MEMORY, 2012, 20 (7), 717–727

Psychology Press
Taylor & Francis Group

Production benefits learning: The production effect endures and improves memory for text

Jason D. Ozubko[1], Kathleen L. Hourihan[2], and Colin M. MacLeod[3]

[1]Department of Psychology, Rotman Research Institute, Toronto, Ontario, Canada
[2]Department of Psychology, Memorial University of Newfoundland, St John's, Newfoundland, Canada
[3]Department of Psychology, University of Waterloo, Waterloo, Ontario, Canada

The production effect is the superior retention of material read aloud relative to material read silently during an encoding episode. Thus far it has been explored using isolated words tested almost immediately. The goal of this study was to assess the efficacy of production as a study strategy, addressing: (a) whether the production benefit endures beyond a short session, (b) whether production can boost memory for more complex material, and (c) whether production transfers to educationally relevant tests. In Experiment 1 a 1-week retention interval was included, and a production effect was observed. In Experiment 2 a production effect was observed for both word pairs and sentence stimuli. In Experiment 3 educationally relevant essays were read and tested with a fill-in-the-blanks test: Memory was superior for questions that probed information that had been read aloud relative to information that had been read silently. We conclude that the production benefit is enduring and generalises to text and different test formats, indicating that production constitutes a worthwhile study strategy.

Keywords: Memory; Recognition; Production effect; Learning; Distinctiveness.

A laudable recent trend in cognitive psychology has been the extension of well-known laboratory phenomena to applied settings outside the laboratory, notably in the realm of education. These extensions include applications of overlearning (Rohrer, Taylor, Pashler, Wixted, & Cepeda, 2005) and of spacing (Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008) to improve learning. A signal example is the testing effect (e.g., Bjork, 1975), lately the subject of considerable research on its application to more realistic classroom testing conditions (e.g., Chan, McDermott, & Roediger, 2006; Kornell, Hays, & Bjork, 2009; McDaniel, Roediger, & McDermott, 2007; Roediger & Karpicke, 2006). Researchers have even begun to adapt laboratory techniques to explore common test formats, like multiple choice testing (Roediger & Marsh, 2005; but see Butler & Roediger, 2008). That these applied extensions have proven useful provides added value to both the basic and the applied research enterprises.

A recently (re)introduced memory phenomenon that would seem to have significant potential for educational purposes is the *production effect*, so named by MacLeod, Gopie, Hourihan, Neary, and Ozubko (2010; see also Conway & Gathercole, 1987; Gathercole & Conway, 1988; Hopkins & Edwards, 1972; Hourihan & MacLeod, 2008;

Address correspondence to: Jason D. Ozubko, Rotman Research Institute, Baycrest Centre for Geriatric Care, 3560 Bathurst Street, Toronto, Ontario M6A 2E1, Canada. E-mail: jozubko@rotman-baycrest.on.ca

This research was supported by Natural Sciences and Engineering Research Council of Canada (NSERC) discovery grant A7459. We thank Clark Amistad, Grace Hsiao, Jingjing Ji, Dushianthini Kenthirarajah, and Shelbie Sutherland for their assistance in collecting the data. We also thank Jason Chan for providing the materials used in Experiment 3.

http://www.psypress.com/memory http://dx.doi.org/10.1080/09658211.2012.699070

Lin & MacLeod, in press; MacDonald & MacLeod, 1998; MacLeod, 2011; Ozubko, Gopie, & MacLeod, 2012; Ozubko & MacLeod, 2010). The production effect is the robust benefit in memory that words read aloud (i.e., produced) show relative to words read silently during the same encoding episode. Inspired by the *relational distinctiveness* account of Conway and Gathercole (1987), MacLeod et al. have argued that the production effect results from enhanced distinctiveness: Words read aloud have associated with them additional unique information that is useful at test for discriminating produced words from other words, and hence for certifying produced words as having been studied.

MacLeod et al. (2010; see also Forrin, Ozubko, & MacLeod, in press) showed the production benefit to be robust across a range of contexts, but only when production was manipulated within participants in a mixed list design. Reading all study words aloud in a between-participants pure list design was not superior to reading all study words silently (see also Hopkins & Edwards, 1972). Distinctiveness is always a relative term (e.g., Hunt, 2006)—words produced aloud are only distinctive *relative to* words read silently—and therefore production is only beneficial when both types of encoding are present during study. Of course, learning the more-important material better than the less-important material is critical to effective learning. Thus the relative nature of the production effect could be considered a strength.

Because the production effect relies on a relative difference, it is important to address the issue of whether production could actually be a cost as opposed to a benefit. That is, given that fact that aloud items are remembered better than silent items, the act of production might either be improving memory for aloud items or interfering with memory for silent items. In their original study of production Hopkins and Edwards (1972) did find a cost of production. As with most studies of production, there was no between-participants production effect (see MacLeod et al., 2010). Thus the recognition rates of aloud and silent items in pure lists were equivalent. However, in a mixed list recognition of the aloud items was also equivalent to that of aloud and silent items in pure lists, whereas recognition of the silent items was actually lower than that of aloud and silent items in pure lists. In this original work, then, production seemingly impaired memory for silent items rather than enhancing memory for aloud items.

Subsequent examinations of the production effect have failed to replicate this apparent cost. MacLeod and colleagues (2010; Ozubko & MacLeod, 2010) have consistently found that the recognition rates of aloud items in a mixed list are greater than the recognition rates of aloud or silent items in pure lists. Furthermore, recognition rates of silent items in a mixed list are often statistically equivalent to those of aloud or silent items in pure lists. Although there is sometimes a small decrease in memory for silent items in the mixed list compared to the pure lists, this change is not large enough to account for the production effect.

Other evidence that the production effect is indeed a benefit for aloud items comes from the fact that it does not rely on poor processing of silent items. In the experiments of MacLeod et al. (2010) silent items were presented for longer during study than were aloud items: Whereas silent item trials lasted 2 seconds, aloud item trials ended as soon as the participant vocalised a response. Thus there was actually more opportunity to encode silent items than aloud items in these experiments, and yet a production effect was consistently found.[1] More compelling, MacLeod et al. also conducted two experiments where both aloud and silent words were first either generated (see Slamecka & Graf, 1978) or semantically encoded (see levels of processing; Craik & Lockhart, 1972) prior to production. If the production effect hinged on a selective impairment of memory for silent items, then generation or deep processing of silent items should have undermined this effect, as in both cases memory for silent items would be relatively substantial. Yet in both of these experiments a clear production effect was observed. Thus the bulk of the evidence to date suggests that the production effect is driven mostly by a memory benefit for aloud items as opposed to by a memory cost for silent items.

Although the production effect is very clearly a powerful laboratory phenomenon, to be useful as a pedagogical tool several aspects of the effect must be confirmed. First, the production benefit

[1] Note that subsequent studies that present both aloud and silent study items for a set amount of time also lead to a production effect, so there is no concern that the effect somehow rests on shorter presentation rates of the aloud items at study.

has never been demonstrated after a delay. That is, to date, all examinations of the production effect have had participants study a list of materials and then take an immediate test. To be useful as a study aid the production effect needs to be reliable at some reasonable delay. Consequently Experiment 1 tests the production effect at a 1-week delay.

Second, all of the literature to date has examined the production effect in the context of single-word stimuli. Obviously, a mnemonic technique that only assisted in memorising individual words would not be very useful outside the laboratory. Therefore Experiment 2 tested the generalisability of the production effect to more complex written stimuli. In Experiment 2A word pairs were tested; in Experiment 2B full sentences were tested. The goal was to generalise incrementally to a broader array of study materials.

Experiments 1 and 2 still constrain production to the laboratory list-learning setting. Experiment 3 served both as a conceptual replication and as an extension of Experiments 1 and 2. Specifically, Experiment 3 tested production in an educationally relevant context, combining the type of materials most often encountered when studying with one of the most common types of test. Participants read short, textbook-like essays. Some paragraphs of the essays were read aloud and some paragraphs were read silently. Subsequently, participants completed a fill-in-the-blanks test.

Ultimately, across these three experiments, we will show that production not only is a robust laboratory encoding manipulation, but that it generalises readily to a variety of written materials; that it is not dependent on the type of test; and that it has a lasting benefit. Based on these findings we will argue that production constitutes a useful addition to the student's toolbox for studying.

EXPERIMENT 1: DELAYED TESTING

To be useful in an academic context outside the laboratory the production effect must last beyond the short experimental sessions that typify laboratory-based memory research. With this as our goal we conducted a basic production experiment in which participants studied a list of words by reading half of them aloud and half of them silently. In the same session we tested their memory with yes/no recognition for only half of the studied words, and then dismissed them. This initial test should replicate the usual production advantage. In addition, however, we asked participants to return for "an unrelated experiment" 1 week later, at which point we tested them on the previously untested half of the studied words. Participants were uninformed as to the purpose of the second session to prevent any additional rehearsal (especially possible preferential rehearsal of the better-remembered produced words) during the retention interval. We also reasoned that the initial test would help to persuade them that the second session was unrelated. Indeed, when asked in the second session, no participant reported being suspicious of another test of the words studied in their initial session.

Method

Participants. A total of 18 undergraduates at the University of Waterloo participated in the first session of the experiment in exchange for payment. Of these, 14 returned to complete the second, surprise test session 1 week later. We report the results only for these 14 individuals, although including the additional four participants in the analysis of the first session had a negligible effect.

Stimuli. The item pool consisted of the same 120 words used in most of the experiments in MacLeod et al. (2010). From these a random 80 were selected for study, half to be read aloud (40 printed in blue) and half to be read silently (40 printed in white). All words were presented on a black background with condition (aloud vs silent) in random order. Half of the words studied aloud and half of the words studied silently were presented on each test (initial and 1-week delayed), each intermingled with 20 new words from the pool.

Apparatus. A PC-compatible computer with a 15-inch monitor was used for testing. The controlling program to display stimuli and record key-press responses was written in E-Prime (version 1.1.4.4; Schneider, Eschman, & Zuccolotto, 2002).

Procedure. The procedure was modelled after Experiment 1 in MacLeod et al. (2010). The first session began with the study phase, in which participants were instructed to read the words printed in blue aloud and the words printed in

white silently. They were told that their memory for the words would be tested, but they were not told the exact nature of the test. Each study trial began with the word in the centre of the screen for 2000 ms. Following this a blank screen was displayed for 500 ms and then the next study word appeared.

Immediately following the study phase participants were given instructions for the yes/no recognition test, conducted one word at a time. They were told to press the "m" key if they recognised a word as either read aloud or read silently and to press the "c" key if they did not recognise a word. A total of 20 words studied aloud, 20 words studied silently, and 20 new words from the pool were randomly selected and presented in random order on this test. All test words were presented in yellow on a black background to prevent colour–word associations from study affecting the results at test. Words remained visible at the centre of the screen until the participant pressed a response key. A blank screen was presented between test trials for 500 ms. Following the test, participants were dismissed but asked to return 1 week later for an "unrelated" experiment.

In the second session 1 week later, participants were again instructed that they would be tested on the words that they had read aloud or silently the week before. The delayed test procedure was identical to the initial test except that the remaining 20 aloud, 20 silent, and 20 new words that had not appeared on the initial test were used.

Results and discussion

An alpha level of .05 was used as the criterion for significance in all inferential tests. Effect size estimates were computed using partial eta-squared (η_p^2) or Cohen's d as appropriate. Table 1 displays the mean hit and false alarm rates to aloud and silent words both on the initial test and on the delayed test. A 2 (production: aloud vs silent) $\times$ 2 (session: initial vs delay) repeated-measures ANOVA was conducted on the recognition hits. A main effect of session revealed the expected reduction in hits after the delay, $F(1, 13) = 123.50$, $MSe = 0.02$, $\eta_p^2 = .91$. As expected, a main effect of production was also observed, with aloud words recognised better than silent words overall, $F(1, 13) = 17.36$, $MSe = 0.02$, $\eta_p^2 = .57$. Critically, the interaction was not significant, $F(1, 13) = 1.11$, $MSe = 0.02$, $p = .31$, $\eta_p^2 = .08$. Follow-up comparisons found that a production effect was present in the hit rates both on the immediate test, $t(13) = 3.78$, $d = 1.06$, and on the delayed test, $t(13) = 2.24$, $d = 0.58$.

Note that false alarms did significantly increase from the first to the second testing session, $t(13) = 2.94$, $d = 0.80$, as is to be expected with the addition of a retention interval. Relying solely on hit rate analyses when false alarm rates are changing can sometimes be misleading. A more appropriate measure to consider in this case might be *d'*, which takes into account both hit and false alarm rates. Repeating our analysis using *d'* instead of hit rates, we replicated our findings: There was a main effect of session, $F(1, 13) = 93.82$, $MSe = 0.37$, $\eta_p^2 = .88$, a main effect of production, $F(1, 13) = 27.72$, $MSe = 0.37$, $\eta_p^2 = .68$, and no significant interaction, $F(1, 13) = 3.32$, $MSe = 0.37$, $p = .09$, $\eta_p^2 = .20$. Indeed, even if this interaction had been significant, the production effect was still significant in *d'* both on the immediate test and on the delayed test, $t(13) = 3.72$, $d = 1.37$, and $t(13) = 2.41$, $d = 0.61$, respectively. Thus the increase in false alarm rate between the first and second sessions did not undermine the fact that the production benefit was present both on an immediate test and on a delayed test. The benefit of producing words aloud compared to reading them silently then, was the same immediately following study and after a 1-week delay.

The goal of Experiment 1 was to determine whether the production effect would persist after an extended delay, as any useful study strategy should. The production effect observed

TABLE 1
Experiments 1 and 2

	Immediate test			*1-week delayed test*		
Condition	*Aloud*	*Silent*	*New*	*Aloud*	*Silent*	*New*
Exp 1 (Words)	.90 (.03)	.71 (.06)	.23 (.04)	.59 (.05)	.48 (.06)	.36 (.05)
Exp 2A (Word pairs)	.73 (.04)	.54 (.03)	.12 (.02)	–	–	–
Exp 2B (Sentences)	.78 (.03)	.63 (.04)	.12 (.02)	–	–	–

Mean proportions of "old" responses on the recognition test for words studied aloud and words studied silently (hits), and new words (false alarms). Standard errors are shown in parentheses below the corresponding means.

in immediate testing did indeed carry through—apparently undiminished—into testing after a 1-week retention interval. This indicates that production, possibly the simplest encoding manipulation to implement for written materials, produces a memory benefit that it is not only robust but enduring.

EXPERIMENT 2: BEYOND SINGLE-WORD STIMULI

Experiment 1 demonstrated that the production effect persists over time. However, to date, the production effect has been examined only as it influences single-word study. Obviously, if the production effect is to be a useful mnemonic outside the laboratory, it must transfer to more complex materials. In Experiment 2 we examined this issue for the first time. Specifically, in Experiment 2A, participants studied random word pairs (e.g., TOWER–GLASSES, PHONE–MIRROR, etc.) and in Experiment 2B they studied short sentences (e.g., ''Take it back for a refund'' and ''Don't go there tonight''). Although there is no theoretical reason to expect production not to generalise to these more complex materials, it was conceivable that studying multiple words, especially in a coherent sentence structure, could somehow interfere with, obscure, or undermine the production effect. Thus, we decided to move incrementally from one word to two words to sentences.

Method

Participants. A total of 41 undergraduates at the University of Waterloo participated, 22 in Experiment 2A and 19 in Experiment 2B.

Stimuli. For Experiment 2A the stimulus pool consisted of 348 words drawn from the MRC Psycholinguistic Database (http://www.psy.uwa.edu.au/MRCDataBase/uwa_mrc.htm). The words were nouns from 5 to 10 letters long with Kučera and Francis (1967) frequencies between 30 and 847 and a mean frequency of 114.71 ($SD = 120.59$). From these words 160 were randomly selected and randomly paired to create 80 study pairs. Of the remaining words, 80 more were randomly selected and randomly paired to create 40 foils for test. At study, participants saw 80 pairs of words (40 pairs printed in blue to be read aloud; 40 pairs printed in white to be read silently). At test, those 80 studied pairs were randomly intermixed with the 40 foil pairs.

For Experiment 2B, 1004 sentences were obtained from Google answers (http://answers.google.com/answers/threadview/id/745189.html). These sentences were drawn from the public domain by a user known as ''eiffel-ga'' for research purposes, and are relatively random and nondescript. This stimulus pool was edited to eliminate especially long sentences. In addition sentences that could be considered objectionable or confusing were eliminated. The final stimulus pool consisted of 548 sentences. Examples of sentences included in the final pool are ''Take it back for a refund'', ''The roses were in bloom'', and ''Don't go there tonight''.

For each participant 160 sentences were randomly selected to act as study items (again, half were designated to be read aloud by appearing in blue and half were designated to be read silently by appearing in white). An additional 80 sentences were randomly selected to be used as foils at test. At study, then, participants saw 160 sentences, and at test 80 new sentences were randomly intermixed with those 160 studied sentences, yielding a final test list of 240 sentences.

Apparatus. The same apparatus as in Experiment 1 was used.

Procedure. The procedures in Experiments 2A and 2B were nearly identical: Participants were instructed to read each blue pair or blue sentence aloud, and to read each white pair or white sentence silently. In Experiment 2A pairs were shown for 4000 ms; in Experiment 2B sentences were shown for 5000 ms. A 500 ms inter-stimulus interval was used in both cases.

Immediately following the study phase participants were given instructions for the yes/no recognition test, conducted one item at a time. As in Experiment 1, participants pressed ''m'' to indicate that an item had been studied and ''c'' to indicate that an item was new. As well, all test items were presented in yellow on a black background to prevent colour–word associations from study affecting the results at test. Unlike Experiment 1, however, all items from the study list were shown during the test phase, randomly intermixed with new items. In Experiment 2A this meant that participants saw 120 pairs at test (80 of which were studied and 40 of which were new); in Experiment 2B this meant that partici-

pants saw 240 sentences at test (160 of which were studied and 80 of which were new).

Results and discussion

Table 1 displays the mean hit and false alarm rates for aloud and silent items in both Experiments 2A and 2B. A production effect was observed both in Experiment 2A and in Experiment 2B, with more hits to items studied aloud than to items studied silently, $t(21) = 5.72$, $d = 1.22$, and $t(18) = 6.26$, $d = 1.55$, respectively. Additionally, a 2 (production: aloud vs silent) $\times$ 2 (experiment: 2A vs 2B) mixed ANOVA revealed a significant effect of production, $F(1, 39) = 64.43$, $MSe = 0.38$, $\eta_p^2 = .62$, and only a borderline effect of experiment, $F(1, 39) = 2.96$, $MSe = 0.04$, $p = .09$, $\eta_p^2 = .23$, Critically, there was no interaction, $F(1, 39) = 1.15$, $MSe = 0.38$, $p = .29$, $\eta_p^2 = .03$. Although some caution is always warranted when comparing across experiments, at the least, these results do not support the notion that the production effect differed between experiments.[2] Thus it appears that the production benefit generalises readily beyond single-word stimuli, aiding memory for both word pairs and sentences alike.

EXPERIMENT 3: EDUCATIONALLY RELEVANT STUDY AND TEST MATERIALS

Experiments 1 and 2 both tested the generalisability of the production effect. Experiment 1 demonstrated that the mnemonic benefit of production persists over time; Experiment 2 demonstrated that production benefits memory not just for single words, but also for word pairs and sentences. From these two experiments it appears that production could well be a useful mnemonic study technique for students to employ during learning. Although Experiments 1 and 2 are consistent with this notion, the goal of Experiment 3 was to test it more directly. Specifically, would production generalise to paragraphs of connected discourse, and would the benefit be sustained on a fill-in-the-blanks test?

To this end we modified our basic paradigm to require participants to study a series of related paragraphs, akin to a short essay, a segment of a textbook, or an article that a student might study and be tested on. In their research Chan et al. (2006, Expt 2) examined how an initial test can affect later retention for both tested and untested material. The articles that they created were designed to be educationally relevant and similar to college-level textbook content, and they showed a testing effect using text materials and tests much more like those that students actually experience in an academic setting—specifically, a fill-in-the-blanks test. Thus the articles and associated test questions from Chan et al. (2006) seemed ideal for determining whether the production effect can be a useful strategy to improve memory when applied to realistic text material. Therefore, in Experiment 3, we presented participants with one of the Chan et al. (2006) short articles. Participants read half of the paragraphs in the article aloud and half silently. If production is a useful study strategy, then a clear advantage should be seen on the fill-in-the-blanks test for information read aloud relative to information read silently.

Method

Participants. A total of 61 undergraduates from the University of Waterloo took part in Experiment 3A for payment or partial course credit. Of these, 30 were assigned to the Hong Kong article and 32 were assigned to the Toucan article. From the same pool, 42 participants participated in Experiment 3B; 21 were assigned to each article.

Materials. The materials used were the Hong Kong and Toucan articles from Chan et al. (2006, Expt 2). These two articles were selected for their interestingness and clarity. Each was approximately 1900 words in length. The Toucan article described the biological characteristics and living habits of toucans. The Hong Kong article described the history of Hong Kong, detailing its origins, its development into a British colony, and the transition to Chinese rule.

Each article was accompanied by a corresponding fill-in-the-blanks test of 24 questions, the identical test used by Chan et al. Throughout the articles each paragraph contained the answers to approximately two questions from the relevant test. Study booklets were printed which contained a page of instructions followed by an article. Articles were printed in 13-point font, leading to approximately three paragraphs being visible on

[2] Note that false alarms did not differ between experiments, $t(39) = 0.04$, $p = .97$.

each page. Test booklets containing the relevant fill-in-the-blanks questions were printed separately with their own instruction page, and were given to participants following completion of the study phase.

In constructing the study materials for Experiment 3 half of the paragraphs were randomly assigned to be presented on a light grey background, and half on a dark grey background. The light and dark grey backgrounds would be used to identify whether those paragraphs were to be read aloud or silently during study; the actions corresponding to the background colours were counterbalanced across participants.

Procedure. Instructions emphasised that participants were to read the article for the purpose of learning its content for an upcoming test. In Experiment 3A half of the participants who read each article were instructed to read the paragraphs with light backgrounds (Set A) aloud and those with dark backgrounds (Set B) silently; the other half of the participants had this assignment reversed. During the study phase a researcher was present in the room with participants to ensure that they read appropriate information aloud or silently. Participants were not allowed to re-read or to review already read information. Experiment 3B constituted a close replication of Experiment 3A but with the addition of a 1-day retention interval to further generalise the delay result of Experiment 1. In addition, in Experiment 3B the experimenter used a computer program to record the total time that participants spent reading aloud vs reading silently.

Following the study phase participants were given the 24-question fill-in-the-blanks test. In Experiment 3A this test was given immediately; in Experiment 3B participants were asked to return to the lab the next day and the test was given then. This difference again permitted us to examine the consistency of the production benefit over retention interval. No time limit was enforced during the test, but all participants completed both the study and test sessions within a cumulative time of 30 minutes. When the tests were scored, participants were awarded 1 point for the correct answer or an extremely close approximation. For answers that were partly correct or generally correct, but lacking the specific level of detail expected, a half point was awarded. Answers that garnered full versus half points were tracked during the scoring process to ensure that tests were scored consistently across all participants.

Results and discussion

Table 2 displays the proportion of items answered correctly for paragraphs that were read aloud vs silently in Experiments 3A and 3B. Given the fact that nearly identical materials and procedures were used in the two experiments, a 2 (production: aloud vs silent) $\times 2$ (experiment: 3A vs 3B) mixed ANOVA was conducted on the proportion of items correct.

Similar to Experiment 1, overall performance was lower in Experiment 3B where a 24-hour delay was inserted, compared to Experiment 3A where there was no delay, $F(1, 99) = 35.55$, $MSe = 0.05$, $\eta_p^2 = .26$. More important, however, a production effect was observed, with more test items being filled in correctly when the relevant information came from paragraphs studied aloud than from paragraphs studied silently, $F(1, 99) = 19.33$, $MSe = 0.02$, $\eta_p^2 = .16$. No interaction was observed, $F(1, 99) = 2.15$, $MSe = 0.02$, $p = .15$, $\eta_p^2 = .02$. Despite this non-significant interaction, the production effect looked smaller in Experiment 3B than in Experiment 3A. However, separate analyses of Experiments 3A and 3B each showed a significant production effect: $t(60) = 4.17$, $d = 0.54$, for Experiment 3A, and $t(41) = 2.21$, $d = 0.34$, for Experiment 3B. Thus we can be sure that the production effect was present in both Experiments 3A and 3B.

The results of Experiment 3 look very promising. There is, however, the remaining question of whether the benefit of production occurs as a

TABLE 2
Experiments 3A and 3B

	p(correct)		*Total reading time*	
Condition	*Aloud*	*Silent*	*Aloud*	*Silent*
Exp 3A (Immediate test)	.47 (.03)	.35 (.02)	–	–
Exp 3B (1-day delayed test)	.26 (.02)	.20 (.02)	392 (13.22)	295 (15.41)

Mean proportion of items answered correctly on the fill-in-the-blanks test based on whether the information was read aloud or silently at study. In Experiment 3B mean total time to read paragraphs aloud vs silently is provided in seconds. Standard errors are shown in parentheses below their respective means.

result of increased exposure. That is, if reading aloud takes longer than reading silently, then the production effect could be driven by the fact that participants are exposed to material longer when it is read aloud (cf. the total-time hypothesis; Cooper & Pantle, 1967). If this were the case the production benefit would be less compelling. MacLeod et al. (2010) dealt with this exposure hypothesis by having study trials end after 2000 ms, or when a word was spoken aloud. Hence words read aloud were exposed to participants for less time than words read silently. Despite this bias against a production effect, a clear production benefit was observed. In our Experiment 3B, although exposure times were not controlled, total reading times were recorded for paragraphs read aloud and for paragraphs read silently (see Table 2).

Examining total reading times in Experiment 3B, it is clear that participants did spend longer reading aloud than reading silently, $t(41) = 4.96$, $d = 0.79$. Despite this being consistent with a total-time hypothesis, upon closer inspection we see that reading time was actually unrelated to the production effect. That is, there actually was no significant relation between the size of the production effect and reading time: Proportion correct aloud did not correlate with aloud paragraph reading times, $r(40) = -.02$, $p = .91$, nor did proportion correct silent correlate with silent paragraph reading times, $r(40) = -.18$, $p = .28$. More pointedly, the size of the production effect (i.e., aloud proportion correct – silent proportion correct) did not correlate with the difference in reading times between aloud and silent paragraphs, $r(40) = -.09$, $p = .60$. If the production effect were driven solely by reading time, this last correlation should have been significant and positive; if anything there is a negative trend, the opposite of what a total-time hypothesis would require. Thus reading time appears to be unrelated to either the presence or the size of the production effect.

Overall, then, the results of Experiment 3 are consistent with those of MacLeod et al. (2010) in showing that the production effect occurs independent of reading time. Although not surprisingly participants do read aloud more slowly than they read silently, this was not the driving force behind the production effect observed in Experiment 3. Experiment 3 supports the findings of Experiment 2 in showing that production generalises beyond single-word stimuli to more complex material. Experiment 3 also supports Experiment 1 in showing that the effect of production can survive a delay. Most importantly, Experiment 3 demonstrates that production generalises to educationally relevant materials and tests: Production improved memory for short essay and textbook-like material and this benefit was observed on an educationally relevant test.

GENERAL DISCUSSION

The goal of this article was to test the generalisability of the production effect, specifically with the goal of assessing whether production could be a viable study strategy for students. Experiment 1 demonstrated that production can have a lasting effect, and Experiment 2 demonstrated that production does generalise beyond single-word stimuli—to word pairs and sentences. Both of these findings were confirmed in Experiment 3. Critically, Experiment 3 also demonstrated that production can aid in the memory of complex, paragraph material on a type of test widely used in educational settings. Furthermore this effect could not be explained by increased reading time for material read aloud vs material read silently. Thus production does indeed appear to be a viable encoding strategy for educational material.

One aspect of production that makes it an appealing study strategy is that it requires minimal effort. Although we certainly would not argue that production should replace other effective studying strategies, such as note-making and self-quizzes, it does provide another technique that students can employ when trying to learn complex material, and especially when the emphasis of some material over other material is important. Of course, this reminds us that the benefit of production is a relative benefit. Namely, in laboratory studies no memory benefit is seen if all of the studied words are spoken aloud (Hopkins & Edwards, 1972; MacLeod et al., 2010; Ozubko & MacLeod, 2010). Hence production relies on the fact that some information is read aloud and other information is not. Production is an effective technique to boost memory but it is still up to the student (or the teacher) to decide which information should be read aloud and which should be read silently.

Although the within-participants nature of production could be viewed as a limitation, from this perspective production would appear to bear some resemblance to highlighting or underlining. Highlighting or underlining text is a study

strategy that is almost standard practice for undergraduates. In fact, studies such as that of Fowler and Barker (1974) have shown that highlighting does help memory for the highlighted portions, and the same benefit has been reported for underlining (e.g., Cashen & Leicht, 1970). Like highlighting or underlining, production offers a quick and effective method for boosting memory for information judged to be important. Indeed, although the need to select material for production could be viewed as a limitation, it could also be an indirect benefit of production. Namely students seeking to use production to aid in memorisation would first need to evaluate and select important information to focus on. This act alone encourages review of material, and could lead to a deeper understanding and comprehension of important material.

Researchers have recently advocated testing as an effective method for improving retention of academic information (e.g., Roediger & Karpicke, 2006). Yet McDaniel, Howard, and Einstein (2009) have pointed out that most studies demonstrating the benefits of testing do so by comparison to a re-read/re-study control group. Simply re-reading an article is likely substantially less involved than the processes that students actually use when studying for an exam. McDaniel et al. examined the effectiveness of the "3R" study strategy, which involves Reading an article, Reciting (aloud) all that can be remembered, and then Reviewing the article a second time. They compared the 3R strategy to re-reading and note-taking strategies, and found superior free recall performance on immediate and delayed tests for the 3R group; the advantage extended to multiple-choice testing with more complex study material in Experiment 2.

In addition to its simplicity McDaniel et al. (2009) advocated the 3R strategy for its combination of testing (i.e., during the Recitation phase one essentially carries out a self-test) and feedback (i.e., during the Review stage one can easily see how much was remembered and forgotten during Recitation), both of which are known to improve retention. We suggest that production would fit well into the Review stage of the 3R strategy. Once a subset of the material is selected as requiring further learning during the Review phase, this material can then be read aloud (whereas the well-known material already recalled in the Recitation phase can be read silently). We think it likely that the material identified as requiring further study would benefit from the distinctiveness of having been read aloud, and that the previously well-learned material would remain well learned, resulting in an overall improvement in retention.

Finally, although production has primarily been examined in the context of recognition memory, several studies have found a recall benefit for production (Conway & Gathercole, 1987; Gathercole & Conway, 1988; Lin & MacLeod, in press). Hence the finding in Experiment 3 that production can help on short-answer, fill-in-the-blank tests is not without previous support. Given that production can help on recognition (whether Yes/No or forced-choice), recall, and fill-in-the-blank tests, we would expect it to help on other retention tests as well.

As an aside, Experiment 3 also serves to rule out potential methodological explanations for the production effect. For example, in many of the production experiments reported by MacLeod and colleagues (e.g., Experiments 1 and 2 here; Hourihan & MacLeod, 2008; MacLeod et al., 2010; Ozubko et al., 2012), two thirds of the items on the recognition test have been studied and one third have been new. Although this design ensures an equal number of aloud, silent, and new items at test, it produces an unequal number of old and new items at test. We now know, however, from Experiment 3 here and other recall experiments (Conway & Gathercole, 1987; Gathercole & Conway, 1988; Lin & MacLeod, in press) and even from other experiments using equal numbers of old and new items on recognition tests (Forrin, Ozubko, & MacLeod, Expt 2, in press) that this methodological feature plays no role in the effect.

Similarly some production studies have not counterbalanced encoding instruction and study colour (e.g., MacLeod et al., 2010; Ozubko & MacLeod, 2010). Given that blue words (the colour typically paired with instructions to read aloud) are perceptually less discriminable on a black background than are white words (the colour typically paired with instructions to read silently) because of the difference in relative luminance of the colours, it is possible that the additional attention required to perceive blue words contributed to the production effect produced when those words are read aloud. However, in Experiment 3 not only were the typical blue/white colour associations for aloud/silent not used, but also the colour coding during study was counterbalanced with encoding instruction, and yet a production effect still emerged. Moreover,

using the more standard blue/white encoding cue distinction with individual words, Lin and MacLeod (in press) counterbalanced colour–response mapping and observed no effect of mapping on recall or recognition. The extension of production to a fill-in-the-blanks test in Experiment 3 helps to further demonstrate that production is indeed a mnemonic benefit and not somehow an artefact of certain procedural features of existing studies.

Why did production help in Experiment 3? One possibility is that production provided a relatively shallow boost to memory, by helping with the rote memorisation of the material at study or simply by selectively focusing attention on some material over other material. Another possibility is that production caused participants to more deeply process the produced material, leading to better understanding of its meaning. Although either alternative still renders production a useful study strategy, it is this second alternative that would make it an especially valuable study strategy. For now, all that can be said is that production can help learning of educationally relevant materials, possibly by increasing the distinctiveness of the materials read aloud. Especially, although rote memorisation would be helpful in the basic recognition tasks of Experiments 1 and 2, it should be less helpful in the short-answer fill-in-the-blank questions of Experiment 3 where the form of the phrasing of the material is not constant between study and test. That a clear production effect was observed in Experiment 3 suggests that production may indeed offer some advantages to comprehension, above pure rote memorisation. For now, whether this result holds for more complex assessments of knowledge (such as essay responses or critical thinking extensions) remains to be seen, but we certainly see it as worthy of exploration.

In closing, we have demonstrated that the simple act of reading some text material aloud results in better memory for that material, relative to material read silently. It is difficult to imagine a simpler technique for improving retention during studying. We have also shown that this benefit can endure over time and that it is not limited to lists of unrelated words but can readily be applied to more complex material as well. And we have shown that the production benefit appears even on educationally relevant tests. Production is simple for participants to implement in the laboratory and it is simple for students to implement in studying for exams. We view its role in learning as similar to that of highlighting, and recommend it as a quick and effective technique for emphasising and better encoding important information.

Manuscript received 17 February 2012
Manuscript accepted 28 May 2012
First published online 26 July 2012

REFERENCES

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

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

Cashen, V. M., & Leicht, R. L. (1970). Role of the isolation effect in a formal educational setting. *Journal of Educational Psychology*, *61*, 484–486.

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science*, *19*, 1095–1102.

Chan, J. C. K., McDermott, K. B., & Roediger, H. L. III. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General*, *135*, 553–571.

Conway, M. A., & Gathercole, S. E. (1987). Modality and long-term memory. *Journal of Memory and Language*, *26*, 341–361.

Cooper, E. H., & Pantle, A. J. (1967). The total-time hypothesis in verbal learning. *Psychological Bulletin*, *68*, 221–234.

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning & Verbal Behavior*, *6*, 671–684.

Forrin, N. D., Ozubko, J. D., & MacLeod, C. M. (in press). Widening the boundaries of the production effect. *Memory & Cognition*.

Fowler, R. L., & Barker, A. S. (1974). Effectiveness of highlighting for retention of text material. *Journal of Applied Psychology*, *59*, 358–364.

Gathercole, S. E., & Conway, M. A. (1988). Exploring long-term modality effects: Vocalisation leads to best retention. *Memory & Cognition*, *16*, 110–119.

Hopkins, R. H., & Edwards, R. E. (1972). Pronunciation effects in recognition memory. *Journal of Verbal Learning & Verbal Behavior*, *11*, 534–537.

Hourihan, K. L., & MacLeod, C. M. (2008). Directed forgetting meets the production effect: Distinctive processing is resistant to intentional forgetting. *Canadian Journal of Experimental Psychology*, *62*, 242–246.

Hunt, R. R. (2006). The concept of distinctiveness in memory research. In R. R. Hunt & J. B. Worthen

(Eds.), *Distinctiveness and memory* (pp. 3–25). New York: Oxford University Press.

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 989–998.

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

Lin, O. Y. H., & MacLeod, C. M. (in press). Aging and the production effect: A test of the distinctiveness account. *Canadian Journal of Experimental Psychology.*

MacDonald, P. A., & MacLeod, C. M. (1998). The influence of attention at encoding on direct and indirect remembering. *ActaPsychologica*, *98*, 291–310.

MacLeod, C. M. (2011). I said, you said: The production effect gets personal. *Psychonomic Bulletin & Review*, *18*, 1197–1202.

MacLeod, C. M., Gopie, N., Hourihan, K. L., Neary, K. R., & Ozubko, J. D. (2010). The production effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Learning, Memory, Cognition*, *36*, 671–685.

McDaniel, M. A., Howard, D. C., & Einstein, G. O. (2009). The read-recite-review study strategy: Effective and portable. *Psychological Science*, *20*, 516–522.

McDaniel, M. A., Roediger, H. L. III, & McDermott, K. B. (2007). Generalising test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review*, *14*, 200–206.

Ozubko, J. D., Gopie, N., & MacLeod, C. M. (2012). Production benefits both recollection and familiarity. *Memory & Cognition. 40*, 326–338.

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

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

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

Rohrer, D., Taylor, T., Pashler, H., Wixted, J. T., & Cepeda, N. J. (2005). The effect of overlearning on long-term retention. *Applied Cognitive Psychology*, *19*, 361–374.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime reference guide*. Pittsburgh, PA: Psychology Software Tools.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, 592–604.