Journal of Experimental Psychology:
Learning, Memory, and Cognition
2010, Vol. 36, No. 3, 671–685

0278-7393/10/$12.00 DOI: 10.1037/a0018785

The Production Effect: Delineation of a Phenomenon

Colin M. MacLeod, Nigel Gopie, Kathleen L. Hourihan, Karen R. Neary, and Jason D. Ozubko
University of Waterloo

In 8 recognition experiments, we investigated the *production effect*—the fact that producing a word aloud during study, relative to simply reading a word silently, improves explicit memory. Experiments 1, 2, and 3 showed the effect to be restricted to within-subject, mixed-list designs in which some individual words are spoken aloud at study. Because the effect was not evident when the same repeated manual or vocal overt response was made to some words (Experiment 4), producing a subset of studied words appears to provide additional unique and discriminative information for those words—they become distinctive. This interpretation is supported by observing a production effect in Experiment 5, in which some words were mouthed (i.e., articulated without speaking); in Experiment 6, in which the materials were pronounceable nonwords; and even in Experiment 7, in which the already robust generation effect was incremented by production. Experiment 8 incorporated a semantic judgment and showed that the production effect was not due to "lazy reading" of the words studied silently. The distinctiveness that accrues to the records of produced items at the time of study is useful at the time of test for discriminating these produced items from other items. The production effect represents a simple but quite powerful mechanism for improving memory for selected information.

Keywords: memory, encoding, production, generation, distinctiveness

In 1978 Slamecka and Graf reported a thorough set of experiments demonstrating that producing a word from a cue (e.g., producing *fast* from the cue *rapid-f*) leads to considerably better memory for that word than does simply reading the word. The phenomenon that they called the "generation effect" has subsequently become one of the most widely used manipulations in memory research, leading to their article becoming a citation classic (see Slamecka, 1992), now having been cited more than 500 times and having spawned hundreds of directly related articles (for a meta-analysis, see Bertsch, Pesta, Wiscott, & McDaniel, 2007).

The generation effect is a member of a very select club: manipulations that have consistent, reliable effects on retention of a stimulus that was presented only once. Other such manipulations include imagery (Paivio, 1971) and elaboration (Craik & Lockhart, 1972). All three of these involve some recoding of the stimulus, and all produce quite substantial memory benefits, often an improvement of 10% or even more relative to a "standard" baseline of simply reading the word. Enriching the encoding of the stimulus definitely enhances memory for it, in line with what we know generally about mnemonic techniques (see Higbee, 1988).

Improving Memory by Saying a Word Aloud

In the same period, another factor was first reported to benefit memory but, unlike the others, it failed to attract subsequent research attention. In the study in question, Hopkins and Edwards (1972) tested a key assumption of frequency theory (Ekstrand, Wallace, & Underwood, 1966)—that recognition should be better for pronounced than for unpronounced words because pronunciation would increment the item's frequency (see Hopkins, Boylan, & Lincoln, 1972, for evidence that pronunciation does increase judged frequency). To test this prediction, Hopkins and Edwards used two recognition tests: two-alternative forced choice (Experiment 1) and yes/no (Experiment 2). In both experiments, three groups of subjects studied 100-word lists. There were two pure-list groups—one read all 100 words aloud, and one read all 100 words silently—and one mixed-list group, which read 50 of the words aloud and 50 silently.

In comparing the two pure-list conditions, Hopkins and Edwards (1972) found no between-subjects benefit to reading words aloud. But in the mixed-list condition, words read aloud were better recognized than those read silently. This pattern held for both types of recognition test, with the within-subject benefit of reading aloud being approximately 10%. Hopkins and Edwards suggested that the effect was at encoding.

Colin M. MacLeod, Nigel Gopie, Kathleen L. Hourihan, Karen R. Neary, and Jason D. Ozubko, University of Waterloo, Waterloo, Canada.

Nigel Gopie is now at the Rotman Research Institute, Toronto, Canada. Kathleen L. Hourihan is now at the University of Illinois at Urbana-Champaign.

This research was supported by Natural Sciences and Engineering Research Council of Canada Discovery Grant A7459. This article is dedicated to the memory of Norman J. Slamecka, longtime colleague of Colin M. MacLeod at the University of Toronto. In tribute, the title of this article is a (slight) variation on that of the original article on the generation effect (Slamecka & Graf, 1978); we thank Peter Graf for his permission. We thank Keehan Bailey, Emily Bryntwick, Samantha Chaves, Domenica De Pasquale, Nicole Deckert, Grace Hsiao, Jingjing Ji, Marita Partanen, and Molly Pottruff for assistance with collecting the data and Amy Criss, Neil Mulligan, and Hubert Zimmer for their advice concerning the original version of this article.

Correspondence concerning this article should be addressed to Colin M. MacLeod, Department of Psychology, University of Waterloo, 200 University Avenue West, Waterloo, Ontario N2L 3G1, Canada. E-mail: cmacleod@uwaterloo.ca

No further work appeared until 15 years later, when Conway and Gathercole (1987) and Gathercole and Conway (1988) reported two series of experiments aimed at investigating modality effects in long-term retention. Conway and Gathercole had their subjects study 30 words in a mixed-list procedure: 10 read silently, 10 mouthed silently, and 10 read aloud. In Experiments 1 and 2, there followed a "batch" recognition test with all 30 studied words plus 30 lures in view at once. In Experiment 3, the test was switched to free recall. They observed a 15%–25% advantage for read aloud over read silently in each experiment. The "mouthed" condition was consistently intermediate to the aloud and silent conditions and did not differ reliably from either, perhaps a result of limited power given the few items per condition.

In line with Hopkins and Edwards (1972), Conway and Gathercole (1987) argued that the advantage of reading words aloud happened at encoding, adding the suggestion that it was due to enhanced distinctiveness. Because distinctiveness can operate only relatively (see Hunt, 2006; Murdock, 1960)—a word must be distinctive with respect to others that are not—this explains why the benefit of speaking a word aloud is seen only in mixed-list, within-subject designs. Gathercole and Conway (1988) reported five more experiments extending those of Conway and Gathercole. They showed a consistent advantage of 14%–20% for reading aloud over reading silently and also showed reading aloud to be superior to hearing (Experiments 1 and 5), mouthing (Experiments 2 and 5), and writing (Experiments 3–5).

Eleven years on, MacDonald and MacLeod (1998) reported two experiments in which subjects read 40 words aloud and 40 silently. Their focus was on how this additional attention to some studied words influenced an explicit memory test versus an implicit memory test. For explicit memory, MacDonald and MacLeod observed in both experiments a clear benefit on a recognition test for the words that had been read aloud at study. It is interesting to note that they found no benefit for words read aloud on their implicit test, speeded reading (also known as "naming").

One other study rounds out the literature on the benefit of reading words aloud.[1] To investigate how people manage to reject potential false memories in the Deese–Roediger–McDermott false-memory paradigm (Roediger & McDermott, 1995), Dodson and Schacter (2001) had subjects study semantically related lists by saying the words aloud or by hearing them. In Experiment 1, this manipulation was done between subjects and did not differentially affect memory for the studied words. In Experiment 2, however, the manipulation was done within subject, and now there was a reliable advantage for words studied by reading them aloud. Thus, their results conceptually replicated those of Hopkins and Edwards (1972). In line with Conway and Gathercole (1987) and Gathercole and Conway (1988), Dodson and Schacter proposed that subjects used a "distinctiveness heuristic," using recollection of having said a word aloud as evidence that it was studied (see also Schacter & Wiseman, 2006, for a more extended treatment).

These five studies constitute the literature investigating the advantage of reading a word aloud at study. All were conducted for purposes other than to directly study the value of saying words aloud. None of the later studies cited the Hopkins and Edwards (1972) study, which has been cited fewer than 10 times, remarkable given the huge literature on the generation effect. Yet the advantage of reading aloud typically is 10%–20% or more, by any standards a fairly large effect of a processing operation (and apparently comparable in size to the generation effect; cf. Bertsch et al., 2007). Moreover, production is so very simple, straightforward, and effortless that it should certainly be better recognized among manipulations that benefit memory. Our research took this as the point of departure. Our first goal was, thus, to raise consciousness about this phenomenon. We recognized that to do so would require naming the phenomenon, and so we have chosen to call it, by analogy to the generation effect, the *production effect*.

Exploring the Production Effect

As Slamecka and Graf (1978) did in reporting a series of experiments delineating the generation effect, so we do for the production effect. We first examined in Experiments 1, 2, and 3 whether the production effect is indeed limited to within-subject, mixed-list designs. If, because distinctiveness is relative, producing only some words during study makes them distinctive, then the production effect should be limited to within-subject designs. Moreover, the effect should arise only with unique responses during encoding: Distinctiveness is inherently item based. In Experiment 4, we tested this idea by having subjects respond to some of the words with a consistent response. If the production effect relies simply on having made a response, then even consistent responses should bring about the advantage. But if distinctiveness is crucial, then without unique responses to each item there should be no production effect. In Experiment 5, we reexamined the mouthing manipulation of Conway and Gathercole (1987). Mouthing should also make items distinctive, so we expected a production effect.

In Experiment 6, we used pronounceable nonwords to examine whether existing representations are important to the production effect. We anticipated a production effect for nonwords because saying them aloud should enhance their distinctiveness. In Experiment 7, subjects generated all the items but did so aloud for only half of them. If production makes words distinctive, it should do so whether they are generated or read. This would generalize the production effect beyond encoding via reading and would show that the production enhancement is not limited to weak encodings. Finally, in Experiment 8, we evaluated the alternative notion that the produced items might not benefit from production but that the items not produced might suffer a cost. Do subjects simply pay less attention to the items not produced, becoming "lazy readers" (cf. Begg & Snider, 1987)? To test this view, we had subjects make a semantic ("deep") judgment about every item before reading it aloud or silently. The lazy reading hypothesis would predict that the production effect should disappear; the distinctiveness account would still expect there to be a production advantage.

Explaining the Production Effect

We champion the idea of distinctiveness as providing a basis for explaining the production effect. This idea has long been an important one in understanding memory (see Hunt, 2006, for an

[1] There is also an advantage for producing items aloud in the short-term memory literature, first laid out by Crowder (1970). However, as Conway and Gathercole (1987) pointed out, this advantage appears to be restricted to the recency portion of short lists and so probably is not related to the benefit of production that appears in recognition of long lists.

overview). Burnham (1888) even credits it to Aristotle. In Murdock's (1960) theory, he highlighted that "the concept of distinctiveness refers to the relationship between a given stimulus and one or more comparison stimuli, and if there are no comparison stimuli the concept of distinctiveness is simply not applicable" (p. 21). As an explanatory construct, distinctiveness received considerable impetus from levels of processing research (Jacoby & Craik, 1979; Lockhart, Craik, & Jacoby, 1976). More recently, Hunt and McDaniel (1993, p. 423) defined distinctiveness as "the processing of differences among the items of an episode," emphasizing that this does not have to be intentional but is routine in the processing of items. As the chapters in the recent Hunt and Worthen (2006) book ably illustrate, distinctiveness covers considerable ground as an explanatory construct.

Put simply, we see production as bringing about unique processing of an item at the time of study, conferring distinctiveness upon the item. A record of that processing is created in memory, as laid out in the proceduralist account of remembering (Kolers, 1973; Kolers & Roediger, 1984; for a review, see Roediger, Gallo, & Geraci, 2002). This processing record can then be "replayed" at the time of an explicit test of memory to help determine whether a test item was indeed studied, possibly a form of reconstruction (see, e.g., Zimmer, Mecklinger, & Lindenberger, 2006, with respect to the enactment effect). Only an item that was produced at study will be able to benefit from such a replay. This corresponds to what Dodson and Schacter (2001) called a distinctiveness heuristic (see also the earlier idea of Conway & Gathercole, 1987). This same heuristic does not serve to discriminate an unproduced studied item from an unproduced (novel) distractor—both are "silent."

We do not subscribe to a tagging account wherein "aloud" becomes a kind of label attached to an item specifying its study format. Rather, at the time of retrieval, the processing record is recovered and can contribute to the decision about the item's study status, much as has been argued in the source memory literature (see Foley, Johnson, & Raye, 1983). In our situation, judgment of prior study replaces attribution of source. The availability of a record that the item was said aloud is definitive and hence is used preferentially in the form of a distinctiveness heuristic. This is also the argument made by Dodson and Schacter (2001) in the context of their false-memory research, although they place more emphasis on the decision, whereas we place more on the match between processing at study and reprocessing at test.

A question that immediately arises is why this heuristic would be operative in a within-subject design but not in a between-subjects design. Our hypothesis is that without item differentiation during study, subjects are unlikely to think of applying the heuristic at the time of test. Distinctive information needs to be experienced at encoding for it to be seen as valuable at retrieval. Dodson and Schacter (2001) also considered the encoding-retrieval relation to be crucial for the deployment of the distinctiveness heuristic.

Without doubt, theoretical refinement will go hand in hand with increasingly better understanding of the phenomenon. Our hope is to initiate that refinement, which is why, like Slamecka and Graf (1978), we have named the phenomenon without reference to any particular theory. We turn now to the experiments that demonstrate and explore the production effect.

Experiment 1

To begin, our goal was to reproduce the production effect in recognition, the explicit test previously used (Conway & Gathercole, 1987; Dodson & Schacter, 2001; Hopkins & Edwards, 1972; MacDonald & MacLeod, 1998). Toward that end, we first used a within-subject design, the design under which all four previous studies had observed the effect. Experiment 1 was conducted twice, providing replication and generalization.

Method

Subjects. Students from Introductory Psychology at the University of Waterloo received bonus credit in the course for taking part. There were 23 students in Experiment 1A and 21 students in Experiment 1B. An entirely nonoverlapping sample of subjects was used in every experiment reported in this article.

Stimuli. The item pool consisted of the same 120 words that appeared in the appendix of MacDonald and MacLeod (1998). They were nouns from five to 10 letters long with frequencies of greater than 30 per million (Thorndike & Lorge, 1944). All stimuli were presented in an 80-character lowercase font against a black background.

From the 120 words, a random 80 were selected for study, with 40 presented in blue and 40 in white, in random order. Twenty words that had been presented in each color, plus 20 unused words from the pool, were assigned to the recognition test, where they were presented in a new random order. The remaining 20 words that had been presented in each color, and the remaining 20 unused words from the pool, were assigned to an implicit test administered between study and the explicit recognition test. This implicit test and the results observed on it are described in the Appendix. In most of the experiments in this article (not including Experiments 3 and 8), we used this same basic procedure of testing half the items with an implicit test and half with an explicit test.

Apparatus. An IBM-compatible microcomputer with a 15-in. (38.10-cm) color monitor was used for testing. The controlling program was written in QuickBASIC 4.5 and used the routines of Graves and Bradley (1991) to achieve millisecond accuracy timing.

Procedure. The procedure was modeled closely after that of MacDonald and MacLeod (1998), except that learning was intentional, not incidental. In the study phase, subjects were instructed to read words presented in blue aloud and words presented in white silently. They were informed that a memory test would follow but were not made aware of its nature. Study trial presentation began with a 500-ms blank preceding each word's appearance at the center of the screen. If the word was in blue, the subject read it aloud into a microphone, which initiated the next 500-ms blank; if the word was in white, it stayed on the screen for 2,000 ms, followed by the 500-ms blank. Thus, available study time was longer for words read silently than for words read aloud, a bias in study time against the anticipated production benefit.

The only difference between the two versions of Experiment 1 was a change in study instructions. In Experiment 1A, subjects were asked simply to say the blue words aloud, but no mention was made of speed of response. In Experiment 1B, we inserted the word "quickly" or the phrase "as quickly as you can" seven times in the short sequence of instructions and also added the sentence "You are being timed."

The second phase, the indirect, implicit test of memory, immediately followed study and is described in the Appendix. The third phase was the direct, explicit recognition test—a yes/no recognition test. (This test order was constant for all subjects in each experiment in which both tests were administered, to minimize contamination between the tests.) Here the remaining 20 blue words, 20 white words, and 20 unstudied new words were shown one at a time, and the subject responded *yes* (the *slash* key) or *no* (the *Z* key). Subjects were told that none of the items from the speeded reading test would be re-presented. All test items were presented in yellow font to avoid overlap with study color. On the recognition test, there was no emphasis on speed; indeed, response latency was not recorded. There was a 500-ms blank before each word, and the word offset with the subject's key response. The next trial began immediately.

Results

Table 1 presents the yes/no recognition data expressed as proportions of *yes* responses, accompanied by their respective standard errors. For the words that had been studied either aloud or silently, these are hit rates; for the unstudied words, this is the false-alarm rate. The same descriptive statistic format is used for recognition in all subsequent experiments except Experiment 3, in which the recognition test was switched to two-alternative forced choice.

For simplicity, all the analyses in this article focus on comparison of hit rates. However, every analysis was also carried out with d' as the dependent measure; in every case, the conclusion reached from examining d' was identical to that reached from examining hit rates.

Experiment 1A. A one-way analysis of variance (ANOVA) comparing aloud, silent, and unstudied showed a significant overall effect, $F(2, 44) = 156.98$, $MSE = 0.012$, $p < .001$. The first planned comparison showed good memory for the studied words, with the contrast between studied (aloud plus silent) and unstudied (new) significant, $F(1, 22) = 297.86$, $MSE = 0.073$, $p < .001$. Most important, the second planned comparison revealed the presence of the production effect, with hits for words read aloud exceeding by .124 those for words read silently during study, $F(1, 22) = 14.68$, $MSE = 0.024$, $p < .001$.

Experiment 1B. A one-way ANOVA showed a significant overall effect, $F(2, 40) = 105.81$, $MSE = 0.014$, $p < .001$. The first planned comparison showed good memory for studied (aloud plus silent) words, $F(1, 20) = 219.03$, $MSE = 0.071$, $p < .001$. Most important, the second planned comparison showed a reliable production effect: Hits for words read aloud exceeded by .204 those for words read silently during study, $F(1, 20) = 26.11$, $MSE = 0.034$, $p < .001$.

Table 1
Proportion of Yes Responses in Recognition as a Function of Study–Test Condition for Experiments 1A and 1B (Within Subject)

	Read aloud		Read silently		Not studied	
Condition	*M*	*SE*	*M*	*SE*	*M*	*SE*
1A: Read at own pace	.739	.034	.615	.030	.191	.021
1B: Read quickly	.783	.025	.579	.036	.250	.022
1A and 1B combined	.760	.022	.598	.023	.219	.016

Because the designs of Experiments 1A and 1B are identical, apart from the speed emphasis at encoding, we conducted a 2×2 ANOVA dropping the unstudied condition and treating experiment as a between-subjects variable. As expected, neither the main effect of experiment ($F < 1$) nor the Condition $\times$ Experiment interaction, $F(1, 42) = 2.50$, $MSE = 0.014$, $p > .10$, was reliable. Only the main effect of condition was significant, $F(1, 42) = 41.37$, $p < .001$, confirming the same production effect pattern in the two experiments.

To preface, we also compared the patterns in Experiments 5, 6, 7, and 8 with that of Experiment 1. For this purpose, because Experiments 1A and 1B had virtually identical outcomes, we combined them into a single Experiment 1. The combined Experiment 1A and 1B data are also shown in Table 1.

Discussion

This first experiment replicated and extended the production effect observed in within-subject comparisons. As in previous studies (Conway & Gathercole, 1987; Dodson & Schacter, 2001; Hopkins & Edwards, 1972; MacDonald & MacLeod, 1998), there was a robust recognition advantage—on the order of a 12%–20% increase in hits—for words read aloud at study over those read silently. This is true whether learning is intentional (here the memory test was announced prior to study) or incidental (the test was not announced in the MacDonald & MacLeod, 1998, experiments). Only on an explicit test, where studied–unstudied discrimination is necessary, should there be a benefit from the distinctiveness information afforded words studied aloud, and that is precisely what we have observed: The Appendix reports extensive data, including data from Experiment 1, showing that the production effect is consistently absent under implicit testing conditions.

Experiment 2

Hopkins and Edwards (1972) demonstrated their production effect in a within-subject, mixed-list design, but so far only they and Dodson and Schacter (2001) have demonstrated that the effect did not occur in a between-subjects, pure-list design. And there is the fact that in their Experiment 5, Gathercole and Conway (1988) did report a between-subjects advantage of aloud over silent reading. So, it is important to confirm whether there is a design difference because it may well be one of the defining characteristics of the production effect.

Method

Subjects. Thirty students from the same source received bonus course credit for taking part in the experiment. Fifteen were randomly assigned to each study condition.

Procedure. The stimuli and apparatus were identical to those in Experiment 1. Except for the change to a between-subjects design, the procedure was also very similar. Subjects again studied 80 items, the only difference being that for both groups all items were studied in white (i.e., there was no color cue to differentiate

two subsets of words at study). As in Experiment 1, all words—both studied and unstudied—were tested in yellow. Subjects in the two groups received instructions at study either to read all words aloud or to read all words silently.

Results

Table 2 presents the yes/no recognition data for each group. A 2 × 2 ANOVA showed only a significant effect of response type, with more *yes* responses to studied items (hits; .666) than to new items (false alarms; .225), $F(1, 28) = 165.92$, $MSE = 0.033$, $p < .001$. The main effect of study condition was not significant, $F(1, 28) = 1.97$, $MSE = 0.018$, $p > .15$, nor was the interaction ($F < 1$), together indicating no production effect in the between-subjects situation. This was buttressed by two *t* tests. The first showed that the false-alarm rates for the two groups did not differ reliably, $t(28) = 1.38$, $p > .15$. More important, the second showed no reliable production effect in the hit rates, $t(28) = 0.95$, $p > .30$. The .062 difference in recognition hits favoring the group that studied the words aloud was almost exactly offset by the .070 difference in false-alarm rate, with the aloud group again having the higher value.

Discussion

This second experiment reproduced the other principal finding of Hopkins and Edwards (1972): the absence of a production effect on an explicit recognition test in a between-subjects design (see also Dodson & Schacter, 2001). The main purpose of Experiment 2 was achieved: to examine whether the production effect is consistently limited to within-subject study manipulations. It is, with the lone exception of a difference in the single between-subjects experiment of Gathercole and Conway (1988, Experiment 5). With this outlier in mind, we decided to go one step farther and replicate the two-alternative forced-choice recognition findings of Hopkins and Edwards (Experiment 1).

Experiment 3

Hopkins and Edwards (1972) showed the same pattern in their two experiments: Words spoken aloud were better recognized than words read silently in within-subject conditions but not in between-subjects conditions, and this was true regardless of recognition test format—two-alternative forced choice (Experiment 1) or yes/no (Experiment 2). Because the forced-choice procedure ensures that the influence of any manipulation is not purely on bias, it is worthwhile to examine it in the production situation. We therefore carried out two parallel experiments using two-alternative forced-choice recognition, one with a within-subject design (Experiment 3A) and one with a between-subjects design (Experiment 3B). We fully expected to observe a production benefit within but not between subjects, replicating the results of Hopkins and Edwards.

Table 2
Proportion of Yes Responses in Recognition as a Function of Study–Test Condition for Experiment 2 (Between Subjects)

	Read aloud		Read silently		Not studied	
Condition	*M*	*SE*	*M*	*SE*	*M*	*SE*
Aloud	.697	.044			.260	.041
Silent			.635	.047	.190	.031

Method

Subjects. Experiment 3A was conducted within subject, and 17 subjects took part. Experiment 3B was conducted between subjects with 15 subjects in each condition.

Stimuli. An enlarged set of 160 words was created, consisting of the previously used 120 words and 40 added words that followed the same selection rules. Because there was no implicit test in Experiments 3A and 3B, all studied items appeared on the test, increasing experimental power. In Experiment 3A, a random 88 words were selected for study, with 44 presented in blue and 44 in white, in random order. To construct test pairs, we paired 36 of the words studied in blue with 36 unstudied words and 36 of the words studied in white with 36 unstudied words. In each subset, the studied word appeared half the time on the left and half the time on the right, with the order of pairs randomized. We also included eight test trials on which both items were previously studied, one aloud and one silently. These "catch trials" appeared on every ninth trial during test, with half the words studied aloud on the left and half the words studied aloud on the right. We were curious whether subjects would show a preference for words studied aloud when pitted against words studied silently. Experiment 3B was simpler, with 40 words studied in blue and 40 in white (subjects were told to ignore color), and 80 test pairs constructed as for the majority of trials in Experiment 3A; catch trials are not possible with production manipulated between subjects.

Apparatus. The apparatus was the same as in Experiments 1 and 2. In this experiment, however, the controlling program was written in E-Prime (Version 1.1; Schneider, Eschman, & Zuccolotto, 2002).

Procedure. All aspects of the procedure closely followed those of the preceding experiments, with two minor changes: (a) Both colors of words were presented for 2,000 ms, and (b) all stimuli were presented in a 16-point lowercase font against a black background. The only significant changes were in the test phase. First, there was no implicit speeded reading test; rather, the recognition test began immediately after study, with only a short set of instructions intervening. In Experiment 3A, subjects were not told of the catch trials. In both experiments, subjects were told that they must choose one of the two words displayed on each test trial, guessing when necessary. Test trial display consisted of two lowercase words in yellow centered on the screen with three spaces between them. Responding was subject paced with no speed emphasis and was done by pressing the *C* key if the word on the left was judged to be the studied word and the *M* key if the word on the right was judged to be the studied word.

Results

Table 3 presents the two-alternative forced-choice recognition data for Experiments 3A and 3B, which represent the proportion of times that the correct word was selected from a test pair as a function of study condition.

Experiment 3A. For the subjects tested under the within-subject design, words studied aloud were recognized significantly

Table 3
Proportion of Studied Words Correctly Selected in Recognition as a Function of Study–Test Condition for Experiments 3A and 3B (Forced Choice Recognition)

	Read aloud		Read silently	
Condition	*M*	*SE*	*M*	*SE*
3A: Within subject	.872	.020	.739	.025
3B: Between subjects	.843	.021	.811	.027

better than words studied silently, $t(16) = 4.77$, $p < .001$, with recognition of words studied aloud exceeding by .133 that of words studied silently. Subjects also preferred words studied aloud (.664) when pitted against words studied silently (.336) on the catch trials, as revealed by a single-sample *t* test, $t(16) = 4.12$, $p < .001$.

Experiment 3B. For the subjects tested under a between-subjects design, words studied aloud were not recognized better than words studied silently, $t(16) = 1.52$, $p > .15$, with recognition of words read aloud only slightly (.032) above that for words read silently during study.

Comparing designs. If we treat Experiment 3B as if it were a within-subject design, then we can directly compare the two sub-experiments by treating experiment as a between-subjects variable and encoding condition as a within-subject variable in a 2 × 2 mixed ANOVA. This analysis showed no effect of experiment ($F < 1$). There was an overall significant effect of encoding condition, $F(1, 30) = 21.41$, $MSE = 0.005$, $p < .001$, but it is crucial that this was modified by a significant interaction, $F(1, 30) = 8.19$, $p < .01$. This interaction is as anticipated: Consistent with the separate analyses, the production effect was present within subject (Experiment 3A) but absent between subjects (Experiment 3B).

Discussion

The results of Experiments 3A and 3B are entirely consistent with those of Experiments 1 and 2. Regardless of the recognition test format—yes/no or forced choice—the production effect was present when the manipulation at study was done within subject but absent when the manipulation was done between subjects. That this occurs in the forced-choice paradigm removes any concern that the production effect might simply be a bias favoring the aloud study condition. For the catch trials, because aloud words benefit from their distinctiveness, subjects tend to select aloud over silent on these trials rather than just guess randomly. Having now reproduced the two major findings of previous work on the production effect, and having completely replicated the Hopkins and Edwards (1972) original study, we report a series of experiments exploring the boundaries of this very replicable, robust phenomenon.

Experiment 4

In memory experiments, investigators typically do not worry about whether words are produced during study, sometimes having subjects study words silently and sometimes having them say words aloud as they study. Of course, this aloud–silent manipulation is generally done between experiments, so, on the basis of Experiments 2 and 3B, and of the experiments of Hopkins and Edwards (1972) and Dodson and Schacter (2001), no production effect should be observed even were an investigator to make the direct comparison. Yet reading some of the words aloud in a within-subject design, as in Experiments 1 and 3A, turns out to be a simple manipulation with a quite powerful effect, increasing recognition by 10%–20%.

Treatment of words read aloud differs from that of words read silently in that the response is overt and vocal and uniquely identifies the item. We suspected that the third factor was the important one, providing a word-specific differentiating record for each word read aloud at study. We consequently sought to eliminate the other two factors in Experiment 4. Thus, in Experiment 4A, subjects pressed a key for each blue word but not for each white word, in a test of the overt response possibility. In Experiment 4B, subjects said "yes" aloud to each blue word but not to each white word, in a test of the vocal response possibility. If simply responding to some words, or doing so vocally, is the crucial factor, then we should see a production effect in one or both of these experiments despite the responses not being unique. If producing individual words provides additional distinctive information about them, and if this is the source of the production effect, then we should not observe the effect in either of these experiments.

Method

Subjects. Forty-eight students from the same source received bonus course credit for taking part. Twenty-four were randomly assigned to each of the experiments.

Procedure. The stimuli, apparatus, and procedure were virtually identical to those in Experiments 1 and 2. The only change was that instead of saying blue words aloud, subjects pressed the space bar for each blue word in Experiment 4A, and they said "yes" aloud for each blue word in Experiment 4B. Subjects were to read each word silently before responding.

Results

Experiment 4A. Table 4 presents the data for keypress responses. A one-way ANOVA showed a significant overall effect, $F(2, 46) = 85.08$, $MSE = 0.017$, $p < .001$. The first planned comparison showed that there was good memory for the studied words, $F(1, 23) = 193.36$, $MSE = 0.087$, $p < .001$. Most important, the second planned comparison showed no reliable produc-

Table 4
Proportion of Yes Responses in Recognition as a Function of Study–Test Condition for Experiments 4A and 4B

	Responded overtly		Read silently		Not studied	
Condition	*M*	*SE*	*M*	*SE*	*M*	*SE*
4A: Keypress	.673	.031	.608	.034	.223	.025
4B: Vocal yes	.625	.029	.625	.026	.319	.033

tion effect, $F(1, 23) = 2.64$, $MSE = 0.038$, $p > .10$, with two subjects responsible for half the apparent small difference.

Experiment 4B. Table 4 presents the recognition data for vocal "yes" responses. A one-way ANOVA showed a significant overall effect, $F(2, 46) = 53.56$, $MSE = 0.014$, $p < .001$. The first planned comparison showed that there was good memory for studied words, $F(1, 23) = 96.04$, $MSE = 0.094$, $p < .001$. Most important, the second planned comparison showed no reliable production effect ($F < 1$).

Together. Treating Experiments 4A and 4B as a between-subjects variable, we analyzed the hit data with a 2 × 2 mixed ANOVA. There were no significant effects (all Fs < 1.60), supporting the claim that making the same overt response to half the items does not result in a production effect, regardless of the type of repeated overt response.

Discussion

The production effect disappeared in Experiment 4, despite use of a within-subject design. Therefore, the effect is not ordinarily the result either of making an overt response (Experiment 4A) or of making specifically a vocal response (Experiment 4B) to a subset of the words. Instead, it appears that the individual words themselves must be produced for the production effect to appear. This act serves to differentiate the individual words spoken aloud from other words studied in the same context.

Experiment 5

If the addition of distinctive information is the basis of the production effect, then it should not be necessary to have spoken production at study to obtain the production advantage. Any item-specific response might well produce the effect, if that unique response can be recollected later. We were reminded of the experiments of Conway and Gathercole (1987), which suggested a possible advantage for mouthed words relative to silent words. This advantage was never reliable in their work, likely due in part to small numbers of subjects and of items per condition, resulting in limited power. If the key to the production effect is the production not of a spoken response but of a unique, distinctive response, then mouthing should produce the effect, too, simultaneously ruling out acoustic processing as necessary to the effect.

Method

Subjects. Twenty-four students from the same source received bonus course credit for taking part.

Procedure. The stimuli and apparatus were the same as in Experiments 1, 2, and 4. As in Experiment 3, the controlling program was written in E-Prime; however, the recognition test was now yes/no. Subjects were instructed to silently mouth words presented in blue and to silently read words presented in white, without mouthing.

Results

Table 5 displays the recognition data. A one-way ANOVA showed a significant overall effect, $F(2, 46) = 97.91$, $MSE = 0.018$, $p < .001$. The first planned comparison showed good memory for studied words, $F(1, 23) = 154.75$, $MSE = 0.121$, $p < .001$. Most important, the second planned comparison showed a reliable production effect, with hits for mouthed words exceeding by .179 those for words read silently during study, $F(1, 23) = 24.62$, $MSE = 0.031$, $p < .001$.

Because the designs of Experiments 1 and 5 are identical, apart from the switch from reading aloud to mouthing, we conducted a 2 × 2 ANOVA dropping the unstudied condition and treating experiment as a between-subjects variable. The main effect of experiment was not significant ($F < 1$), indicating comparable levels of overall recognition in the two experiments. As expected, the main effect of condition was significant, $F(1, 66) = 59.97$, $MSE = 0.015$, $p < .001$, but the interaction of experiment and condition was not ($F < 1$), confirming the identical production effect pattern in Experiments 1 and 5.

Discussion

As in Experiment 1, in which subjects produced a spoken response under a within-subject design, in Experiment 5 the production of a mouthed, nonvocal response for some studied items also brought about a production effect. This is consistent with the idea that it is the production of an item-specific, unique response that underlies enhanced memory due to production. Vocalization is not necessary to make produced items distinctive from nonproduced items.

Experiment 6

Thus far, the production effect has been restricted to words. Is it more general? If producing a word makes it distinctive at study, and that is what results in the production benefit, then the effect should hold for other stimuli and for other encoding tasks as well. In Experiment 6, we changed the stimuli to pronounceable nonwords (e.g., *datch*). These items have neither meaning nor prior memory representation. Nonwords typically do not produce generation effects (e.g., Mulligan, 2002c), although there is some debate on this (Johns & Swanson, 1988; Payne, Neely, & Burns, 1986). This debate may rest in part on the fact that rules for the

Table 5
Proportion of Yes Responses in Recognition as a Function of Study–Test Condition for Experiments 5 and 6

	Mouthed		Read aloud		Read silently		Not studied	
Condition	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
5: Mouthing at study	.779	.036			.600	.032	.248	.021
6: Nonwords			.714	.036	.517	.039	.279	.031

generation of nonwords are relatively complicated. In contrast, production of nonwords is straightforward (e.g., say *manty* aloud). Will nonwords show a production effect? We think that they should, revealing one potential difference between production and generation.

Method

Subjects. Twenty-one students from the same source received bonus course credit for taking part.

Stimuli and apparatus. The stimuli were a new set of 120 pronounceable nonwords four to six letters in length (e.g., *hest*, *slass*, *prench*). In each case, these were composed by altering a single letter—almost always a consonant—in a word such that the resulting nonword remained readily pronounceable. The apparatus was the same as that used in the previous experiments. The controlling program again was written in E-Prime.

Procedure. Subjects were instructed to read aloud nonwords in blue and to read silently nonwords in white. All other procedural aspects were identical to those in Experiment 5.

Results

Table 5 presents the recognition data. A one-way ANOVA showed a significant overall effect, $F(2, 40) = 45.13$, $MSE = 0.022$, $p < .001$. The first planned comparison showed good memory for studied nonwords, $F(1, 20) = 77.10$, $MSE = 0.124$, $p < .001$. Most important, the second planned comparison revealed a robust production effect: Hits for nonwords read aloud exceeded by .197 those for nonwords read silently, $F(1, 20) = 17.31$, $MSE = 0.047$, $p < .001$.

Because the designs of Experiments 1 and 6 are identical, apart from the switch from words to nonwords, we conducted a 2×2 ANOVA dropping the unstudied condition and treating experiment as a between-subjects variable. There was a marginally significant main effect of experiment, $F(1, 63) = 3.67$, $MSE = 0.031$, $p = .06$, reflecting the unsurprising overall better recognition in Experiment 1 (words) than in Experiment 6 (nonwords). As expected, the main effect of condition was significant, $F(1, 63) = 52.25$, $MSE = 0.018$, $p < .001$, but the interaction of experiment and condition was not ($F < 1$), confirming the identical production effect pattern in Experiments 1 and 6.

Discussion

There clearly was a strong production effect for nonwords. Indeed, the effect was at least as large as that seen in previous production experiments with words, despite the reduced overall recognition due to the unfamiliar stimuli. An existing representation is, therefore, not a necessary condition for the production effect to occur, nor must words be the stimuli that are studied. As well, the production effect can benefit relatively weak encodings. Having shown that the effect is robust to a material change, in the final two experiments we turn to processing changes during encoding.

Experiment 7

In Experiment 7, our goal was to examine a different encoding task, in particular a richer task that leads to superior memory performance. All of the previous experiments involved stimuli that were single words (or nonwords) to be produced versus read silently. Here we brought together generation and production. Subjects studied via generation of response words from definitional cues, such as "the white drink extracted from a dairy cow—m?" generating the response aloud (saying "milk") or saying "next" aloud when they had generated the response covertly. We expected a production effect in recognition here as well. Despite the much richer encoding fostered by generation than by reading, additional distinctiveness should still be provided by oral production of the word that fits the cue. It is important to note that such a result would also demonstrate that production can increment memory even for words that are very well remembered, as is the case following generation.

Method

Subjects. Thirty-five students from the same source received bonus course credit for taking part.

Stimuli and apparatus. The apparatus was identical to that in prior experiments. The stimuli were switched from single words to generation cues for words (e.g., "the tiny infant commonly put in a cradle—b?" for the word "baby"). There were 120 cues, treated otherwise just as the 120 words in prior experiments were treated. These generation cues were an extension of the set of 90 cues used by Masson and MacLeod (2002). In their experiments, probability of generating the intended target word during study was over 90%.

Procedure. The procedure was also virtually identical to that of Experiment 1, except that the study materials were now generation cues, with those presented in blue to be generated aloud and those presented in white to be generated silently. Each cue stayed on the screen until the subject either generated aloud (blue) or responded "next" aloud (white). Overt responding to the white cues should not influence performance (see Experiment 4B in the present article and the contrast between Experiments 1 and 2 in MacDonald and MacLeod, 1998) and was incorporated to ensure that subjects generated items even on the silent trials.

Results

Table 6 presents the recognition data. A one-way ANOVA showed a significant overall effect, $F(2, 68) = 460.39$, $MSE = 0.016$, $p < .001$. The first planned comparison showed good memory for words generated during study, $F(1, 34) = 824.52$, $MSE = 0.106$, $p < .001$. Most important, the second planned comparison revealed the presence of the production effect, with hits for words generated aloud exceeding by .096 those for words generated silently, $F(1, 34) = 11.22$, $MSE = 0.029$, $p < .01$.

Because the designs of Experiments 1 and 7 are identical, apart from the switch from reading to generating, we conducted a 2×2 ANOVA dropping the unstudied condition and treating experiment as a between-subjects variable. As expected, the main effect of experiment was significant, $F(1, 77) = 55.84$, $MSE = 0.024$, $p < .001$, with better overall recognition of words that were generated (Experiment 7) than of words that were read (Experiment 1). Also as expected, the main effect of condition was significant, $F(1, 77) = 44.53$, $MSE = 0.015$, $p < .001$, demonstrating the production effect. The marginally significant interaction of experiment with condition, $F(1, 77) = 2.98$, $p = .09$,

Table 6
Proportion of Yes Responses in Recognition as a Function of Study–Test Condition for Experiments 7 and 8

	Generated aloud		Generated silently		Read aloud		Read silently		Not studied	
Condition	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
7: Generation at study	.913	.012	.817	.028					.076	.017
8: Semantic encoding at study					.868	.020	.804	.029	.143	.015

reflects a larger production effect in Experiment 1 (0.16) than in Experiment 7 (0.10), but this may well have resulted from a limitation on the size of the effect in Experiment 7, in which overall recognition was very high due to generation. The important finding here is that there was a reliable production effect in Experiment 7, despite all the items having been generated, which resulted in excellent overall recognition.

Discussion

Remarkably, the production effect persisted even when the base encoding task had already promoted exceptionally good memory. In prior experiments in this series, hit rate was .60–.75 and false-alarm rate was .20 or higher. Here hit rate rose to .80–.90, and false-alarm rate fell to under .10, yet the production effect remained. Moreover, the improvement was equivalent for the aloud and silent conditions: From Experiment 1 (base of reading) to Experiment 7 (base of generating), the aloud condition rose .153, and the silent condition rose a comparable—indeed, slightly greater—.219. So, subjects clearly were generating good encodings even in the silent condition. One might have imagined that the generation effect would overwhelm the production effect, but it did not. Apparently, production benefits relatively weak encoding (simply reading) and relatively strong encoding (generating), possibly even to similar extents.

This outcome makes sense if the production effect relies on the distinctiveness available at the time of test for items that were studied aloud. To ascertain whether a test item was studied, subjects have an extra source of confirming information if they can recall having produced that item. This is true regardless of how the word was encoded—by reading or by generating—providing diagnostic information in favor of judging the item to have been studied. Thus, production is as beneficial to a word encoded strongly as to a word encoded more weakly. Consequently, we would expect production to benefit other types of encoding as well. As just one example, using a levels-of-processing manipulation, recognition of both the structurally and elaboratively encoded items should benefit from production; indeed, the strong view would hold that they should benefit equivalently.

Experiment 8

We have been arguing throughout this series of experiments that production benefits the produced items. But there is another possibility: Perhaps production hurts the read items! This alternative view was considered in work on the generation effect, notably by Begg and his collaborators (Begg & Roe, 1988; Begg & Snider, 1987; Begg, Vinski, Frankovich, & Holgate, 1991). Begg and Snider (1987) referred to this idea as the lazy reading hypothesis—that generating some words in a mixed-list design might result not in superior encoding of the generated words but in inferior encoding of the words that were not generated and hence received less attention.

Because on generation trials the word is not physically presented, it is difficult to try to equate initial reading in the generation and read conditions. This problem does not arise in production experiments, in which the whole word is presented on both production and read trials. Therefore, we decided to ensure that the word was processed for meaning even in the read condition. To accomplish this, we incorporated a quite standard levels-of-processing orienting task: Subjects were required to indicate as each word appeared whether it represented something living or nonliving; only after making that judgment were they to say the word aloud or silently.

If the production effect hinges on lazy reading of the words read silently, then this added semantic judgment should eliminate the production effect by enforcing semantic analysis of all words. If, however, the production effect derives from having an additional encoded feature to remember—that the aloud words were in fact said aloud—then the production effect should persist, and the lazy reading hypothesis will be refuted. We note as well that this experiment permits another test of whether the production effect occurs even when item encoding is very strong, similar to that seen in Experiment 7 with generation as the encoding task.

Method

Subjects. Twenty-seven students from the same source received bonus course credit for taking part.

Stimuli and apparatus. The stimuli were 120 of the 128 words from Experiment 2 of Hourihan and MacLeod (2007), assembled such that 60 were living and 60 nonliving. The apparatus was the same as in Experiment 1. The controlling program was written in E-Prime.

Procedure. Each trial began with the word presented at the center of the screen in white. At the bottom of the screen in yellow were two choices: "not living" with *Z* below it (on the left) and "living" with *M* below it (on the right). When the subject pressed a key indicating his or her animacy decision, the instructions disappeared and the word turned either green (to be read aloud) or red (to be read silently) for an additional 1,000 ms. A 1,000-ms blank then preceded the next trial. At test, words were presented in yellow, and subjects responded *yes* with *M* and *no* with *Z*. All items were tested with recognition (i.e., there was no implicit test). Subjects also had four practice trials at the start of the experiment, one trial in each condition (living vs. not living by aloud vs. silent).

These four trials were the same (but in random order) for all subjects. For all experimental study trials, items and colors were randomly assigned.

Results

Table 6 presents the recognition data. A one-way ANOVA showed a significant overall effect, $F(2, 52) = 732.00$, $MSE = 0.006$, $p < .001$. The first planned comparison showed very good memory for studied words, $F(1, 26) = 1{,}097.87$, $MSE = 0.047$, $p < .001$. Most important, the second planned comparison revealed a reliable production effect, with hits for words read aloud exceeding by .064 those for words read silently, $F(1, 26) = 13.49$, $MSE = 0.008$, $p < .001$.

Because the designs of Experiments 1 and 8 are identical, apart from the addition of the living–nonliving judgment, we conducted a 2 × 2 ANOVA dropping the unstudied condition and treating experiment as a between-subjects variable. There was a significant main effect of experiment, $F(1, 69) = 28.33$, $MSE = 0.029$, $p < .001$, reflecting the overall better recognition in Experiment 8 than in Experiment 1, no doubt due to the enhanced semantic processing in Experiment 8. As expected, the main effect of condition was significant, $F(1, 69) = 39.67$, $MSE = 0.011$, $p < .001$, indicative of the overall production effect. The significant interaction of experiment with condition, $F(1, 69) = 7.65$, $p < .01$, with a larger production effect in Experiment 1 than in Experiment 8, may well have resulted from a limitation on the size of the effect in Experiment 8, in which overall recognition was high due to the added semantic judgment.[2]

Discussion

The important finding in Experiment 8 is the reliable production effect despite enforced semantic processing of all the words, most notably of the words read silently. Once again, as was the case with generation in Experiment 7, a strong encoding that substantially raised overall recognition performance did not eliminate the production effect. Therefore, the lazy reading hypothesis was shown not to underlie the production effect because the effect persisted even when lazy reading was no longer possible. There clearly is a systematic benefit to production.

General Discussion

Across these eight experiments, we have shown the benefit of having produced a word at study to be robust, even more than might have been expected based on the five isolated prior studies (Conway & Gathercole, 1987; Dodson & Schacter, 2001; Gathercole & Conway, 1988; Hopkins & Edwards, 1972; MacDonald & MacLeod, 1998). We have confirmed that the production effect is limited to within-subject, mixed-list designs in which distinctive responses are made to individual items during study. The effect is clearly a benefit for the items that are produced, not a cost for the items that are not produced. The effect does not require vocal production of the word (mouthing is sufficient), nor does it require that the stimuli be meaningful (nonwords show a robust benefit). And it is important to note that the production effect increments even study via generation or deep, semantic processing, where performance is already strong. We have also shown in the Appendix that relative to simply reading a word silently, production does not enhance priming in speeded reading, an optimally configured implicit test. This entire pattern is stitched together theoretically by the idea that production at study makes the produced items distinctive, thereby providing a source of discrimination that can be used heuristically on an explicit test to determine whether a word was in fact studied.

The production effect would appear to be a member of a class of variables that increase distinctiveness, thereby enhancing explicit memory. These variables include the enactment effect, in which performing an action oneself in response to a brief instruction (e.g., "break the match") leads to better memory than simply reading the instruction (e.g., Engelkamp & Dehn, 2000; Engelkamp & Jahn, 2003; see Engelkamp, 1998, for a review), and of course the generation effect, in which producing a word from a cue leads to better memory than simply reading the word (e.g., Mulligan, 2002c; Slamecka & Graf, 1978; see Bertsch et al., 2007, for a review and meta-analysis). All three encoding manipulations involve producing an overt response. All three also share three notable attributes. First, with infrequent exceptions, they all work primarily in within-subject, mixed-list designs (for enactment, see Engelkamp & Dehn, 2000; for generation, see Begg & Snider, 1987; Hirshman & Bjork, 1988; Mulligan, 1999; Slamecka & Katsaiti, 1987; but see Mulligan, 2002b).[3] Second, with rare exceptions, all three typically express themselves strongly on explicit tests but not on implicit tests (for enactment, see Engelkamp, Zimmer, & Kurbjuweit, 1995; Nyberg & Nilsson, 1995; for generation, see Gardiner, 1988; Masson & MacLeod, 1992; Toth, Reingold, & Jacoby, 1994). And third, all three ordinarily require a unique response to produce an advantage (for generation, this is simply assumed in the method; for enactment, see Zimmer & Engelkamp, 2003).

Before returning to our theoretical framework, we would be remiss if we did not consider an apparently related phenomenon: the list-strength effect. The list-strength effect refers to the finding that increasing the strength of some items in a list—by manipulation of repetition or study duration or level of processing, for example—decreases the strength of the other items in that list (Ratcliff, Clark, & Shiffrin, 1990; Verde, 2009). Measurement of the effect involves comparison of pure strong lists with pure weak lists and particularly of strong versus weak items in a mixed-list design. Especially noteworthy is the fact that although the list-strength effect is quite robust in free recall, it is absent in cued

[2] A 2 × 2 ANOVA treating experiment as a variable to directly compare Experiments 7 and 8 showed a significant main effect only of encoding condition, $F(1, 60) = 19.59$, $MSE = 0.010$, $p < .001$. The main effect of experiment, $F(1, 60) = 1.14$, $MSE = 0.022$, and the interaction ($F < 1$) were both not significant. This analysis indicates an entirely consistent pattern across these two experiments in which memory performance was overall very high due to elaborative encoding, yet the production effect was consistently present.

[3] It is noteworthy that in a recent review, McDaniel and Bugg (2008) have identified a set of five memory phenomena—the generation effect, the enactment effect, the word frequency effect (see, e.g., MacLeod & Kampe, 1996), the perceptual interference effect (see, e.g., Mulligan, 1996), and the bizarreness effect (see, e.g., McDaniel & Einstein, 1986)—all of which typically reveal differential performance within subject but not between subjects (see also Mulligan & Peterson, 2008).

recall and in recognition (e.g., Verde, 2009; Yonelinas, Hockley, & Murdock, 1992).

The list-strength effect and the production effect look superficially similar. Both show better memory for strong items than for weak items in a mixed list, and this advantage is reduced or eliminated when pure lists are compared. So, could the production effect simply be an instance of the list-strength effect? We think not—for two reasons. First, we have not seen a reliable decrease for silent items in a mixed list as opposed to a pure list. The production effect seems to be more an enhancement of the aloud items. The second reason is more compelling. We are convinced of their difference because of a striking dissociation: The list-strength effect does not occur in recognition, yet the production effect has been observed primarily in recognition, where it is large and easily obtained. This is important because the fact that relative strength does not affect recognition despite the production effect being solid in recognition indirectly suggests that the production effect is not due to relative strength. This is at least consistent with our argument that the production effect hinges on distinctiveness, not strength.

As set out in the introduction, the idea of distinctiveness has had a long history in explanations of memory (see Hunt, 2006, for an overview). From our perspective, it is the active ingredient in the production effect: The produced words are differentiated by being processed distinctively against the backdrop of the silently read (unproduced) words. This is consistent with the general account of Conway and Gathercole (1987; see also Gathercole & Conway, 1988) and with the more developed framework of Dodson and Schacter (2001), and it traces back to Murdock's (1960) theory and earlier. From our perspective, it is not surprising that generation has also been posited to make words distinctive (see Begg, Snider, Foley, & Goddard, 1989; Gardiner & Hampton, 1988).

Our view is that, at the time of test, a word that was produced at study has an additional source of discrimination relative to a word that was not produced. We think of this in terms of the proceduralist framework of Kolers (1973; Kolers & Roediger, 1984; for a review, see Roediger et al., 2002). The memory record of a produced item includes that it was produced during study. Replaying this record at the time of retrieval can be used to diagnose whether an item was in fact studied: Successfully retrieving that the word was produced necessarily indicates that it was studied. This source is not available for words that were read silently, which essentially is the default state and therefore not distinctive, particularly in view of the fact that on a recognition test the distractors also were not produced at study. So discrimination is better for words that were produced than for those that were read silently.

In the introduction we asked why this benefit would accrue to produced words only when mixed with words read silently. Remembering that a word was produced at study is diagnostic even if all the words were studied aloud. Our view is that subjects do not use study mode as a source of distinctiveness when there is only one study mode, as in a between-subjects design. Dodson and Schacter (2001) offered a similar argument, and Murdock's (1960) theory certainly emphasizes that the two types of processing must coexist for distinctiveness to be operative. Perhaps, if the subjects who produced all the words were coached that it was potentially helpful to try to remember whether a test word had been produced, a production benefit might emerge (relative to the subjects who read all the words silently). Our claim is that subjects in within-subject experiments routinely run this diagnostic test. As a result, performance benefits from production regardless of the type of encoding, explaining why production even improves recognition of items that were generated or semantically processed at study. However, for an item to be distinctive, it must be produced uniquely at study; hence pressing a key or saying yes to a subset of the items fails to produce the benefit.

Additional distinctive information can presumably benefit any test in which differentiation of studied from unstudied items is relevant or important at study and can contribute to reprocessing at test. Thus, recognition should certainly benefit from prior production, as we, Hopkins and Edwards (1972), Conway and Gathercole (1987), MacDonald and MacLeod (1998), and Dodson and Schacter (2001) have shown. Under a simple generate–recognize model (see, e.g., Kintsch, 1970), production should also benefit recall in that discrimination of studied from unstudied items will be better in the recognition stage for words spoken aloud at study. In their Experiment 3, Conway and Gathercole reported just such a production effect in recall.

More generally, any explicit memory test should show a production effect because all such tests are characterized—indeed, defined—by emphasis on distinguishing studied from unstudied items. In contrast, implicit tests expressly do not invoke such discrimination, making no reference to the study episode, and therefore should not show a production effect. That is precisely what we have repeatedly observed, and it is comforting that this pattern coincides with that observed for enactment and generation.

The production effect is robust and substantial. Consequently, we intend to pursue the empirical and theoretical issues that arise out of it, further examining its boundary conditions and further testing the distinctiveness explanation. Moreover, the effect may well turn out to be of considerable practical import. Studying some information aloud and other information silently may prove to be a useful strategy in highlighting the important information and making it more accessible to retrieval, and perhaps even in diminishing the accessibility of information deemed to be less important and hence not studied aloud.

References

Begg, I., & Roe, H. (1988). On the inhibition of reading by generating. *Canadian Journal of Psychology, 42,* 325–336.

Begg, I., & Snider, A. (1987). The generation effect: Evidence for generalized inhibition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 553–563.

Begg, I., Snider, A., Foley, F., & Goddard, R. (1989). The generation effect is no artifact: Generating makes words distinctive. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 977–989.

Begg, I., Vinski, E., Frankovich, L., & Holgate, B. (1991). Generating makes words memorable, but so does effective reading. *Memory & Cognition, 19,* 487–497.

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition, 35,* 201–210.

Burnham, W. H. (1888). Memory, historically and experimentally considered. I. *American Journal of Psychology, 2,* 39–90.

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

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A

framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11,* 671–684.

Crowder, R. G. (1970). The role of one's own voice in immediate memory. *Cognitive Psychology, 1,* 157–178.

Dodson, C. S., & Schacter, D. L. (2001). "If I had said it I would have remembered it": Reducing false memories with a distinctiveness heuristic. *Psychonomic Bulletin & Review, 8,* 155–161.

Ekstrand, B. R., Wallace, W. P., & Underwood, B. J. (1966). A frequency theory of verbal-discrimination learning. *Psychological Review, 73,* 566–578.

Engelkamp, J. (1998). *Memory for actions.* Hove, England: Psychology Press.

Engelkamp, J., & Dehn, D. M. (2000). Item and order information in subject-performed tasks and experimenter-performed tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 671–682.

Engelkamp, J., & Jahn, P. (2003). Lexical, conceptual and motor information in memory for action phrases: A multi-system account. *Acta Psychologica, 113,* 147–165.

Engelkamp, J., Zimmer, H. D., & Kurbjuweit, A. (1995). Verb frequency and enactment in implicit and explicit memory. *Psychological Research, 57,* 242–249.

Foley, M. A., Johnson, M. K., & Raye, C. L. (1983). Age-related changes in confusion between memories for thoughts and memories for speech. *Child Development, 54,* 51–60.

Gardiner, J. M. (1988). Generation and priming effects in word-fragment completion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 495–501.

Gardiner, J. M., & Hampton, J. A. (1988). Item-specific processing and the generation effect: Support for a distinctiveness account. *American Journal of Psychology, 101,* 495–504.

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

Geraci, L., & Rajaram, S. (2006). The distinctiveness effect in explicit and implicit memory. In R. R. Hunt & J. B. Worthen (Eds.), *Distinctiveness and memory* (pp. 211–234). New York, NY: Oxford University Press.

Graves, R. E., & Bradley, R. (1991). Millisecond timing on the IBM PC/XT/AT and PS/2: A review of the options and corrections for the Graves and Bradley algorithm. *Behavior Research Methods, Instruments, & Computers, 23,* 377–379.

Higbee, K. L. (1988). *Your memory: How it works and how to improve it* (2nd ed.). New York, NY: Prentice Hall.

Hirshman, E., & Bjork, R. A. (1988). The generation effect: Support for a two-factor theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 484–494.

Hopkins, R. H., Boylan, R. J., & Lincoln, G. L. (1972). Pronunciation and apparent frequency. *Journal of Verbal Learning and Verbal Behavior, 11,* 105–113.

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

Hourihan, K. L., & MacLeod, C. M. (2007). Capturing conceptual implicit memory: The time it takes to produce an association. *Memory & Cognition, 35,* 1187–1196.

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, NY: Oxford University Press.

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

Hunt, R. R., & Worthen, J. B. (Eds.). (2006). *Distinctiveness and memory.* New York, NY: Oxford University Press.

Jacoby, L. L., & Craik, F. I. M. (1979). Effects of elaboration of processing at encoding and retrieval: Trace distinctiveness and recovery of initial context. In L. S. Cermak & F. I. M. Craik (Eds.), *Levels of processing in human memory* (pp. 1–22). Hillsdale, NJ: Erlbaum.

Johns, E. E., & Swanson, L. G. (1988). The generation effect with nonwords. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 180–190.

Kintsch, W. (1970). Models for free recall and recognition. In D. A. Norman (Ed.), *Models of human memory* (pp. 333–370). New York, NY: Academic Press.

Kolers, P. A. (1973). Remembering operations. *Memory & Cognition, 1,* 347–355.

Kolers, P. A., & Roediger, H. L., III. (1984). Procedures of mind. *Journal of Verbal Learning and Verbal Behavior, 23,* 425–449.

Lockhart, R. S., Craik, F. I. M., & Jacoby, L. (1976). Depth of processing, recognition and recall. In J. Brown (Ed.), *Recall and recognition* (pp. 75–102). Oxford, England: Wiley.

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

MacLeod, C. M. (1996). How priming affects two speeded implicit tests of remembering: Naming colors versus reading words. *Consciousness and Cognition, 5,* 73–90.

MacLeod, C. M. (2008). Implicit memory tests: Techniques for reducing conscious intrusion. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of metamemory and memory* (pp. 245–263). New York, NY: Psychology Press.

MacLeod, C. M., & Daniels, K. A. (2000). Direct versus indirect tests of memory: Directed forgetting meets the generation effect. *Psychonomic Bulletin & Review, 7,* 354–359.

MacLeod, C. M., & Kampe, K. E. (1996). Word frequency effects on recall, recognition, and word fragment completion tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 132–142.

MacLeod, C. M., & Masson, M. E. J. (2000). Repetition priming in speeded word reading: Contributions of perceptual and conceptual processing episodes. *Journal of Memory and Language, 42,* 208–228.

Masson, M. E. J., & MacLeod, C. M. (1992). Reenacting the route to interpretation: Enhanced perceptual identification without prior perception. *Journal of Experimental Psychology: General, 121,* 145–176.

Masson, M. E. J., & MacLeod, C. M. (2002). Covert operations: Orthographic recoding as a basis for repetition priming in word identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 858–871.

McDaniel, M. A., & Bugg, J. M. (2008). Instability in memory phenomena: A common puzzle and a unifying explanation. *Psychonomic Bulletin & Review, 15,* 237–255.

McDaniel, M. A., & Einstein, G. O. (1986). Bizarre imagery as an effective memory aid: The importance of distinctiveness. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12,* 54–65.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16,* 519–533.

Mulligan, N. W. (1996). The effects of perceptual interference at encoding on implicit memory, explicit memory, and memory for source. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1067–1087.

Mulligan, N. W. (1999). The effects of perceptual interference at encoding on organization and order: Investigating the roles of item-specific and relational information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 54–69.

Mulligan, N. W. (2002a). The effects of generation on conceptual implicit memory. *Journal of Memory and Language, 47,* 327–342.

Mulligan, N. W. (2002b). The emergent generation effect and hypermnesia: Influences of semantic and nonsemantic generation tasks. *Journal of*

Experimental Psychology: Learning, Memory, and Cognition, 28, 541–554.

Mulligan, N. W. (2002c). The generation effect: Dissociating enhanced item memory and disrupted order memory. *Memory & Cognition, 30,* 850–861.

Mulligan, N. W. (2006). Conceptual implicit memory and the item-specific–relational distinction. In R. R. Hunt & J. B. Worthen (Eds.), *Distinctiveness and memory* (pp. 183–209). New York, NY: Oxford University Press.

Mulligan, N. W., & Peterson, D. (2008). Assessing a retrieval account of the generation and perceptual-interference effects. *Memory & Cognition, 36,* 1371–1382.

Mulligan, N. W., & Stone, M. (1999). Attention and conceptual priming: Limits on the effects of divided attention in the category–exemplar production task. *Journal of Memory and Language, 41,* 253–280.

Murdock, B. B., Jr. (1960). The distinctiveness of stimuli. *Psychological Review, 67,* 16–31.

Nyberg, L., & Nilsson, L.-G. (1995). The role of enactment in implicit and explicit memory. *Psychological Research, 57,* 215–219.

Paivio, A. (1971). *Imagery and verbal processes.* Oxford, England: Holt, Rinehart & Winston.

Payne, D. G., Neely, J. H., & Burns, D. J. (1986). The generation effect: Further tests of the lexical activation hypothesis. *Memory & Cognition, 14,* 246–252.

Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). List-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 163–178.

Roediger, H. L., III. (1990). Implicit memory: Retention without remembering. *American Psychologist, 45,* 1043–1056.

Roediger, H. L., III, Gallo, D. A., & Geraci, L. (2002). Processing approaches to cognition: The impetus from the levels-of-processing framework. *Memory, 10,* 319–332.

Roediger, H. L., III, & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 803–814.

Scarborough, D. L., Gerard, L., & Cortese, C. (1979). Accessing lexical memory: The transfer of word repetition effects across task and modality. *Memory & Cognition, 7,* 3–12.

Schacter, D. L., & Wiseman, A. L. (2006). Reducing memory errors: The distinctiveness heuristic. In R. R. Hunt & J. B. Worthen (Eds.), *Distinctiveness and memory* (pp. 89–107). New York, NY: Oxford University Press.

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

Slamecka, N. J. (1992, May 4). Delineation of the generation effect. *Current Contents/Social & Behavioral Sciences, 18,* 10.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 592–604.

Slamecka, N. J., & Katsaiti, L. T. (1987). The generation effect as an artifact of selective displaced rehearsal. *Journal of Memory and Language, 26,* 589–607.

Smith, R. E., & Hunt, R. R. (2000). The effects of distinctiveness require reinstatement of organization: The importance of intentional memory instructions. *Journal of Memory and Language, 43,* 431–446.

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words.* New York, NY: Teachers College, Columbia University.

Toth, J. P., Reingold, E. M., & Jacoby, L. L. (1994). Toward a redefinition of implicit memory: Process dissociations following elaborative processing and self-generation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 290–303.

Verde, M. F. (2009). The list-strength effect in recall: Relative-strength competition and retrieval inhibition may both contribute to forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 205–220.

Weldon, M. S., & Coyote, K. C. (1996). Failure to find the picture superiority effect in implicit conceptual memory tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 670–686.

Yonelinas, A. P., Hockley, W. E., & Murdock, B. B. (1992). Tests of the list-strength effect in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 345–355.

Zimmer, H. D., & Engelkamp, J. (2003). Signing enhances memory like performing actions. *Psychonomic Bulletin & Review, 10,* 450–454.

Zimmer, H. D., Mecklinger, A., & Lindenberger, U. (2006). Levels of binding: Types, mechanisms, and functions of binding in remembering. In H. D. Zimmer, A. Mecklinger, & U. Lindenberger (Eds.), *Handbook of binding and memory: Perspectives from cognitive neuroscience* (pp. 3–24). Oxford, England: Oxford University Press.

Appendix

Speeded Reading as an Implicit Test

In most of the experiments reported in this article—except Experiments 3 and 8—the studied list materials were split in half, with one half of the items assigned to the recognition test and the other half assigned to a speeded reading test (also known as a naming or pronunciation test). In speeded reading, subjects simply read words aloud into a microphone as quickly as possible (e.g., MacLeod, 1996; MacLeod & Daniels, 2000; MacLeod & Masson, 2000; Scarborough, Gerard, & Cortese, 1979), and implicit memory is demonstrated by faster reading of studied than of unstudied words. The aim was to further explore the influence of production on both an explicit test and an implicit test, as begun by MacDonald and MacLeod (1998). But because our emphasis in the present article centered on explicit test performance, we present the implicit test results here, briefly and without disrupting the flow of the recognition findings.

Two views of implicit memory make opposite predictions regarding the influence of production. A transfer-appropriate processing view (Morris, Bransford, & Franks, 1977; Roediger, 1990) would appear to predict more priming for words read aloud at study. Because the implicit test was speeded reading (aloud),

(*Appendix continues*)

reading aloud at study should produce optimal transfer and hence a priming advantage. But an account that views production as making items studied aloud distinct from those studied silently would predict no such differential priming: Such distinctiveness should be useful only on a conscious memory test, where assistance in recovering whether an item was studied is pertinent to the goal.

In accord with the latter account, MacDonald and MacLeod (1998) observed no additional priming on an implicit speeded reading test for words read aloud over those read silently at study. Because this result appears to constitute a violation of transfer-appropriate processing, we chose to obtain further evidence here, and so we included a parallel speeded reading test in most experiments. This test has the virtues that (a) it is less vulnerable than other implicit tests to the intrusion of conscious retrieval processes (see MacLeod, 2008; MacLeod & Masson, 2000) and (b) it is highly similar to the encoding operation of interest—reading a word aloud—and therefore should optimize transfer-appropriate processing (see Morris et al., 1977; Roediger, 1990).

The implicit test, on an independent set of items, constituted Phase 2 and was inserted between study and the explicit test in Experiments 1, 2, 4, 5, 6, and 7. Twenty words that had been presented in each color at study (i.e., half the studied items in each condition), together with 20 words not shown at study (and not used as recognition test distractors), were assigned to the speeded reading test and presented in a new random order. This test was represented as a "filler task" before the actual memory test. All words were presented in yellow to avoid color overlap between study and test. Subjects were to read each word aloud into the microphone as quickly as possible, avoiding errors. On each trial, following a 250-ms blank, the word appeared and remained on the screen until the subject responded. A 250-ms blank followed the response, and finally the word "Ready?" appeared until the experimenter pressed a key to indicate trial acceptability. Response timing used a voice key to measure the time between stimulus presentation and oral response onset; accuracy was scored online by the experimenter.

The speeded reading test descriptive statistics are presented in Table A1. To save space, we do not present the results of formal analyses here because—with the exception of Experiments 6 and 7—all of these analyses had identical outcomes: Relative to the unstudied baseline, there was always reliable priming for words studied either aloud or silently, but there was never any reliable differential priming between the two study conditions. We now briefly summarize these results and their implications.

Experiment 1A replicated the pattern observed by MacDonald and MacLeod (1998)—no differential priming for words produced aloud versus read silently at study. Even when processing match was optimized by requiring speeded reading at both study and test, Experiment 1B also displayed no production effect. It is interesting that the transfer-appropriate processing view would predict more priming for words produced aloud in both within-subject and between-subjects designs, but Experiment 2 also showed no differential priming in a between-subjects design. Experiment 4, despite returning to the within-subject design, would not have been expected to show a production effect on the implicit test given that there was no effect of repeated overt responses on the explicit test. Experiment 5 showed no differential priming for a different form of production (mouthing), which had shown a strong production effect on the corresponding explicit recognition test.

In Experiments 6 and 7, because the studied materials were no longer single words, the pattern of results in speeded reading was expected to change. Experiment 6 switched from words to nonwords. As a consequence, the speeded reading task became much harder, as indicated by considerably higher means accompanied by standard errors on the order of twice as large as in the other experiments in this series. Either for this reason or because nonwords have no preexisting representations in memory, or both, there was no priming at all in this experiment. Finally, in Experiment 7, in which all the studied words were generated, there also was no reliable priming at all, presumably because these words were not actually seen at study and therefore there was no perceptual contribution to produce priming. Most important, though, is that no experiment produced differential priming favoring the items produced at study, a set of results inconsistent with the prediction of the transfer-appropriate processing framework.

Table A1
Implicit Test Results: Mean Speeded Reading Latency (ms)

	Overt response		No overt response		Not studied	
Experiment	*M*	*SE*	*M*	*SE*	*M*	*SE*
1: Read aloud (within subject)						
A: Read at own pace	525	13.26	523	13.12	548	14.93
B: Read quickly	502	15.08	504	13.94	520	17.75
2: Read aloud (between subjects)	527	23.80	517	19.99	554/532[a]	28.63/18.39
4: Repeated overt response						
A: Keypress	530	14.24	534	13.08	546	13.04
B: Say "yes"	543	20.22	541	19.43	551	19.70
5: Mouth	523	14.17	530	14.44	541	16.43
6: Nonwords	566	33.03	582	36.57	594	37.14
7: Generation	497	11.30	506	11.66	500	11.94

[a] Because Experiment 2 used a between-subjects design, there are two independent values for the not-studied condition; the value for the aloud condition precedes the value for the silent condition.

The distinctive information encoded for items produced during study can be seen as providing a basis for differentiating those items at the time of an explicit test, but such differentiation with respect to study status is irrelevant for an implicit test. In this regard, our results fit with others in the literature. For example, Smith and Hunt (2000) previously demonstrated that distinctiveness influences performance on an explicit test but not on an implicit test of memory, a conclusion in line with that of other researchers (Mulligan, 1996, 2002a, 2006; Mulligan & Stone, 1999; Weldon & Coyote, 1996; but see Geraci & Rajaram, 2006, for an exception).

Received March 17, 2008
Revision received March 4, 2009
Accepted March 6, 2009 ■