13421-012-0210-8

Frisbie, D. A., & Becker, D. F. (1991). An analysis of textbook advice about true–false tests. *Applied Measurement in Education*, *4*(1), 67–83. https://doi.org/10.1207/s15324818ame0401_6

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, *81*(3), 392–399. https://doi.org/10.1037/0022-0663.81.3.392

Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, *16*(1), 107–112. https://doi.org/10.1016/S0022-5371(77)80012-1

Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior*, *10*(5), 562–567. https://doi.org/10.1016/S0022-5371(71)80029-4

Jersild, A. T. (1929). Examination as an aid to learning. *Journal of Educational Psychology*, *20*(8), 602–609. https://doi.org/10.1037/h0070530

Kang, S. H., McDermott, K. B., & Roediger, H. L., III. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *The European Journal of Cognitive Psychology*, *19*(4–5), 528–558. https://doi.org/10.1080/09541440601056620

Karpicke, J. D., Butler, A. C., & Roediger, H. L., III. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory*, *17*(4), 471–479. https://doi.org/10.1080/09658210802647009

Kellogg, W. N., & Payne, B. (1938). The true-false question as an aid in studying. *Journal of Educational Psychology*, *29*(8), 581–589. https://doi.org/10.1037/h0060751

Keys, N. (1934). The influence of true-false items on specific learning. *Journal of Educational Psychology*, *25*(7), 511–520. https://doi.org/10.1037/h0070696

LaPorte, R. E., & Voss, J. F. (1975). Retention of prose materials as a function of postacquisition testing. *Journal of Educational Psychology*, *67*(2), 259–266. https://doi.org/10.1037/h0076933

Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science*, *23*(11), 1337–1344. https://doi.org/10.1177/0956797612443370

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*(5), 519–533. https://doi.org/10.1016/S0022-5371(77)80016-9

Miyatsu, T., Nguyen, K., & McDaniel, M. A. (2018). Five popular study strategies: Their pitfalls and optimal implementations. *Perspectives on Psychological Science*, *13*(3), 390–407. https://doi.org/10.1177/1745691617710510

Ozubko, J. D., & Macleod, C. M. (2010). The production effect in memory: Evidence that distinctiveness underlies the benefit. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(6), 1543–1547. https://doi.org/10.1037/a0020604

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin*, *144*(7), 710–756. https://doi.org/10.1037/bul0000151

Pashler, H., Bain, P. M., Bottge, B. A., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). *Organizing instruction and study to improve student learning (NCER 2007–2004)*. National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.

Pressley, M., McDaniel, M. A., Turnure, J. E., Wood, E., & Ahmad, M. (1987). Generation and precision of elaboration: Effects on intentional and incidental learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(2), 291–300. https://doi.org/10.1037/0278-7393.13.2.291

Putnam, A. L., Sungkhasettee, V. W., & Roediger, H. L., III. (2016). Optimizing learning in college: Tips from cognitive psychology. *Perspectives on Psychological Science*, *11*(5), 652–660. https://doi.org/10.1177/1745691616645770

Rawson, K. A., & Zamary, A. (2019). Why is free recall practice more effective than recognition practice for enhancing memory? Evaluating the relational processing hypothesis. *Journal of Memory and Language*, *105*, 141–152. https://doi.org/10.1016/j.jml.2019.01.002

Remmers, H. H., & Remmers, E. M. (1926). The negative suggestion effect on true-false examination questions. *Journal of Educational Psychology*, *17*(1), 52–56. https://doi.org/10.1037/h0070067

Rittle-Johnson, B., & Loehr, A. M. (2017). Instruction based on self-explanation. In R. Mayer & P. Alexander (Eds.), *Handbook of research on learning and instruction* (pp. 349–365). Routledge.

Roberts, H. M., & Ruch, G. M. (1928). Minor studies on objective examination methods. *The Journal of Educational Research*, *18*(2), 112–116. https://doi.org/10.1080/00220671.1928.10879866

Roediger, H. L., III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.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*(5), 1155–1159. https://doi.org/10.1037/0278-7393.31.5.1155

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. https://doi.org/10.1037/a0037559

Schaap, L., Verkoeijen, P., & Schmidt, H. (2014). Effects of different types of true–false questions on memory awareness and long-term retention. *Assessment & Evaluation in Higher Education*, *39*(5), 625–640. https://doi.org/10.1080/02602938.2013.860422

Smith, M. A., Roediger, H. L., III, & Karpicke, J. D. (2013). Covert retrieval practice benefits retention as much as overt retrieval practice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(6), 1712–1725. https://doi.org/10.1037/a0033569

Sproule, C. E. (1934). Suggestion effects of the true–false test. *Journal of Educational Psychology*, *25*(4), 281–285. https://doi.org/10.1037/h0070917

Standlee, L. S., & Popham, W. J. (1960). Quizzes' contribution to learning. *Journal of Educational Psychology*, *51*(6), 322–325. https://doi.org/10.1037/h0048442

Storey, A. G. (1966). A review of evidence or the case against the true–false item. *The Journal of Educational Research*, *59*(6), 282–285. https://doi.org/10.1080/00220671.1966.10883357

Szpunar, K. K., Khan, N. Y., & Schacter, D. L. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(16), 6313–6317. https://doi.org/10.1073/pnas.1221764110

Toppino, T. C., & Brochin, H. A. (1989). Learning from tests: The case of true–false examinations. *The Journal of Educational Research*, *83*(2), 119–124. https://doi.org/10.1080/00220671.1989.10885940

Toppino, T. C., & Luipersbeck, S. M. (1993). Generality of the negative suggestion effect in objective tests. *The Journal of Educational Research*, *86*(6), 357–362. https://doi.org/10.1080/00220671.1993.9941229

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*(4), 240–246. https://doi.org/10.1111/j.1467-9280.1992.tb00036.x

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*(3), 441–517. https://doi.org/10.1006/jmla.2002.2864

Zeidner, M. (1987). Essay versus multiple-choice type classroom exams: The student's perspective. *The Journal of Educational Research*, *80*(6), 352–358. https://doi.org/10.1080/00220671.1987.10885782

Received September 20, 2020
Revision received April 1, 2021
Accepted April 7, 2021 ■