Canadian Journal of Experimental Psychology / Revue canadienne de psychologie expérimentale
2016, Vol. 70, No. 2, 116–124

1196-1961/16/$12.00 http://dx.doi.org/10.1037/cep0000092

Auditory Presentation at Test Does Not Diminish the Production Effect in Recognition

Noah D. Forrin and Colin M. MacLeod
University of Waterloo

Three experiments investigated whether auditory information at test would undermine the relational distinctiveness of vocal production at study, diminishing the production effect. In Experiment 1, with visual presentation during study, the production effect was equivalently large regardless of whether participants read each test word out loud prior to making their recognition decision. In Experiment 2, incorporating auditory presentation during study, the production effect was unaltered by whether recognition test words were presented visually or auditorily. In Experiment 3, the authors manipulated whether presentation was visual or auditory both at study and at test. Once again, presentation modality at test did not affect the size of the production effect, although the effect was significantly smaller when words were presented auditorily at study. These experiments demonstrate that production at the time of study stands out as distinct above and beyond auditory information. Moreover, this distinct aloud information need not "stand out" against a background of silent unstudied words on a recognition test. Consistent with the distinctiveness account, encoding via production enhances later recognition consistently, regardless of study or test modality.

Keywords: production effect, memory, recognition, auditory, translation hypothesis

There is a robust memory advantage for words read aloud over those read silently. Although this was recognized a century ago (Gates, 1917, p. 67), Hopkins and Edwards (1972) were the first to investigate this encoding technique and its benefit, which they referred to as an effect of pronunciation. This effect was periodically revisited over the years (e.g., Conway & Gathercole, 1987; Dodson & Schacter, 2001; Gathercole & Conway, 1988; MacDonald & MacLeod, 1998) but did not gain traction in the memory literature compared to related phenomena such as the generation effect (see Bertsch, Pesta, Wiscott, & McDaniel, 2007). However, MacLeod, Gopie, Hourihan, Neary, and Ozubko (2010) have cast a spotlight on this neglected mnemonic, renaming it more generally as the production effect and outlining its boundaries.

MacLeod et al. (2010) demonstrated that the production effect is not limited to speech. Even nonvocal responses such as mouthing (their Experiment 5) boosted memory as long as the responses were unique and item-specific. In contrast, when participants responded uniformly to each word—by saying "yes" or by pressing the spacebar (their Experiment 4)—the production effect was eliminated. Similarly, Richler, Palmeri, and Gauthier (2013) have recently showed that the effect is attenuated when participants are limited to only two responses—"table" or "chair"—based on the item's category membership. Recent research has widened the boundaries of the phenomenon, finding production effects for spelling, writing, and typing (Forrin, MacLeod, & Ozubko, 2012), as well as singing (Quinlan & Taylor, 2013). Moreover, the benefit of production occurs regardless of the initial encoding to which production is added: Forrin, Jonker, and MacLeod (2014) found the production effect to be as robust for words already deeply encoded (generated or imagined) as for words simply read (see also MacLeod et al., 2010, Experiments 7 and 8). Thus, the production effect appears to be consistent across a variety of manipulations at encoding.

Despite early production experiments not individually showing a significant between-participants effect in recognition memory (e.g., Hopkins & Edwards, 1972; MacLeod et al., 2010; cf. Gathercole & Conway, 1988), a recent meta-analysis (Fawcett, 2013) did reveal a small but reliable overall effect (g = 0.37). Notably, however, the vocal production effect in recognition appears to be larger in a within-participant design (in which a mixed list of aloud and silent words is studied) than in a between-participants design (in which a pure list of either aloud or silent words is studied), particularly for hit rates. For example, Bodner, Taikh, and Fawcett (2014, Experiment 1) directly compared the within- to between-participants production effects and found a within-participant effect in hits that was substantially larger in magnitude than the between-participants effect (20% vs. 4%). Although they did not compare the sizes of these two effect statistically, a reanalysis of their data using Erlebacher's (1977) method for comparing designs revealed a significantly larger within-participant effect, $F(1, 94) = 18.81$, $MSE = 0.02$, $p < .001$, $\eta^2 = 0.06$.

Forrin, Groot, and MacLeod (2016) also found evidence of a larger within-participant production effect for hits across three experiments that employed different variations of a within-versus

Noah D. Forrin and Colin M. MacLeod, Department of Psychology, University of Waterloo.

This research was supported by Natural Sciences and Engineering Research Council of Canada Discovery Grant A7459. We thank Emily Cyr, Tyler Good, and Madison Stange for their assistance in collecting the data.

Correspondence concerning this article should be addressed to Noah D. Forrin, Department of Psychology, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada N2L 3G1. E-mail: nforrin@uwaterloo.ca

between-participants design.[1] Standard within-versus between-participants designs cannot suitably compare the within-versus between-participants production effects for false alarms (FAs) because recognition tests cannot dissociate between aloud and silent FAs following a mixed study list. Forrin and colleagues, however, were able to obtain separate mixed-list FA rates using variations of a typical within-participant design (see their Experiments 2 and 3). These designs yielded significantly larger within-versus between-participants production effect in terms of FAs, and also d' (a measure of memory discrimination which takes into account both hits and FAs).

This larger within-participant production effect in recognition has been shown to reflect both a cost to reading silently and a benefit to reading aloud in a mixed study list. In terms of hits, significant costs and benefits were obtained both in Bodner et al. (2014; as analysed in Forrin et al., 2016) and across three experiments in Forrin et al. (2016). In terms of d', Bodner and colleagues only found evidence of significant mixed-list costs, both in an experiment and in a meta-analysis. Forrin and colleagues, on the other hand, primarily found evidence of benefits (accompanied by statistically weaker costs) when they used designs that dissociated mixed-list false alarm rates (see their Experiments 2 and 3).[2] Overall, then, the available evidence suggests that the vocal production effect in recognition is significantly larger in a within-participant design than in a between-participants design, and that this larger effect is at least partially attributable to a memorial benefit of reading aloud in a mixed list.

Inspired by Hunt's (2006, 2013) distinctiveness research, MacLeod and colleagues (2010) have argued that speech confers its benefit via distinctive encoding in within-participant designs. More specifically, the processes involved in speech—articulation and audition—stand out in a mixed study list relative to silently read items, and therefore recruit distinctive processing (Conway & Gathercole, 1987, referred to this as *relational distinctiveness*). In comparison, in a between-participants design, aloud items do no benefit from distinctive encoding because speech no longer stands out as a distinct processing dimension, given that all items are spoken or no items are spoken. Jamieson, Mewhort, and Hockley (2016) present a computational model that successfully implements this explanation. In sum, it has been argued that distinct, item-specific processing is a necessary component of the production effect (MacLeod et al., 2010; Ozubko & MacLeod, 2010; Ozubko, Major, & MacLeod, 2014). Production may be particularly memory-enhancing in a within-participant design because speech stands out as relationally distinct compared to silent reading (Conway & Gathercole, 1987; Forrin et al., 2012; Forrin et al., 2016).

Notably, previous production studies focused on manipulations at the time of encoding. Little is known about how the character of the test phase may influence the production effect. MacLeod et al. (2010) argued that the distinctively encoded speech information in a mixed-list production experiment is retained as part of the processing record (cf. the proceduralist framework of Kolers, 1973; Kolers & Roediger, 1984), and that this information can be retrieved strategically at test. The episodic recollection of speech information ("I remember saying it") can help individuals recognize studied words (a type of distinctiveness heuristic; cf. Dodson & Schacter, 2001).

The majority of production experiments have used recognition tests. Inherent to a recognition test, participants read each word on the test silently—whether the word is a target or a lure—immediately prior to making each old/new recognition judgment. MacLeod et al. (2010; Ozubko & MacLeod, 2010) noted that distinctive speech information should retain its distinctiveness against a backdrop of lures. That is, a record of speech constitutes evidence that a word was studied, given that none of the lures were said aloud. Moreover, in line with Dodson and Schacter's (2001) distinctiveness heuristic, participants are able to reject lures on the ground that they lack distinct speech information.

Potentially, then, were the lures to be read aloud, the distinctiveness of speech would be undermined, and consequently the production effect might be attenuated. Ozubko and MacLeod (2010) found evidence consistent with this assumption using a list-discrimination test. They asked participants to study two lists of words—a mixed list and a pure list (either all aloud or all silent). At test, participants were shown studied words and had to indicate the list to which each word belonged. When the pure list was studied silently, there was a production effect for the mixed list: Participants were more accurate at identifying the correct list for aloud words than for silent words. When the pure list was studied aloud, however, the production effect was eliminated for the mixed list, presumably because speech information could not be used to uniquely identify list membership, because words were read aloud in both lists.[3]

Although this previous research suggests that reading lures aloud can interfere with the production effect, this has only been demonstrated in the context of a list-discrimination test. In fact, when Bodner and Taikh (2012; Experiment 4) substituted a standard forced-choice recognition test for the list-discrimination test, the production effect was equally large regardless of whether the pure study list had been read aloud or silently. Moreover, both the Ozubko and MacLeod (2010) and the Bodner and Taikh experiments had the distracting aloud items in the study phase. It is not yet known whether saying the lures aloud at test would diminish the production effect: The present article addresses this gap in the production literature.

Our primary goal in here was to examine the generality of the production effect at test. MacLeod et al. (2010) contended that memory for words that are studied aloud benefits not only from distinctive processing at study (relative to silent study items), but also at test (relative to silent distractors). In the present article, we tested the second part of MacLeod et al.'s claim—that produced

[1] In Experiments 1 and 2 of this research, the larger within-subject effect was found using a blocked design in which participants studied and were tested on three different lists (pure-aloud, pure-silent, and mixed). When only participants' first block data were analyzed, a comparable pattern of data was obtained. Although we did not compare these first-block production effects using Erlebacher's (1977) method in that article, a subsequent analysis demonstrated that the within-subject effect was significantly larger in each of these experiments ($ps < 0.01$).

[2] Conversely, when recall performance is tested, there has been no evidence of a between-participants production effect, and the larger mixed-list effect appears to be driven primarily by costs rather than benefits (see Forrin et al., 2016; Jones & Pyc, 2014; Jonker, Levene, & MacLeod, 2014; Lambert et al., 2016).

[3] Using this design, Bodner & Taikh (2012) found a reverse production effect when the pure list was all aloud.

words benefit from standing out against a "background" of silent distractors.

Across three experiments, we investigated whether auditory processing at test would mitigate the size of the production effect in recognition. In Experiment 1, the auditory processing involved participants reading each visually presented word aloud immediately prior to their recognition decision. In Experiments 2 and 3, words were presented auditorily at test. In all three experiments, we did not expect that the production effect would be diminished by auditory processing at test. Participants would encode a distinct record of speech for words that they read aloud during study and would be able to distinguish this speech record from the auditory processing performed at test.

In Experiments 2 and 3, we also examined whether auditory study presentation would diminish the production effect. In line with the relational distinctiveness account (Conway & Gathercole, 1987; MacLeod et al., 2010), we expected that the production effect would endure when study presentation was auditory (see Mama & Icht, 2016a, 2016b). Speech involves motor processing (articulation) that stands out as distinct relative even to an auditory baseline (MacLeod, 2011). Nevertheless, speech involves fewer relationally distinct processes when study presentation is auditory (articulation) versus visual (articulation and audition), which may reduce the production effect for auditory study presentation.

Experiment 1: Visual Presentation at Study

The purpose of Experiment 1 was to test the prediction that there would still be a production effect even when every word was read aloud at test prior to its recognition judgment. Essentially, this would constitute producing all words at test, only half of which had been produced during study. Would producing words at test undermine the value of study-based production? It could do so if the benefit of production is due simply to having some kind of auditory encoding in addition to the visually based encoding. However, we predicted that the production effect would persist because the distinctiveness of producing words at study would remain intact and accessible—and we hypothesize it that this distinctiveness underlies the production effect. Consequently, we also expected that the size of the production effect would not interact with whether words were read aloud versus silently at test. Therefore, to permit all of the comparisons to occur within a single study, we also included a condition in which all words were read silently at test.

Method

Participants. Forty-four undergraduate students at the University of Waterloo participated in the experiment and were reimbursed with bonus course credit. Twenty-two students were randomly assigned to each condition (silent test or aloud test).

Apparatus. A PC-compatible computer with a 17-inch color monitor was used for testing. The controlling program was written in E-prime 1.2.

Stimuli. The item pool consisted of the 120 words in the Appendix of MacDonald and MacLeod (1998). The words were nouns that were five to 10 letters long (and one to four syllables long). They had frequencies of greater than 30 per million (Thorndike & Lorge, 1944). Words were presented in 12-point Courier New bold font, in lowercase against a black background.

Procedure. For the study phase, 80 words were randomly selected from the 120-item pool. Of these 80, 40 were presented in blue print and 40 in white print. The words were randomly intermixed and presented individually in the centre of the screen for 3,000 ms, with a 500-ms blank between successive words. Participants were instructed to read the blue words out loud, and to read the white words silently to themselves without moving their lips or saying anything out loud.

A recognition test immediately followed the study phase and consisted of all 120 words from the pool: 80 studied (40 aloud, 40 silent) and 40 unstudied. Words were presented individually in the center of the screen and were shown in yellow font to avoid color overlap between study and test. Participants responded to each item at their own pace by pressing either the 'm' key to classify the word as old (studied either aloud or silently) or the 'c' key to classify the word as new. These key press instructions appeared on the screen for the duration of the recognition test. There was a 500-ms blank screen between successive test items.

The silent test condition duplicated the design of previous production experiments (e.g., MacLeod et al., 2010, Experiment 1). The aloud test condition incorporated one crucial difference: Participants read each test word aloud at test immediately prior to making their recognition judgment. Each word was initially presented by itself without the accompanying recognition instruction labels. After 1,250 ms, the instruction labels appeared. Participants were instructed not to respond with a keypress until they had said the word and the instruction labels had appeared, at which point they could make their recognition judgment at their own pace; the test word stayed on the screen until they responded. Through pilot testing, we determined that 1,250 ms gave participants ample time to say a word aloud before the recognition instruction labels appeared. As with the silent test condition, participants' recognition judgments were self-paced. Indeed, apart from the instruction to read each word aloud first, the procedure was identical for the two test conditions. An experimenter (Noah D. Forrin) was in the room with each participant. In the aloud test condition, all participants complied with the instruction to read each word prior to making their recognition judgment.[4]

Results and Discussion

Table 1 displays the mean hit rates and false alarm rates for each cell of the experiment. We first note that the 4% difference in false alarms shown in Table 1 for aloud versus silent tests was not significant, $t(42) = 1.28$, $p = .21$. Hit rates were analysed using a 2×2 repeated measures analysis of variance (ANOVA), with study condition (aloud vs. silent) as the within-participant factor and test condition (aloud vs. silent) as the between-participants factor.[5]

Unsurprisingly, given the robust nature of the production effect, there was a main effect of whether words had been produced at study, as measured by hit rates on the recognition test. Words read

[4] It is possible that subjects made their recognition judgments covertly prior to speaking a test word aloud and pressing the decision key, but given the task instructions this seems unlikely.

[5] We do not analyze d' here because, as we have argued previously (see Forrin et al., 2016; cf. Bodner et al., 2014), a suitable measure of d' cannot be obtained in a standard mixed-list design, which we used in this experiment.

Table 1
Experiment 1 (Visual Presentation at Study): Mean Proportions of Hits for Produced (Read Aloud) and Unproduced (Read Silently) Words That Were Read Aloud Versus Read Silently at Test

Test	Read aloud	Read silently	False alarm
Aloud test	.85 (.02)	.63 (.04)	.13 (.02)
Silent test	.89 (.02)	.67 (.03)	.17 (.03)

Note. Standard errors are shown in parentheses. The rightmost column shows the proportion of false alarms made to new items.

aloud were better remembered than those read silently, $F(1, 42) = 94.33$, $MSE = 0.01$, $p < .001$, $\eta^2 = 0.69$. The main effect of test condition, however, was nonsignificant, $F(1, 42) = 1.80$, $MSE = 0.02$, $p = .19$, $\eta^2 = 0.04$, indicating that overall memory performance was unaffected by whether the test items were read aloud (as opposed to silently) at test.

Most important was the nonsignificant interaction between study condition and test condition, $F < 1$. There was an equally robust production effect regardless of whether each word was processed silently at test, $t(21) = 7.94$, $p < .001$, or was read aloud at test, $t(21) = 6.14$, $p < .001$. Indeed, the two production effects were numerically equivalent (21.93%). We further analyzed this null interaction using Wagenmakers' (2007) Bayesian approximation procedure. Posterior odds were calculated from the sum of squares from ANOVA (see Masson, 2011, for a tutorial) and were converted to p_{BIC}, which quantifies the support for the null relative to the alternative hypothesis on a scale from 0 to 1 (with 1 indicating full support for the null). The analysis yielded "positive" evidence in favor of a null interaction, $p_{\mathrm{BIC}} = 0.87$, according to Raftery's (1995) labelling system.

The results were straightforward: The test format, whether it required production versus not, did not matter. All that mattered was whether a word had been read aloud versus silently during initial study. The production effect was equivalent whether all test words were first spoken aloud or not. The simple view that the production effect is a boost due to the presence at test of an auditory code is incompatible with this result. The view that distinctiveness at study underlies the production benefit is, however, entirely consistent with our results.

Experiment 2: Auditory Presentation at Study

In Experiment 1, the production effect was equivalently large regardless of whether words were read aloud on the recognition test. Thus, speech at test did not undermine the production effect. However, it is still possible that presenting words auditorily at study could attenuate the production effect by rendering words that are read aloud less distinctive against an auditory baseline. We tested this possibility in Experiment 2 by presenting all words auditorily at study.

Previous production experiments have typically presented words visually during study, with the exception of Mama and Icht (2016a, 2016b), who found a production effect in recall with auditory presentation at study. A second alteration from Experiment 1, and from prior studies, was that participants heard recordings of each word spoken at test—half of the targets and half of the lures in each voice—without any accompanying visual presentation. Because the distinctiveness account emphasizes what the subject does during the study phase, we did not expect this change in test presentation format to influence recognition.

Thus, our research extended Mama and Icht (2016a) by examining whether their finding of a production effect following auditory study would (a) replicate when a recognition test was used instead of recall and (b) differ in magnitude depending on whether words were presented visually or auditorily at test. We expected to find a production effect for words studied auditorily because production during study still involves distinct motor processing that would stand out relative to baseline, whether that baseline is visual or auditory. Moreover, the active nature of self-initiated speech may make spoken words particularly memorable relative to the passive baselines of silent reading and listening (MacLeod, 2011; Mama & Icht, 2016a). We expected that this record of a distinct, active encoding would aid memory equivalently at test, regardless of whether words were presented visually or auditorily. To make a direct comparison in a single experiment, we included both visual and auditory test conditions in Experiment 2, as was done in Experiment 1.

Method

Participants. Fifty-five undergraduate students at the University of Waterloo participated and were reimbursed with bonus course credit. Participants were randomly assigned to the visual test condition ($N = 30$) or to the auditory test condition ($N = 25$).

Apparatus. The equipment was the same as in Experiment 1. Auditory words were presented via the computer speakers and did not appear on the monitor.

Stimuli. Because all study words were presented auditorily, homophones had to be excluded from the study list. The item pool therefore consisted of 120 words, 90 from Masson and MacLeod (1992) and an additional 30 from the MRC Psycholinguistic Database (http://www.psy.uwa.edu.au/mrcdatabase/uwa_mrc.htm). Audio files of these 120 words were created by two research assistants, one male and one female, who recorded themselves saying each word; Audacity 2.0 was used to save each word in .wav format with a bit rate of 1,411 kbps.

Procedure. At study, words were presented auditorily. Sixty words were randomly selected from the item pool. Words played one at a time in random order, 30 spoken by a male and 30 spoken by a female. Words were played every 3,500 ms. The screen was blank during the entire study phase. Participants were instructed to say aloud each word that they heard in the male's voice and to listen to each word that they heard in the female's voice, or the reverse (i.e., these instructions were counterbalanced). Because both the main effect for voice and the Voice × Production Effect interaction were nonsignificant, we do not consider the voice variable further.

In the visual test condition, the recognition test was entirely visual. It consisted of all 60 studied words as well as 60 lure words. As was the case in Experiment 1, participants were instructed to press 'm' if they had studied the word—either by hearing it and saying it or simply by hearing it—or to press 'c' if the word was new (these instructions remained on the screen throughout the test). Test words were printed in yellow and a 500-ms blank interval occurred between a response and the next test word.

In the auditory test condition, recordings of the 120 test words were played individually. Participants responded to each by pressing 'm' (studied) or 'c' (unstudied). The 60 studied words comprised the same 60 recordings (30 male, 30 female) that were played at study. Participants were pressed 'm' regardless of how they studied the word (i.e., whether they had heard it and said it or simply had heard it). Of 60 new words, 30 were in the male voice and 30 in the female voice.[6]

Results and Discussion

In the visual test condition, we excluded three participants' data, leaving 27 participants in this condition. Two of these three participants had false alarm rates exceeded their hit rates, and one participant had an extremely high false alarm rate (0.70) that was over 3 standard deviations above the mean false alarm rate (M = 0.20, SE = 0.03). Including these participants' data did not impact the statistical significance of any of the results. We did not exclude any participants' data from the auditory test condition.

Means for hits and false alarms are shown in Table 2. The 3% difference in false alarms for auditory versus visual tests was not significant, $t(50) = 0.82$, $p = .42$. Mean hit rates were analyzed using a 2 × 2 repeated measures ANOVA, with study condition (speak vs. listen) as the within-participant factor and test condition (visual vs. auditory) as the between-participants factor.

The results were entirely consistent with Experiment 1. There was the anticipated production main effect: The hit rate was greater for words that were read aloud at study than for words that were read silently, $F(1, 50) = 52.51$, $MSE = 0.01$, $p < .001$, $\eta^2 = 0.51$. There was not, however, a main effect of test condition. Words that were presented auditorily at test were just as likely to be recognized as those that were presented visually, $F(1, 50) = 2.15$, $MSE = 0.02$, $p = .15$, $\eta^2 = 0.04$. Of primary interest, the interaction between production and test condition was once again nonsignificant, $F < 1$. The production effect was comparable whether the test phase was visual, $t(26) = 4.99$, $p < .001$, or auditory, $t(24) = 5.28$, $p < .001$. Wagenmakers's (2007) Bayesian approximation procedure revealed "positive" evidence in favor of this null interaction.

Combined Experiment 1 and 2 Analysis

We combined the data from both experiments to increase our statistical power in testing whether auditory processing at test influenced the size of the production effect in terms of hit rates. Study condition (production vs. no production) was a within-participant factor and test condition (auditory processing vs. no auditory processing) was a between-participants factor. Of main interest, the interaction was nonsignificant, $F < 1$, signifying that the size of the production effect was not influenced by the presence of auditory information at test.

Table 2
Experiment 1 (Auditory Presentation at Study): Mean Proportions of Hits for Produced (Read Aloud) and Unproduced (Heard) Words That Were Presented Either Visually or Auditorily at Test

Test	Read aloud	Heard	False alarm
Auditory test	.81 (.02)	.64 (.03)	.17 (.03)
Visual test	.84 (.02)	.69 (.03)	.20 (.03)

Note. Standard errors are shown in parentheses. The rightmost column shows the proportion of false alarms made to new items.

Experiment 3: Visual Versus Auditory Presentation at Study and Test

In Experiments 1 and 2, the size of the production effect in recognition was not influenced by the presence of auditory information at test. Further, the production effect remained statistically significant when words were presented auditorily at study in Experiment 2 (see also Mama & Icht, 2016a, 2016b). However, because we did not include a visual study condition in Experiment 2, it is still possible that auditory study presentation could diminish the size of the production effect relative to visual study presentation.

We investigated this possibility in Experiment 3 by using an experimental design in which both study and test presentation modalities (visual vs. auditory) were manipulated between-participants. This design provided a more conclusive test of whether the production effect was smaller when words are presented auditorily at study, insofar as the same pool of study words and the same old/new ratio on the recognition test was used for all participants. This experiment also allowed for a final test of whether the production effect was influenced by auditory presentation at test.

An overall smaller production effect for auditory versus visual study presentation would be consistent with MacLeod et al.'s (2010) distinctiveness account, which asserts that processing dimensions that stand out as relationally distinct in a mixed study list can be retrieved on a recognition test to confirm prior study (see also Conway & Gathercole, 1987). When words are presented visually, speech confers two distinct processing dimensions relative to silent reading (i.e., articulation and audition). Comparatively, when words are presented auditorily, speech only confers one distinct processing dimension (i.e., articulation). In short, speech involves a greater number of distinct processing dimensions when the study presentation is visual (vs. auditory), which, according to the distinctiveness account, ought to bestow the larger boost to recognition. Prior research has shown that the number of relationally distinct processing dimensions is positively related to recognition performance (Forrin et al., 2012; Quinlan & Taylor, 2013).

The "translation hypothesis" also predicts a larger production effect when words are presented visually versus auditorily (Conway & Gathercole, 1990). This hypothesis presumes that the act of "translating" information from one modality (e.g., auditory) to another modality (e.g., visual), at encoding, facilitates memory

[6] Participants were equally likely to FA to words in the male or female voice, $t(24) = 1.24$, $p = 0.23$. Moreover, FA rates were not influenced by whether the new words were presented in what had been the "speak" condition voice vs. the "listen" condition voice at study, $t(24) = 0.71$, $p = 0.50$. These results suggest that this auditory information (male voice vs. female voice) was not used by subjects to help guide their recognition at test. Rather, we contend that subjects drew upon recollections of their own distinct speech to discriminate between old and new words, consistent with the distinctiveness account of the production effect.

relative to cases in which a modality translation does not occur. That is, producing a word in a different modality than its presentation modality should have an added memorial benefit above and beyond the other mechanisms that enhance memory for produced words (e.g., distinctiveness). Several experiments have supported the translation hypothesis: Writing words that are auditorily presented (an auditory to visual translation) is more memorable than writing words that are visually presented (Conway & Gathercole, 1990; Mama & Icht, 2016b; Rackie et al., 2015), and writing words that are auditorily presented is more memorable than speaking words that are auditorily presented (Mama & Icht, 2016a). These results support half of the translation hypothesis—that an auditory to visual translation enhances memory. The other half of the translation hypothesis—that a visual to auditory translation also enhances memory—has yet to be investigated. Experiment 3 provided an initial test of this heretofore unexplored dimension of the translation hypothesis by comparing memory for spoken words that are presented visually versus auditorily at study.

Of primary interest in Experiment 3 was whether visual versus auditory study presentation affected the size of the production effect. Both the distinctiveness account (MacLeod et al., 2010) and the translation hypothesis (Conway & Gathercole, 1990) predict that the production effect will be larger when words are presented visually versus auditorily at study. We again hypothesized that the test modality would not have a significant influence on the size of the production effect.

Method

Participants. One hundred and sixty undergraduate students at the University of Waterloo participated and were reimbursed with bonus course credit. Forty participants were randomly assigned to each of the four conditions: visual study-visual test, visual study-auditory test, auditory study-visual test, and auditory study-auditory test.

Apparatus. This was identical to Experiment 2.

Stimuli. The item pool consisted of 120 words. Ninety-five were from the appendices of Masson and MacLeod (1992), and an additional 25 were taken from the MRC Psycholinguistic Database. Audio files of these 120 words were created by recording a female research assistant saying each word, following the same protocol as in Experiment 2.

In this experiment, small black and white icons were used in both the visual and the auditory study conditions to indicate to participants how each word should be studied. In the visual study condition, words were presented in the centre of the screen in 18-point bold Courier New font. A "speech" icon (depicting a talking head) indicated that the word should be read aloud and an "eye" icon indicated that the word should be read silently. An icon appeared above each study word at the same time as the word was presented.

In the auditory study condition, a "speech" icon again was used to denote that the word should be read aloud, and an "ear" icon indicated that participants should only listen to the word. These icons appeared in the same above-center location on the screen as in the visual presentation conditions and at the same time as the recording of each word was played.

Procedure. The auditory study conditions were identical to Experiment 2 except icons above each word indicated the study condition (speak vs. listen). Words were played every 3,500 ms while the screen remained blank (white). In the visual study conditions, icons were also used to indicate study condition (speak vs. read silently). Words were presented in black font on a white screen for 3000 ms with a 500-ms ISI.[7] The visual and auditory test conditions were identical to Experiment 2.

Results and Discussion

One participant in the visual study-auditory test condition was excluded from the analysis, leaving 159 participants. This individual's hit rate in the aloud condition (0.20) was 3.68 standard deviations below the mean for that condition. Including this individual's data did not impact the statistical significance of any of the results.

Means for hits and false alarms are shown in Table 3. False alarm rates did not differ significantly based on the study modality (visual vs. auditory), $F(1, 155) = 1.21$, $p = .27$, $\eta^2 = 0.01$, or test modality (visual vs. auditory), $F(1, 155) = 1.81$, $p = .18$, $\eta^2 = 0.01$, and the interaction was also nonsignificant ($F < 1$). Nonsignificant main effects were also observed in Experiments 1 and 2—false alarms tended to be slightly lower when study presentation was visual and when test presentation was auditory. It appears that study modality and test modality have, at most, very modest effects on false alarms.

We analyzed hit rates using a $2 \times 2 \times 2$ mixed-model ANOVA, in which study modality (visual vs. auditory) and test modality (visual vs. auditory) were between-participants factors and study condition (production vs. no production) was a within-participant factor. This ANOVA yielded the expected main effect of study condition, $F(1, 155) = 327.63$, $p < .001$, $\eta^2 = 0.68$, with production enhancing memory relative to silent reading/listening. There was also a significant main effect of study modality, $F(1, 155) = 10.95$, $p = .001$, $\eta^2 = 0.07$, which indicated that words that were presented auditorily at study were better recognised than words presented visually at study. This was true for words that were read aloud, $t(157) = 2.08$, $p = .04$, and for words that were not produced (i.e., read silently or listened to), $t(157) = 3.49$, $p = .001$.

The main effect of test modality was nonsignificant, $F(1, 155) = 1.36$, $p = .25$, $\eta^2 = 0.01$. And, notably, the interaction between test modality and study condition was also nonsignificant, $F < 1$, replicating the finding from Experiments 1 and 2 that test modality does not influence the size of the production effect. The Bayesian approximation procedure yielded "positive" evidence in support of the null, $p_{BIC} = 0.91$.

Of main interest, the ANOVA revealed a significant Study Modality $\times$ Study Condition interaction, $F(1, 155) = 4.55$, $p = .03$, $\eta^2 = 0.03$. There were robust production effects in all four conditions of this experiment (all ps < 0.001), but controlling for

[7] In previous research we have also presented words visually for 3,000 ms with a 500-ms ISI (e.g., Forrin et al., 2016), so for consistency we chose to use the same time parameters here in the visual condition. For the auditory version, then, to equate the total time available for studying each word in the two conditions, we presented a word every 3.5 s. Of course, this meant that actual stimulus presentation time varied. With these presentation times, one might expect an overall advantage in the visual encoding condition due to its longer presentation time; the opposite in fact occurred.

Table 3
Experiment 3 (Visual/Auditory Presentation at Study and Test): Mean Proportions of Hits for Produced (Read Aloud) and Unproduced (Read Silently/Heard) Words That Were Presented Either Visually or Auditorily at Study and Either Visually or Auditorily at Test

Study/test	Read aloud	Read silently/ heard	False alarm
Visual study, visual test	.79 (.02)	.57 (.03)	.15 (.02)
Visual test, auditory test	.75 (.02)	.51 (.03)	.13 (.02)
Auditory study, visual test	.80 (.02)	.63 (.03)	.17 (.02)
Auditory study, auditory test	.81 (.02)	.63 (.02)	.15 (.02)

Note. Standard errors are shown in parentheses. The rightmost column shows the proportion of false alarms made to new items.

test modality, the production effect was significantly larger when words were presented visually ($M = 0.23$, $SE = 0.02$) versus auditorily at study ($M = 0.18$, $SE = 0.02$). The Study Modality × Test Modality interaction, and the three-way interaction, were nonsignificant ($ps > 0.10$).[8]

In summary, Experiment 3 replicated the result from Experiments 1 and 2 that the size of the production effect is unaffected by the presence of auditory information at test. Experiment 3 also generated reliable evidence that, relative to visual presentation at study, the production effect is reduced by auditory presentation at study. This result is in keeping with both the relational distinctiveness account: Reading aloud invokes two distinct processes when words are presented visually but only one distinct process when they are presented auditorily. It is also consistent with the translation account: Saying visually presented words entails a modality translation that saying auditorily presented words aloud does not. These two accounts could not be parsed apart in Experiment 3 given that, relative to the auditory study conditions, the visual study conditions had more distinct processes and involved a translation. Either (or both) of these factors could have enhanced the production effect. It should be noted, however, that the effect of study presentation modality was modest ($\eta^2 = 0.03$). The more important messages from this experiment are that the production effect is robust regardless of study modality, test modality, and how study and test modalities are combined.

General Discussion

Although many studies have demonstrated a production effect in recognition, all of them have presented items visually at test. Our primary goal, therefore, was to test the boundaries of the production effect at test, analogous to how previous research has examined the boundaries of the effect at encoding (Forrin et al., 2012). But our goal was not simply to generalize the phenomenon. It is conceivable that the benefit of production rests on the unique auditory encoding that is created for each produced word rather than, as we have previously argued (e.g., Forrin et al., 2016), on the distinctiveness of production during study. By presenting words auditorily at study and test, we were able to refute this hypothesis.

In Experiment 1, using standard visual presentation during study, we showed that the production effect was equivalent whether the test phase was only visual or involved participants reading each word aloud—a process that had been distinctive at encoding—prior to making their recognition decision. In Experiments 2 and 3, we again showed that the production effect was not diminished by auditory processing at test, even when words had been presented auditorily at study.

Clearly, the production effect is more than just an auditory advantage: Speech at the time of study stands out as distinct above and beyond auditory information (MacLeod, 2011), and this distinctiveness is preserved at test even when auditory information (including speech) is prevalent. Indeed, in Experiments 2 and 3, we found that auditory study does not eliminate the production effect in recognition, consistent with a parallel result obtained by Mama and Icht (2016a) in recall.

The production effect occurs regardless of whether words are studied visually or auditorily; notwithstanding, Experiment 3 yielded evidence that the effect may be slightly larger when words are presented visually during study. This larger visual production effect is consistent with both the relational distinctiveness account (Conway & Gathercole, 1987; MacLeod et al., 2010) and the translation hypothesis (Conway & Gathercole, 1990). It may be challenging, however, to tease apart the influence of these two mechanisms on the production effect. For example, to examine whether the process of "modality translation" enhances memory for produced words above and beyond the benefit that may arise from distinctive processing, one would have to hold constant the number of relationally distinct processes involved in the production while manipulating whether the production involved a modality translation. Reading aloud visually presented words involves two distinct processes (articulation and audition) and one modality translation (visual to auditory), whereas reading aloud auditorily presented words involves one distinct process (articulation) and no modality translation. It is not clear, therefore, whether the larger production effect that arises from visual study occurs because of the additional distinctive process or because of the modality translation (or both).

Importantly, the present research provides evidence that the distinctiveness of speech during study—which arises regardless of whether the presentation mode is visual or auditory—is not undermined by an auditory test. Moreover, these results suggest that the speech distinctiveness heuristic (Dodson & Schacter, 2001) is quite resilient insofar as its effectiveness does not appear to be diminished by the presence of aloud information at test—even when each word at test is spoken (as in our Experiment 1). Participants seem quite capable of distinguishing between words that they read aloud at study and those that they read aloud at test: Reading each word aloud at test did not reduce the size of the production effect on hits, nor was there an increase in false alarms.

Although consistent with a distinctiveness account, the results of Experiment 1 appear to be somewhat at odds with Morris, Bransford, and Franks (1977) transfer-appropriate processing frame-

[8] Although participants' memory for silently read words appeared to be particularly low when an auditory test followed visual study, the mean hit rate in the visual-auditory condition was not significantly lower than that of the visual-visual condition, $t(77) = 1.71$, $p = 0.09$. The overall pattern of results suggests that recognition is lower overall for visual study than for auditory study. There was no reliable evidence that, controlling for the influence of study modality, recognition was lower for an auditory test than for a visual test.

work, which postulates that memory is enhanced when the processes enacted at test match those performed at encoding. Contrary to this framework, we did not find that requiring participants to read words aloud at test improved recognition for words that had been read aloud at study. It is possible that a ceiling effect occurred: Hit rates for words studied aloud in the silent recognition condition were already quite high ($M = 0.89$), leaving little room for improvement in the aloud recognition condition. It is also possible that transfer-appropriate processing is less influential when only some words were studied using the same process that is engaged at test, as was the case for the mixed study list used in our Experiment 1. Future research could address these issues by using a longer study list or adding a delay between the study and test phases, and by using a between-participants Study Type (aloud vs. silent) × Test Type (aloud vs. silent) design.

Prior to concluding, we note that Mama and Icht (2016b) report results that partly run counter to ours. Mama and Icht had participants either read words aloud or write words in a mixed study list. In contrast to our recognition tests, they tested participants' recall—using either a written test or an oral test. In comparing their results to ours, we will focus only on the common encoding condition: reading aloud. In their Experiment 1 (in which words were presented visually at study), they found that the recall of words that were studied aloud was equivalent regardless of whether there was auditory information at test (i.e., an oral vs. a written recall test). Thus, their results were consistent with ours when study was visual.

However, where results appear to diverge is in the case of auditory study presentation. In their Experiment 2, recall was poorer for auditorily presented words that were studied aloud on a written recall test, relative to an oral recall test. Conversely, we did not find evidence that words that were studied auditorily were more poorly recognised when the test involved auditory versus visual information. There are many differences between respective experiments (e.g., word lists, design, procedure), making it difficult to surmise what underlies the diverging results. We do agree, however, with Mama and Icht's position that the most influential difference is likely to be the type of test—recall versus recognition. Specifically, they attribute their results to differing levels of retrieval interference that build up when the modality of the recall test overlaps with the mode in which certain words were studied. As they point out, retrieval interference would not influence recognition test performance (Dennis & Humphreys, 2001), in keeping with our null effect of test modality.

Harking back to Conway and Gathercole (1987), and as recently emphasized by MacLeod et al. (2010), the distinctiveness account has thus far provided the best explanation of the production effect (see, e.g., Ozubko, Major, & MacLeod, 2014). Defined by Hunt as "the processing of difference in the context of similarity" (e.g., Hunt, 2013, p. 10), distinctive processing has proven to be a powerful way to improve memory (see, e.g., Hunt, 2006, 2013). When some items are singled out to be uniquely acted on and others are not, the advantage of that action—that production—is substantial. Remembering that a word was produced during study constitutes evidence that it was studied. As part of the memory record, production is diagnostic information based on what one did and not simply on what one passively experienced. By extending both study and test to auditory presentation—and showing that the production effect is robust in both presentation modes—the present study generalizes the phenomenon and also demonstrates that the enhancement does not rest on the mere presence of an auditory code. Instead, when retrieved, the distinctiveness of our own actions forms a compelling basis for remembering.

Résumé

Trois expériences ont été réalisées afin de déterminer si l'information auditive à l'essai nuirait au caractère distinctif relationnel de la production vocale à l'étude, ayant pour effet de réduire l'effet de production. Dans l'expérience 1, avec présentation visuelle au cours de l'étude, l'effet de production a eu la même importance peu importe si les participants lisaient chaque mot-test à voix haute ou non avant de prendre leur décision de reconnaissance. Dans l'expérience 2, où l'on incorporait la présentation auditive au courant de l'étude, l'effet de production n'a pas varié, peu importe que les mots tests à reconnaître aient été présentés visuellement ou auditivement. Dans l'expérience 3, les auteurs ont comparé les données selon que la présentation était visuelle ou auditive ou un mélange des deux durant l'étude ainsi que lors des tests. De nouveau, la modalité de présentation à l'essai n'a pas eu d'incidence sur l'ampleur de l'effet de production, bien que l'effet ait été significativement plus faible quand les mots ont été présentés auditivement au coursant de l'étude. Ces expériences démontrent que la production au moment de l'étude se démarque comme étant distincte bien au-delà de l'information auditive. De plus, cette information distincte lue à voix haute ne doit pas ressortir sur un fond de mots lus silencieusement non-étudiés suite à un test de reconnaissance. En conformité avec le postulat du caractère distinctif, l'encodage via la production améliore la reconnaissance ultérieure de manière constante, indépendamment de la modalité d'étude ou d'essai.

Mots-clés : effet de production, mémoire, reconnaissance, auditif, hypothèse de traduction.

References

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition, 35,* 201–210. http://dx.doi.org/10.3758/BF03193441

Bodner, G. E., & Taikh, A. (2012). Reassessing the basis of the production effect in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38,* 1711–1719. http://dx.doi.org/10.1037/a0028466

Bodner, G. E., Taikh, A., & Fawcett, J. M. (2014). Assessing the costs and benefits of production in recognition. *Psychonomic Bulletin & Review, 21,* 149–154. http://dx.doi.org/10.3758/s13423-013-0485-1

Conway, M., & Gathercole, S. (1987). Modality and long-term memory. *Journal of Memory and Language, 26,* 341–361. http://dx.doi.org/10.1016/0749-596X(87)90118-5

Conway, M. A., & Gathercole, S. E. (1990). Writing and long-term memory: Evidence for a "translation" hypothesis. *The Quarterly Journal of Experimental Psychology, 42,* 513–527. http://dx.doi.org/10.1080/14640749008401235

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review, 108,* 452–478. http://dx.doi.org/10.1037/0033-295X.108.2.452

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. http://dx.doi.org/10.3758/BF03196152

Erlebacher, A. (1977). Design and analysis of experiments contrasting the within-and between-participants manipulation of the independent variable. *Psychological Bulletin, 84,* 212–219. http://dx.doi.org/10.1037/0033-2909.84.2.212

Fawcett, J. M. (2013). The production effect benefits performance in between-subject designs: A meta-analysis. *Acta Psychologica, 142,* 1–5. http://dx.doi.org/10.1016/j.actpsy.2012.10.001

Forrin, N. D., Groot, B., & MacLeod, C. M. (2016, January 28). The d-prime directive: Assessing costs and benefits in recognition by dissociating mixed-list false alarm rates. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. http://dx.doi.org/10.1037/xlm0000214

Forrin, N. D., Jonker, T. R., & MacLeod, C. M. (2014). Production improves memory equivalently following elaborative vs non-elaborative processing. *Memory, 22,* 470–480. http://dx.doi.org/10.1080/09658211.2013.798417

Forrin, N. D., Macleod, C. M., & Ozubko, J. D. (2012). Widening the boundaries of the production effect. *Memory & Cognition, 40,* 1046–1055. http://dx.doi.org/10.3758/s13421-012-0210-8

Gates, A. I. (1917). Recitation as a factor in memorizing. In R. S. Woodworth (Ed.), *Archives of psychology* (vol. 40, pp. 1–104). New York, NY: The Science Press.

Gathercole, S. E., & Conway, M. A. (1988). Exploring long-term modality effects: Vocalization leads to best retention. *Memory & Cognition, 16,* 110–119. http://dx.doi.org/10.3758/BF03213478

Hopkins, R. H., & Edwards, R. K. (1972). Pronunciation effects in recognition memory. *Journal of Verbal Learning & Verbal Behavior, 11,* 534–537. http://dx.doi.org/10.1016/S0022-5371(72)80036-7

Hunt, R. R. (2006). The concept of distinctiveness in memory research. In R. R. Hunt & J. B. Worthen (Eds.), *Distinctiveness and memory* (pp. 2–25). New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780195169669.003.0001

Hunt, R. R. (2013). Precision in memory through distinctive processing. *Current Directions in Psychological Science, 22,* 10–15. http://dx.doi.org/10.1177/0963721412463228

Jamieson, R. K., Mewhort, D. J. K., & Hockley, W. E. (2016). (in press). A computational account of the production effect: Still playing twenty questions. *Canadian Journal of Experimental Psychology*.

Jones, A. C., & Pyc, M. A. (2014). The production effect: Costs and benefits in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40,* 300–305. http://dx.doi.org/10.1037/a0033337

Jonker, T. R., Levene, M., & Macleod, C. M. (2014). Testing the item-order account of design effects using the production effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40,* 441–448. http://dx.doi.org/10.1037/a0034977

Kolers, P. A. (1973). Remembering operations. *Memory & Cognition, 1,* 347–355. http://dx.doi.org/10.3758/BF03198119

Kolers, P. A., & Roediger, H. L., III. (1984). Procedures of mind. *Journal of Verbal Learning & Verbal Behavior, 23,* 425–449. http://dx.doi.org/10.1016/S0022-5371(84)90282-2

Lambert, A. M., Bodner, G. E., & Taikh, A. (2016). *Evaluating the basis of the production effect in recall*. Manuscript submitted for publication.

MacDonald, P. A., & MacLeod, C. M. (1998). The influence of attention at encoding on direct and indirect remembering. *Acta Psychologica, 98,* 291–310. http://dx.doi.org/10.1016/S0001-6918(97)00047-4

MacLeod, C. M. (2011). I said, you said: The production effect gets personal. *Psychonomic Bulletin & Review, 18,* 1197–1202. http://dx.doi.org/10.3758/s13423-011-0168-8

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, and Cognition, 36,* 671–685. http://dx.doi.org/10.1037/a0018785

Mama, Y., & Icht, M. (2016a). Auditioning the distinctiveness account: Expanding the production effect to the auditory modality reveals the superiority of writing over vocalising. *Memory, 24,* 98–113. http://dx.doi.org/10.1080/09658211.2014.986135

Mama, Y., & Icht, M. (2016b). *The role of retrieval in the production effect: Evidence for costs in free recall*. Manuscript submitted for publication.

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods, 43,* 679–690. http://dx.doi.org/10.3758/s13428-010-0049-5

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. http://dx.doi.org/10.1037/0096-3445.121.2.145

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.

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. http://dx.doi.org/10.1037/a0020604

Ozubko, J. D., Major, J., & MacLeod, C. M. (2014). Remembered study mode: Support for the distinctiveness account of the production effect. *Memory, 22,* 509–524. http://dx.doi.org/10.1080/09658211.2013.800554

Quinlan, C. K., & Taylor, T. L. (2013). Enhancing the production effect in memory. *Memory, 21,* 904–915. http://dx.doi.org/10.1080/09658211.2013.766754

Rackie, J. M., Brandt, K. R., & Eysenck, M. W. (2015). Interaction between mode of learning and subjective experience: Translation effects in long-term memory. *Memory, 23,* 318–328. http://dx.doi.org/10.1080/09658211.2014.886701

Raftery, A. E. (1995). Bayesian model selection in social research. In P. V. Marsden (Ed.), *Sociological methodology* (pp. 111–196). Cambridge, MA: Blackwell.

Richler, J. J., Palmeri, T. J., & Gauthier, I. (2013). How does using object names influence visual recognition memory? *Journal of Memory and Language, 68,* 10–25. http://dx.doi.org/10.1016/j.jml.2012.09.001

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

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review, 14,* 779–804. http://dx.doi.org/10.3758/BF03194105

Received June 29, 2015
Accepted March 10, 2016 ■