# All of the Above: When Multiple Correct Response Options Enhance the Testing Effect

Anthony J. Bishara and Lauren A. Lanzo

College of Charleston

## Author Note

Anthony J. Bishara, Department of Psychology, College of Charleston.

Lauren Lanzo, Department of Psychology, College of Charleston.

We thank Darius Becker-Krail, Joshua Edelson, and Caroline Gardner for help with both data collection and coding. We also thank Branden Abushanab, Dorothy Alice Blake, Marino Mugayar-Baldocchi, Giao Nguyen, and Elizabeth Schilb for help with data collection. Experiments 1A and 2 were conducted as part of the requirements for Lauren Lanzo's Bachelor's Essay at the College of Charleston. Lauren Lanzo is now at East Carolina University.

Correspondence concerning this article should be addressed to Anthony J. Bishara, Dept. of Psychology, College of Charleston, 66 George St., Charleston, SC 29424. E-mail: BisharaA@cofc.edu

Word Count: 7,802 (excluding cover, references, tables, and figures)

Abstract

Previous research has shown that multiple choice tests often improve memory retention. However, the presence of incorrect lures often attenuates this memory benefit. The current research examined the effects of "all of the above" (AOTA) options. When such options are correct, no incorrect lures are present. In the first three experiments, a correct AOTA option on an initial test led to a larger memory benefit than no test and standard multiple choice test conditions. The benefits of a correct AOTA option occurred even without feedback on the initial test; for both 5-minute and 48-hour retention delays; and for both cued recall and multiple choice final test formats. In the final experiment, an AOTA question led to better memory retention than did a control condition that had identical timing and exposure to response options. However, the benefits relative to this control condition were similar regardless of the type of multiple choice test (AOTA or not). Results suggest that retrieval contributes to multiple choice testing effects. However, the extra testing effect from a correct AOTA option, rather than being due to more retrieval, might be due simply to more exposure to correct information.

KEYWORDS: testing effect, multiple choice, all of the above, memory

# All of the Above: When Multiple Correct Response Options Enhance the Testing Effect

Testing often enhances memory retention, a finding sometimes referred to as the "testing effect" (Roediger & Karpicke, 2006; for recent reviews, see Carpenter, 2012; Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013). Indeed, testing is so powerful that it can often cause better memory retention than equivalent restudy opportunities (Carrier & Pashler, 1992; Kuo & Hirshman, 1996, 1997). Such mnemonic benefits of testing have produced a fertile ground for research, perhaps due to the testing effect's obvious implications for educational settings (e.g., Butler & Roediger, 2007; Glass & Sinha, 2013; McDaniel, Roediger, & McDermott, 2007; Rawson & Dunlosky, 2012). The present research examines a testing situation common to many classrooms and lecture halls, a situation where multiple correct responses are possible within a single test question. Specifically, the current research examines the effect of "all-of-the-above" multiple choice questions on subsequent memory retention.

Previous research on the testing effect has shown that initial multiple choice tests often enhance memory retention. Initial multiple choice tests often improve accuracy on a later test even after extended delays, and regardless of whether the later test is cued recall or multiple choice (Butler & Roediger, 2008; Fazio, Agarwal, Marsh, & Roediger, 2010; Roediger & Marsh, 2005; Spitzer, 1939). However, as a negative side effect of multiple choice testing, this accuracy benefit is often attenuated as the number of lures on the initial test increases. Additionally, a larger number of lures on the initial test often leads to increased lure responses (i.e., intrusions) on later tests (Roediger & Marsh, 2005; also see, e.g., Fazio et al., 2010). It should be noted that such negative side effects of additional lures do not always occur, particularly if initial multiple choice performance is near ceiling (Butler, Marsh, Goode, & Roediger, 2006; Marsh, Roediger,

Bjork, & Bjork, 2007). Overall, though, multiple choice testing tends to enhance memory retention, but a large number of incorrect lures often dampens this retention benefit.

Incorrect lures could have several possible effects on memory, and these effects are not necessarily mutually exclusive. Incorrect lures could increase the fluency of incorrect information, leading to misattributions about their veracity. Indeed, several studies have found that exposure to statements or true-false questions increases later ratings of their truthfulness. This effect occurs both for true and false ideas, and it occurs even when participants are forewarned that such information may be false (Hasher, Goldstein, & Toppino, 1977; Parks & Toth, 2006; Toppino & Brochin, 1989; Toppino & Luipersbeck, 1993). Such findings could explain why exposure to incorrect lures often decreases the mnemonic benefits of testing, and leads to greater acceptance of these lures on later tests. Importantly, if the presence of incorrect lures increases the perceived truth of those lures, an "all of the above" (AOTA) question should show an enhanced testing effect only in certain circumstances. When AOTA is the correct option, there are no incorrect lures present. Thus, when AOTA is the correct option, fluency should increase for the correct options above it rather than for incorrect options. In contrast, when AOTA is the incorrect option, the participant will be exposed to the same number of specific incorrect lures as with standard multiple choice questions, and so when it is the incorrect option, it should result in later memory performance that is similar to standard multiple choice tests (without an AOTA option).

The negative effects of incorrect lures could also be due commitment effects, either simplistic or sophisticated. In terms of the simplistic version of the commitment effect, students often report that changing answers is ill-advised because changing answers is believed (incorrectly) to lead to lower scores (Benjamin, Cavell, & Shallenberger, 1984). Thus, once a

student has committed to particular response, even if it is wrong, he or she may be unlikely to correct it when later given the opportunity. By this simplistic commitment account, a participant commits to the response option as a surface cue, without necessarily committing to other ideas that are logically consistent with it. This type of commitment effect could cause a commitment to the AOTA option without commitment to the individual options listed in "a" through "d" above it. However, commitment to an answer could also be more sophisticated. For example, one may commit to an answer on an initial test because one has found consistencies between that answer and one's background knowledge (Marsh et al., 2007), or perhaps because one has discovered two or more correct options and concludes that AOTA must be true. In such cases, one may be committing not just to an option in a superficial manner, but to the full meaning of that option. The simplistic and sophisticated commitment effects make different suggestions regarding the circumstances in which AOTA should enhance the benefits of testing. The simplistic account suggests that AOTA benefits should be restricted to final recognition tests, specifically when the AOTA option reappears on the final test. In contrast, by the sophisticated commitment effect account, if the AOTA response is chosen through a logical reasoning process, the participant is likely to come to believe all of the answers above it, not just the surface cue of "all-of-the-above." Thus, by the sophisticated account, AOTA should enhance the benefits of testing whenever AOTA is the correct response on the initial test, and by this account, the benefits should appear more generally, even in later cued recall tests (where "all of the above" is no longer an option).

The theoretical accounts described above – fluency and commitment effects – have been useful in explaining the effects of traditional multiple choice tests. Additional accounts may also be relevant when there is more than one correct response option on an initial test. For example,

AOTA questions may sometimes aid memory through induction of a more general category representation. In support of this idea, learning to classify members into categories via repeated testing can result in better classification than if the learning occurred via repeated studying (Jacoby, Wahlheim, & Coane, 2010). An AOTA test question with multiple correct response options may especially aid the induction process, encouraging learners to develop a more accurate representation of the category, a representation that can fit all correct response options rather than only one. This should result in improved memory performance on later tests, particularly when the AOTA option is correct on the initial test.

The mnemonic effects of AOTA options have not been examined in the existing literature, but closely related work has been done on "none of the above" options (Odegard & Koen, 2007). In those experiments, participants took an initial memory test that had standard multiple choice test questions (options A through D), questions with a correct "none of the above" option (option E), questions with an incorrect "none of the above" option (also option E), and also no-test control items. The correct "none of the above" option caused the benefits of testing to disappear, or even reverse. That is, the correct "none of the above" condition led to final memory test accuracy that was similar to or even lower than that of the no-test control condition. Thus, the presence of solely incorrect lures led to worse memory performance (also see Jang, Pashler, & Huber, in press). Conversely, it is possible that the presence of solely correct response options, as occurs with correct AOTA options, may enhance memory performance.

The present research is the first to examine the effects of AOTA questions on memory retention. Experiments 1A, 1B, and 2 were intended to establish the effects of AOTA options on

later memory. Experiment 3 was conducted to determine whether such mnemonic effects of AOTA options were due to exposure, due to retrieval, or due to both.

## Experiment 1A

In Experiment 1A, participants read non-fiction passages and took an initial multiple-choice test. On this initial test, there were four within-participant conditions: No-Test, Standard-Test, AOTA-right, and AOTA-wrong (see Figure 1 for an example). After the initial multiple choice memory test, participants solved math problems for 5 minutes, and then took a final cued-recall memory test.

On the initial test, when the AOTA option was the correct option, all specific responses above it (A, B, C, and D) were correct, and thus there was no exposure to specific incorrect responses. However, when the AOTA option was the incorrect one, only one specific response option above it was correct, much like in standard multiple choice tests. Thus, it was expected that the AOTA-right initial test format would lead to higher accuracy and higher confidence on the final test as compared to the standard-test or AOTA-wrong conditions. Additionally, due to the more basic testing effect, it was also expected that the standard-test condition would lead to higher accuracy and confidence than the no-test condition.

### Method

#### Participants

There were 108 participants (30 male, 78 female). Participants were from the College of Charleston, a medium sized liberal arts university in South Carolina. In exchange for participating, participants received credit toward a research requirement in an introductory psychology course. Participants were tested in groups of no more than 14 individuals. Each participant was tested on a separate computer and had a private view of his or her own screen.

The number of recruited participants was determined primarily by the availability of participants in the participant pool.

**Design and Materials**

The experiment had a within-subjects design with 4 levels for the initial test type: No-Test (control), Standard-Test, AOTA-right, and AOTA-wrong. Figure 1 shows an example. In the No-Test condition, the question did not appear on the initial test. In the Standard-Test condition, participants selected from options "a" through "d". In the AOTA-right condition, participants selected from options "a" through "e," where "e" was "all of the above," and it was correct. The AOTA-wrong condition was structured in the same way as the previous condition, except that "all of the above" was incorrect.

Study materials consisted of 10 passages adapted from the Encyclopedia Britannica. Passage length ranged from 167 to 226 words, which is similar in length to passages used in other testing effect research (e.g., McDaniel, Howard, & Einstein, 2009; Roediger & Karpicke, 2006). The passage topics were chosen to be distinctive from one another (e.g., Jazz, Nightshade, Wildebeests, etc.). Of the 10 passages, there were 2 practice passages and 8 critical passages. The 8 critical passages were divided into 4 sets of 2 passages each, with each set assigned to a specific initial test condition. Assignment of passages to conditions was counterbalanced across participants.

Each critical passage had 3 questions related to that passage (e.g., "European instrument(s) that influenced Jazz include"). Each question was constructed such that there could be 4 possible correct options and 3 possible lure options. The appearance of correct and/or lure options on the initial multiple choice test depended on the condition in which the question appeared (see Figure 1 for examples). Note that, because questions were nested within passages,

and passages were assigned to conditions, in no case did a question from one condition relate to a passage in a different condition. In other words, a given passage received only one initial test type (e.g., Standard-Test) on the initial test, and for the No-Test condition, the whole passage was not tested on the initial test (in pilot testing, we found that it did not matter whether random assignment was used to assign questions to conditions or to assign entire passages to conditions).

The initial multiple choice test consisted of 4 practice questions about the practice passages, followed by 18 critical questions about the critical passages. The 18 critical questions were evenly divided among 3 initial test conditions (i.e., all except for the "No-Test" condition), resulting in 6 critical questions per cell. Assignment of correct and lure options to locations "a." through "d." was random on each trial. The AOTA option, when available, was always in location "e."

The final test consisted of 3 practice questions, followed by 24 critical questions about the critical passages. The final test also had 6 critical questions per cell, but because up to 4 correct responses were possible per question, each cell had 24 measurements per participant.

**Procedure**

In the Study Phase, participants were instructed to read each passage carefully because their memory would be tested later. Each passage was presented on the screen, with the title of the passage (e.g., "Jazz") in bold at the top center of the screen. The length of time each passage remained on the screen was determined by the number of syllables in the passage multiplied by 0.3 seconds (pilot testing determined that this pace allowed unrushed reading for most participants).

Next, in the Initial Test Phase, participants answered multiple choice questions. For each question, the passage title appeared at the top of the screen, with the question below it, and the

four or five answer options below that. Participants responded by pressing a letter on the keyboard ("a" through "e"). Participants had an unlimited amount of time to answer each question. The order of critical questions was random.

Next, in the Retention Phase, participants were instructed to solve as many math problems as they could within 5 minutes. They made responses on the keyboard.

In the Final Test Phase, on each trial, the topic and question appeared at the top of the screen, and below it were four spaces for responses. Participants were instructed to fill in each blank space and to guess if necessary. Participants responded by typing on the keyboard and pressing enter to proceed to the next blank space. After pressing enter for the final blank space, an arrow appeared on the screen pointing to the typed response for first blank space. Participants were then asked how confident they were in that response on a scale from 1 to 7. The numbers 1 through 7 were shown at the bottom of the screen, along with the label "Purely Guessing" below "1," and "Absolutely Certain" below "7." Participants pressed a number on the keyboard. The process was repeated, with the arrow moving to the 2$^{nd}$ typed response, and so on, until all 4 typed responses had received a confidence rating. Participants had an unlimited amount of time to respond to all questions, and the order of critical questions was random.

**Coding Cued Recall Responses on the Final Test**

Two coders decided whether each cued recall response counted as a correct response, a lure response, or neither. Coders made decisions independently of one another and were blind to condition. Coders agreed in 91.1% of cases. For the purpose of analyzing correct and lure responses, disagreements among coders were averaged (e.g., if only 1 of 2 coders coded a response as correct, the response was counted as .5 correct). For the purpose of analyzing

confidence conditional on the correctness of a response, in order to assure that conditionalization was strict, disagreements among coders were excluded from the analyses.

## Results and Discussion

For ANOVAs, when the sphericity assumption was violated, the Greenhouse-Geisser correction was used.

### Final Test Proportion Correct

As shown in Figure 2A, accuracy on the final test was highest for items that had an all-of-the-above option on the initial test, but only when that option had been correct. Accuracy on the final test was analyzed with a one-way repeated measures ANOVA with initial test type as the independent variable. Supporting the idea that the AOTA-right condition led to better memory retention, there was a significant main effect of initial test type, $F(2.4,253.8)=34.74$, $p<.001$, $\eta_p^2=.25$. Post hoc paired t-tests showed that the AOTA-right condition had significantly higher accuracy than No-Test, Standard-Test, and AOTA-wrong conditions, $p$s < .001, $d$=.80, .87, and .64, respectively. There was a trend for the Standard-Test condition and the AOTA-wrong conditions to be higher than the No-Test control condition, but these differences did not reach significance, $p$s = .07, $d$=.18 and .17.

### Final Test Confidence

Confidence paralleled the mean proportion correct across conditions on the final test, and it did so more generally across all experiments. For the interested reader, confidence ratings can be found in the Appendix.

### Final Test Lure Responses

Participants occasionally ($M$=.07) produced a lure response (i.e., an incorrect response option that could appear on the initial multiple choice test) on the final cued recall test. Just as

AOTA-right led to the highest accuracy, it also led to the lowest probability of a lure response. Of course, this is not surprising given that the AOTA-right condition did not expose participants to lures in the first place. The probability of a lure response was lowest in the AOTA-right condition (*M*=.03, *SE*=.003), next lowest in the No-Test condition (*M*=.04, *SE*=.004), higher in the Standard-Test condition (*M*=.09, *SE*=.007), and highest in the AOTA-wrong condition (*M*=.12, *SE*=.007). A one-way repeated measures ANOVA on lure response probabilities showed a significant main effect of initial test type, $F(2.5,262.2)=61.65$, $p<.001$, $\eta_p^2=.37$. Post-hoc paired t-tests showed that all conditions were significantly different from one another, all *p*s < .01. Of particular interest, lure responses were rarer in the AOTA-right condition than in the No-Test, Standard-Test, and AOTA-wrong conditions, *d*=.31, .83, and 1.06. Lure responses were more frequent in the AOTA-wrong condition than in the No-Test or Standard Test conditions, *d*=.89 and .28.

**Initial Test Response Time**

If participants spent more time on AOTA-right questions on the initial test, that pattern could provide a rather uninteresting explanation for the superior memory retention of that condition. To rule out this explanation, response times on the initial test were examined. Responses times longer than 30 seconds were excluded to reduce the influence of outliers. (To avoid a biased analysis, this cut-off was established by examining cumulative frequencies blind to condition, and this cut-off eliminated only 1.3% of the trials.) If anything, response times were shorter for AOTA-right initial test trials (*M*=8.5, *SD*=2.4 seconds) than for the Standard-Test (*M*=8.9, *SD*=2.2) and the AOTA-wrong trials (*M*=10.0, *SD*=2.8). In other words, the superior memory retention observed in the AOTA-right condition could not be explained as due simply to increased exposure time on the initial test.

### Initial Test Proportion Correct

Could the benefits of the AOTA-right condition on the final test be due simply to more successful retrieval on the initial test? To address this question, proportion correct on the initial test was examined. Proportion correct on the initial test was similar for the AOTA-right (*M*=.71, *SD*=.21) and the Standard-Test conditions (*M*=.71, *SD*=.20), which were in turn higher than the AOTA-wrong condition (*M*=.61, *SD*=.25), $F(2,214)=10.25$, $p<.001$, $\eta_p^2= .09$. In other words, differences in initial test accuracy could not account for the mnemonic benefits of an initial AOTA-right relative to an initial Standard-Test question, though they might account for the benefits relative to the AOTA-wrong condition.

## Experiment 1B

Experiment 1B was a small follow-up experiment meant to address two issues. First, the previous experiment examined a retention period of only 5 minutes, which may not be considered practically relevant for memory retention. To address this issue, Experiment 1B used a 2-day retention period to determine whether the mnemonic benefits of a correct all-of-the-above option endured for a more meaningful period of time. Instead of a 5-minute math distractor task, participants left the lab and returned 2 days later before taking the final test.

The second issue addressed in Experiment 1B was that, in Experiment 1A, the basic testing effect appeared as a trend in the means (Standard-Test > No-Test condition for the proportion correct), but was not statistically significant. In order to have more assurance about our findings, we hoped to replicate the basic testing effect, while still showing an additional benefit for the AOTA-right condition. In Experiment 1B, the 2-day retention period was expected to help distinguish the Standard-Test condition from the No-Test condition because testing effects are sometimes more apparent after longer delays (Carpenter, Pashler, Wixted, &

Vul, 2008; Congleton & Rajaram, 2012; Hogan & Kintsch, 1971; Kornell, Bjork, & Garcia, 2011; Meyer & Logan, 2013; Roediger & Karpicke, 2006).

With a 2-day retention period, we expected to again find that the AOTA-right condition led to the highest proportion of accurate responses. Additionally, we expected to replicate the basic testing effect, as shown by significantly higher accuracy in the Standard-Test condition than the No-Test condition.

## Method

Only the differences in methodology relative to Experiment 1A are reported here. There were 29 participants (10 male, 19 female). After taking the initial test, participants left the lab, and then returned 48-hours later to take the final test. Participants were tested individually. When examining responses on the final test, coders agreed in 95.3% of cases.

## Results and Discussion

### Final Test Proportion Correct

As shown in Figure 2B, there was both a general testing effect, whereby Standard-Test condition had higher accuracy than the No-Test condition, and also an extra accuracy benefit in the AOTA-right even beyond that of the other tested conditions. There was a significant main effect of initial test type, $F(3,84)=11.80$, $p<.001$, $\eta_p^2=.30$. Consistent with a general testing effect, post hoc paired t-tests showed that the No-Test condition had significantly lower accuracy than the Standard Test condition, $p<.05$, $d=.41$, as well as the both the AOTA-right and AOTA-wrong conditions, $p$s $< .05$, $d=.97$ and .38. Additionally, the AOTA-right condition had significantly higher accuracy than the Standard-Test and AOTA-wrong conditions, $p$s $< .01$, $d=.64$ and .76. The Standard-Test and AOTA-wrong conditions were not significantly different from one another, $p=.46$, $d=.14$.

**Final Test Lure Responses**

As before, the AOTA-right condition had fewer lure responses than the other tested conditions, though its lure rates no longer fell below the No-Test condition. The probability of a lure response was lowest in the No-Test condition (*M*=.09, *SE*=.02), next lowest in the AOTA-right condition (*M*=.12, *SE*=.03), higher in the Standard-Test condition (*M*=.20, *SE*=.03), and highest in the AOTA-wrong condition (*M*=.27, *SE*=.04). A one-way repeated measures ANOVA on lure response probabilities showed a significant main effect of initial test type, $F(3,84)=6.43$, $p<.001$, $\eta_p^2=.19$. Post-hoc paired t-tests showed that the No-Test condition had significantly fewer lure responses than the Standard-Test and AOTA-wrong conditions, *p*s < .01, *d*=.53 and .70. Additionally, the AOTA-right condition had significantly fewer lure responses than the AOTA-wrong condition, $p < .01$, *d*=.54.

**Initial Test Response Time**

Just as with the previous experiment, initial test response time could not explain the superior memory retention of the AOTA-right condition. Response time was no greater for the AOTA-right condition (*M*=8.4, *SD*=2.8 seconds) than for the Standard-Test condition (*M*=8.4, *SD*=2.4) or AOTA-wrong condition (*M*=9.9, *SD*=2.8).

**Initial Test Proportion Correct**

As with Experiment 1A, initial test accuracy could not account for the benefits of AOTA-right relative to the Standard-Test condition. If anything, proportion correct on the initial test was lower for the AOTA-right (*M*=.74, *SD*=.25) than for the Standard-Test condition (*M*=.78, *SD*=.14), but the lowest initial test accuracy was again in the AOTA-wrong condition (*M*=.63, *SD*=.25), $F(2,56)=4.74$, $p<.05$, $\eta_p^2= .15$.

## Experiment 2

Experiment 1A and 1B showed that AOTA response options – when correct – led to better memory performance on a final cued recall test. In Experiment 2, the final cued recall test was replaced with a final multiple choice test. Given the prevalence of multiple choice tests in educational settings, it was hoped that the benefits of a correct AOTA option would generalize to later multiple choice tests.

Specifically, Experiment 2 used the same 4 types of initial tests as before. There was a 5-minute retention period after the initial test. On the final test, all multiple choice questions had an "e. all of the above" option. This AOTA option was right for half of the questions on the final test, and wrong for the other half. It was expected that the AOTA-right condition (for the initial test) would again lead to superior memory retention, regardless of the question type on the final test.

### Method

Only changes from the previous experiment's method are reported below.

#### Participants

Initially, 100 participants were recruited, but data from three were lost due to computer malfunction during the experiment, leaving 97 total (19 male, 78 female). Participants were tested in groups of no more than 14 individuals.

#### Design and Materials

The experiment had a 4 (initial test type) X 2 (final test type) within-participant design. The final memory test type had 2 levels: the AOTA option was either right or wrong on the final test.

In the current experiment, participants were able to respond faster on a multiple choice than a cued recall final test, and so it was possible to expand the materials list without increasing the experiment's duration. There were 2 practice passages and 16 critical passages. The 16 critical passages were divided into 8 sets of 2 passages each, with each set assigned to a specific combination of initial test and final test type (counterbalanced across participants).

The initial multiple choice test consisted of 4 practice questions about the practice passages, followed by 48 critical questions about the critical passages. The 48 critical questions were evenly divided among 3 initial test conditions (i.e., all except for the "No-Test" condition) and 2 final test conditions, resulting in 8 critical questions per cell. The final multiple choice test consisted of 4 practice questions, followed by 64 critical questions about the critical passages. The final multiple choice test also had 8 critical questions per cell.

### Procedure

In the Study Phase, the duration of each passage was determined by the number of syllables in the passage multiplied by 0.26. In the Retention Phase, participants solved math problems for 5 minutes. In the Final Test Phase, on each trial, a participant would first answer a multiple choice question by pressing the corresponding letter on the keyboard ("a" through "e"). After answer the multiple choice question, for the confidence rating, an arrow appeared on the screen pointing to the response option that had been chosen, and then the participant gave a confidence rating for that option.

## Results and Discussion

### Final Test Proportion Correct

As shown in Figure 3, whenever there was only one correct answer on the final test (i.e., AOTA was wrong on the final test), performance closely resembled that of the earlier

experiments (left half of Figure 3). However, when AOTA was the correct answer on the final test (right half of Figure 3), the pattern was altered somewhat. Importantly, just as in previous experiments, accuracy on the final test tended to be highest for items that had an all-of-the-above option on the initial test, but only when that option had been correct (AOTA-right on initial test).

Accuracy on the final test was analyzed with a 4 (initial test type) X 2 (final test type) repeated measures ANOVA. Supporting the idea that the AOTA-right condition led to better memory retention, there was a significant main effect of initial test type, $F(3,288)=13.31$, $p<.001$, $\eta_p^2=.12$. The main effect of final test type was not significant, $F(1,96)=3.24$, $p=.08$, $\eta_p^2=.03$. However, this pattern was qualified by a significant initial test by final test type interaction, $F(3,288)=7.14$, $p<.001$, $\eta_p^2=.07$.

Results were further examined with follow-up paired t-tests. Examining first the simplest final test, where there was only one correct option (left half of Figure 3), results appeared similar to those of previous experiments. The initial test AOTA-right condition led to significantly higher accuracy than all three other initial test types, $p$s < .05, $d$=.50, .20, and .46 for No-Test, Standard-Test, and AOTA-wrong conditions respectively. A basic testing effect also appeared, with the Standard-Test condition having significantly higher accuracy than the No-Test condition, $t(96) = 2.67$, $p < .01$, $d$=.27. The AOTA-wrong condition was not significantly different from the No-Test condition, $t(96)=.34$, $p=.73$, $d=.03$.

Examining the final test condition where all items were correct (right half of Figure 3), the No-Test condition was unusually high, perhaps because participants were inclined to guess "all-of-the-above" when they had not seen a similar test item earlier in the experiment. Such a guess would have been correct here. The AOTA-right condition again led to significantly higher accuracy than the Standard-Test and AOTA-wrong conditions, $p$s < .001, $d$=.45 and .36.

However, the AOTA-right condition was not significantly different from the No-Test condition, $p$=.25, $d$=.12.

**Initial Test Response Time**

As with previous experiments, initial test response time could not explain the enhanced memory retention in the AOTA-right condition. Initial test response times were no slower for the AOTA-right conditions ($M$=9.3, $SD$=2.2 seconds) than for the Standard-Test ($M$=9.3, $SD$=2.3) or AOTA-wrong conditions ($M$=10.2, $SD$=2.1) during the initial test.

**Initial Test Proportion Correct**

Again, initial test accuracy could not account for the benefits of AOTA-right relative to the Standard-Test condition. If anything, proportion correct on the initial test was lower for the AOTA-right ($M$=.65, $SD$=.20) than for the Standard-Test condition ($M$=.67, $SD$=.18), but the lowest initial test accuracy was again in the AOTA-wrong condition ($M$=.58, $SD$=.16), $F(2,192)=11.34$, $p<.001$, $\eta_p^2= .11$.

## Experiment 3

Experiments 1 and 2 showed that an AOTA option, when it was correct, led to enhanced memory retention. However, the cause of this enhanced memory retention was unclear. One possibility is that the memory enhancement was due to special retrieval processes involved in such questions. A second possibility is that the memory enhancement was due to the exposure to additional correct response options. To illustrate, in the example shown in Figure 1, in the AOTA-right condition, the initial test provided exposure to 4 correct response options (saxophone, trumpet, piano, and string bass). In contrast, in the Standard-test and AOTA-wrong conditions, the initial test provided exposure to only 1 correct option (trumpet). Finally, it is

possible that both retrieval and exposure contribute to enhanced memory retention in the AOTA-right condition.

There is no previous research comparing exposure to retrieval for AOTA questions. Indeed, more generally, it is rare for multiple choice testing effect experiments to include an exposure control condition. In the few experiments that have done so (Butler & Roediger, 2007; Kang, McDermott, & Roediger, 2007), multiple choice testing has not produced significant improvements in retention relative to the exposure control. It is possible that the benefits of multiple choice testing were hard to detect because, in those experiments, the exposure control condition was not equivalent to a multiple choice test in terms of duration or presentation format, but instead consisted of summary paragraphs or sentences that reiterated the study material.

To help distinguish the effects of exposure and retrieval, we adapted a procedure from Carrier and Pashler (1992). Their experiments compared a *Test-Study* condition to a nearly identical *Pure-Study* condition. In the Test-Study condition, a cue word was presented alone for the first half of a trial, followed by the cue word plus its response word answer for the last half of the trial. As a basis for comparison, in the Pure-Study condition, participants were exposed to the cue word plus its response for the entire trial. In both conditions, participants were encouraged to try to recall the correct response as quickly as possible. This procedure provides a stringent test of retrieval's effects on memory retention. In both conditions, participants are exposed to the correct answer, and total trial time is equated. The Pure-Study condition even involves more feedback about the correct answer (the entire trial rather than half). However, in the Pure-Study condition, presenting the correct response immediately at the start of the trial spoils the need for retrieval. Thus, if the Test-Study condition leads to better performance on the final test, then there is evidence that retrieval has benefited memory. Indeed, Carrier and Pashler

found that the Test-Study condition led to better memory retention than the Pure-Study condition, supporting the idea that retrieval enhances retention (for other examples based on this procedure, see Bishara & Jacoby, 2008; Carpenter & Pashler, 2007).

In Experiment 3, we compared a Test-Study condition to a Pure-Study condition. The experiment was similar to Experiment 1B, with a 48-hour retention period and a final cued recall test. However, after the Study Phase, the Initial Test phase included a manipulation of answer appearance (Test-Study versus Pure-Study). In the Test-Study condition, the question and multiple choice options were shown on the screen for a fixed period of time. During the last half of the trial, the correct option (e.g., "b. trumpet") was highlighted by a different font color. In the Pure-Study condition, the procedure was identical except that the correct option was highlighted for the entire trial, thereby obviating the need for retrieval. If retrieval contributes to memory retention, then performance on the final test should be better for items in the Test-Study condition as compared to the Pure-Study condition.

The Test-Study versus Pure-Study manipulation was crossed with the initial test type (Standard-Test, AOTA-right, and AOTA-wrong). The no-test control condition was omitted because the Pure-Study condition provides a basis for comparison, and also because there was no clean way to manipulate Pure-Study versus Test-Study without an initial test. Consistent with the literature on the testing effect, it was expected that retrieval would generally enhance memory retention, which would be shown by a significant main effect of answer appearance (Test-Study > Pure-Study). If AOTA-right enhances memory via exposure, then there should also be a significant main effect of initial test type. Finally, if AOTA-right enhances memory via extra retrieval processes that are unique to that condition, then there should be a significant interaction between answer appearance and the initial test type.

## Method

Only the differences in methodology relative to Experiment 1B are reported here.

### Participants

There were 77 participants, but only 70 attended both sessions (19 male, 51 female).

### Design and Materials

The experiment had a 3 (initial test type: Standard-Test, AOTA-right, and AOTA-wrong) X 2 (answer appearance: Test-Study and Pure-Study) within-participant design. The 24 critical multiple choice questions were divided into 6 sets of 4 questions each, with each set assigned to a specific combination of initial test type and answer appearance (counterbalanced across participants). Questions from a single passage could occur in different initial test type or answer appearance conditions for the same participant.

Answer appearance was manipulated on the initial test. In both the Test-Study condition and the Pure-Study condition, the question and multiple choice response options appeared on the screen in black font. Each initial test trial lasted 22 seconds. In the Test-Study condition, the correct answer option (e.g., "b. trumpet") changed from black to green font after 11 seconds had elapsed, and remained green for the remaining 11 seconds. In the Pure-Study condition, the correct answer option appeared in green font immediately at the start of the trial, remaining green for all 22 seconds.

### Procedure

At the beginning of the initial test phase, participants were instructed to type a letter for each multiple choice question. Participants were told that the correct answer would appear in green font at various points in time, and that sometimes the answer would appear immediately, even before pressing a letter. They were instructed to type a letter as soon as possible, regardless

of whether or not the answer had already appeared in green. Participants returned to the lab for the final test 48-hours later.

### Coding Cued Recall Responses on the Final Test

Coders agreed in 95.8% of cases.

## Results and Discussion

### Final Test Proportion Correct

As shown in Figure 4, just like in previous experiments, a correct all-of-the-above option on an initial test improved memory performance on a final test. Perhaps more importantly, results suggest that retrieval enhanced retention, but retrieval did so regardless of the initial test type. Accuracy on the final test was analyzed with a 3 (initial test type) X 2 (answer appearance) repeated measures ANOVA. Supporting the idea that the AOTA-right condition led to better memory retention, there was a significant main effect of initial test type, $F(2,138)=10.24$, $p<.001$, $\eta_p^2=.13$. Follow-up t-tests confirmed that retention was higher in the AOTA-right condition ($M=.26$) than either the Standard-Test or AOTA-wrong conditions (both .20), $p$s < .001, $d=.48$ and .42. Supporting the idea that retrieval enhanced retention, memory retention was significantly higher in the Test-Study condition (.23) than the Pure-Study condition (.21), as shown by a significant main effect of answer appearance, $F(1,69)=4.30$, $p<.05$, $\eta_p^2=.06$. The interaction between initial test type and answer appearance was not significant, $F(2,138)=.20$, $p=.82$, $\eta_p^2=.00$. In other words, retrieval appeared to enhance memory in all types of multiple choice tests, but there was no evidence to suggest special retrieval-related memory improvements in the AOTA-right condition that differed from those in other conditions.

**Final Test Lure Responses**

As expected, the AOTA-right condition reduced lure responses. However, the Test-Study condition slightly increased lure responses. Lure responses were rarer in the AOTA-right condition (.02) than in the Standard-Test condition (.04) or AOTA-wrong condition (.05), $F(1.8,123.8)=15.45$, $p<.001$, $\eta_p^2=.18$. Lure responses were slightly more frequent in the Test-Study condition (.04) than the Pure-Study condition (.03), $F(1,69)=6.33$, $p<.05$, $\eta_p^2=.08$. The interaction was not significant, $p=.34$.

## General Discussion

The results presented here extend knowledge of the testing effect by showing that "all of the above" options, when correct, enhance the mnemonic benefits of testing. The memory benefits of AOTA were observed both with final cued recall (Experiments 1A, 1B, & 3) and multiple-choice tests (Exp. 2), and they were observed both after 5-minute (Exp. 1A and 2) and 2-day delays (Exp. 1B & 3). Additionally, the memory benefits were observed even when there was no feedback on the initial test (Experiments 1A, 1B, & 2). It was not the AOTA option in general that enhanced retention, but rather the correct AOTA option, accompanied by only correct options above it, that did so. This pattern is consistent with the general finding that multiple choice testing benefits memory, and also that such benefits are attenuated as a result of the presence of incorrect lures (Fazio et al., 2010; Odegard & Koen, 2007; Roediger & Marsh, 2005).

Multiple choice questions generally expose learners to at least one correct option, and so a testing effect from multiple choice questions could potentially be due simply to exposure rather than retrieval. As shown by the comparison between Pure-Study and Test-Study conditions (Exp. 3), retrieval did indeed help retention. However, it did so to similar degrees regardless of

whether the initial test question involved standard multiple choice or AOTA multiple choice. Thus, Experiment 3 suggested that the extra mnemonic benefits of a correct AOTA option could be due to the extra exposure to correct options, rather than due to special or extra retrieval processes involved in evaluation of such questions.

The size of the retrieval benefit in Experiment 3, as shown by the comparison between Pure-Study and Test-Study conditions, might seem rather modest, but it is meaningful for at least two reasons. First, the benefit of retrieval in the Test-Study condition persisted after a 48-hour delay. Second, and more importantly, the Test-Study condition provided a challenging test of the retrieval hypothesis. The Test-Study condition provided feedback for only half of the trial, whereas the Pure-Study condition provided feedback for the entire trial. Thus, any retrieval benefit in the Test-Study condition must be sufficiently large enough to overcome the extra feedback time in the Pure-Study condition. The fact that we observed any advantage of the Test-Study condition under such circumstances is a testament to the power of retrieval for enhancing retention. Indeed, to our knowledge, these results are the first to show that multiple choice testing effects are not solely due to exposure; retrieval contributes to multiple choice testing effects.

The AOTA-right condition produced a memory retention advantage over other conditions, but there was no evidence to suggest that it gained that advantage through retrieval. Rather, the extra mnemonic benefits in the AOTA-right condition may be due to mechanisms that rely on exposure to several correct options, mechanisms such as fluency effects and sophisticated commitment effects. Both of these mechanisms would predict that positive effects of testing are enhanced by the presence of correct information at the initial test, and dampened by the presence of incorrect information. However, in contrast to the simplistic commitment

effect, the benefits of the AOTA-right condition were not restricted to situations where participants could make the exact same "all-of-the-above" response on a later test. Rather, the benefits occurred both on later cued recall and multiple choice tests. Additionally, they occurred even when the correct response on an initial multiple choice test did not match that of a final multiple choice test (Exp. 2). Thus, the pattern of results is unlikely to be due simply to commitment to the surface cue "all of the above," but perhaps to a more sophisticated commitment to the specific options listed above the AOTA option.

The present research did not show evidence of special or extra retrieval processes for AOTA questions, but other contexts might show different results. The retrieval benefits of AOTA could depend on the construction of the response options. For example, if the options are related to one another, then retrieval of one correct option in AOTA-right condition might mediate retrieval of another correct option (see Little, Bjork, Bjork, & Angello, 2012). Additionally, retrieval benefits might depend on the proportion of correct AOTA options (i.e., the proportion of trials in which "e" is the correct option). A high proportion of correct AOTA options might encourage mindless guessing of "e" and discourage retrieval of previously studied information.

Although the present research showed benefits of AOTA questions for a later multiple choice test (Exp. 2), the later multiple choice test used here was unusual in that it had an AOTA option in every question. This later test was somewhat similar to a typical multiple choice test, though, whenever the AOTA option was wrong because, in that case, one of the simple options (a. through d.) was correct, just like in a traditional multiple-choice test. The only difference was that there was an additional lure ("e. all of the above"). In this situation, final test performance was still improved by AOTA-right on the initial test. This pattern suggests that the benefits of

AOTA on an initial test will likely generalize to final tests that involve more traditional multiple choice formats. Of course, future research should more directly examine this issue.

In some ways, the effects of an AOTA option appear to be the converse of those of a "none-of-the-above" (NOTA) option. In the present experiments, when AOTA was the correct option, all specific options above it had to be correct, and so later memory performance was enhanced. In contrast, when NOTA was the correct option (Odegard & Koen, 2007), all specific options above it had to be incorrect, and so later memory performance was impaired.

A comparison of AOTA and NOTA testing effects might suggest that, when educators are constructing tests, NOTA questions should be avoided, but AOTA questions should be encouraged. Of course, AOTA questions only enhanced the testing effect when AOTA was the correct option. If instructors were to use only correct AOTA options, students would likely learn to choose the AOTA option mindlessly, perhaps diminishing the testing effect. However, wrong AOTA options could also be used. In the present experiments, even when the AOTA option was wrong, it sometimes benefited memory retention as compared to the No-Test control condition, or at least it did no harm. This pattern contrasts with the one for NOTA: when NOTA was the correct option, not only was it less beneficial, it sometimes significantly reduced retention performance below baseline level (see Odegard & Koen, 2007, Exp. 2). Thus, a correct NOTA option is one of the few situations where testing reduces rather than increases correct responses on a later memory test (see also Bishara & Jacoby, 2008; Peterson & Mulligan, 2013). Unlike NOTA, AOTA appears to help, or at worse, be benign, at least in terms of its mnemonic benefits. There was sometimes a slight increase in lure responses in the AOTA-wrong condition relative to the Standard-Test condition. However, this increase in lure responses was more than made up for by the AOTA-right condition's decrease in lures responses and increase in accurate

responses, as well as by the finding that the AOTA-wrong condition's accuracy tended to be similar to that of the Standard-Test condition. Overall, the pedagogical benefits of including AOTA questions on tests appear to outweigh the costs.

Despite these mnemonic advantages of AOTA questions, there may be a psychometric cost. Textbooks on educational testing typically discourage the use of AOTA questions because such questions allow students to answer with partial knowledge, and so these questions do not reliably measure understanding of all correct response options (Haladyna, Downing, & Rodriguez, 2002; also see Albanese, 1993). That is, even though AOTA questions may have desirable memory properties, they may also have undesirable measurement properties. In contrast to AOTA questions, questions that allow any number of correct options (e.g., "mark all that are correct") tend to have higher reliability and validity (Harasym, Leong, Violato, Brant, & Lorscheider, 1998). Such "mark all" questions could potentially serve both mnemonic and measurement purposes. Like AOTA questions, they could still re-expose learners to more correct responses and fewer lures – on average - than standard multiple choice questions do. Additionally, "mark all" questions might lead to interrogation of each individual response option, which could lead to more elaborative retrieval processes.

The present research was the first to examine the effects of "all of the above" multiple choice questions on later memory, and there are several conclusions that can be drawn from it. First, an "all of the above" option, when correct, produces a testing effect above and beyond that of standard multiple choice questions. Second, multiple choice tests enhance memory retention at least partly through retrieval. Third, the extra testing effect from an "all of the above" question, rather than being due to more retrieval, might be due simply to more exposure to correct information. Finally, even though "all of the above" questions have been discouraged for

measurement purposes, these question formats and similar ones could provide potent learning opportunities for students.

## References

Albanese, M. A. (1993). Type K and other complex multiple-choice items: An analysis of research and item properties. *Educational Measurement: Issues and Practice*, *12*(1), 28-33. doi:10.1111/j.1745-3992.1993.tb00521.x

Benjamin, L. T., Cavell, T. A., & Shallenberger, W. R. (1984). Staying with initial answers on objective tests: Is it a myth? *Teaching of Psychology*, *11*(3), 133-141.

Bishara, A. J., & Jacoby, L. L. (2008). Aging, spaced retrieval, and inflexible memory performance. *Psychonomic Bulletin & Review*, *15*(1), 52-57. doi:10.3758/PBR.15.1.52

Butler, A. C., Marsh, E. J., Goode, M. K., & Roediger, H. (2006). When additional multiple-choice lures aid versus hinder later memory. *Applied Cognitive Psychology*, *20*(7), 941-956. doi:10.1002/acp.1239

Butler, A. C., & Roediger, H. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology*, *19*(4-5), 514-527. doi:10.1080/09541440701326097

Butler, A. C., & Roediger, H. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*, *36*(3), 604-616. doi:10.3758/MC.36.3.604

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science*, *21*(5), 279-283. doi:10.1177/0963721412452728

Carpenter, S. K., & Pashler, H. (2007). Testing beyond words: Using tests to enhance visuospatial map learning. *Psychonomic Bulletin & Review*, *14*(3), 474-478. doi:10.3758/BF03194092

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition, 36*(2), 438-448. doi:10.3758/MC.36.2.438

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20*(6), 633-642. doi:10.3758/BF03202713

Congleton, A., & Rajaram, S. (2012). The origin of the interaction between learning method and delay in the testing effect: The roles of processing and conceptual retrieval organization. *Memory & Cognition, 40*(4), 528-539. doi:10.3758/s13421-011-0168-y

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. doi:10.1177/1529100612453266

Fazio, L. K., Agarwal, P. K., Marsh, E. J., & Roediger, H. (2010). Memorial consequences of multiple-choice testing on immediate and delayed tests. *Memory & Cognition, 38*(4), 407-418. doi:10.3758/MC.38.4.407

Glass, A. L., & Sinha, N. (2013). Multiple-choice questioning is an efficient instructional methodology that may be widely implemented in academic courses to improve exam performance. *Current Directions in Psychological Science, 22*, 471-477. doi: 10.1177/0963721413495870

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-334. doi:10.1207/S15324818AME1503_5

Harasym, P. H., Leong, E. J., Violato, C., Brant, R., & Lorscheider, F. L. (1998). Cueing effect of "all of the above" on the reliability and validity of multiple-choice test items. *Evaluation and the Health Professions, 21*(1), 130-133.

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. doi: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 & Verbal Behavior*, *10*(5), 562-567. doi:10.1016/S0022-5371(71)80029-4

Jacoby, L. L., Wahlheim, C. N., & Coane, J. H. (2010). Test-enhanced learning of natural concepts: Effects on recognition memory, classification, and metacognition. *Journal of Experimental Psychology: Learning, Memory, And Cognition*, *36*(6), 1441-1451. doi:10.1037/a0020636

Jang, Y., Pashler, H., & Huber, D. E. (in press). Manipulations of choice familiarity in multiple-choice testing support a retrieval practice account of the testing effect. *Journal of Educational Psychology*.

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

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory And Language*, *65*(2), 85-97. doi:10.1016/j.jml.2011.04.002

Kuo, T., & Hirshman, E. (1996). Investigations of the testing effect. *The American Journal of Psychology, 109*(3), 451-464. doi:10.2307/1423016

Kuo, T., & Hirshman, E. (1997). The role of distinctive perceptual information in memory: Studies of the testing effect. *Journal of Memory And Language*, *36*(2), 188-201. doi:10.1006/jmla.1996.2486

Little, J. L., Bjork, E., 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. doi:10.1177/0956797612443370

Marsh, E. J., Roediger, H., Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review, 14*(2), 194-199. doi:10.3758/BF03194051

McDaniel, M. A., Howard, D. C., & Einstein, G. O. (2009). The read-recite-review study strategy: Effective and portable. *Psychological Science, 20*(4), 516-522. doi:10.1111/j.1467-9280.2009.02325.x

McDaniel, M. A., Roediger, H., & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14*(2), 200-206. doi:10.3758/BF03194052

Meyer, A. D., & Logan, J. M. (2013). Taking the testing effect beyond the college freshman: Benefits for lifelong learning. *Psychology and Aging*, *28*(1), 142-147. doi:10.1037/a0030890

Odegard, T. N., & Koen, J. D. (2007). 'None of the above' as a correct and incorrect alternative on a multiple-choice test: Implications for the testing effect. *Memory, 15*(8), 873-885. doi:10.1080/09658210701746621

Parks, C. M., & Toth, J. P. (2006). Fluency, familiarity, aging, and the illusion of truth. *Aging, Neuropsychology, and Cognition*, *13*(2), 225-253. doi:10.1080/138255890968691

Peterson, D. J., & Mulligan, N. W. (2013). The negative testing effect and multifactor account. *Journal of Experimental Psychology: Learning, Memory, And Cognition*, *39*(4), 1287-1293. doi:10.1037/a0031337

Rawson, K. A., & Dunlosky, J. (2012). When is practice testing most effective for improving the durability and efficiency of student learning? *Educational Psychology Review*, *24*(3), 419-435. doi:10.1007/s10648-012-9203-1

Roediger, H., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249-255. doi:10.1111/j.1467-9280.2006.01693.x

Roediger, H., & 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. doi:10.1037/0278-7393.31.5.1155

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology*, *30*(9), 641-656. doi:10.1037/h0063404

Toppino, T. C., & Brochin, H. (1989). Learning from tests: The case of true-false examinations. *The Journal of Educational Research*, *83*(2), 119-124.

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. doi:10.1080/00220671.1993.9941229

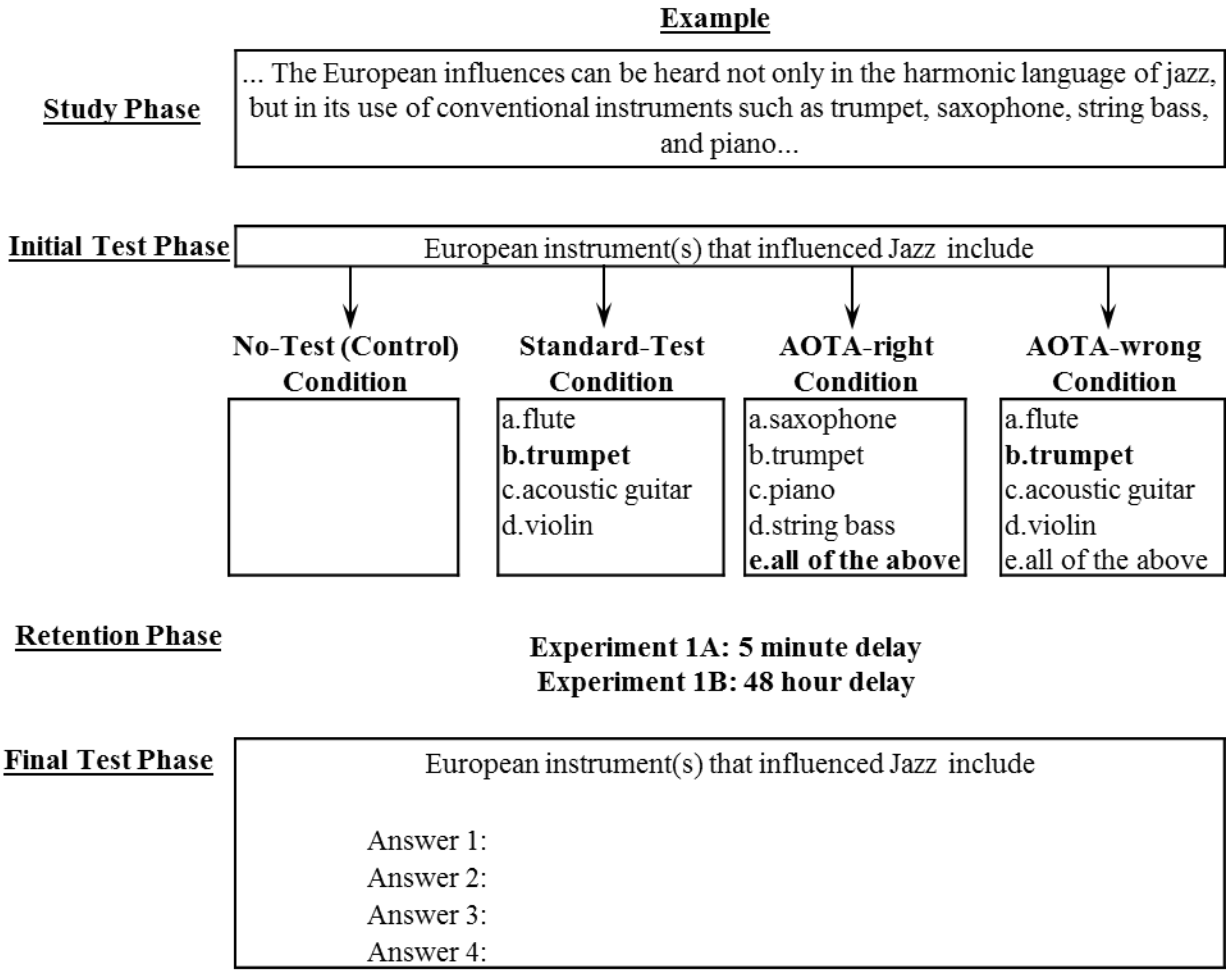
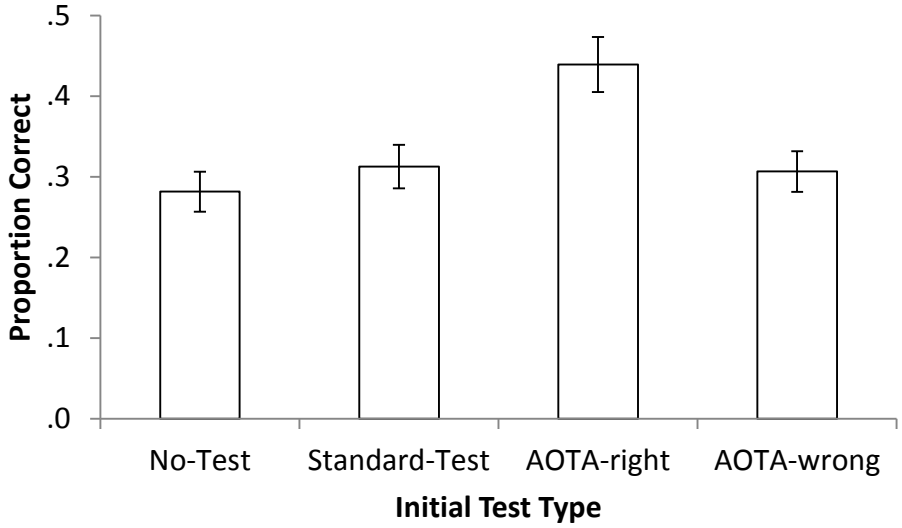



*Figure 1*. General method in Experiments 1A and 1B, as illustrated through an example. The correct answer to each multiple choice question is highlighted in bold (though it was not bold in the actual experiments). AOTA=All Of The Above.

**A. Experiment 1A: After 5-minute Retention Delay**




**B. Experiment 1B: After 48-hour Retention Delay**

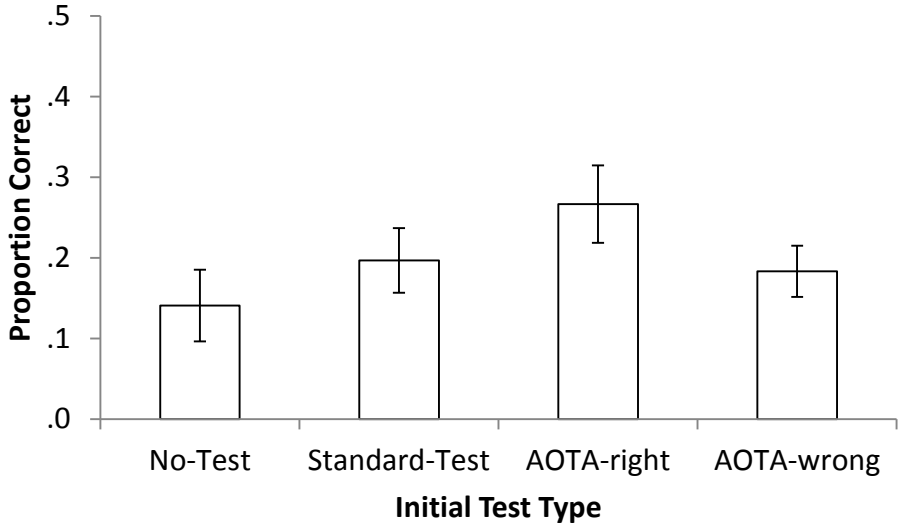



*Figure 2*. In Experiments 1A and 1B, proportion correct on the final cued recall test as a function of the initial test format. Error bars show 95% confidence intervals of the mean. AOTA=All Of The Above.

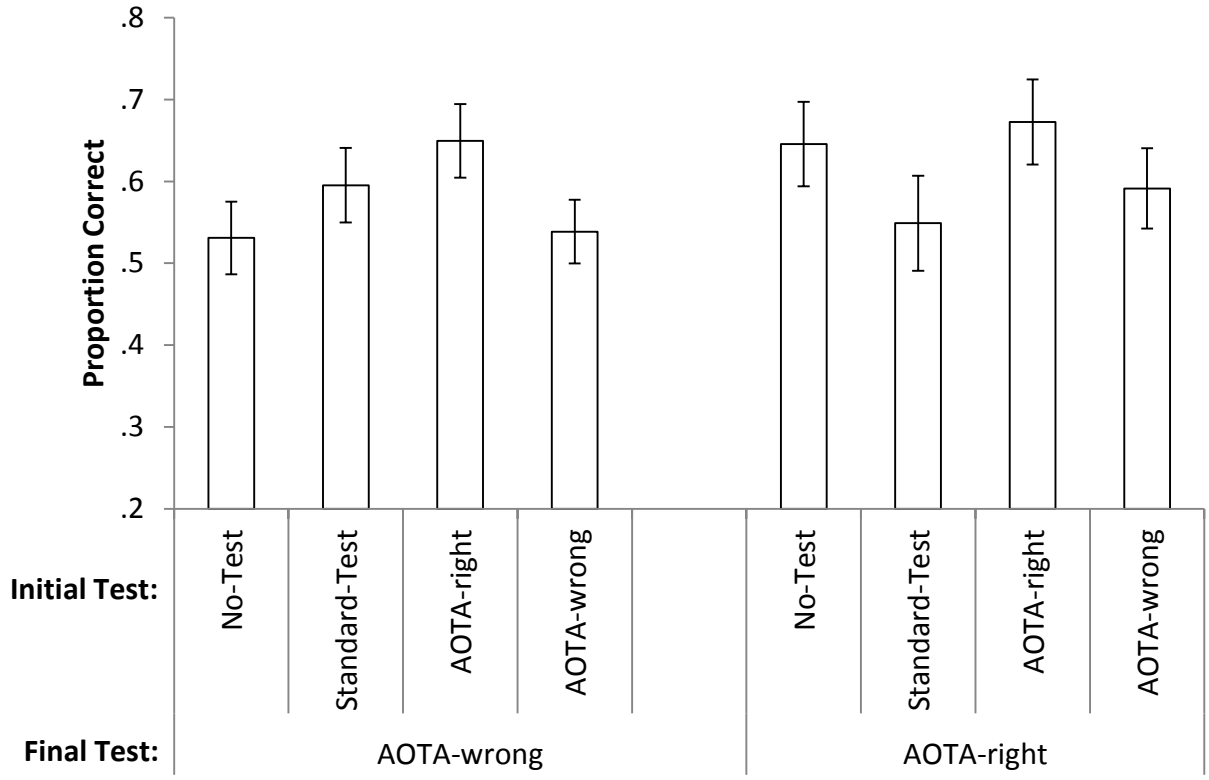



*Figure 3*. In Experiment 2, proportion correct on the final test multiple choice test as a function of the initial test and final test formats. Error bars show 95% confidence intervals of the mean. AOTA=All Of The Above.

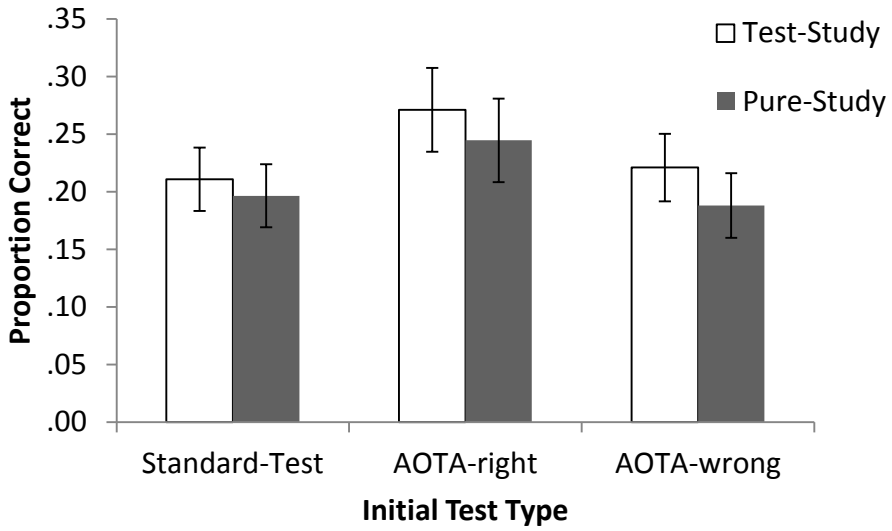



*Figure 4*. In Experiment 3, proportion correct on the final test as a function of the initial test type and answer appearance (Test-Study versus Pure-Study) on the initial test, which occurred 48-hours earlier. Error bars show 95% confidence intervals of the mean. AOTA=All Of The Above.

# Appendix

Table A1

*Confidence Ratings on the Final Test*

| | | Initial Test Type | | | |
|---|---|---|---|---|---|
| | | No-Test | Standard-Test | AOTA-right | AOTA-wrong |
| | | | | *M* | |
| Exp. 1A | | 3.72 | 3.89 | 4.30 | 3.92 |
| Exp. 1B | | 3.13 | 3.57 | 3.66 | 3.33 |
| Exp. 2 | AOTA-wrong on final | 4.59 | 4.71 | 4.87 | 4.75 |
| | AOTA-right on final | 5.28 | 5.10 | 5.21 | 5.14 |
| Exp. 3 | Test-Study | | 3.49 | 3.98 | 3.64 |
| | Pure-Study | | 3.47 | 3.52 | 3.37 |
| | | | | *SD* | |
| Exp. 1A | | 1.19 | 1.17 | 1.11 | 1.06 |
| Exp. 1B | | 1.50 | 1.08 | 1.14 | 1.25 |
| Exp. 2 | AOTA-wrong on final | 1.08 | 1.25 | 1.23 | 1.06 |
| | AOTA-right on final | 1.13 | 1.26 | 1.29 | 1.17 |
| Exp. 3 | Test-Study | | 1.34 | 1.39 | 1.14 |
| | Pure-Study | | 1.34 | 1.37 | 1.45 |

*Note*. AOTA=All of the Above.