Educational Psychology Review
https://doi.org/10.1007/s10648-020-09546-w

INTERVENTION STUDY

True-False Testing on Trial: Guilty as Charged or Falsely Accused?

Jordan Andrew Brabec[1] · Steven C. Pan[1] · Elizabeth Ligon Bjork[1] · Robert A. Bjork[1]

Published online: 06 July 2020

Abstract

Although widely used, the true-false test is often regarded as a superficial or even harmful test, one that lacks the pedagogical efficacy of more substantive tests (e.g., cued-recall or short-answer tests). Such charges, however, lack conclusive evidence and may, in some cases, be false. Across four experiments, we investigated how true-false testing of studied passages (e.g., on Yellowstone National Park) might enhance—or be optimized to enhance—performance on subsequent cued-recall tests. In Experiments 1–2, relative to control performance that did not benefit from any additional exposure, we found that (a) the evaluation of true statements enhanced the recall of tested (but not related) content and that (b) the evaluation of false statements enhanced the recall of related (but not tested) content, a differential pattern of benefits that did not depend on the syntactic structure of the test items. Moreover, when competitive clauses were embedded within the true-false items of Experiment 3 (e.g., *True or false? Castle Geyser (not Steamboat Geyser) is the tallest geyser*), we found that the evaluation of both types of statements enhanced the recall of both types of content. Finally, in Experiment 4, these holistic benefits proved robust to a retention interval of 48 h and were comparable with the benefits of a restudy condition in which learners restudied all of the propositions that could have been retrieved in the evaluation of the true-false items. Accordingly, although it was not uncommon for participants to misremember information as a consequence of true-false practice, our findings broadly indicate that, especially when carefully constructed, true-false tests can elicit beneficial, not superficial, processes that belie their poor reputation.

Keywords True-false · Learning · Memory · Retrieval practice · Educational psychology

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10648-020-09546-w) contains supplementary material, which is available to authorized users.

✉ Jordan Andrew Brabec
jbrabec@g.ucla.edu

[1] Department of Psychology, University of California, 502 Portola Plaza, 1285 Franz Hall, Box 951563, Los Angeles, CA 90095-1563, USA

For over a century, the true-false test—wherein test-takers must evaluate statements as true or false—has been a perennial staple of educational and other settings (e.g., Cocks 1929; Venn 1884), including classrooms, licensure exams, and even popular game shows. The ubiquity of true-false tests may stem from the fact that, despite their reputation as superficial tests, they are easy to construct, explain, and grade objectively. Ironically, however, very little is known about true-false tests beyond their capacity for assessment. In particular, it is unclear how such tests affect learning, which seems remarkable considering their widespread use in many contexts wherein learning is a primary objective.

Although it is not known whether true-false tests do so, other tests (e.g., short-answer or cued-recall tests) are known to stimulate retrieval processes that enhance long-term learning (for reviews, see Bjork 1975; Roediger and Butler 2011; Roediger and Karpicke 2006b) and, in select cases, promote transfer (e.g., enhance learning beyond that which is explicitly tested), the latter of which is often prioritized by educators and learners (Druckman and Bjork 1994; Pan and Rickard 2018). For instance, Little et al. (2012) demonstrated that multiple-choice tests—which, like true-false tests, are also often regarded as superficial—can be designed to elicit broader retrieval processes than other tests and, in turn, improve the recall of tested and untested information.

Specifically, in Little et al. (2012), participants studied educational text passages and then completed either a multiple-choice or cued-recall practice test. Critically, the multiple-choice items (e.g., *What is the area between the A and F Rings of Saturn?*) were constructed to feature competitive (i.e., highly plausible) alternatives (e.g., *(A) Cassini Division, (B) Encke Gap, (C) Maxwell Gap, or (D) Roche Division*), which were all drawn from the same preceding passage as the correct answer. Relative to no additional practice, both conditions demonstrated enhanced recall of explicitly tested content on a final cued-recall test, but only the multiple-choice condition demonstrated enhanced recall of related, untested content (i.e., information cued by the competitive alternatives).

These findings reveal, importantly, that (a) tests need not elicit retrieval in obvious ways to enhance learning and that (b) the incorporation of competitive information within tests can promote transfer to related information that is not explicitly tested. Thus, it seems that true-false tests, which do not elicit retrieval in obvious ways and can readily incorporate competitive information, might be more effective for learning than currently believed. Moreover, because the construction of optimal multiple-choice tests can often be prohibitively difficult, it is of both theoretical and practical interest to examine whether true-false tests, which are more easily constructed and modified, might achieve similar results.

On the Docket: Prior Investigations of True-False Testing

In a literature that has focused predominantly on how the true-false test fares as an assessment (for reviews, see Downing 1992; Ebel 1970; Storey 1966), however, the few studies that have examined its pedagogical effectiveness are either inconclusive or suggest little promise. In an early and influential investigation, Jersild (1929) manipulated whether undergraduate students completed true-false pretests prior to 3 weeks of lectures and readings. Such testing did not affect final exam performance, a result that Jersild interpreted as “an indictment of the true-false test” and an indication of its “dubious value as a pedagogical instrument” (pp. 607–608). In contrast, Hertzberg et al. (1932; see also Keys 1934) found that students who were given a series of study aids, including true-false practice tests, outperformed their peers on subsequent exams. Given, however, that the specific effects of testing and the exact nature of the control conditions were not reported, it is impossible to draw firm conclusions about the benefits of true-false testing per se from this early research.

Glover (1989) contributed to the resolution of this issue in an exploration of the mnemonic value of various practice test formats, including a true-false recognition test. Relative to participants given no intervening practice, participants who completed an intervening free-recall, cued-recall, or true-false recognition test 48 h after reading an educational essay then performed better on a final test administered an additional 48 h later. Irrespective of the format of the final test, Glover observed that performance increased as a function of the amount of generation required by the intervening tests, indicating that true-false tests might, if only to a limited degree, be beneficial for learning. Again, however, because it is not entirely clear how the true-false items were constructed, the conclusions that might be drawn from such work, although more promising, are limited.

To compound matters, however, true-false tests have also been accused of enhancing the learning of misinformation. Although initial studies yielded limited indications of this pattern, often described as the influence of *negative suggestion* (e.g., Sproule 1934; Roberts and Ruch 1928; Remmers and Remmers 1926), more recent work has lent credence to this concern (Toppino and Luipersbeck 1993; Toppino and Brochin 1989). Toppino and Brochin (1989), for instance, manipulated whether learners, after reading educational passages, were exposed to false information on a true-false practice test. One week later, these learners endorsed false information that they had previously been assigned to evaluate as truer than other false information to which they had not been exposed. Accordingly, true-false tests are sometimes thought to exemplify the “dangers” of testing (Roediger et al. 2010, pp. 31–32), and it is important to consider such costs when evaluating the benefits of true-false tests.

Have True-False Tests Been Convicted of Charges that Are False?

Although much of the work that has examined the effects of true-false testing on learning paints a fairly dismal picture, there remain several reasons to investigate whether such testing might be more beneficial than commonly believed. Not only are detailed, well-controlled studies designed to examine the specific effects of this test format scarce, but it also seems plausible, as noted previously, that true-false tests might elicit productive retrieval processes in a manner not dissimilar to competitive multiple-choice tests and, moreover, either promote—or be constructed to promote—transfer to related information that is not explicitly tested. Indeed, given these considerations, and despite the concerns regarding the influence of negative suggestion, we hypothesized that true-false tests can be effective learning devices. Moreover, we predicted that false items, if and when multiple pieces of information could be retrieved to evaluate such items, might elicit broader retrieval processes than true items.

Generating New Evidence: The Current Research

Across four experiments, we investigated whether true-false testing might enhance the learning of educational text passages. Specifically, we explored how testing with true-false items might enhance the learning of both tested and related content (or be optimized to do so). We designed Experiments 1, 2, and 3 to examine, respectively, whether any learning fostered by true-false testing is moderated by the truth or falsity of the test items, the syntactic structure of the test items, and the inclusion of competitive clauses within the test items. Thereafter, Experiment 4 compared the value of true-false testing with the value of restudying information at different retention intervals. In supplementary analyses, we also investigated whether true-false testing yielded evidence of negative suggestion.

Experiment 1

In Experiment 1, participants studied two educational passages and then evaluated a series of true-false items referring to one of the passages. After a retention interval of 5 min, a final cued-recall test measured the learning of tested and related content.

Method

Participants

Sixty-three undergraduate students from a large research university on the west coast of the USA were recruited in exchange for course credit. Data from four participants were excluded because of their prior experience with the study materials, leaving 59 participants in the final sample.

Design

A 3 (type of prior practice: true practice vs. false practice vs. control) × 2 (type of content: tested content vs. related content) within-subjects design was used. Regarding the type of prior practice, it should be noted that *true practice* refers to the evaluation of test items that are true, *false practice* refers to the evaluation of test items that are false, and *control* refers to no true-false practice. Furthermore, with respect to the type of content, it should be noted that the distinction between *tested* and *related* simply refers to whether the cues that were explicitly presented during the true-false practice test were also explicitly presented during the final cued-recall test.

Materials

The materials were as follows.

Text Passages The educational passages were the two 1100-word passages previously used by Little et al. (2012). One passage described the history and features of Yellowstone National Park while the other described the history and features of the planet Saturn. Each passage included at least eight categories of information with a minimum of four distinct propositions per category. For example, the four propositions pertaining to the category of famous geysers from the Yellowstone passage were (a) *Castle Geyser is the oldest geyser*, (b) *Steamboat Geyser is the tallest geyser*, (c) *Old Faithful is the most popular geyser*, and (d) *Daisy Geyser is the most reliable geyser*. Such propositions were interwoven non-systematically within the narrative of their respective passage.

True-False Practice Test The practice test drew from a list of 64 true-false items, examples of which can be reviewed in Table 1 (for this experiment and subsequent experiments). This list encompassed one set of four test items for each of the eight categories per passage. Each four-item set was created using two propositions from a given category of information. Two test items of a given set were true propositions with minimal modification (e.g., *True or false?*

Castle Geyser is the oldest geyser and *True or false? Steamboat Geyser is the tallest geyser*), and the remaining two test items were false propositions constructed by joining all or part of the subject from one of the true propositions with all or part of the predicate of the other true proposition (e.g., *True or false? Steamboat Geyser is the oldest geyser* and *True or false? Castle Geyser is the tallest geyser*).

Cued-Recall Final Test The final test drew from a list of 32 cued-recall test items, examples of which can be reviewed in Table 1 (for this experiment and subsequent experiments). These test items encompassed one set of two test items for each of the eight categories per passage. Each two-item set was created using the same two propositions per category of information that were used to create the true-false items. Instead of requiring true-false evaluation, however, the final test items were designed to require the retrieval of the subject or primary referent (e.g., *Castle Geyser* or *Steamboat Geyser*) of the queried propositions (e.g., *What is the oldest geyser?* or *What is the tallest geyser?*).

Procedure

The three phases of the experiment—initial study, practice test, and final test—all occurred during a single computer-based session in a laboratory room and used the open-source software platform, Collector. During the initial study phase, participants read the two passages for 9 min each. The passages were presented in counterbalanced orders such that approximately half of the participants read the Yellowstone passage first while the remaining participants read the Saturn passage first. The instructions encouraged them to study each passage as if they were preparing for a class.

Next, during the practice-test phase, the participants answered eight true-false items on one of the studied passages, the selection of which was counterbalanced across participants. (The content of the control passage was not practice tested, and test items regarding this passage only appeared on the final test as control items.) For each participant, the practice test featured one randomly selected true-false item for each of the eight categories of information described in the selected passage. Constraints were set such that each participant encountered, in a random order, four test items that were true and four test items that were false. Each item could only be answered one at a time, and participants were allotted 24 s per item. No feedback was provided. The instructions encouraged participants to think deeply about the test items and to consider any and all reasons why each item might be true or false.

Following a 5-min distractor task in which participants played Tetris, participants completed the final test. The final test consisted of cued-recall items for previously tested information and related (but not previously tested) information, which were presented in separate blocks of 12 items each. The order of items was randomized within each block, and, to account for possible order effects (e.g., output interference), the order of blocks was counterbalanced across participants.

In the tested block, 8 of the test items (e.g., *What is the tallest geyser?*) featured descriptions that were previously tested (i.e., explicitly featured) during the true-false practice test. Given the nature of the true-false test, half of these items corresponded to practice test items that were true (e.g., *True or false? Steamboat Geyser is the tallest geyser*) and half to practice test items that were false (e.g., *True or false? Castle Geyser is the tallest geyser*).

In the related block, 8 of the test items (e.g., *What is the oldest geyser?*) featured descriptions that were not previously tested during the true-false practice test but were importantly related to (i.e., belonged to the same categories of information as) the descriptions that were explicitly tested. As with the tested block, half of these items corresponded to practice test items that were true (e.g., *True or false? Steamboat Geyser is the tallest geyser*) and half to practice test items that were false (e.g., *True or false? Castle Geyser is the tallest geyser*).

The remaining 4 final test items in each block assessed the control passage (i.e., the passage that was not practice tested). As such, the distinction between tested and related was artificial (given the lack of prior testing) and merely indicated in which of the two blocks these control items were presented.

Final test items could only be answered one at a time, and participants were allotted 20 s per item. No feedback was provided. Afterwards, participants answered a series of metacognitive questions, were thanked for their participation, and were then dismissed.

Results

True-False Practice Test

Mean accuracy (i.e., proportion correct across participants) was $M = 0.73$, $SE = 0.03$, 95% CI [0.68, 0.78], and $M = 0.69$, $SE = 0.03$, 95% CI [0.64, 0.74], on the true and false practice test items, respectively.

Table 1 Example practice and final test items for the famous geysers category of the Yellowstone National Park passage used in Experiments 1–3

	Practice test items		Final test items	
Expt.	Type	Example	Content	Example (answer)
1	True	*True or False? Castle Geyser is the oldest geyser.*	Tested	*What is the oldest geyser?* (Castle Geyser)
			Related	*What is the tallest geyser?* (Steamboat Geyser)
	False	*True or False? Steamboat Geyser is the oldest geyser.*	Tested	*What is the oldest geyser?* (Castle Geyser)
			Related	*What is the tallest geyser?* (Steamboat Geyser)
2	True	*True or False? The oldest geyser is Castle Geyser.*	Tested	*What is the oldest geyser?* (Castle Geyser)
			Related	*What is the tallest geyser?* (Steamboat Geyser)
	False	*True or False? The oldest geyser is Steamboat Geyser.*	Tested	*What is the oldest geyser?* (Castle Geyser)
			Related	*What is the tallest geyser?* (Steamboat Geyser)
3	True	*True or False? Castle Geyser (not Steamboat Geyser) is the oldest geyser.*	Tested	*What is the oldest geyser?* (Castle Geyser)
			Related	*What is the tallest geyser?* (Steamboat Geyser)
	False	*True or False? Steamboat Geyser (not Castle Geyser) is the oldest geyser.*	Tested	*What is the oldest geyser?* (Castle Geyser)
			Related	What *is the tallest geyser?* (Steamboat Geyser)

Test items have been shortened to conserve space (in particular, the phrase “in Yellowstone National Park” has been omitted). Expt., experiment

Table 2 Mean proportions correct (SEs) with 95% CIs on the final test in Experiments 1–3

Experiment	Prior practice	Tested content		Related content	
		M (*SE*)	95% CI	*M* (*SE*)	95% CI
1	True	0.52 (0.04)	[0.45, 0.59]	0.32 (0.03)	[0.26, 0.38]
	False	0.25 (0.03)	[0.18, 0.31]	0.41 (0.04)	[0.34, 0.48]
	Control	0.27 (0.03)	[0.22, 0.33]	0.25 (0.03)	[0.19, 0.32]
2	True	0.58 (0.04)	[0.51, 0.66]	0.34 (0.03)	[0.27, 0.41]
	False	0.26 (0.03)	[0.19, 0.33]	0.44 (0.04)	[0.36, 0.52]
	Control	0.35 (0.04)	[0.27, 0.42]	0.30 (0.04)	[0.23, 0.37]
3	True	0.47 (0.04)	[0.40, 0.55]	0.36 (0.04)	[0.28, 0.44]
	False	0.45 (0.04)	[0.37, 0.54]	0.42 (0.04)	[0.33, 0.51]
	Control	0.28 (0.04)	[0.20, 0.36]	0.27 (0.04)	[0.20, 0.34]

Final Cued-Recall Test

The means for each condition are presented in Table 2 and, for ease of interpretation, are depicted in terms of the mean difference between each experimental condition and the corresponding control condition in Fig. 1. With respect to the primary questions that the present experiment was designed to explore, the pattern of results that we obtained suggests that, somewhat contrary to our hypotheses, the true-false practice test enhanced final recall in a strikingly differential manner. That is, the pattern suggests that (a) true practice enhanced the recall of tested (but not related) information and that (b) false practice enhanced the recall of related (but not tested) information.

To test the statistical validity of these apparent effects, we analyzed response accuracy (i.e., proportion correct on the final test) using a 3 (type of prior practice: true practice vs. false practice vs. control) × 2 (type of content: tested content vs. related content) repeated-measures analysis of variance (ANOVA). In this and all subsequent analyses, α was set at 0.05. The significant main

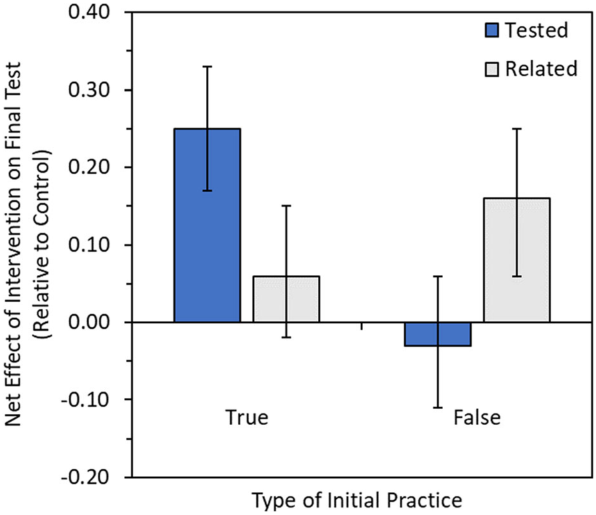

Fig. 1 Net proportion differences in final test performance for experimental conditions, relative to control performance, in Experiment 1. "True" and "False" indicate final test items preceded by true and false practice items, respectively. "Tested" and "Related" specify the type of content assessed on the final test. Error bars represent 95% confidence intervals of the difference

effect of prior practice, F (2, 116) = 10.59, $\eta_p^2 = 0.15$, $p < 0.001$, revealed that performance was affected by whether participants experienced true practice, false practice, or no practice, and the lack of a significant main effect of content ($p = 0.316$) indicated that, overall, participants performed comparably on tested and related items. Importantly, however, the two-way interaction of prior practice and content was significant, F (2, 116) = 22.99, $\eta_p^2 = 0.28$, $p < 0.001$, broadly indicating that true practice, false practice, and no practice affected performance differently on tested and related items. As described next, we interpreted this pattern using tests of simple effects.

Tested Items The simple main effect of prior practice at the tested level of content was significant, F (2, 57) = 25.36, $\eta_p^2 = 0.47$, $p < 0.001$, indicating that the level of recall for tested content depended on the type of prior practice. Indeed, there was no significant difference in performance on tested items when comparing the false practice condition ($M = 0.25$, $SE = 0.03$, 95% CI [0.18, 0.31]) with the control condition ($M = 0.27$, $SE = 0.03$, 95% CI [0.22, 0.33]), $p = 0.536$, but there was a significant advantage for the true practice condition ($M = 0.52$, $SE = 0.04$, 95% CI [0.45, 0.59]) relative to the control condition, F (1, 58) = 41.73, $\eta_p^2 = 0.42$, $p < 0.001$. Thus, only prior practice with true items enhanced final test performance on tested items.

Related Items The simple main effect of prior practice at the related level of content was also significant, F (2, 57) = 5.55, $\eta_p^2 = 0.16$, $p = 0.006$, indicating that the level of recall for related content depended on the type of prior practice. There was no significant difference in performance on related items when comparing the true practice condition ($M = 0.32$, $SE = 0.03$, 95% CI [0.26, 0.38]) with the control condition ($M = 0.25$, $SE = 0.03$, 95% CI [0.19, 0.32]), $p = 0.129$, but there was a significant advantage for the false practice condition ($M = 0.41$, $SE = 0.04$, 95% CI [0.34, 0.48]) relative to the control condition, F (1, 58) = 11.23, $\eta_p^2 = 0.16$, $p = 0.001$. Thus, only prior practice with false items enhanced final test performance on related items.

Discussion

The results of Experiment 1 indicate that true-false tests can enhance learning in particular ways that depend on whether the administered test items are true or false. Specifically, while the evaluation of true items seems to enhance the learning of explicitly tested (but not related) content, the evaluation of false items seems to enhance the learning of related (but not explicitly tested) content. Although puzzling, these patterns suggest that learners might retrieve no more information than necessary during the evaluation of either type of statement, even if—as was the case for the false items—multiple pieces of information might be retrieved to enable evaluation. The factors that might govern this “one-and-done” phenomenon, however, are unclear, at least regarding the evaluation of false items (e.g., *True or false? Castle Geyser is the tallest geyser*). Specifically, it is not obvious why information pertaining to the subjects or primary referents of the false items (e.g., *Castle Geyser*) should be retrieved when information pertaining to their predicated descriptions (e.g., *the tallest geyser*) is not. We investigated this issue in the next experiment.

Experiment 2

Experiment 2 explored whether the syntactic structure of true-false items affects the retrieval of information as the items are evaluated. If retrieval (a) terminates once evaluation is possible, per the "one-and-done" hypothesis, and (b) is elicited in an order parallel to that of the syntactic sequence, then the evaluation of false items (e.g., *True or false? Castle Geyser is the tallest geyser*) wherein the subject or primary referent always precedes the predicated description—as was the case in Experiment 1—would be expected to enhance the learning of related but not tested content. Accordingly, if the syntactic orders of the referents and their descriptions were reversed (e.g., *True or false? The tallest geyser is Castle Geyser*), then a different pattern might be observed.

Method

Participants Experiment 2 was conducted in the same academic term as Experiment 1. Fifty-nine undergraduate students, recruited in the same manner as in Experiment 1, participated for course credit. Data from three participants who did not complete the experiment as prescribed were excluded, resulting in a final sample of 56 participants.

Design, Materials, and Procedure Except for the aforementioned changes to the syntactic structure of the true-false items, the design, materials, and procedure were identical to those of Experiment 1.

Results

True-False Practice Test

Mean accuracy (i.e., proportion correct across participants) was $M = 0.70$, $SE = 0.04$, 95% CI [0.63, 0.77], and $M = 0.72$, $SE = 0.03$, 95% CI [0.66, 0.79], on the true and false practice test items, respectively.

Final Cued-Recall Test

The means for each condition are presented in Table 2 and, for ease of interpretation, are depicted in terms of the mean difference between each experimental condition and the corresponding control condition in Fig. 2. With respect to the primary questions that the present experiment was designed to explore, the pattern of results suggests that the true-false practice test again enhanced final recall but, contrary to our hypotheses, did so in a differential manner that was nearly identical to that observed in Experiment 1. That is, although true practice, as expected, seems to have again enhanced the recall of tested (but not related) information, false practice, in a manner that was unexpected, seems to have again enhanced the recall of related (but not tested) information.

To test the statistical validity of these apparent effects, an ANOVA identical to that performed for Experiment 1 revealed a significant main effect of prior practice, $F\ (2,\ 110) = 8.27$, $\eta_p^2 = 0.13$, $p < 0.001$, no significant main effect of content ($p = 0.131$), and a significant two-way interaction of prior practice and content, $F\ (2,\ 110) = 27.74$, $\eta_p^2 = 0.34$, $p < 0.001$, a pattern that was identical to that of Experiment 1. As described next, we interpreted this pattern using tests of simple effects.

Tested Items The simple main effect of prior practice at the tested level of content was significant, $F\ (2,\ 54) = 29.51$, $\eta_p^2 = 0.52$, $p < 0.001$, indicating that the recall of

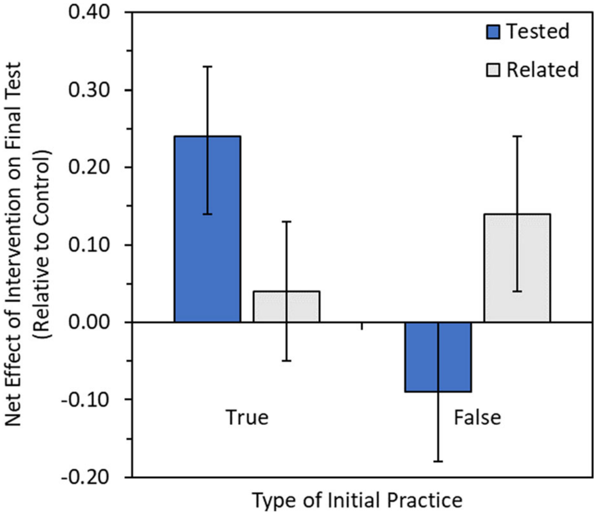

Fig. 2 Net proportion differences in final test performance for experimental conditions, relative to control performance, in Experiment 2. “True” and “False” indicate final test items preceded by true and false practice items, respectively. “Tested” and “Related” specify the type of content assessed on the final test. Error bars represent 95% confidence intervals of the difference

tested content depended on the type of prior practice. As in Experiment 1, there was a significant performance advantage for the true practice condition ($M = 0.58$, $SE = 0.04$, 95% CI [0.51, 0.66]) relative to the control condition ($M = 0.35$, $SE = 0.04$, 95% CI [0.27, 0.42]), F (1, 55) = 24.02, $\eta_p^2 = 0.30$, $p < 0.001$. In a potential departure from the findings of Experiment 1, it initially appeared that there could have been a performance deficit for the false practice condition ($M = 0.26$, $SE = 0.03$, 95% CI [0.19, 0.33]) relative to the control condition, F (1, 55) = 4.32, $\eta_p^2 = 0.07$, $p = 0.042$, but this possible deficit does not survive a Bonferroni correction.

Related Items The simple main effect of prior practice at the related level of content was significant, F (2, 54) = 4.20, $\eta_p^2 = 0.14$, $p = 0.020$, indicating that the recall of related content depended on the type of prior practice. There was no significant difference in performance on related items when comparing the true practice condition ($M = 0.34$, $SE = 0.03$, 95% CI [0.27, 0.41]) with the control condition ($M = 0.30$, $SE =$ 0.04, 95% CI [0.23, 0.37]), $p = 0.380$, but a significant performance advantage for the false practice condition ($M = 0.44$, $SE = 0.04$, 95% CI [0.36, 0.52]) relative to performance in the control condition was observed, F (1, 55) = 8.13, $\eta_p^2 = 0.13$, $p = 0.006$. Thus, this pattern fully replicated that of Experiment 1.

Discussion

The results of Experiment 2 indicate that syntactic structure does not govern the retrieval processes elicited by true-false items. Rather, the nearly identical patterns across Experiments 1 and 2 suggest that, irrespective of their syntactic placement, the names of studied referents (e.g., *Steamboat Geyser*) elicit retrieval while the descriptions of those referents (e.g., *the tallest geyser*) do not. It seems possible, then, that true-false items might be optimized with the inclusion of additional names of competitive referents. That is, such modification might undo the “one-and-done” phenomenon and, akin to competitive multiple-choice items (e.g., Little et al. 2012), elicit broader, more productive retrieval processes. We explored this possibility in the next experiment.

Experiment 3

Experiment 3 explored whether the inclusion of competitive clauses within true-false items might affect the retrieval processes elicited by the evaluation of such test items. To assess this possibility, we developed test items (revised from Experiment 1) wherein the subjects or primary referents were followed by parenthetical, dissociative clauses that featured a competitive referent from the relevant categories of information (e.g., *True or false? Castle Geyser (not Steamboat Geyser) is the tallest geyser*). If such competitive clauses elicit broader retrieval processes, then the revised true-false items might enhance the learning of both tested and related content, irrespective of whether the test items are true or false.

Method

Experiment 3 was preregistered with AsPredicted.org at the following link:https://aspredicted.org/vg8mg.pdf.

Participants

Sixty-nine undergraduate students, recruited in the same manner as in the preceding experiments, participated for course credit. Data from twelve participants that either (a) had prior experience with the materials or (b) did not complete the experiment as prescribed were excluded, resulting in a final sample of 57 participants.

Design, Materials, and Procedure

With the sole exception of the incorporation of competitive clauses within the true-false items, the design, materials, and procedure were identical to those of Experiment 1. It should be emphasized that, although the final cued-recall items were the same as those used in the preceding experiments, the correct answers for both tested and related items were displayed during the practice test of Experiment 3 (as either a primary referent or a dissociated alternative within a competitive clause on the true-false practice test) as a consequence of the “this-not-that” construction.

Results

True-False Practice Test

Mean accuracy (i.e., proportion correct across participants) was $M = 0.69$, $SE = 0.04$, 95% CI [0.62, 0.77], and $M = 0.76$, $SE = 0.03$, 95% CI [0.70, 0.82], on the true and false practice test items, respectively.

Final Cued-Recall Test

The means for each condition are presented in Table 2 and, for ease of interpretation, are depicted in terms of the mean difference between each experimental condition and the corresponding control condition in Fig. 3. With respect to the primary questions that the present experiment was designed to explore, the pattern of results suggests that, in a manner consistent with our hypotheses, both true practice and false practice enhanced final recall of both tested and related content.

To test the statistical validity of these apparent effects, an ANOVA identical to those performed for the preceding experiments revealed a significant main effect of prior practice, F (2, 112) = 13.45, $\eta_p^2 = 0.19$, $p < 0.001$, a significant main effect of content, F (1, 56) = 5.15, $\eta_p^2 = 0.08$, $p = 0.027$, and, in a departure from the findings of Experiments 1 and 2, no significant interaction between prior practice and content ($p = 0.170$), broadly indicating that the extent to which true practice, false practice, and no practice affected performance did not depend significantly on whether tested or related content was assessed.

The main effect of content indicated that tested performance ($M = 0.40$, $SE = 0.03$, 95% CI [0.34, 0.46]) exceeded related performance ($M = 0.35$, $SE = 0.03$, 95% CI [0.29, 0.41]), and further investigation of the main effect of prior practice revealed significant overall performance advantages for both the true practice condition ($M = 0.42$, $SE = 0.03$, 95% CI [0.35, 0.48]) and the false practice condition ($M = 0.44$, $SE = 0.04$, 95% CI [0.36, 0.51]) relative to the control condition ($M = 0.27$, $SE = 0.03$, 95% CI [0.21, 0.34]), F (1, 56) = 20.26, $\eta_p^2 = 0.27$, $p < 0.001$, and F (1, 56) = 19.64, $\eta_p^2 = 0.26$, $p < 0.001$, respectively.

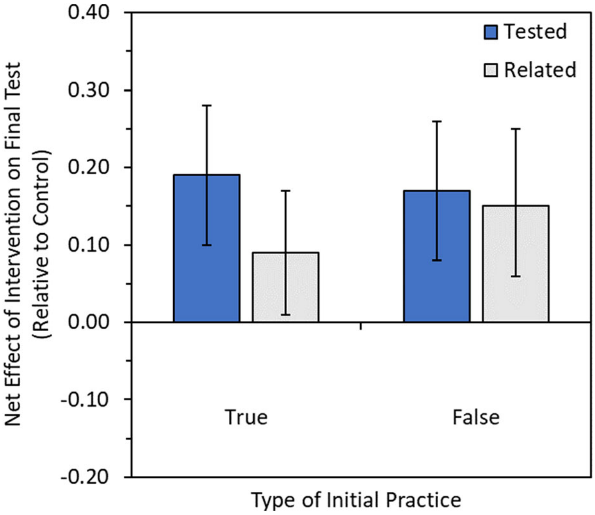

Fig. 3 Net proportion differences in final test performance for experimental conditions, relative to control performance, in Experiment 3. "True" and "False" indicate final test items preceded by true and false practice items, respectively. "Tested" and "Related" specify the type of content assessed on the final test. Error bars represent 95% confidence intervals of the difference

Discussion

The results of Experiment 3 reveal that the incorporation of competitive clauses within true-false items elicits broader retrieval processes than would otherwise occur during the evaluation of conventional true-false items. That is, unlike in Experiments 1–2, the evaluation of both true and false statements on a practice test enhanced performance on a later cued-recall test for both explicitly tested and related content. Thus, irrespective of whether true-false items are true or false, the inclusion of competitive clauses seems to undo the “one-and-done” phenomenon and yields more holistic learning benefits. Whether these benefits might be retained following greater retention intervals that are more representative of practical contexts, however, is unclear, as is how these benefits might compare to the benefits of simply restudying the target propositions (cf. Carrier and Pashler 1992; Pan and Rickard 2017). We explored these questions in the final experiment.

Experiment 4

To explore the value of competitive true-false tests across different retention intervals and relative to the value of restudying information, Experiment 4 compared the benefits of true-false testing with the benefits of rereading to-be-tested facts at retention intervals of 5 min and 48 h. To this end, we developed items for restudy that, instead of soliciting true-false evaluations (e.g., *True or false? Castle Geyser (not Steamboat Geyser) is the tallest geyser*), re-presented the same information for restudy that participants, as implied by the results of our three previous experiments, might have retrieved in their evaluations of the true-false items (e.g., *Castle Geyser is the oldest geyser. Steamboat Geyser is the tallest geyser*). If the evaluation of true-false items elicits productive retrieval processes (as indicated by these previous experiments), then, at least relative to the effect of no additional practice or exposure, the value of true-false practice should still emerge after 48 h. Furthermore, if such true-false evaluation elicits productive processes that are more effective than those elicited by restudying, then the expected decrease in performance from 5 min to 48 h should be less drastic in the true-false practice condition than in the restudy condition.

Method

Experiment 4 was preregistered with AsPredicted.org at the following link: https://aspredicted.org/zw3pi.pdf.

Participants

One hundred fifty-five undergraduate students, recruited from two large research universities on the west coast of the USA, participated for course credit. Data from 34 participants that either (a) had prior experience with the materials or (b) did not complete the experiment as prescribed were excluded, resulting in a final sample of 121 participants.

Design

A 2 (prior re-exposure (within-subjects): re-exposed vs. control) × 2 (method of prior re-exposure (between-subjects): true-false practice test vs. restudy intervention) × 2 (retention interval (between-subjects): 5 min vs. 48 h) mixed design was used. With respect to the first two factors, all participants were re-exposed to information from one of the studied passages—either via a true-false practice test or via a restudy intervention—and were not re-exposed to information from the other studied passage until, with respect to the third factor, they completed the final test after a retention interval of either 5 min or 48 h. Furthermore, in contrast with the prior experiments, and given that the holistic benefits of true-false tests observed in Experiment 3 did not depend significantly on whether the practice test items were true or false, we elected a priori to forego analyzing the distinctions between (a) true and false practice and (b) tested and related information in the primary analysis of Experiment 4. Moreover, it should be noted that, while these distinctions remained meaningful for the true-false practice condition, these distinctions were not meaningful for the restudy condition.

Materials

Except for the addition of 32 restudy items for the restudy intervention, the materials were identical to those of Experiment 3. The restudy items were designed to correspond to the

Table 3 Example practice test items, restudy items, and final test items for the famous geysers category of the Yellowstone National Park passage used in Experiment 4

Practice test items (answer)	Restudy items	Final test items (answer)
True or False? Castle Geyser (not Steamboat Geyser) is the oldest geyser. (True)	*Castle Geyser is the oldest geyser. Steamboat Geyser is the tallest geyser.*	*What is the oldest geyser?* (Castle Geyser) *What is the tallest geyser?* (Steamboat Geyser)
True or False? Steamboat Geyser (not Castle Geyser) is the tallest geyser. (True)	*Steamboat Geyser is the tallest geyser. Castle Geyser is the oldest geyser.*	*What is the oldest geyser?* (Castle Geyser) *What is the tallest geyser?* (Steamboat Geyser)
True or False? Steamboat Geyser (not Castle Geyser) is the oldest geyser. (False)	*Castle Geyser is the oldest geyser. Steamboat Geyser is the tallest geyser.*	*What is the oldest geyser?* (Castle Geyser) *What is the tallest geyser?* (Steamboat Geyser)
True or False? Castle Geyser (not Steamboat Geyser) is the tallest geyser. (False)	*Steamboat Geyser is the tallest geyser. Castle Geyser is the oldest geyser.*	*What is the oldest geyser?* (Castle Geyser) *What is the tallest geyser?* (Steamboat Geyser)

Items have been shortened to conserve space (in particular, the phrase "in Yellowstone National Park" has been omitted) competitive true-false items, and examples that illustrate this relationship are shown in Table 3. The list of restudy items included one set of two restudy items for each of the eight categories per passage. Each two-item set was created using two propositions from a given category of information and, importantly, both items of each set featured both propositions. To counterbalance the order in which participants restudied these within-item propositions, however, each item presented the propositions in a different order. (For example, in the set of restudy items corresponding to the famous geysers of Yellowstone National Park, one item featured the proposition regarding Castle Geyser followed by the proposition regarding Steamboat Geyser, and the other item featured these propositions in the reverse order.)

Procedure

Except for the type of intervention that participants experienced and the retention interval before the final test, the three phases of the procedure—initial study, intervention, and final test—were otherwise identical to those of Experiment 3.

During the intervention phase, participants were randomly assigned to answer a series of eight true-false items (in a manner identical to Experiment 3) or to re-read a series of eight restudy items pertaining to one of the studied passages, the selection of which was counterbalanced across participants. (The content of the control passage was not revisited, and test items regarding this passage only appeared on the final test as control items.) For each participant in the restudy condition, the restudy intervention featured one randomly selected restudy item for each of the eight categories of information described in the selected passage. Because each item consisted of two propositions, however, each participant in the restudy condition restudied sixteen propositions across eight trials. Participants were allotted 24 s per item, and only one item could be viewed at a time. The instructions encouraged participants to review the propositions and to wait patiently for the next trial.

Critically, for a given trial during the intervention phase, it should be emphasized that participants who restudied statements were always presented with both pieces of complete information about which they would later, on the final test, be tested. Participants who instead completed the true-false practice test, however, could only have been similarly re-exposed to such information if they had successfully retrieved both complete pieces of information on their own. Accordingly, it can be reasonably argued that participants who did not experience the restudy intervention—and instead experienced the true-false practice test—were at a considerable disadvantage on the final test.

With respect to the retention interval before the final test, participants were randomly assigned to complete the final test after 5 min of Tetris (as in the previous experiments) or after 48 h. Immediately following the final test, participants answered a series of metacognitive questions, were thanked for their participation, and were then dismissed.

Results

True-False Practice Test

Nearly identical to the observations of Experiment 3, mean accuracy (i.e., proportion correct across participants) was $M = 0.69$, $SE = 0.03$, 95% CI [0.63, 0.75], and $M = 0.75$, $SE = 0.03$, 95% CI [0.70, 0.81], on the true and false practice test items, respectively.

Table 4 Mean proportions correct (SEs) with 95% CIs on the final test in Experiment 4

Experiment	Condition	5-min retention interval		48-h retention interval	
		M (*SE*)	95% CI	*M* (*SE*)	95% CI
4	True-false	0.48 (0.03)	[0.42, 0.55]	0.31 (0.03)	[0.25, 0.37]
	Control	0.32 (0.03)	[0.25, 0.39]	0.19 (0.03)	[0.12, 0.25]
	Restudy	0.54 (0.03)	[0.47, 0.60]	0.28 (0.03)	[0.21, 0.35]
	Control	0.28 (0.04)	[0.22, 0.35]	0.15 (0.04)	[0.08, 0.22]

Final Cued-Recall Test

The means for each condition are presented in Table 4 and, for ease of interpretation, are depicted in Fig. 4 in terms of the mean difference between each experimental condition and the corresponding control condition. With respect to the primary questions that the present experiment was designed to explore, the pattern of results suggests that (a) the benefits of true-false practice, relative to no additional practice, emerged after both 5 min and 48 h, (b) the benefits of restudying are superior to those of true-false practice after 5 min but not after 48 h, and (c) performance in the restudy condition decreased as a function of the retention interval to a greater extent than did performance in the true-false practice condition.

Check of Key Assumptions To reiterate our a priori decision, we elected not to distinguish between (a) the effects of true and false practice on (b) the recall of tested and related information in the primary analysis of Experiment 4. To confirm the basis for the rationale that we previously articulated, however, we initially subjected the accuracy of the final test responses of the participants in the true-false practice condition to a supplementary 3 (type of prior practice (within-subjects): true practice vs. false practice vs. control) × 2 (type of content (within-subjects): tested content vs. related content) × 2 (retention interval (between-subjects): 5 min vs. 48 h) repeated-measures ANOVA, the primary contents of which, in addition to the associated means, can be found in Tables 5 and 6. Critically, this supplementary analysis not

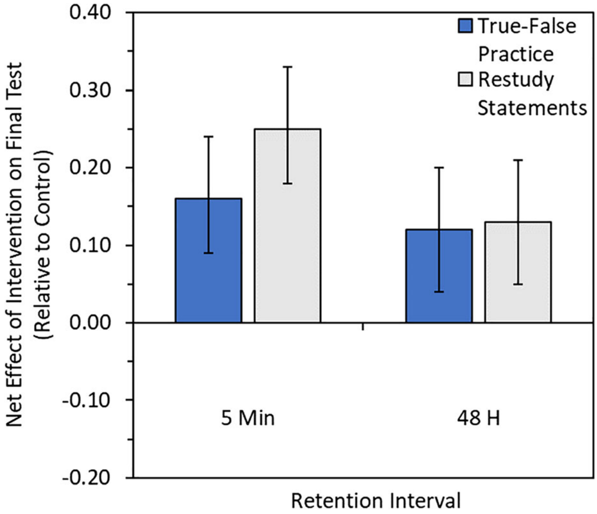

Fig. 4 Net proportion differences in final test performance for experimental conditions, relative to control performance, in Experiment 4. "True-False Practice" and "Restudy Statements" indicate final test items preceded by true-false practice items and restudy items, respectively. "5 Min" and "48 H" specify the retention interval before the final test. Error bars represent 95% confidence intervals of the difference

Table 5 Supplementary ANOVA of the true-false practice conditions from Experiment 4

Factor	*df*	*F*	η_p^2	*p*
Type of Practice	(2, 124)	14.14	0.19	<.001
Type of Content	(1, 62)	3.93	0.06	.052
Retention Interval	(1, 62)	20.61	0.25	<.001
Type of Practice * Type of Content	(2, 124)	0.88	0.01	.419
Type of Practice * Retention Interval	(2, 124)	0.44	0.01	.645
Type of Content * Retention Interval	(1, 62)	0.05	0.00	.862
Type of Practice * Type of Content * Retention Interval	(2, 124)	0.49	0.01	.613

only replicated the pattern of the significant holistic benefits observed in Experiment 3 but also, given that the retention interval did not moderate the effect of prior practice, extended it across a more practical retention interval.

Primary Omnibus Analysis Given the confirmation of our a priori assumptions, and to test the statistical validity of the apparent pattern of effects observed in Experiment 4, we analyzed the accuracy of responses on the final cued-recall test using a 2 (prior re-exposure (within-subjects): re-exposed vs. control) × 2 (method of prior re-exposure (between-subjects): true-false practice test vs. restudy intervention) × 2 (retention interval (between-subjects): 5 min vs. 48 h) repeated-measures ANOVA. Thereafter, we followed this omnibus analysis with planned tests to examine, specifically, (a) whether the value of true-false practice relative to no additional exposure emerged after both 5 min and 48 h, (b) whether true-false practice enhanced performance to a different extent than did the restudy intervention, and (c) whether the method of re-exposure moderated any decreases in performance as a function of the retention interval.

True-False Practice vs. No Prior Re-exposure Consistent with (a) the significant main effect of prior re-exposure, F (1, 117) = 73.24, $\eta_p^2 = 0.39$, $p < 0.001$, (b) the significant simple main effect of prior re-exposure at the true-false level of the method of re-exposure, F (1, 117) = 28.44, $\eta_p^2 = 0.20$, $p < 0.001$, and (c) the findings of the supplementary analysis, two significant tests of simple effects re-confirmed that, after both 5 min and 48 h, F (1, 117) = 18.83, $\eta_p^2 =$ 0.14, $p < 0.001$, and F (1, 117) = 10.26, $\eta_p^2 = 0.08$, $p = 0.002$, respectively, participants who had completed the true-false practice test scored significantly better on the corresponding final test items ($M_{5\ \mathrm{MIN}} = 0.48$, $SE = 0.03$, 95% CI [0.42, 0.55]; $M_{48\ \mathrm{H}} = 0.31$, $SE = 0.03$, 95% CI [0.25, 0.37]) than they did on the control items whose contents had not been previously re-exposed ($M_{5\ \mathrm{MIN}} = 0.32$, $SE = 0.03$, 95% CI [0.25, 0.39]; $M_{48\ \mathrm{H}} = 0.19$, $SE = 0.03$, 95% CI

Table 6 Supplementary table of means of the true-false practice conditions from Experiment 4

Retention interval	Type of practice	Tested content		Related content	
		M (*SE*)	95% CI	*M* (*SE*)	95% CI
5-min	True	0.48 (0.05)	[0.39, 0.58]	0.45 (0.05)	[0.36, 0.54]
	False	0.54 (0.05)	[0.44, 0.64]	0.47 (0.05)	[0.38, 0.56]
	Control	0.33 (0.04)	[0.24, 0.41]	0.31 (0.04)	[0.23, 0.40]
48-h	True	0.36 (0.05)	[0.26, 0.45]	0.25 (0.05)	[0.16, 0.34]
	False	0.34 (0.05)	[0.24, 0.43]	0.29 (0.05)	[0.20, 0.38]
	Control	0.19 (0.04)	[0.10, 0.27]	0.19 (0.04)	[0.11, 0.27]

[0.12, 0.25]). These analyses further indicate that Experiment 4 replicated and extended the significant holistic benefits of true-false practice across both retention intervals.

True-False Practice vs. Restudy Intervention With respect to the comparison of the true-false practice and the restudy conditions on final test performance, however, and consistent with both (a) the nonsignificant main effect of the method of re-exposure, F (1, 117) = 0.18, $\eta_p^2 = 0.00$, $p = 0.675$, and (b) the nonsignificant simple main effect of the method of re-exposure on final test performance for items whose contents had been previously re-exposed, F (1, 117) = 0.14, $\eta_p^2 = 0.00$, $p = 0.707$, tests of simple effects confirmed that—whether after 5 min or 48 h—neither of the numerical differences in performance between the true-false practice condition ($M_{5\text{ MIN}} = 0.48$, $SE = 0.03$, 95% CI [0.42, 0.55]; $M_{48\text{ H}} = 0.31$, $SE = 0.03$, 95% CI [0.25, 0.37]) and the restudy condition ($M_{5\text{ MIN}} = 0.54$, $SE = 0.03$, 95% CI [0.47, 0.60]; $M_{48\text{ H}} = 0.28$, $SE = 0.03$, 95% CI [0.21, 0.35]) on the corresponding final test items was significant, F (1, 117) = 1.26, $\eta_p^2 = 0.01$, $p = 0.264$, and F (1, 117) = 0.34, $\eta_p^2 = 0.00$, $p = 0.562$, respectively. Although such analyses might suggest parity between the effects of the two interventions, it should be emphasized that, despite the methodological advantage afforded to the restudy condition and the associated numerical advantage in performance after 5 min, the true-false practice condition achieved a modest numerical advantage in performance on the corresponding final test items after 48 h, an observation that is obscured by the subtraction of control performance in Fig. 4.

Interactions with Retention Interval Despite the implied interaction between the method of re-exposure and retention interval, however, neither the two-way interaction between these factors nor the three-way interaction reached significance, F (1, 117) = 0.52, $\eta_p^2 = 0.00$, $p = 0.474$, and F (1, 117) = 1.05, $\eta_p^2 = 0.01$, $p = 0.308$, respectively, broadly indicating that the changes in final test performance across retention intervals were not significantly affected by whether participants had previously completed the true-false practice test or the restudy intervention. That is, with particular respect to the final test items whose contents had been previously re-exposed, although performance appeared to decrease as a function of the retention interval to a notably greater extent in the restudy condition ($M_{5\text{ MIN}} = 0.54$, $SE = 0.03$, 95% CI [0.47, 0.60]; $M_{48\text{ H}} = 0.28$, $SE = 0.03$, 95% CI [0.21, 0.35]), F (1, 117) = 28.17, $\eta_p^2 = 0.19$, $p < 0.001$, than in the true-false practice condition ($M_{5\text{ MIN}} = 0.48$, $SE = 0.03$, 95% CI [0.42, 0.55]; $M_{48\text{ H}} = 0.31$, $SE = 0.03$, 95% CI [0.25, 0.37]), F (1, 117) = 14.99, $\eta_p^2 = 0.11$, $p < 0.001$, the lack of a significant three-way or two-way interaction—in addition to the nonsignificant simple two-way interaction between the method of re-exposure and retention interval on the final test items whose contents had been previously re-exposed, F (1, 117) = 1.45, $\eta_p^2 = 0.01$, $p = 0.231$—might suggest parity between these significant decreases.

The two-way interaction of prior re-exposure and retention interval, however, did reach significance, F (1, 117) = 4.52, $\eta_p^2 = 0.04$, $p = 0.036$, which broadly indicated that (a) the level of disparity between performance on final test items whose contents had been previously re-exposed and control performance depended significantly on when the final test was completed and that (b) the extent to which the retention interval affected performance depended significantly on whether the contents of the final test items had been previously re-exposed. Thus, although (a) the simple main effect of prior re-exposure was significant after both 5 min, F (1, 117) = 57.61, $\eta_p^2 = 0.33$, $p < 0.001$, and 48 h, F (1, 117) = 20.49, $\eta_p^2 = 0.15$, $p < 0.001$, and (b) the simple main effect of retention interval was significant both when items whose contents had been previously re-exposed were examined, F (1, 117) = 42.48, $\eta_p^2 = 0.27$, $p < 0.001$, and when control items were examined, F (1, 117) = 14.71, $\eta_p^2 = 0.11$, $p < 0.001$, the benefit of prior re-exposure was greater after 5 min ($M_{\mathrm{DIFF}} = 0.21$, $SE = 0.03$, 95% CI [0.15, 0.26]) than after 48 h ($M_{\mathrm{DIFF}} = 0.13$, $SE = 0.03$, 95% CI [0.07, 0.18]), and the decrease in performance as a function of the retention interval was greater for the items whose contents had been previously re-exposed ($M_{\mathrm{DIFF}} = -0.22$, $SE = 0.03$, 95% CI [− 0.15, − 0.28]) than for the control items ($M_{\mathrm{DIFF}} = -0.13$, $SE = 0.04$, 95% CI [− 0.06, − 0.20]).

Discussion

Overall, the results of Experiment 4 reveal that the holistic benefits of true-false tests can be retained following more practical retention intervals and imply, moreover, that these benefits are not only competitive with the benefits of restudy interventions but might also be more durable across greater retention intervals. That is, relative to the effect of no additional practice, the evaluation of true-false statements on a practice test enhanced performance on a later cued-recall test after both 5 min and 48 h. Although restudying the propositions of interest produced a similar pattern of benefits relative to no additional exposure and resulted in numerically superior recall after 5 min, the evaluation of true-false items, as shown in Table 4, resulted in numerically superior recall on the corresponding final test items after the more practical retention interval of 48 h. Moreover, it should be re-emphasized that (a) learners in the true-false practice condition were only truly re-exposed to all experimental propositions of interest that were restudied if these propositions were, in fact, successfully retrieved during true-false practice and that, (b) despite this disadvantage, the true-false practice condition remained competitive with the restudy condition on the final test.

Furthermore, given the implied interaction with retention interval that our primarily between-subjects design was, perhaps, underpowered to detect, it is possible that the significant reduction of the size of the overall benefit of either intervention (relative to no additional practice or exposure) as a function of the retention interval was predominantly—although not significantly—driven by the greater decrease in performance in the restudy condition than in the true-false practice condition from 5 min to 48 h. This interpretation of the present findings, although somewhat tentative, is consistent with other examinations of the testing effect (e.g., Roediger and Karpicke 2006a; Toppino and Cohen 2009) and is strengthened by the methodological advantage afforded to participants within the restudy condition, who, presumably, would not have demonstrated the same levels of performance without the explicit re-exposure to all experimental propositions of interest.

Did True-False Testing Yield Evidence of Negative Suggestion?

To investigate whether true-false testing yielded evidence of negative suggestion, we analyzed the frequencies with which learners answered final test items with competitive, incorrect answers (e.g., *Castle Geyser* when *Steamboat Geyser* was correct) when under experimental and control conditions. To that end, we conducted a 3 (type of prior practice: true practice vs. false practice vs. control) × 2 (type of content: tested content vs. related content) repeated-measures ANOVA of the final test responses for each of the first three experiments and a 2 (prior re-exposure (within-subjects): re-exposed vs. control) × 2 (method of prior re-exposure (between-subjects): true-false practice test vs. restudy intervention) × 2 (retention interval (between-subjects): 5 min vs. 48 h) ANOVA of the final test responses for the fourth experiment. The means for each condition in Experiments 1–3 and Experiment 4 are presented in Tables 7 and 8, respectively.

Experiments 1–2

The two-way interaction of prior practice and content was significant in Experiment 1, F (2, 116) = 26.64, $\eta_p^2 = 0.32$, $p < 0.001$, and Experiment 2, F (2, 54) = 16.62, $\eta_p^2 = 0.38$, $p < 0.001$. These interactions, which we investigated further with tests of simple effects, broadly indicated that the extent to which true practice, false practice, and no practice yielded evidence of negative suggestion depended on the content that was assessed.

For tested items, the simple main effect of prior practice was significant in Experiment 1, F (2, 57) = 17.40, $\eta_p^2 = 0.38$, $p < 0.001$, and Experiment 2, F (2, 54) = 17.59, $\eta_p^2 = 0.39$, $p < 0.001$, indicating that true practice, false practice, and no practice yielded different levels of evidence of negative suggestion when later recall of previously tested content was assessed on the final test. In both experiments, true practice did not yield significant evidence of negative suggestion relative to the control condition (ps ≥ 0.536), but false practice did (ps < 0.001).

For related items, the simple main effect of prior practice was significant in Experiment 1, F (2, 57) = 6.27, $\eta_p^2 = 0.18$, $p = 0.003$, and Experiment 2, F (2, 54) = 12.01, $\eta_p^2 = 0.31$, $p < 0.001$, indicating that true practice, false practice, and no practice yielded different levels of evidence of negative suggestion when later recall of related content was assessed on the final test. In Experiment 1, true practice did not yield evidence of negative suggestion relative to the control condition ($p = 0.678$), and false practice seemed to protect against the influence of negative suggestion relative to the control condition ($p =$ 0.005). In Experiment 2, however, true practice yielded significant evidence of negative suggestion relative to the control condition ($p < 0.001$), and the false practice condition did not differ from the control condition ($p = 0.859$).

Table 7 Mean proportions of incorrect, competitive responses (SEs) with 95% CIs on the final test in Experiments 1–3

Experiment	Prior practice	Tested content		Related content	
		M (*SE*)	95% CI	*M* (*SE*)	95% CI
1	True	0.06 (0.01)	[0.04, 0.09]	0.14 (0.02)	[0.09, 0.18]
	False	0.24 (0.03)	[0.18, 0.29]	0.05 (0.01)	[0.02, 0.07]
	Control	0.05 (0.02)	[0.02, 0.08]	0.12 (0.02)	[0.08, 0.17]
2	True	0.05 (0.02)	[0.02, 0.09]	0.21 (0.03)	[0.15, 0.26]
	False	0.22 (0.03)	[0.17, 0.28]	0.07 (0.02)	[0.04, 0.11]
	Control	0.06 (0.02)	[0.03, 0.09]	0.07 (0.02)	[0.03, 0.10]
3	True	0.11 (0.02)	[0.07, 0.15]	0.11 (0.02)	[0.06, 0.15]
	False	0.12 (0.02)	[0.08, 0.16]	0.11 (0.02)	[0.07, 0.16]
	Control	0.08 (0.02)	[0.04, 0.11]	0.09 (0.02)	[0.05, 0.13]

Table 8 Mean proportions of incorrect, competitive responses (SEs) with 95% CIs on the final test in Experiment 4

Experiment	Condition	5-min retention interval		48-h retention interval	
		M (*SE*)	95% CI	*M* (*SE*)	95% CI
4	True-false	0.14 (0.01)	[0.11, 0.16]	0.14 (0.01)	[0.11, 0.17]
	Control	0.07 (0.02)	[0.05, 0.10]	0.05 (0.02)	[0.02, 0.08]
	Restudy	0.08 (0.02)	[0.06, 0.12]	0.10 (0.02)	[0.07, 0.13]
	Control	0.09 (0.02)	[0.06, 0.12]	0.05 (0.02)	[0.02, 0.08]

Thus, Experiments 1 and 2 yielded some evidence of negative suggestion as a result of false practice when later recall of tested content was assessed on the final test. When later recall of related content was assessed on the final test, however, false practice either did not stimulate negative suggestion (per Experiment 2) or else significantly prevented its influence (per Experiment 1). Meanwhile, true practice never prevented the influence of negative suggestion but, strangely, did seem to stimulate it in Experiment 2 when later recall of related content was assessed on the final test. Intriguingly, this pattern suggests that the manipulation of the syntactic sequence of the true-false items between Experiments 1 and 2, which did not affect the learning of accurate information, might have affected the influence of negative suggestion.

Experiment 3

In a departure from the findings of Experiments 1 and 2, Experiment 3 produced no significant evidence of negative suggestion. That is, the analysis of Experiment 3 yielded no significant main effects of prior practice or content (ps ≥ 0.249) and no significant interaction of prior practice and content ($p = 0.887$), indicating that evidence of negative suggestion was not significantly more apparent on any measure in any of the experimental conditions than in the control conditions, although it should be emphasized that the observed means nevertheless present numerical evidence of negative suggestion that we might not have been sufficiently powered to detect.

Experiment 4

In Experiment 4, however, the two-way interaction between prior re-exposure and the method of prior re-exposure reached significance, F (1, 117) = 6.78, $\eta_p^2 = 0.06$, $p =$ 0.010, broadly indicating that the extent to which participants demonstrated significant evidence of negative suggestion depended on the experimental intervention and whether items whose contents had been previously re-exposed or control items were examined. Indeed, a significant simple main effect of prior re-exposure at the true-false level of the method of re-exposure confirmed that, across retention intervals, the true-false practice condition demonstrated significantly greater evidence of negative suggestion on the corresponding final test items than on the control items, F (1, 117) = 29.83, $\eta_p^2 = 0.20$, $p < 0.001$. Moreover, a significant simple main effect of the method of re-exposure on items whose contents had been previously re-exposed confirmed that, on such items, the true-false practice condition demonstrated significantly greater evidence of negative suggestion than the restudy condition across retention intervals, F (1, 117) = 9.06, $\eta_p{}^2 = 0.07$, $p = 0.003$.

Metacognitive Measures from Experiments 1–4

At the conclusion of Experiments 1–3, participants answered three metacognitive questions that assessed (a) how helpful participants thought that the true-false practice test was, (b) the general studying preferences of the participants, and (c) whether participants were aware that the false items that they had previously evaluated were correctible in more than one way. For each question, the proportion of respondents per response did not differ significantly across experiments, χ^2 (4, N = 172) = 1.47, $p = 0.833$, χ^2 (4, N = 172) = 2.90, $p = 0.575$, and χ^2 (2, N = 172) = 0.55, $p = 0.761$, respectively. Overall, the true-false practice test was more commonly regarded as helpful (36–46%) or neither helpful nor harmful (33–42%) rather than harmful (21–23%). Although approximately one-third of participants indicated no preference, participants were fairly evenly split regarding whether they generally preferred rereading (27–42%) or practice testing (26–38%) as a primary studying strategy, which was similar to other research (e.g., Blasiman et al. 2017). Finally, a somewhat narrow majority of participants (58–64%) claimed to be aware that the false items that they had previously evaluated could have been corrected in more than one way.

Participants also answered three similar metacognitive questions at the conclusion of Experiment 4, and the proportion of respondents per response for the final two items depended significantly on whether participants had experienced the true-false practice test or the restudy intervention, χ^2 (2, N = 121) = 1.20, $p = 0.549$, χ^2 (2, N = 121) = 11.82, $p = 0.003$, χ^2 (1, N = 121) = 8.53, $p = 0.003$, respectively. Consistent with the first three experiments, most participants regarded their respective interventions as helpful (51–61%) or neither helpful nor harmful (28–36%) rather than harmful (11–13%). In an intriguing twist that bears similarities with some prior research (e.g., Baddeley and Longman 1978; Yue et al. 2013), however, when participants were informed of the full nature of the experiment, the true-false practice test was more commonly regarded as the superior intervention in the restudy condition (74%) than in the true-false practice condition (44%), the restudy intervention was more commonly regarded as the superior intervention in the true-false condition (53%) than in the restudy condition (23%), and very few participants expressed indifference in this regard (3–4%). Finally, while most participants indicated that they were aware that the false items could have been corrected in more than one way (or generally believe that false items can be corrected in multiple ways), this endorsement was less common in the true-false practice condition (63%) than in the restudy condition (86%).

General Discussion

Far from supporting a comprehensive indictment of true-false testing, the present findings reveal that, despite their potential costs, true-false tests can provide meaningful opportunities for learning. In Experiments 1 and 2, we found that the processes underlying the evaluation of conventional true and false items enhanced the subsequent recall of tested and related content, respectively, and that this differential pattern was robust to variations of syntactic structure. The significance of these findings, which reveal that the poor reputation of true-false testing is not entirely warranted, is heightened by the fact that such benefits were observed on final cued-recall tests, which required the recall of studied content. Furthermore, Experiment 3 revealed that the separate benefits of testing with true and false items—retention of explicitly tested information and transfer to related information—could both be obtained with the incorporation of competitive clauses (e.g., *Castle Geyser (not Steamboat Geyser) is the tallest geyser*), which, moreover, might reduce (but not eliminate) the influence of negative suggestion. Finally, Experiment 4 confirmed that the benefits of true-false tests with such competitive clauses are robust to more practical retention intervals and, while not found to be significantly superior to the benefits of restudying targeted propositions, are at least comparable and perhaps more durable over greater retention intervals.

Potential Costs of Negative Suggestion

With respect to the influence of negative suggestion, however, and in an extension of prior research (e.g., Toppino and Luipersbeck 1993; Toppino and Brochin 1989), the present findings also reveal that the evaluation of true-false items without feedback can significantly increase the extent to which learners mistakenly recall incorrect answers that are competitive with correct answers. Although such costs are important to consider, the results of our experiments nevertheless indicate that the influence of negative suggestion does not overshadow the significant benefits of true-false tests, perhaps particularly when test items are thoughtfully constructed with competitive clauses. Moreover, it should be emphasized that the benefits of true-false evaluation emerged despite the fact that learners never received corrective feedback and, accordingly, could never have known with certainty whether their evaluations were correct, let alone whether any retrieved propositions were accurate. Furthermore, the influence of negative suggestion within true-false items might, as is the case within multiple-choice items (Butler and Roediger 2008), be significantly reduced with such feedback, the provision of which might also serve to enhance the benefits of true-false tests. Although additional research is critical, such prospects seem particularly encouraging when considering the relative ease with which true-false tests, like multiple-choice tests, can be graded.

True-False vs. Multiple-Choice Tests

To the extent that the inclusion of competitive clauses makes a true-false item similar to a two-alternative multiple-choice item, one might argue that the findings of Experiments 3 and 4 merely amount to a re-demonstration of the benefits of competitive multiple-choice tests (e.g., Little et al. 2012). The evaluation of true-false items, however, differs from that of multiple-choice items in at least three important ways. First, unlike multiple-choice answer options, true-false answer options convey minimal semantic information. Second, whereas multiple-choice tests require learners to retrieve associations between question-stems and correct answers, true-false tests, presumably, do not. Third, and perhaps most importantly, the typical expectation with multiple-choice tests is that there are, in fact, correct associations to be found within the test items. Such expectations cannot occur with true-false tests, however, because any given true-false item, even those with competitive clauses, might not feature the components of any correct associations.

Furthermore, from a practical standpoint, true-false items with competitive clauses might often be easier to write than, for example, multiple-choice items with four competitive alternatives. While it is one thing to be aware that the incorrect alternatives on a multiple-choice item should be competitive with the correct response, it is—as any experienced educator will readily agree—another matter entirely to generate such test items. Not only might true-false items with competitive clauses save time for educators to devote to other matters, but it also seems plausible that such test items might be particularly suitable for educators to identify or otherwise focus on distinctions known to pose difficulties for students.

Future Directions

In accordance with the suggested examination of whether the provision of feedback might enhance the benefits of true-false tests and reduce the influence of negative suggestion, future research might profitably expand upon the findings of the present study and address its potential limitations through several lines of investigation. Given that the highest proportion of correct answers demonstrated within any of our experimental conditions was 0.58, for instance, researchers might choose to examine the effects of true-false tests with materials and manipulations that promote higher levels of performance that are more representative of educational contexts. Moreover, to determine whether the detection of these effects is moderated by various aspects of the materials, researchers might also alter or manipulate the content that is learned, the organization of this content, the relative amounts of true and false practice, and the characteristics of the final tests that are used to assess retention and transfer (e.g., assessing cue-target combinations that differ from those on an initial test). Finally, with respect to the ecological limitations of experimental laboratories, systematic investigations might also be set within authentic educational contexts (e.g., Foss and Pirozzolo 2017; Jones et al. 2016) to further inform the development and optimization of true-false tests as effective learning interventions.

Reaching a Verdict: Concluding Comments

We discovered that true-false tests can elicit productive retrieval processes that resemble those of other types of tests. Such benefits, however, might not be as consistent or powerful as those elicited by other tests (cf. Rowland 2014) and can depend critically on how true-false items are constructed and the conditions of reference with which the effects of true-false tests are compared. Our conclusions, which represent a more nuanced perspective than complete exoneration, are supported by four key points. First, conventional true-false items seem to elicit the retrieval of information that is highly specific and, irrespective of the syntactic placement of key terms, tethered to those terms. Second, although such focused retrieval might have represented a limitation of true-false testing, broader retrieval processes can be elicited when competitive clauses—and possibly other modifications—are incorporated. Third, we confirmed that evidence of these broader processes is still detectable after retention intervals that are more representative of practical contexts. Fourth, and finally, the effects of true-false testing were highly competitive with, although not clearly superior to, those of a restudy condition. Collectively, the present results thus imply that, despite their potential costs, the true-false tests of the past century might have been more beneficial for learning than previously assumed and, although concerned educators might reasonably prefer other test formats, promote what could be a promising instrument for the benefit of educators and learners alike.

Author Contributions All authors were involved in the conception and design of the study. J.A.B. generated initial hypotheses, programmed all experiments, coordinated their execution, and analyzed and interpreted the

data. J.A.B. and S.C.P. drafted the manuscript with input from E.L.B. and R.A.B. All authors approved the manuscript for submission.

Compliance with Ethical Standards

All procedures performed in studies involving human participants were in accordance with the ethical standards of the Office of the Human Research Protection Program at the University of California, Los Angeles (IRB# 11-002880) and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study.

References

Baddeley, A. D., & Longman, D. J. A. (1978). The influence of length and frequency of training session on the rate of learning to type. *Ergonomics, 21*(8), 627–635.

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

Blasiman, R. N., Dunlosky, J., & Rawson, K. A. (2017). The what, how much, and when of study strategies: comparing intended versus actual study behaviour. *Memory, 25*(6), 784–792.

Butler, A. C., & Roediger III, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36*(3), 604–616.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20*(6), 633–642.

Cocks, A. W. (1929). *The pedagogical value of the true-false examination*. Baltimore: Warwick & York.

Downing, S. M. (1992). True-false, alternate-choice, and multiple-choice items. *Educational Measurement: Issues and Practice, 11*(3), 27–30.

Druckman, D., & Bjork, R. A. (1994). *Learning, remembering, believing: enhancing human performance*. Washington, DC: National Academy Press.

Ebel, R. L. (1970). The case for true-false test items. *The School Review, 78*(3), 373–389.

Foss, D. J., & Pirozzolo, J. W. (2017). Four semesters investigating frequency of testing, the testing effect, and transfer of training. *Journal of Educational Psychology, 109*(8), 1067–1083.

Glover, J. A. (1989). The "testing" phenomenon: not gone but nearly forgotten. *Journal of Educational Psychology, 81*(3), 392–399.

Hertzberg, O. E., Heilman, J. D., & Leuenberger, H. W. (1932). The value of objective tests as teaching devices in educational psychology classes. *Journal of Educational Psychology, 23*(5), 371–380.

Jersild, A. T. (1929). Examination as an aid to learning. *Journal of Educational Psychology, 20*(8), 602–609.

Jones, A. C., Wardlow, L., Pan, S. C., Zepeda, C., Heyman, G. D., Dunlosky, J., & Rickard, T. C. (2016). Beyond the rainbow: retrieval practice leads to better spelling than does rainbow writing. *Educational Psychology Review, 28*(2), 385–400.

Keys, N. (1934). The influence of true-false items on specific learning. *Journal of Educational Psychology, 25*(7), 511–520.

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.

Pan, S. C., & Rickard, T. C. (2017). Does retrieval practice enhance learning and transfer relative to restudy for term-definition facts? *Journal of Experimental Psychology: Applied, 23*(3), 278–292.

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: meta-analytic review and synthesis. *Psychological Bulletin, 144*(7), 710–756.

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

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

Roediger III, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*(1), 20–27.

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

Roediger III, H. L., & Karpicke, J. D. (2006b). The power of testing memory: basic research and implications for educational practice. *Perspectives on Psychological Science, 1*(3), 181–210.

Roediger III, H. L., Agarwal, P. K., Kang, S. H. K., & Marsh, E. J. (2010). Benefits of testing memory: best practices and boundary conditions. In G. M. Davies & D. B. Wright (Eds.), *New frontiers in applied memory* (pp. 13–49). Brighton, U.K.: Psychology Press.

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.

Sproule, C. E. (1934). Suggestion effects of the true-false test. *Journal of Educational Psychology, 25*(4), 281–285.

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.

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.

Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval: questions and answers. *Experimental Psychology, 56*(4), 252–257.

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.

Venn, J. (1884). Studies and exercises in formal logic, including a generalisation of logical processes in their application to complex inferences. *Mind, 9*(34), 301–304.

Yue, C. L., Bjork, E. L., & Bjork, R. A. (2013). Reducing verbal redundancy in multimedia learning: an undesired desirable difficulty ? *Journal of Educational Psychology, 105*(2), 266–267.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.