APPLIED COGNITIVE PSYCHOLOGY
Appl. Cognit. Psychol. **20**: 941–956 (2006)
Published online 19 June 2006 in Wiley InterScience
(www.interscience.wiley.com). DOI: 10.1002/acp.1239

When Additional Multiple-Choice Lures Aid Versus Hinder Later Memory

ANDREW C. BUTLER[1]*, ELIZABETH J. MARSH[2], MICHAEL K. GOODE[1] and HENRY L. ROEDIGER, III[1]

[1]*Washington University in St. Louis, St. Louis, USA*
[2]*Duke University, Durham, USA*

SUMMARY

Three experiments were conducted to investigate whether increasing the number of lures on a multiple-choice test helps, hinders or has no effect on later memory. All three patterns have been reported in the literature. In Experiment 1, the stimuli were unrelated word lists, and increasing the number of lures on an initial multiple-choice test led to better performance on later free recall and cued recall tasks. In contrast, in Experiments 2 and 3, stimuli were facts from prose materials, and increasing the number of multiple-choice lures led to robust costs in cued recall and smaller costs in free recall. The results indicate that performance on the initial multiple-choice test is a critical factor. When initial multiple-choice performance was near ceiling, testing with additional lures led to superior performance on subsequent tests. However, at lower levels of multiple-choice performance, testing with additional lures produced costs on later tests. Copyright © 2006 John Wiley & Sons, Ltd.

Multiple-choice exams are frequently used in educational settings because they are easy to score objectively. Taking a test generally aids performance on later exams, a phenomenon called the testing effect (Bjork, 1975; Carrier & Pashler, 1992; Cooper & Monk, 1976; Glover, 1989; Hogan & Kintsch, 1971; Izawa, 1970; Kuo & Hirshman, 1996; Runquist, 1986; Spitzer, 1939; Thompson, Wenger, & Bartling, 1978; Tulving, 1967; Wheeler & Roediger, 1992; Whitten & Bjork, 1977). Testing can be more powerful than restudying material, even when testing only permits recovery of part of the material and restudying involves re-presentation of the entire set (McDaniel & Masson, 1985; Roediger & Karpicke, 2006).

Whereas initial tests generally produce superior retention on tests given later, evidence also exists that multiple-choice exams can have negative consequences. This is most clear when considering multiple-choice questions that do not contain the correct answer (Schooler, Foster, & Loftus, 1988). Such multiple-choice questions serve as a source of misinformation in the same way as do exposures to incorrect spellings (Brown, 1988; Jacoby & Hollingshead, 1990) or false facts embedded within a short story (Marsh, Meade, & Roediger, 2003). However, negative consequences are possible even when the questions

*Correspondence to: A. C. Butler, Department of Psychology, Washington University, Campus Box 1125, St. Louis, MO 63130-4899, USA. E-mail: butler@wustl.edu

Contract/grant sponsor: Collaborative Activity Award from the James S. McDonnell Foundation.

are more typical ones that include the correct answer among the alternatives (e.g. Roediger & Marsh, 2005). In this case the multiple-choice question requires students to carefully examine the stem and four or five plausible alternatives, only one of which is completely correct. Therefore, the multiple-choice question potentially allows the creation of both correct and erroneous associations. Consistent with these ideas, Toppino and Luipersbeck (1993) have shown that when students judge the truth of statements after taking a multiple-choice test, they increase their ratings for both the correct answers and for the multiple-choice lures. Similar increases in truth ratings have been found after reading false items on true-false tests (Toppino & Brochin, 1989). Likewise, a re-test with novel lures yields better performance than a re-test with old lures (Rees, 1986), because initial testing increases the perceived truth-value or familiarity of the lures (Jacoby, Shimizu, Daniels, & Rhodes, 2005). In a more direct demonstration, Roediger and Marsh (2005) showed that taking a multiple-choice test increased production of lures as answers on a final general knowledge test, even though subjects were warned against guessing. That is, if participants endorsed an erroneous answer on the multiple-choice test, they essentially learned it and expressed the same knowledge on the final cued recall test. The upshot is that having many alternatives on multiple-choice tests can actually be harmful to students' knowledge.

The current paper aims to resolve a puzzle that exists in the literature. The evidence reviewed in the previous paragraph (mostly with educationally relevant materials, like short prose passages) shows that the lures on multiple-choice tests can create interference and lead to erroneous knowledge. However, other patterns of results also exist in somewhat different research traditions. Most experiments have manipulated whether or not questions were tested, rather than the number of incorrect alternatives paired with the correct answer. But at least three patterns of data have been published: increasing the number of lures can increase the negative consequences of testing as just discussed (Roediger & Marsh, 2005), increase the positive benefits of testing (Whitten & Leonard, 1980) or have no effect on later tests (Brown, 1988). The aim of the present paper is to replicate and reconcile these disparate results.

A theoretical argument can be made for all three patterns of outcome, so it may well be that each outcome occurs under certain situations. First, consider the hypothesis that increasing the number of lures should increase negative consequences of testing. In general, interference theory would predict that increasing the number of competing associations would reduce access to the target (e.g. Postman & Underwood, 1973). In the prose literature, this is generally referred to as a fan effect, and has been documented with true and false facts about famous people (Lewis & Anderson, 1976). Supporting these predictions, Roediger and Marsh (2005) showed that increasing the number of multiple-choice lures increased lure intrusions on a final cued recall test of general knowledge (initially presented in prose passages), and decreased the positive testing effect.

In contrast, a general class of studies demonstrates that challenging learning conditions often lead to superior long-term retention relative to easy learning conditions ('desirable difficulties in learning'; e.g. Bjork, 1994, 1999), allowing the prediction that a more difficult multiple-choice test may aid memory. That is, several findings support the notion that increasing the difficulty or cognitive effort during encoding enhances retention (Jacoby, 1978; McDaniel, Einstein, Dunay, & Cobb, 1986; McNamara, Kintsch, Songer, & Kintsch, 1996; Tyler, Hertel, McCallum, & Ellis, 1979). Extending the idea to multiple-choice tests, one might argue that additional lures on an initial test would increase the difficulty (and thus enhance the processing) of the question and answer, which would lead to better retention of the information on a later test. Supporting this prediction, Whitten and

Leonard (1980) used a word-list learning paradigm and found that increasing the number of lures hindered initial recognition of words, but that the pattern reversed in later free recall. Specifically, the greater the number of lures provided with a target on the initial multiple-choice test, the more likely the target was to be recalled on the final free recall test. False recall was reduced numerically after increasing numbers of lures, but the effect could not be tested statistically due to a floor effect.

Thus, the perspectives of interference theory and the desirable difficulties hypothesis offer conflicting predictions about the effects of increasing the number of multiple-choice lures on later retention. These approaches are not necessarily mutually exclusive, and both may hold under certain conditions. As noted previously, a null effect of number of multiple-choice alternatives on a later test has also been obtained (Brown, 1988). This outcome may have resulted from positive and negative effect balancing one another or that the lures just had no effect on processing.

In light of this empirical and theoretical puzzle, the present research aimed to reconcile the disparate findings obtained by Whitten and Leonard (1980) and Roediger and Marsh (2005). Specifically, we sought to explore whether the different criterial tests (free recall vs. cued recall) and/or different materials (words vs. passages) are responsible for the inconsistency between published results. Three experiments were conducted in order to investigate this issue.

EXPERIMENT 1

Because our lab had previously obtained a negative effect of increasing the number of multiple-choice alternatives (i.e. Roediger & Marsh, 2005), we began by attempting to replicate the results of Whitten and Leonard (1980) using unrelated words as the stimuli and free recall as the criterial test (as they did). The only major change was the addition of a cued recall test (the dependent measure in Roediger & Marsh, 2005), which always occurred after the free recall test. Even though the cued recall test performance would be contaminated by prior free recall (the dependent variable of prime interest), the addition of cued recall was inexpensive and could lead to useful information if a different pattern of results was obtained relative to free recall.

Method

Participants and design

Forty-eight Washington University undergraduates participated for course credit. They were tested in groups of 3–10 people. The design was a single-factor, within-participants design. The number of alternatives on the multiple-choice test was 2, 4 or 8 items and the dependent variables of primary interest were later free recall and cued recall of the studied words.

Materials

In creating the stimuli and counterbalancing the multiple-choice tests, we adhered as closely as possible to Whitten and Leonard's (1980) procedure. All target and lure items consisted of minimally related nouns. First, 36 words were selected as to-be-remembered targets from the updated Battig and Montague (1969) category norms (Van Overschelde, Rawson, & Dunlosky, 2004). Each word was an exemplar from a different category,

ranging between four and seven letters in length (e.g. *topaz*—a type of precious stone, *skirt*—an article of clothing etc.). None of these target words were among the top 10 most frequent responses in their respective categories. The target words were divided into three study lists matched for average word length and frequency. Words within a study list were always studied in the same order, but the order of lists was counterbalanced across participants, so that overall, each list was studied and tested equally often in the 1st, 2nd and 3rd positions. Next, 132 additional words were generated by the experimenters to serve as lures on the multiple-choice tests. Although the lure words did not come from published category norms, they were similar to the target words in that they ranged from four to seven letters and each belonged to a unique category, different from both the target words and the other lure words (e.g. *towel, castle, straw*, *whale,* etc.). Overall, stimuli consisted of 168 words (36 targets and 132 lures), each from a different category.

Six 12-item multiple-choice tests were constructed, two for each study list. Each test question contained the target (old) item and 1, 3 or 7 lures. Thus, participants selected the target from a total of 2, 4 or 8 alternatives. Each test contained an equal number of 2-, 4- and 8-alternative questions, and across participants, targets were tested equally in each of the three testing conditions (i.e. 2, 4 or 8 items). In addition, the position of the target was varied across participants such that on each list participants would have to read (left to right) through an average of 1.5, 2.5 and 4.5 items on the 2-, 4- and 8-alternative questions respectively, before encountering the target. For both the 2- and 4-alternative questions, the target appeared equally in each of the possible positions on a given 12-question test. For example, on the 2-alternative questions, the target appeared twice in Position 1 and twice in Position 2. For the 8-alternative questions, the target item appeared equally in Positions 1, 2, 4, 5, 7 and 8 across all the lists.

Procedure

As described below, the procedure followed that used by Whitten and Leonard (1980). The only exception was the final cued recall test, which occurred after the entire original procedure was completed.

The experiment began with three study-test phases. During each study-test phase, 12 words were presented at a rate of 4 seconds per word on a PC using Microsoft PowerPoint. Immediately after studying each list, participants received a 12-question, multiple-choice test in which they were given a total of 2 minutes to identify the studied word in each of the 12 test items. Participants used a cardboard mask to move down the page, so that only one test item was visible at a time. For each item, they crossed out the unstudied words and left the target word unmarked, which insured that all items were examined to some degree. The third study-test trial was followed by an 8-minute filler task in which participants rated the 'memorability' of 48-colour photographs on a 5-point scale. Next, a 3-minute, surprise free recall test of the word lists was given. Participants were instructed to recall as many of the 36 target words as possible from the original study lists, but only to write down a word if they were reasonably sure it was correct (i.e. not to guess). This test was followed by a 3-minute, surprise cued recall test, and instructions against guessing were reiterated. The 36 critical category labels were provided, and for each, participants recalled the corresponding study word. For example, the cue '*a precious stone*' was presented, and participants were supposed to recall '*topaz*' from the studied list.

The experiment took less than 1 hour. After a full debriefing, participants were thanked and dismissed.

Table 1. Mean proportion correct on the multiple-choice test as a function of the number of response alternatives (including correct)

		Number of MC alternatives				
	Stimuli	Two	Three	Four	Six	Eight
Experiment 1	Words	0.99	—	0.95	—	0.97
Experiment 2	Prose	0.85	0.81	0.79	0.74	—
Experiment 3	Prose	0.88	0.85	0.85	0.79	—

Results

All results were significant at the 0.05 level of confidence unless otherwise noted. Pairwise comparisons were Bonferroni-corrected to the 0.05 level. In the analysis of repeated measures, a Geisser–Greenhouse correction was used for violations of the sphericity assumption (Geisser & Greenhouse, 1958). When linear trend analyses were conducted, the contrast coefficients were customised to adjust for unequal intervals.

Multiple-choice test

Mean proportion correct on the multiple-choice test is shown in the first row of Table 1 and the data are clearly near ceiling. An ANOVA revealed a significant difference among the different testing conditions, $F(2,75) = 5.74$, $MSE = 0.004$, *partial* $\eta^2 = 0.11$. However, the linear trend was not significant. Pair-wise comparisons showed the only significant difference to be between the 2- and 4-alternative conditions, $t(47) = 4.65$, $SEM = 0.009$. Overall, performance ($M = 0.97$) was similar to the extremely high level observed by Whitten and Leonard (1980).

Free recall test: correct answers

The mean proportions of correctly recalled items are shown in the first row of Table 2. An ANOVA revealed a significant linear trend $F(1,47) = 10.60$, $MSE = 0.009$, *partial* $\eta^2 = 0.18$, indicating that correct recall increased as the number of prior multiple-choice alternatives increased. A quadratic trend was also significant, $F(1,47) = 4.31$, $MSE = 0.013$, *partial* $\eta^2 = 0.08$. This finding replicates Whitten and Leonard (1980) in

Table 2. Free recall, as a function of the number of alternatives (including target) on the initial multiple-choice test

			Number of prior MC alternatives				
	Stimuli	DV	Two	Three	Four	Six	Eight
Experiment 1	Words	Correct	0.33	—	0.31	—	0.38
		Intrusions	0.00	—	0.01	—	0.00
Experiment 2	Prose	Correct	0.47	0.46	0.46	0.43	—
		Intrusions	0.06	0.05	0.07	0.11	—
Experiment 3	Prose	Correct	0.47	0.49	0.53	0.51	—
		Intrusions	0.05	0.06	0.06	0.09	—

For each experiment, the 1st dependent variable (DV) is proportion of targets correctly recalled, and the 2nd is proportion of lures intruded.

Table 3. Cued recall, as a function of the number of alternatives (including target) on the initial multiple-choice test

			Number of prior MC alternatives				
	Stimuli	DV	Two	Three	Four	Six	Eight
Experiment 1	Words	Correct	0.44	—	0.43	—	0.52
		Intrusions	0.00	—	0.00	—	0.00
Experiment 2	Prose	Correct	0.75	0.71	0.69	0.67	—
		Intrusions	0.09	0.11	0.13	0.15	—
Experiment 3	Prose	Correct	0.78	0.78	0.77	0.73	—
		Intrusions	0.08	0.10	0.10	0.12	—

For each experiment, the 1st dependent variable (DV) is proportion of targets correctly recalled, and the 2nd is proportion of lures intruded.

that a greater number of multiple-choice alternatives (i.e. 8) led to increased correct free recall relative to fewer prior alternatives (i.e. 2 and 4). There was no difference in free recall after 2- or 4-alternative testing, with a slight difference occurring in the non-predicted direction.

Free recall test: production of the multiple-choice lures

The mean proportion of multiple-choice lures intruded into free recall (i.e. incorrectly recalled items) is shown in the second row of Table 2. Practically no multiple-choice lures were intruded in free recall. No relationship between prior number of lures and false recall was apparent, but floor effects cloud any conclusion. Participants confined their free recall to list items surprisingly well.

Cued recall test: correct answers

The first row of Table 3 displays mean proportions of targets correctly recalled in response to category cues as a function of prior multiple-choice condition. A significant linear trend, $F(1,47) = 11.77$, $MSE = 0.016$, *partial* $\eta^2 = 0.20$, indicated that as the number of initial alternatives increased, the proportion of items correctly recalled also increased. Of course, this effect could be due to carryover effects from the free recall test as almost every item correctly recalled during free recall was correctly recalled on the cued recall test as well. However, note that cued recall levels were much higher than those of free recall. The quadratic trend did not reach significance, $F(1,47) = 1.45$, $MSE = 0.019$, $p > 0.23$.

Cued recall test: production of the multiple-choice lures

No items that were used as lures during the multiple-choice test were reproduced during the cued recall test.

Discussion

Overall, the data from Experiment 1 replicated those of Whitten and Leonard (1980): testing with additional multiple-choice lures facilitated correct responding on the free recall test, increasing from 33% recalled after testing with two alternatives to 38% after testing with eight alternatives. Curiously, correct free recall declined from the two-alternative condition to the four-alternative condition (a pattern also observed in the cued

recall data, albeit not to the same degree). This result is probably related to the lower level of performance on the initial multiple-choice test for the four-alternative condition relative to the other conditions (see Table 1, top row). Nevertheless, the higher level of correct free recall in the eight-alternative condition relative to the two- and four-alternative conditions indicates that that the inclusion of additional multiple-choice lures did facilitate correct responding. In addition, no cost was associated with testing with more lures: although the error rate was low, increasing the number of distracters did not increase intrusions of lures into free recall.

Experiment 1 also yielded a novel result: Whitten and Leonard's result was obtained with cued recall. That is, testing with additional multiple-choice lures facilitated responding on the final cued recall test, increasing by 8% with no intrusions in cued recall. This result must be interpreted with some caution because the cued recall test was confounded by the prior free recall test. However, the result is particularly interesting because one major difference between the original Whitten and Leonard procedure and Roediger and Marsh's was in test-type: Whitten and Leonard used free recall whereas Roediger and Marsh used cued recall. Thus, it may not be that the positive vs. negative effects of increased lures are attributable to free versus cued recall.

EXPERIMENT 2

To ensure that type of test was not the key factor, we next attempted to replicate the results of Roediger and Marsh (2005), who found with prose stimuli that increasing the number of initial multiple-choice lures had a negative effect on subsequent cued recall performance. Of particular interest was whether negative effects of testing would be obtained on both cued recall and free recall tests. The type of test variable was manipulated between-participants so that neither test would be contaminated.

Method

Participants and design

Sixty-four Washington University undergraduates participated for course credit. Participants were tested in groups of one to four people. The experiment had a 2 (final test-type: cued recall, free recall) $\times$ 4 (number of multiple-choice alternatives: 2, 3, 4, 6) mixed design. Only final test-type was manipulated between-participants.

Materials

Stimuli consisted of 12 non-fiction passages used by Roediger and Marsh (2005). Originally taken from Graduate Record Examination (GRE), Scholastic Assessment Test (SAT) and Test of English as a Foreign Language (TOEFL) practice test books, the passages varied in topic and averaged two paragraphs in length. Across participants, the passages were read in different orders randomly generated by the computer. Four multiple-choice questions were constructed for each passage. The 6-alternative version of a question paired the correct answer with five lures; all answers consisted of a single word. To create the 2-, 3- and 4-alternative forms of a given question, lures were randomly removed from the 6-alternative form. Across participants, each item was tested in all four test conditions (2-, 3-, 4- and 6-alternatives) and appeared equally often in the different possible positions. The computer randomised the question order on the multiple-choice test.

Procedure

All the testing was completed on a PC using E-Prime software (Schneider, Eschman, & Zuccolotto, 2002). In the initial learning phase, the 12 text passages were studied for 90 seconds apiece. This study phase was immediately followed by a self-paced multiple-choice test. Participants were asked to respond to every question, even if that necessitated guessing. After the test was completed, a 6-minute filled delay ensued. Finally, depending on condition, participants completed either the cued recall or the free recall test. On the cued recall test, they responded to prompts such as '*Who invented the game of basketball?*' (Answer: *Naismith*). On the free recall test, they were instructed to recall the answers to the original multiple-choice questions. In both conditions, participants were warned against guessing. Finally, participants were debriefed and dismissed.

Results

Multiple-choice test

The second row of Table 1 shows the mean proportion correct on the multiple-choice test as a function of the number of alternatives. As the number of alternatives increased, performance declined. This linear trend was significant, $F(1,62) = 38.15$, $MSE = 0.010$, *partial* $\eta^2 = 0.38$.

Free recall test: correct answers

Table 2 shows the proportion of correct answers recalled, as a function of the number of alternatives on the multiple-choice test. There was a non-significant trend for recall to decrease as a function of the number of multiple-choice alternatives, $F(1,31) = 1.07$, $MSE = 0.024$, $p > 0.30$.

Free recall test: production of the multiple-choice lures

Intrusions of multiple-choice lures into free recall are also shown in Table 2. As the number of alternatives on the multiple-choice test increased, the proportion of lures intruded in free recall increased. This was confirmed by a significant linear trend, $F(1,31) = 7.94$, $MSE = 0.006$, *partial* $\eta^2 = 0.20$. Overall, 87% of intrusions during free recall consisted of lure items that had been incorrectly selected on the initial multiple-choice test. By necessity the remaining free recall lure intrusions stemmed from items that had been correctly selected on the multiple-choice test, but for which incorrect responses were provided on later tests. This remaining proportion (13%) represents a very small number of items and therefore is not further discussed.

Cued recall test: correct answers

Correct responses on the cued recall test are shown in Table 3. As the number of alternatives on the multiple-choice test increased, the proportion of items correctly recalled on the cued recall test decreased. This was confirmed by a significant linear trend, $F(1,31) = 7.21$, $MSE = 0.014$, *partial* $\eta^2 = 0.19$. Performance dropped from 75% correct following testing with two alternatives to 67% following testing with six alternatives.

Cued recall test: production of the multiple-choice lures

Table 3 shows the use of the multiple-choice lures as answers to the cued recall questions. As the number of prior alternatives increased, so did the proportion of questions answered with multiple-choice lures. This observation was confirmed by a significant linear trend,

Table 4. Probability of correct free recall (FR) and correct cued recall (CR) conditional upon correct multiple-choice (MC) test performance; data are from Experiments 2 and 3

		Number of prior MC alternatives			
		Two	Three	Four	Six
Experiment 2	*p* (Recalled FR/Correct MC)	0.54	0.54	0.55	0.59
	p (Recalled CR/Correct MC)	0.86	0.85	0.85	0.87
Experiment 3	*p* (Recalled FR/Correct MC)	0.54	0.58	0.62	0.64
	p (Recalled CR/Correct MC)	0.88	0.91	0.89	0.91

$F(1,31) = 6.86$, $MSE = 0.010$, *partial* $\eta^2 = 0.18$. Overall, 95% of intrusions during cued recall consisted of lure items that were incorrectly selected on the initial multiple-choice test.

Conditional analyses

Of interest was how performance on the initial multiple-choice test related to subsequent free or cued recall test performance. To this end, both free and cued recall data were re-examined as a function of whether questions had been correctly or incorrectly answered on the multiple-choice test.

We began with a re-analysis of the free recall data, dependent on multiple-choice test performance. The first line of Table 4 shows the relationship between correct initial multiple-choice performance and later correct free recall. Given that a question was correctly answered on the multiple-choice test, free recall performance actually improved after testing with additional multiple-choice lures. Selecting the correct answer from among more lures yielded a larger benefit in free recall than did correctly answering a multiple-choice question with fewer lures. However, when a lure was selected on the initial multiple-choice test (i.e. an incorrect response), the probability of producing the correct response on the free recall test was very low ($M = 0.06$) and did not depend on the number of prior lures. When a lure was selected on the initial multiple-choice, the lure was frequently produced on free recall test ($M = 0.30$), but did not differ with regard to the number of prior alternatives.

Next, we re-examined the data from the cued recall test. The second line of Table 4 shows the relationship between correct initial multiple-choice performance and later correct cued recall. Interestingly, when the cued recall data were conditionalised upon initial multiple-choice performance, the effect of number of multiple-choice lures disappeared. If the correct answer was chosen on the initial multiple-choice test, the probability of answering the corresponding cued recall question was high and did not depend on the number of prior multiple-choice lures ($M = 0.86$). When participants selected a lure on the initial multiple-choice test, they rarely answered that item correctly in cued recall ($M = 0.11$) and there was no relationship to the number of prior alternatives. If a multiple-choice lure was chosen, there was a high probability that the lure would be produced on the cued recall test ($M = 0.56$); however, this probability did not depend on the number of prior alternatives.

In summary, conditionalising performance on the multiple-choice responses yielded two very distinct patterns of performance on the free and cued recall tests. When a multiple-choice question was correctly answered, free recall benefited from testing with additional

lures. Cued recall performance was related to whether or not the multiple-choice question was correctly answered, but beyond that it did not matter how many lures had been read.

Discussion

Overall, increasing the number of multiple-choice lures with prose material in Experiment 2 had a negative effect on performance. Additional lures decreased ability to select the correct answers on the multiple-choice test. While the number of multiple-choice lures did not significantly affect correct free recall, including more lures did increase the proportion of intrusions produced. An even stronger pattern was observed for cued recall: as the number of prior lures increased, correct answers decreased and lure answers increased.

The cued recall data replicated those observed by Roediger and Marsh (2005): increasing the number of multiple-choice lures had negative impact on later cued recall performance (decreasing correct answers and increasing intrusions). The free recall data tell a different story. First, increasing the number of multiple-choice lures had no significant effect on correct free recall, a result unlike both Whitten and Leonard (1980) and Roediger and Marsh (2005) but like that of Brown (1988). However, increasing the number of multiple-choice lures did increase intrusions in free recall, a result similar to Roediger and Marsh's (2005) cued recall findings.

The conditional analyses offered additional insight into the relation between performance on the initial multiple-choice test and subsequent performance on the free and cued recall test. Given that the multiple-choice question was correctly answered, testing with additional lures yielded a benefit in free recall but had no effect on cued recall. This finding helps to explain the differences between our unconditional results and those of Whitten and Leonard (1980). In the latter experiment, multiple-choice performance was at ceiling whereas in our experiment it was not. However, when the data from our Experiment 2 were conditionalised upon a correct response on the initial multiple-choice test (i.e. 100% correct), the free recall data showed a similar pattern to that of Whitten and Leonard (1980).

Of course, by their nature, conditional analyses contain item selection artifacts. However, the most straightforward prediction that can be made regarding such an item selection artifact would be that of an overall main effect in both free and cued recall, with increasingly 'easy' items being recognised the more lures that are included on the test. Yet, the conditionalised data yielded a differential effect of the number of multiple-choice alternatives on the type of final test. This result suggests that an item selection artifact does not completely explain the obtained effects.

EXPERIMENT 3

Because the free recall data appeared more similar to Roediger and Marsh (although the effect of increasing the number of multiple-choice lures on correct responses was not significant), one goal of Experiment 3 was to re-examine this effect. We switched to a within-participants procedure (to generalise results) and also to allow us to compare free recall and cued recall within the same participant, as in Experiment 1. This permitted us to link performance on the multiple-choice test to later performance on both the free and cued recall tests.

Method

Participants and design

Fifty-six Washington University undergraduate psychology students participated for course credit. They were tested in groups of two to five people. The experiment had a 2 (final test-type: cued recall, free recall) × 4 (number of multiple-choice alternatives: 2, 3, 4, 6) within-participants design.

Materials

The same passages and questions were used as in Experiment 2. The only difference was that testing was done in paper-and-pencil format. The free recall test consisted of a sheet of paper with 48 blank spaces, whereas the cued recall test consisted of the 48 question stems from the multiple-choice test. Two versions of the cued recall test were created, each with a randomly ordered list of questions.

Procedure

The procedure was the same as in Experiment 2 except that the final test was manipulated within-participants rather than between-participants. That is, participants read 12 passages for 90 seconds apiece, completed the self-paced multiple-choice test, did a filler task for 6 minutes, then spent 5 minutes completing each of the final tests (free and then cued recall). The free recall test always preceded the cued recall test. As before, the instructions for the free recall test were to remember as many correct answers from the multiple-choice test (e.g. Naismith) as possible in no particular order. The cued recall test prompted participants with the question stems from the multiple-choice test, such as '*Who invented the game of basketball?*' Participants were instructed to only answer if they were reasonably sure that their answer was correct (i.e. not to guess). After the experiment was completed, participants were fully debriefed and dismissed.

Results

Multiple-choice test

The third row of Table 1 shows the mean proportion of multiple-choice questions correctly answered. Replicating Experiment 2, increasing the number of alternatives decreased correct answers. This observation was confirmed by a significant linear trend, $F(1,55) = 22.68$, $MSE = 0.010$, *partial* $\eta^2 = 0.29$.

Free recall test: correct answers

Table 2 shows the mean proportion of correct recall. There appears to be a slight upward trend: as the number of alternatives increased, the proportion of targets recalled increased. However, analyses revealed no significant difference among the means, $F(1,55) = 1.89$, $MSE = 0.019$, $p > 0.17$. Even when items were divided in two groups (2- and 3-alternatives vs. 4- and 6-alternatives), the difference only began to approach significance, $t(55) = 1.84$, $SEM = 0.16$, $p > 0.07$.

The trend in correct free recall was opposite to that observed in Experiment 2, so we decided to pool the data from Experiments 2 and 3 to further analyse the effect by conducting a 4 (number of multiple-choice alternatives: 2, 3, 4, 6) × 2 (experiment: 2 vs. 3) ANOVA with experiment as a between-participants factor. Neither the main effect of number of alternatives, $F(3,258) = 0.35$, $MSE = 0.007$, $p > 0.79$, nor the main effect of

experiment, $F(1,86) = 1.90$, $MSE = 0.148$, $p > 0.17$, were significant. In addition, the interaction failed to reach significance, $F(3,258) = 1.21$, $MSE = 0.024$, $p > 0.31$.

Free recall test: production of the multiple-choice lures

The last row in Table 2 shows intrusions of multiple-choice lures in free recall. Prior testing with more lures led to greater intrusions of lures in free recall. This linear trend was significant, $F(1,55) = 11.32$, $MSE = 0.004$, *partial* $\eta^2 = 0.17$. Overall, 89% of intrusions during free recall consisted of lure items that were incorrectly selected on the initial multiple-choice test.

Cued recall test: correct answers

Table 3 shows the mean proportion of correct cued recall. As the number of alternatives on the initial multiple-choice test increased, correct cued recall decreased. A significant linear trend confirmed this observation, $F(1,55) = 10.26$, $MSE = 0.010$, *partial* $\eta^2 = 0.16$.

Cued recall test: production of the multiple-choice lures

The last row of Table 3 depicts intrusions of multiple-choice lures in cued recall. As the number of prior alternatives increased, the proportion of lures produced increased. This observation was confirmed by a highly significant linear trend, $F(1,55) = 15.19$, $MSE = 0.006$, *partial* $\eta^2 = 0.22$. Overall, 86% of the intrusions in cued recall consisted of lures that were incorrectly selected on the initial multiple-choice test.

Conditional analyses

As in Experiment 2, we investigated how performance on the initial multiple-choice test related to subsequent free and cued recall tests. Of interest was whether the within-participants design would yield a similar pattern of conditionalised data to that found in Experiment 2. Again, both free and cued recall data were re-examined as a function of whether questions had been correctly or incorrectly answered on the multiple-choice test as before.

The conditional analysis of the free recall data yielded results similar to those of Experiment 2. The third line of Table 4 shows the probability of correct free recall given a correct response on the initial multiple-choice test. For items that had been answered correctly on the initial multiple-choice test, free recall performance improved as the number of multiple-choice lures increased. The pattern and level of performance are almost identical to that found in the conditional analysis of Experiment 2. In contrast, when a lure was endorsed on the initial multiple-choice test, the correct item was almost never recalled during free recall and did not differ as a function of the number of multiple-choice lures ($M = 0.01$). If an incorrect response was endorsed on the initial multiple-choice test, the lure was often produced on the free recall test ($M = 0.34$), but this proportion did not differ as a function of the number of multiple-choice lures.

The conditional analysis of the cued recall data also produced results that replicate those of Experiment 2. The fourth line of Table 4 the probability of correct cued recall given a correct response on the initial multiple-choice test. Again, the effect of number of multiple-choice lures on cued recall performance disappeared when the cued recall data were conditionalised upon correct initial multiple-choice performance. Similar to the results of Experiment 2, the probability of correct cued recall given a correct answer on the initial multiple-choice test did not depend on the number of prior multiple-choice lures ($M = 0.90$). If a lure was selected on the initial multiple-choice test, the probability of correctly answering the cued recall question was low ($M = 0.06$) and did not depend on the

number of prior multiple-choice lures. Given a multiple-choice lure was selected, there was a high proportion of later lure production in response to the cued recall question ($M = 0.61$), but again this did not depend on the number of prior multiple-choice lures.

In summary, the conditional analysis of Experiment 3 yielded two patterns of data that replicated the conditional results of Experiment 2. Given a correct response on the initial multiple-choice test, free recall performance benefited from prior testing with a greater number of lures, whereas cued recall performance was not affected by the number of lures.

Discussion

In Experiment 3, the cued recall data replicated those observed in Experiment 2: in both experiments with prose materials, testing with more multiple-choice lures decreased correct responding and increased errors on a subsequent cued recall test. The two experiments were less consistent in the free recall data. Although Experiment 2 and 3 showed different numerical trends with regard to correct free recall, neither trend was statistically significant. Furthermore, an analysis pooling the data from Experiments 2 and 3 supported the conclusion that increasing the number of multiple-choice lures did not have a substantial effect on correct free recall given these experimental parameters (prose recall etc.). Nevertheless, increasing the number of multiple-choice lures did serve to increase intrusions in free recall in both studies.

In addition, the conditional analyses for Experiment 3 produced the same general pattern of results as Experiment 2: for items that were correctly answered on the initial multiple-choice test, testing with additional lures led to an increase in correct free recall, but had no effect on correct cued recall. This replication of Experiment 2 using a within-participants design strengthens the conclusion that level of initial multiple-choice performance can help to explain the differences between the unconditionalised results of Experiment 1 and those of Experiments 2 and 3.

It should be clear that these results do *not* contradict our earlier claims about the effect of number of multiple-choice lures on cued recall of prose. Testing with additional multiple-choice lures did increase intrusions on subsequent recall tests in both Experiments 2 and 3, but this effect was due to the increased endorsement of lure items on the multiple-choice test when more alternatives were given. For example, the higher intrusion rate in the 6-alternative condition (0.14) was driven by the higher rate of errors on multiple-choice questions with six options (0.21) (see Tables 1 and 3). Once the data were conditionalised upon making an error on the multiple-choice tests, the effect of number of lures disappeared. As we discuss next, we believe that level of multiple-choice performance is the critical factor in explaining the different patterns of results.

GENERAL DISCUSSION

Three experiments examined the relationship between the number of multiple-choice lures and subsequent performance on free recall and cued recall both with lists of words and prose. Across experiments, we found all three patterns of results previously documented in the literature: increasing the number of multiple-choice lures sometimes benefited performance on a later test, but in other cases there was a cost or no effect. We turn now to integrating these results and to describing the conditions under which the different patterns occur.

We believe that the differential patterns of results obtained in our experiments (as well as previous research) can be largely accounted for by the ease or difficulty of the initial multiple-choice test. When initial multiple-choice performance is almost perfect (as in Experiment 1 and the conditional analysis of Experiments 2 and 3), participants can (by definition) recognise the target without attending too carefully to the lures. That is, when the correct answer is readily identified by a quick scan of alternatives, the decision is based largely on item-specific processing (Hunt & McDaniel, 1993) and the lures may not be related to the question stem. In contrast, when the correct answer is not obvious to subjects, they may resort to a process of careful elimination of lures (Dunning & Stern, 1994). The lures may be strongly considered or even endorsed—precisely the situation in which negative effects of taking multiple-choice tests (and increasingly negative effects of additional lures) occur (Experiments 2 and 3; Roediger & Marsh, 2005). When subjects spend time comparing the lures and the target, they are engaging in relational processing and associating the incorrect answer to the question stem.

Other factors may, of course, be uncovered that mediate the varying effects of the number of multiple-choice lures on later recall. The type of study material and the type of test are two factors we have considered, although the results do not indicate that either is definitive. Still, negative effects of multiple-choice lures seem more easily obtained on a later cued recall test than on a later free recall test with prose passages, whereas positive effects may be more readily obtained with lists of words than with prose passages.

We close with some practical considerations to be drawn from our work. If we consider conditions most like those in education, we can see that there are benefits and costs to giving multiple-choice tests. Although our experiments were not designed to permit us to show this effect, there is an overall positive effect in prose retention from taking a multiple-choice test. That is, compared to a condition in which students study prose passages and then take a delayed cued recall test, having an intervening multiple-choice test generally enhances performance on the final test (Roediger & Marsh, 2005). We did not include this control condition in the current experiments to make this comparison (the inclusion of a free recall test in all three experiments made this impractical), but the general benefit of testing should be borne in mind. However, in educational settings students will be studying prose texts and will often take multiple-choice tests in which performance is far from perfect. Under these conditions, increasing the number of multiple-choice alternatives can be reliably expected to harm students' knowledge. Increasing the number of lures will lower performance on a later cued recall (general knowledge) test. In this sense then, for educational purposes, the outcome of Roediger and Marsh (2005) represents the critical message for education rather than the contrasting positive pattern of Whitten and Leonard (1980), as interesting as that may be for other purposes. Increased numbers of similar lures on multiple-choice tests can impair participants' performance on the multiple-choice test and general knowledge as assessed on a later cued recall test. We have replicated this negative pattern in cued recall of prose in both Experiments 2 and 3 and so consider it the relevant outcome for educators to consider when examining the effects of multiple-choice testing.

ACKNOWLEDGEMENTS

We thank Aurora Steinle for her help in collecting and scoring the data. We also thank Bob Bjork for his insightful comments that prompted this project. This research was supported by a Collaborative Activity Award from the James S. McDonnell Foundation.

REFERENCES

Battig, W. F., & Montague, W. E. (1969). Category norms for verbal items in 56 categories: a replication and extension of the Connecticut norms. *Journal of Experimental Psychology, 80*, 1–46.

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*. New York: Wiley.

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

Bjork, R. A. (1999). Assessing our own competence: heuristics and illusions. In D. Gopher, & A. Koriat (Eds.), *Attention and performance XVII. Cognitive regulation of performance: Interaction of theory and application* (pp. 435–459). Cambridge, MA: MIT Press.

Brown, A. S. (1988). Experiencing misspellings and spelling performance: why wrong isn't right. *Journal of Educational Psychology, 80*, 488–494.

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

Cooper, A. J. R., & Monk, A. (1976). Learning for recall and learning for recognition. In J. Brown (Ed.), *Recall and recognition*. London: Wiley.

Dunning, D., & Stern, L. B. (1994). Distinguishing accurate from inaccurate eyewitness identifications via inquiries about decision processes. *Journal of Personality and Social Psychology, 67*, 818–835.

Geisser, S., & Greenhouse, S. W. (1958). An extension of Box's results on the use of F distribution in multivariate analysis. *Annals of Mathematical Statistics, 29*, 885–891.

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

Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior, 10*, 562–567.

Hunt, R. R., & McDaniel, M. A. (1993). The enigma of organization and distinctiveness. *Journal of Memory and Language, 32*, 421–445.

Izawa, C. (1970). Optimal potentiating effects and forgetting-prevention effects of tests in paired-associate learning. *Journal of Experimental Psychology, 83*, 340–344.

Jacoby, L. L. (1978). On interpreting the effects of repetition: solving a problem versus remembering the solution. *Journal of Verbal Learning and Verbal Behavior, 17*, 649–667.

Jacoby, L. L., Shimizu, Y., Daniels, K. A., & Rhodes, M. G. (2005). Modes of cognitive control in recognition and source memory: depth of retrieval. *Psychonomic Bulletin and Review, 12*(5), 852–857.

Jacoby, L. L., & Hollingshead, A. (1990). Reading student essays may be hazardous to your spelling: effects of reading incorrectly and correctly spelled words. *Canadian Journal of Psychology, 44*, 345–358.

Kuo, T. M., & Hirshman, E. (1996). Investigations of the testing effect. *American Journal of Psychology, 109*, 451–464.

Lewis, C. H., & Anderson, J. R. (1976). Interference with real world knowledge. *Cognitive Psychology, 8*, 311–335.

Marsh, E. J., Meade, M. L., & Roediger, H. L., III. (2003). Learning facts from fiction. *Journal of Memory and Language, 49*, 519–536.

McDaniel, M. A., Einstein, G. O., Dunay, P. K., & Cobb, R. E. (1986). Encoding difficulty and memory: toward a unifying theory. *Journal of Memory and Language, 25*, 545–656.

McDaniel, M. A., & Masson, M. E. J. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*, 371–385.

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14*, 1–43.

Postman, L., & Underwood, B. J. (1973). Critical issues in interference theory. *Memory and Cognition, 1*, 19–40.

Rees, P. J. (1986). Do medical students learn from multiple-choice examinations? *Medical Education, 20*, 123–125.

Roediger, H. L., III, & Karpicke, J. D. (2006). Test-enhanced learning: taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255.
Roediger, H. L., III, & Marsh, E. J. (2005). The positive and negative consequence of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 1155–1159.
Runquist, W. N. (1986). The effect of testing on the forgetting of related and unrelated associates. *Canadian Journal of Psychology, 40*, 65–76.
Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-prime reference guide*. Pittsburgh: Psychology Software Tools, Inc.
Schooler, J. W., Foster, R. A., & Loftus, E. F. (1988). Some deleterious consequences of the act of recollection. *Memory and Cognition, 16*, 243–251.
Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology, 30*, 641–657.
Thompson, C. P., Wenger, S. K., & Bartling, C. A. (1978). How recall facilitates subsequent recall: a reappraisal. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 210–221.
Toppino, T. C., & Brochin, H. A. (1989). Learning from tests: the case of true-false examinations. *Journal of Educational Research, 83*, 119–124.
Toppino, T. C., & Luipersbeck, S. M. (1993). Generality of the negative suggestion effect in objective tests. *Journal of Educational Psychology, 86*, 357–362.
Tulving, E. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning and Verbal Behavior, 6*, 175–184.
Tyler, S. W., Hertel, P. T., McCallum, M. C., & Ellis, H. C. (1979). Cognitive effort and memory. *Journal of Experimental Psychology: Human Learning and Memory, 5*, 607–617.
Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: an updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory & Language, 50*(3), 289–335.
Wheeler, M. A., & Roediger, H. L., III. (1992). Disparate effects of repeated testing: reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science, 3*, 240–245.
Whitten, W. B., & Bjork, R. A. (1977). Learning from tests: effects of spacing. *Journal of Verbal Learning and Verbal Behavior, 16*, 465–478.
Whitten, W. B., & Leonard, J. M. (1980). Learning from tests: facilitation of delayed recall by initial recognition alternatives. *Journal of Experimental Psychology: Human Learning and Memory, 6*(2), 127–134.

