MEMORY
2021, VOL. 29, NO. 2, 168–179
https://doi.org/10.1080/09658211.2020.1868526

Routledge
Taylor & Francis Group

Production between and within: distinctiveness and the relative magnitude of the production effect

Yichu Zhou and Colin M. MacLeod

Department of Psychology, University of Waterloo, Waterloo, Canada

ABSTRACT

The production effect is the memory advantage for items studied aloud over items studied silently. Three experiments examined the influence of (1) the distinctiveness heuristic in a pure-list paradigm and (2) statistical distinctiveness during study. Aloud versus silent processing was manipulated within-subject in a mixed-list procedure and additional pure-list items were alternated with the to-be-remembered words. This arrangement permitted the first examination of the production effect using both within-subject and between-subjects manipulations in the same experiment. The quite large between-subjects production effect observed for the pure-list words is attributed to the distinctiveness of the aloud words being enhanced by the co-occurring within-subject manipulation. In addition, when the pure-list words were all read aloud, they effectively increased the overall proportion of aloud words, thereby decreasing the distinctiveness of the to-be-remembered aloud words in the mixed list. Correspondingly, there was a decrease in the magnitude of the production effect. However, when the pure-list words were all read silently, the magnitude of the production effect was unchanged relative to baseline. These results provide partial support for the influence of statistical distinctiveness on the magnitude of the production effect.

ARTICLE HISTORY
Received 28 August 2020
Accepted 20 December 2020

KEYWORDS
Production effect; distinctiveness account; recognition; memory

The memory benefit of reading words aloud relative to reading words silently has now been demonstrated in numerous studies (see MacLeod & Bodner, 2017, for a brief review). Earlier studies reporting this effect (Conway & Gathercole, 1987; Dodson & Schacter, 2001; Gathercole & Conway, 1988; Hopkins & Edwards, 1972; Kurtz & Hovland, 1953; MacDonald & MacLeod, 1998) had received scant attention until MacLeod et al. (2010) brought it into the spotlight, labelling this memory benefit the *production effect*. Since then, the effect has been demonstrated to be very robust (see the special issue of the *Canadian Journal of Experimental Psychology*; Bodner & MacLeod, 2016) and has been shown to extend beyond speaking words: Writing (Forrin et al., 2012), typing (Forrin et al., 2012; Jamieson & Spear, 2014), and mouthing (MacLeod et al., 2010) all result in a production advantage, although none of these are as great as speaking.[1] In addition, the production effect has been generalised to studying educationally relevant material like text and to longer retention intervals (Ozubko, Hourihan, et al., 2012), as well as to learning new vocabulary in a second language (Icht & Mama, 2019), indicating that production can be effective as a learning strategy beyond the laboratory.

The production effect has been explained primarily as due to the distinctive processing applied to the produced words (Conway & Gathercole, 1987; Forrin et al., 2014; MacLeod et al., 2010; Quinlan & Taylor, 2013). Under the distinctiveness account, it is hypothesised that the aloud words stand out relative to the silent words during study, resulting in an additional dimension of encoding for the produced words. By analogy, the aloud words are seen as figure whereas the silent words are seen as ground. Hunt (2013, p. 10) defines distinctive processing as "the processing of difference in the context of similarity". Thus, the words are all similar in a variety of ways, including being common nouns and being presented in lowercase font at the centre of the screen. But the aloud words differ from the silent ones by virtue of being produced, and that makes their encoding distinctive.

The importance of the distinctive processing during study becomes apparent at the time of test. In addition to trying to remember the word itself, participants can use the strategy – consciously or unconsciously – of retrieving whether a word was produced at study. On a recognition test, then, remembering having spoken a word during the study phase is an additional way to certify that it was indeed studied. Therefore, the distinctiveness dimension provides an additional path for successful retrieval, augmenting memory for the word itself.

CONTACT Yichu Zhou or Colin MacLeod craig.zhou@uwaterloo.ca, cmacleod@uwaterloo.ca Department of Psychology, University of Waterloo, Waterloo, ON, Canada N2L 3G1

Dodson and Schacter (2001) characterised this type of retrieval strategy as a distinctiveness heuristic.

Evidence for the distinctiveness account of the production effect comes from two main types of results. First, the production effect is typically much larger in within-subject mixed list designs, in which a participant studies both aloud and silent words, than in between-subjects pure list designs, in which a participant experiences only one of these conditions (Hopkins & Edwards, 1972; MacLeod et al., 2010; see Fawcett, 2013). This makes sense in that establishing the contrast between aloud and silent items during study is possible only under the within-subject design. Second, Fawcett and Ozubko (2016) used remember-know judgments as well as a signal detection approach to further examine the processes underlying the influence of this design difference. Under a within-subject design, they observed a production effect both for the words that participants "recollected" and for the words that participants indicated were just "familiar" (see also Ozubko, Gopie, et al., 2012). In contrast, under a between-subjects design, they observed a production effect only for familiarity – that is, the between-subjects effect did not involve a recollective-based component.

These findings suggest that experiencing both aloud words and silent words at the time of study plays a critical role in making the aloud words distinctive, thereby producing an enhanced memory benefit via recollection. In addition, studies have demonstrated that participants' source memory for whether a word was produced at study is better for words that were studied aloud than for those that were studied silently (Ozubko, Gopie, et al., 2012) and is independent of manipulating memory strength of the studied items (Ozubko et al., 2014). Older adults, who do not use distinctive information as well as younger adults, also show a smaller production effect (Lin & MacLeod, 2012). These findings support the idea that, in addition to encoding the words themselves, participants also encode whether they studied them aloud; that information can then be retrieved at test to assist remembering.

Our first question pertains to whether, for distinctive processing to be effective, the aloud-silent distinction must be experienced in a mixed-list paradigm, where obvious shifts between the aloud and silent conditions are present such that participants can notice them. Specifically, when studying a pure list of aloud words, would participants' memory for these items improve if, either consciously or unconsciously, they implemented the distinctiveness heuristic for the aloud words just as when studying mixed lists? That is, would they come to use memory for whether items had been produced at study to help their recognition, thereby increasing the magnitude of the production effect in a pure-list between-subjects design?

It has been widely reported that the production effect is considerably smaller using a between-subjects, pure list design than using a within-subject, mixed list design. Indeed, initially (Dodson & Schacter, 2001; Hopkins & Edwards, 1972; MacLeod et al., 2010), the effect was thought to be restricted to within-subject designs, a belief that formed the original foundation for the distinctiveness account. But it has since become clear that the effect is present under both designs, albeit smaller between-subjects (see MacLeod & Bodner, 2017, for a brief review).

To answer our first question, we wanted to design experiments in which we did not explicitly instruct participants that they could use the distinctiveness heuristic to their advantage, since this is not done in a mixed-list paradigm. We accomplished this by inserting a pure-list set of items to be studied either all aloud or all silently, into mixed lists containing both aloud and silent items: One group of participants would study the pure list items all aloud and another group would study the pure list items all silently, and then be tested on all items. We could then compare whether the production effect for the pure-list items was affected given that participants should appreciate the value of the distinctiveness heuristic from experience with the mixed-list manipulation. The final hurdle was to find a way to differentiate the pure-list items from the mixed-list items; otherwise, participants would still be studying mixed lists but simply with the proportion of aloud and silent items altered from the typical equal split in previous production effect studies.

To solve this problem, we gave participants a cover story in which the pure-list items would be presented in a different colour from the mixed-list items, and participants were told that they did not have to remember the pure-list items because they were irrelevant and would not be tested. We predicted, however, that participants would have above-chance memory for these items at least if they were studied aloud, since participants could not avoid processing them. In contrast, the pure-list items could be ignored if they were studied silently, and we were curious to see whether participants would still show above-chance memory for these items in such a case. If participants showed above-chance memory for the pure-list items in both groups, then we would be able to examine the magnitude of the production effect in a between-subjects paradigm where participants should have detected the value of the distinctiveness heuristic because of the co-occurring mixed-list manipulation, to answer our original research question.

Therefore, we designed experiments where presentation of the pure-list items alternated with the mixed-list items, and were presented in a different colour. To ensure that such a design per se does not influence the magnitude of the production effect within-subject, as our predictions rested on having a similar production effect compared to previous studies, we began with a baseline experiment. In Experiment 1, we inserted after every aloud or silent item an irrelevant event that required no processing: a row of Xs in a colour not used for the to-be-remembered items. In Experiments 2 and 3, the row of Xs was replaced with the pure-list items that participants

read either all aloud or all silently. This, then, will be the first time that the within-subject and the between-subjects designs are examined in the same experiment.

This setup also enabled us to examine a second research question. We were interested in examining the influence of relative frequency – or statistical distinctiveness – of the aloud and silent subsets of words at study. Ordinarily, in a mixed list, there are equal numbers of words read aloud and silently, but what happens if this balance is disturbed? There is some limited evidence showing that statistical distinctiveness modulates the size of the production effect. Ozubko and MacLeod (2010) showed that when participants studied one mixed list and one pure list, the production effect for the mixed list was robust for those whose pure list was all silent but was eliminated for those whose pure list was all aloud. Presumably, having an additional pure list of all aloud words reduced the statistical distinctiveness of the aloud words in the mixed list, hence the finding of a reduced production effect. Icht et al. (2014; see also Bodner et al., 2016) directly manipulated the relative frequency of the aloud and silent words in the study list and found a similar pattern of results: The magnitude of the production effect was reduced as the statistical distinctiveness of the aloud words decreased (i.e., as the proportion of aloud words relative to silent words was increased).

As just described, in Experiments 2 and 3, the row of Xs between successive mixed-list words were replaced with the pure-list words in a different colour; the pure-list words were read all aloud by one group and all silently by another group, and participants were told that they did not have to remember these items. Would words that were not relevant and that could be forgotten nevertheless influence the magnitude of the production effect with respect to statistical distinctiveness? Specifically, when the irrelevant words were read aloud, would they lower the statistical distinctiveness of the to-be-remembered aloud words (given our prediction that participants would remember them above chance), thereby differentially affecting the production effect relative to when the irrelevant words were read silently, which should increase the statistical distinctiveness? Or would they – simply because they were defined as irrelevant – not influence the magnitude of the production effect under a statistical distinctiveness hypothesis?

In summary, the goals of the present study were twofold. First, we investigated whether, similar to mixed-list designs, being made aware of the distinctiveness heuristic in a pure-list paradigm – through observing the aloud-silent distinction when pure-list items are inserted into mixed lists – would influence the magnitude of the production effect. Second, our setup allowed us to examine whether changing the relative frequency of the aloud and silent words by the addition of the pure-list words at study would differentially affect the size of the production effect under a statistical distinctiveness account.

Experiment 1

Experiment 1 served as a baseline for the main experiments to follow. Here, each word designated as aloud or silent by its colour (e.g., blue = aloud; white = silent) was followed by a row of Xs in red such that immediate contrasts of aloud and silent items were precluded. No response was required to these red Xs. The goal was simply to prevent the aloud and silent items from being adjacent to each other, and to determine whether these interleaved red events would disrupt the normal production effect by reducing the aloud/silent contrast on a trial-by-trial basis.

Method

Participants

Participants were 24 undergraduate students (9 men, 13 women; age range: 17–25 years, mean age: 20.6 years, SD = 2.9 years; 2 participants declined to provide demographic information) from the University of Waterloo, recruited via the Department of Psychology's research participation system. Ethics approval was obtained from a University of Waterloo Research Ethics Board, and written consent was obtained from all participants (the same applies to Experiments 2 and 3). Participants received either course credit or $5 in exchange for their participation. Based on the existing studies examining the production effect, we believed that this number of participants – frequently used in prior studies – was sufficient to test for a production effect in the current manipulation, and would allow us to compare the magnitude of the production effect obtained in this experiment to previous studies.

Apparatus

The experiment was controlled by a PC-compatible computer running a programme written in E-Prime 3.0 (Psychology Software Tools, Pittsburgh, PA). Study and test trials were presented on a LCD monitor, with responses collected via a standard QWERTY keyboard.

Materials

In both the study and test phases, words were presented in lowercase in the Consolas font, size 36, against a black background. The set of 120 words was the set used in MacDonald and MacLeod (1998); these are listed in Appendix A. Words that formed each condition (i.e., aloud, silent, or new) were selected randomly for each participant. Each participant studied 40 words aloud and 40 words silently, with the sequence of words and conditions randomised. The remaining 40 words were used as distractors on the recognition test, with the resulting 120 test words randomised anew. This 2:1 ratio of targets to distractors has been commonly used in production effect studies (e.g., MacLeod et al., 2010) but so has the 1:1 ratio (e.g.,

MacLeod, 2011): A robust production effect is observed in both arrangements.

Procedure

For the study phase, participants were instructed to read the words presented in blue aloud and the words presented in white silently; silent reading was to be done without moving their lips. Each blue or white study word was presented individually for 3 s at the centre of the screen. Between successive aloud or silent words, a row of five red Xs was presented at the centre of the screen for 3 s. Participants were told that no action was required during these red-X displays. A blank period of 500 ms intervened between successive stimuli.

A recognition test immediately followed the study phase. The 80 studied words, intermingled with the 40 distractor words, were presented one at a time in a random order in yellow, so that colour would not serve as a retrieval cue. Using a key press, participants indicated whether they remembered studying each word (the "Y" key for yes; the "N" key for no), taking as long as they needed for each response. Upon response, the word disappeared and, after a 500-ms blank screen, the next word appeared. The entire procedure took under 30 min.

Results

The recognition test data are shown in the top row of Table 1 as well as the left section of Figure 1: The dependent measure is hit rate. Primary analyses were conducted using IBM SPSS Statistics (Version 26).[2] The single false alarm rate was low, indicating overall good memory for the studied words. As expected, a paired-samples *t*-test revealed that memory was significantly better for the aloud words than for the silent words, $t(23) = 5.77$, $p < .001$, $g_{av} = 1.01$, 95% CI [0.575, 1.51][3], a quite typical production effect of .178. Clearly, then, interleaving the repeated red stimuli among the aloud and silent words at study without requiring any action from participants did not reduce the normally observed production effect. This set the stage for changing the interleaved items from easily-ignored Xs to our pure-list words, thereby allowing us to investigate both of our research questions.

Table 1. Experiments 1-3: Proportions of "Yes" responses, corresponding to hits for studied targets and false alarms for unstudied distractors (standard errors shown in parentheses below each mean).

Experiment	Aloud	Silent	Distractor	Red	Red Distractor
1 – red Xs	.827	.649	.197		
	(.024)	(.043)	(.034)		
2 – red aloud	.642	.554	.218	.644	.157
	(.037)	(.034)	(.031)	(.030)	(.020)
2 – red silent	.799	.520	.233	.510	.229
	(.029)	(.042)	(.041)	(.033)	(.026)
3 – red aloud	.715	.605	.151		
	(.022)	(.030)	(.025)		
3 – red silent	.835	.599	.189		
	(.023)	(.040)	(.031)		

Experiment 2

Experiment 1 demonstrated that the production effect remained robust despite the interleaved irrelevant Xs. Thus, this simple manipulation interrupting the usual switching back and forth between aloud and silent items in a mixed list did not alter the effect. In Experiment 2, the red Xs were replaced with the pure-list words in red. One group read all of these red words aloud whereas the other group read all of them silently.

If the red words are simply compartmentalised as disruptor items, then we would expect no change to occur in terms of the magnitude of the mixed-list production effect. If, however, participants are driven by this manipulation to actually process and remember the red words, this would create a situation where we would be able to examine the between-subjects production effect for the red words, as well as to examine the two different situations in which there now would be more aloud than silent words when the irrelevant words were read aloud, or more silent than aloud words when the irrelevant words were read silently. In the first case, the aloud words would be made less statistically distinctive, and consequently the production effect for the to-be-remembered blue and white words would be reduced. In the second case, the aloud words would be made more statistically distinctive, and the production effect for the to-be-remembered blue and white words would increase.

Given that we wished to test memory for both the mixed-list and pure-list words, there was no ideal order for examining both. Therefore, first in Experiment 2, we examined memory for the pure-list red words before testing memory for the mixed-list words. This order ensured that testing of the mixed-list words did not interfere with the pure-list words while also not hindering any possible effect of the distinctiveness heuristic at test, since our hypothesis was that participants would become aware of the strategy through the study phase alone (i.e., when remembering the aloud pure-list items, they would use the heuristic to distinguish between the aloud and distractor items). In Experiment 3, which partially served as a replication of Experiment 2, participants were correctly informed that they would not be tested on the pure-list red words, so that we could confirm the results from testing only the mixed-list words.

Method

Participants

The participants were 48 undergraduate students (13 men, 34 women; age range: 17–33 years, mean age: 20.3 years, SD = 2.6 years; 3 participants declined to provide their age, 1 participant declined to provide demographic information) from the same source as Experiment 1. Participants received either course credit or $5 in exchange for their participation. Half of these participants were randomly assigned to the red-aloud condition and half to the red-

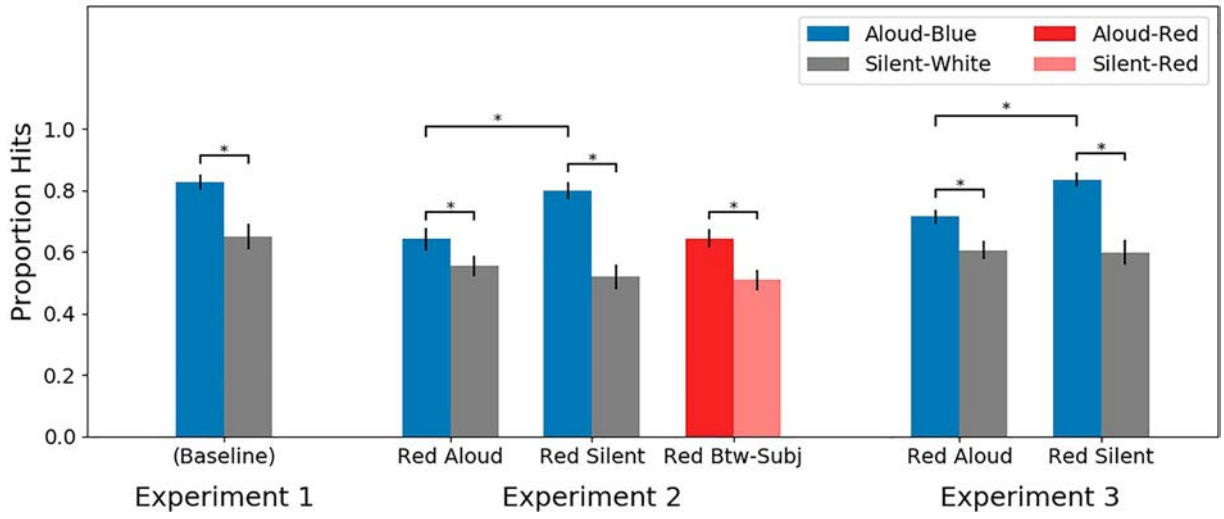

Figure 1. Proportions of hits for studied targets, with error bars representing standard error. Significant differences are indicated with an asterisk.

silent condition. (Although an explicit power analysis was not performed for this experiment, we will use the results of Experiment 2 to perform power analyses for Experiment 3, which partially serves as a replication of Experiment 2.)

Apparatus and stimuli

The apparatus was the same as in Experiment 1. The set of words used in Experiment 1 was augmented for Experiment 2; the additional words are also shown in Appendix A. The study and test lists were constructed as in Experiment 1, but with the addition of 79 red words inserted between successive aloud and silent words during study. An additional 40 words were included in the stimulus set to accommodate the test of the pure-list red words. The entire set of words was randomised anew for each participant.

Procedure

During the study phase, there were blue and white words to be treated as in Experiment 1. Instead of red Xs, however, here there were red words presented between successive blue and white words, creating the red-aloud and red-silent pure list conditions. In the red-aloud condition, participants were told to read all of the red words aloud; in the red-silent condition, participants were told to read all of the red words silently. Participants were explicitly told that only the words presented in blue and white would be tested. In fact, though, an incidental recognition test for the red words occurred immediately after the study phase and prior to the recognition test of the intentionally learned blue aloud and white silent words. The 79 red words and the additional 40 distractor words were presented in a random order in yellow during the test. The test was conducted in the same manner as the blue-white words test. Following this was the test for the blue and white words, which was conducted exactly as in Experiment 1.

Results

The results are shown in rows 2 and 3 of Table 1, as well as the middle section of Figure 1.

Test of the red words

False alarm rates were low, indicating that participants had very good memory for the red words despite being told that they would not be tested. The false alarm rate in the red-silent condition, however, was significantly higher than that in the red-aloud condition, $t(46) = 2.21$, $p = .032$, $g = 0.628$, 95% CI [.053, 1.21]. The key result was the hit rate for the red words. Recognition of them was significantly better when they were read aloud (.644) as opposed to silently (.510), $t(46) = 2.99$, $p = .004$, $g = 0.850$, 95% CI [.267, 1.45], indicating a large between-subjects production effect of .134, occurring in the same experiment as the within-subject production effect to be reported next. It is noteworthy that the magnitude of this effect was similar to that of the effect usually seen under the mixed list design and considerably larger than the effect usually seen under the pure list design.

Test of the blue and white words

False alarm rates were low for both groups – red-aloud and red-silent – and did not differ, $t(46) = 0.303$, $p = .763$, $g = 0.083$, 95% CI [−0.479, 0.652], evidence of good memory for the studied words. We performed a 2 (blue aloud/white silent) × 2 (red aloud/red silent) mixed ANOVA to examine how the magnitude of the production effect was influenced by whether participants read the red words aloud or silently. Recognition of the blue (aloud) words (.720) was better than recognition of the white (silent) words (.537), $F(1, 46) = 59.75$, $MSE = 0.014$, $p < .001$, $\eta_p^2 = 0.565$, 90% CI [0.394, 0.667][4], a quite typical within-subject production effect of .183, and very similar to that of Experiment 1. The main effect of the between-subjects condition (red aloud/red silent) was not

significant, F(1, 46) = 1.89, MSE = 0.048, $p = .176$, $\eta_p^2 = 0.039$, 90% CI [0, 0.160].

Critically, though, the interaction was significant, F(1, 46) = 16.33, MSE = 0.014, $p < .001$, $\eta_p^2 = 0.262$, 90% CI [0.094, 0.412], indicating that the magnitude of the production effect differed according to whether the red words had been read aloud or silently. Given the significant interaction, we next examined the simple main effects. Paired-samples *t*-tests with Bonferroni correction showed a significant production effect both in the red-aloud condition, $t(23) = 2.48$, $p = .021$, $g_{av} = 0.489$, 95% CI [0.076, 0.920], as well as in the red-silent condition, $t(23) = 8.80$, $p < .001$, $g_{av} = 1.53$, 95% CI [1.00, 2.15]. The effect was, however, significantly larger in the red-silent condition (.279) than in the red-aloud condition (.088), $t(46) = 4.04$, $p < .001$, $g = 1.15$, 95% CI [0.547, 1.77]. As the data in Table 1 show, the production effect was significantly larger in the red-silent condition than in the red-aloud condition because memory for the blue aloud words was significantly higher when the red words had been read silently, $t(46) = 3.33$, $p < .001$, $g = 0.943$, 95% CI [0.358, 1.56]. There was, however, no significant difference in memory for the white silent words as a function of whether the red words were read aloud or silently, $t(46) = 0.639$, $p = .526$, $g = 0.180$, 95% CI [−0.384, 0.750].

Comparing Experiments 1 and 2

Because Experiments 1 and 2 were identical in every respect apart from the critical change from red Xs to red words, we compared performance in the two experiments. We begin with the analyses for the to-be-remembered blue aloud items. A one-way ANOVA (with Bonferroni correction) with three levels – red Xs (from Experiment 1), red-aloud, and red-silent (both from Experiment 2) – revealed a significant effect, F(2, 69) = 10.62, MSE = 0.022, $p < .001$, $\eta_p^2 = 0.235$, 90% CI [0.091, 0.353]. Specifically, memory for the blue aloud words was poorer in the red-aloud condition than in the baseline red Xs condition, $t(46) = 4.19$, $p < .001$, $g = 1.18$, 95% CI [0.586, 1.819]. In contrast, there was no difference in memory for the blue aloud words when comparing the red-silent condition with the red Xs baseline, $t(46) = 0.740$, $p = .463$, $g = 0.209$, 95% CI [−0.355, 0.779]. Next, we similarly examined performance on the to-be-remembered white silent words; there was no significant difference, F(2, 69) = 2.85, MSE = 0.038, $p = .064$, $\eta_p^2 = 0.076$, 90% CI [0, 0.174]. These results confirm that the decreased magnitude of the production effect in the red-aloud condition was due to a decrease in memory for only the blue aloud words. Reading the red words silently did not result in a significant change relative to the red Xs baseline.

In summary, in Experiment 2, we observed a quite large between-subjects production effect for the red words relative to what has been typically observed in a pure-list experimental design. This finding supports our prediction that inserting a pure-list paradigm within a mixed-list paradigm allowed participants to observe the aloud-silent distinction and to use the distinctiveness heuristic even in a pure-list design, consequently increasing the magnitude of the pure list production effect compared to previous studies. In addition, Experiment 2 demonstrated that inserting the pure-list red words between successive blue aloud and white silent words differentially affected the magnitude of the within-subject production effect: Relative to the Experiment 1 baseline, the production effect decreased when the red words were read aloud but was not significantly altered when the red words were read silently. This is partially consistent with the hypothesis regarding the influence of statistical distinctiveness on the size of the production effect.

Our data fit the pattern of data reported by Icht et al. (2014) and by Bodner et al. (2016): When the red words are read aloud, there is effectively an overall greater number of aloud words than silent words, thereby making the aloud words less distinctive and diminishing the production effect. It is surprising, then, that when the red words were read silently the production effect was not enhanced despite there being as a result relatively fewer aloud words overall, which would be expected to make the aloud words more distinctive. In Experiment 3, we sought to replicate our latter findings.

Experiment 3

In Experiment 2, to address our first question with respect to pure-list experimental designs, we tested for memory of the red words from the pure lists prior to testing the words from the mixed lists. It is possible that this could have interfered with retrieval of the words from the mixed lists in some way, which would affect our interpretation of the results in terms of representing statistical distinctiveness. To address this issue, in Experiment 3 we tested only memory for the mixed-list words. Experiment 3 should therefore provide a straightforward replication of the findings of Experiment 2 regarding statistical distinctiveness.

Method

Participants

The participants were 48 undergraduate students (12 men, 35 women; age range: 17–23 years, mean age: 19.1 years, SD = 1.3 years; 1 participant declined to provide demographic information) from the University of Waterloo, recruited as previously described. Half of these participants were randomly assigned to the red-aloud condition and half to the red-silent condition. We performed a power analysis using G*Power Version 3.1.9.4 (Faul et al., 2007) to determine a suitable sample size based on the results of Experiment 2. Based on alpha = .05 and power = 0.95, and using the effect sizes from the mixed ANOVA and one-way ANOVA for the aloud items[5] in Experiment 2, we estimated that the appropriate sample size needed to satisfy these parameters was approximately N = 18 per

group. Therefore our sample size of N = 24 per group is adequate for investigating our main question in the current experiment.

Apparatus, stimuli, and procedure

The apparatus, stimuli, and procedure were the same as in Experiment 2 except that participants were only tested on the mixed-list words. To maintain consistency with Experiment 2, participants were now correctly told at the start of the experiment that the words in red would not be tested. Forty words were removed from the set of words used in Experiment 2 as a result (see Appendix A). The set of words remaining was randomised anew for each participant.

Results

The recognition test data are shown in the bottom two rows of Table 1, as well as the right section of Figure 1. False alarm rates were low in both groups (red-aloud and red-silent) and did not differ, $t(46) = 0.946$, $p = .349$, $g = 0.269$, 95% CI [−0.297, 0.840], evidence of good memory for the studied words. As in Experiment 2, we first performed a 2 (blue aloud/white silent; within) × 2 (red aloud/red silent; between) mixed ANOVA to examine the magnitude of the production effect as a function of whether participants read the red words aloud or silently. Recognition of the blue (aloud) words (.775) was better than recognition of the white (silent) words (.602), $F(1, 46) = 78.85$, $MSE = 0.009$, $p < .001$, $\eta_p^2 = 0.632$, 90% CI [0.476, 0.719], a quite typical within-subject production effect of .173, and a pattern very similar to that of Experiments 1 and 2. Similar to Experiment 2, the main effect of the between-subjects condition (red aloud/red silent) was not significant, $F(1, 46) = 2.39$, $MSE = 0.033$, $p = .129$, $\eta_p^2 = 0.049$, 90% CI [0, 0.176].

As in Experiment 2, the interaction was significant, $F(1, 46) = 10.65$, $MSE = 0.033$, $p = .002$, $\eta_p^2 = 0.188$, 90% CI [0.045, 0.341], again indicating that the magnitude of the production effect for the blue and white words differed according to whether the red words were read aloud or silently. We therefore probed the relevant simple main effects. Paired-samples *t*-tests with Bonferroni correction showed a production effect for both conditions, red-aloud: $t(23) = 4.85$, $p < .001$, $g_{av} = 0.828$, 95% CI [0.426, 1.27]; and red-silent: $t(23) = 7.45$, $p < .001$, $g_{av} = 1.42$, 95% CI [0.892, 2.04]. The effect was, however, significantly larger in the red-silent condition (.236) than in the red-aloud condition (.110), $t(46) = 3.26$, $p = .002$, $g = 0.923$, 95% CI [0.341, 1.54]. Again just as in Experiment 2, this difference was entirely due to memory for the blue aloud words being significantly better when the red words were read silently than when they were read aloud, $t(46) = 3.77$, $p < .001$, $g = 1.07$, 95% CI [0.476, 1.69]. There was no significant difference in memory for the white silent words as a function of whether the red words were read aloud or silently, $t(46) = 0.125$, $p = .901$, $g = 0.034$, 95% CI [−0.528, 0.600].

Comparing Experiments 1 and 3

We begin with analyses for the blue aloud items. A one-way ANOVA (with Bonferroni correction) with three levels – red Xs (from Experiment 1), red-aloud, and red-silent (both from Experiment 3) – revealed a significant effect, $F(2, 69) = 8.47$, $MSE = 0.013$, $p < .001$, $\eta_p^2 = 0.197$, 90% CI [0.062, 0.314]. Specifically, this difference resulted from poorer memory for the blue aloud words in the red-aloud condition relative to the red Xs baseline condition, $t(46) = 3.42$, $p < .001$, $g = 0.978$, 95% CI [0.384, 1.58], whereas there was no difference in memory for the blue aloud words when comparing the red-silent condition with the red Xs baseline, $t(46) = 0.249$, $p = .805$, $g = 0.068$, 95% CI [−0.494, 0.638]. In contrast to the effect on the blue aloud words, the difference in memory for the white silent words between Experiments 1 and 3 was not significant, $F(2, 69) = 0.52$, $MSE = 0.034$, $p = .598$, $\eta_p^2 = 0.015$, 90% CI [0, 0.069].

Comparing Experiments 2 and 3

Experiments 2 and 3 differed only in whether the red words were tested. Consequently, it makes sense to compare their results for the blue and white words directly. Toward this end, we carried out a 2 (blue aloud/white silent; within) × 2 (red-aloud/red-silent; between) × 2 (Experiment 2/3; between) mixed ANOVA (with Bonferroni correction). The three-way interaction was not significant, $F(1, 92) = 1.11$, $MSE = 0.011$, $p = .295$, $\eta_p^2 = 0.012$, 90% CI [0, 0.072], demonstrating the consistency of the results in Experiments 2 and 3.[6] As confirmation, this analysis also revealed an overall red-aloud/red-silent by blue aloud/white silent interaction, $F(1, 92) = 26.97$, $MSE = 0.011$, $p < .001$, $\eta_p^2 = 0.227$, 90% CI [0.111, 0.339], reaffirming that the between-subjects manipulation of red-aloud versus red-silent influenced the size of the within-subject production effect in the collapsed data. There were no other significant two-way interactions or main effects.

Both Experiments 2 and 3 showed that changes in the magnitude of the production effect between the red Xs baseline and the experimental conditions were due to changes in memory for only the blue aloud words. Specifically, the changes in the size of the production effect were due to participants having poorer memory for the blue aloud words in the red-aloud condition, which was evident in both experiments. We attribute this difference to the increased relative frequency of aloud words in the red-aloud condition: Participants were remembering the red words even though this was not required. Although participants also showed some memory for the red words in the red-silent condition, which should have correspondingly changed the relative frequency of the silent words in these conditions, as in Experiment 2 we did not find a significant influence on the magnitude of the production effect for the white silent words.

In summary, the results for memory of the blue aloud and white silent words in Experiment 3 replicated those of Experiment 2, producing consistent within-subject production effects. That the way in which the red words were processed also influenced the magnitude of the production effect for the blue versus white words – the effect was smaller when the red words were read aloud – fits nicely with the statistical distinctiveness explanation: Reading the red words aloud essentially makes the blue aloud words less distinctive. In the following, we further elaborate why the magnitude of the production effect may not have increased in the red-silent condition, as well as why the between-subjects production effect observed in Experiment 2 appears to be larger than what has typically been observed.

General discussion

Across three experiments, we integrated the examination of two dimensions of distinctiveness that we hypothesised could influence the magnitude of the production effect. First, we extended the distinctiveness heuristic to the pure-list experimental design, to test whether the magnitude of the production effect in a between-subjects setup would increase when participants are shown that they can use the aloud-silent distinction to their advantage when remembering the study items under this type of design. Here, we also created a situation where both within-subject and between-subjects production effects could be examined in the same experiment. Second, we wanted to examine the effect of statistical distinctiveness using situations in which the number of aloud and silent items at study were unequal: This design is atypical of past production effect studies. To accomplish these goals, we performed experiments in which we inserted additional words that participants read either all aloud or all silently, in between successive to-be-remembered aloud and silent words, and we distinguished these inserted pure-list items from the mixed-list items by providing participants with a cover story that we would not test these additional words. We compared the results between these experimental groups as well as with a baseline group not required to perform any action when additional stimuli were presented.

Our results support the prediction made in our first research question – that the magnitude of the production effect would increase relative to what has been shown in previous production effect studies when participants realise, consciously or unconsciously, that they can use the distinctiveness heuristic even in a pure-list experimental design. We were able to test this hypothesis in our setup given that participants showed above-chance memory for the additional red words both when they were read aloud and when they were read silently. The present study therefore allowed, for the first time, the investigation of the between-subjects and within-subject production effect magnitudes in the same experiment. Aloud versus silent was manipulated within-subject using the blue versus white cues; it was manipulated between-subjects by having one group read all of the red words aloud and the other group read all of the red words silently.

In our data, the production effect was numerically somewhat smaller between-subjects than within-subject (.134 between-subjects, and .183 in Experiment 2/.173 in Experiment 3 within-subject). However, the between-subjects effect here, with an effect size of 0.850, is considerably larger than what has been reported in previous between-subjects studies: In his meta-analysis, Fawcett (2013) reported an average between-subjects effect size of 0.37 with a 95% confidence interval of [0.16, 0.57] across twelve experiments that employed three different testing methods. As we have explained, we suspect that demonstrating the aloud-silent distinction in our experimental design, especially in the group reading the red words aloud, invoked greater use of the distinctiveness heuristic at test (Dodson & Schacter, 2001). This would mean that when participants observe distinctiveness at encoding they become more likely to use the distinctiveness heuristic. They do not need to be explicitly instructed to use this heuristic – and, again, to be clear, we claim that they do not even need to be consciously aware of using this heuristic – to show a robust production effect even in the pure-list paradigm.

Second, our pattern of data partially supports the statistical distinctiveness account presented by Icht et al. (2014; see also Bodner et al., 2016), where it was shown that the size of the production effect was reduced as the statistical distinctiveness of the aloud words decreased (i.e., increasing the number of aloud relative to silent words at study), and vice versa. In our experiments, we found a similar pattern in the groups of participants who read the additional red words aloud. These participants had good memory for the red words despite being told that those words would not be tested. Effectively, reading the red words aloud, even though participants were told that they were irrelevant and would not be tested, increased the proportion of aloud words compared to silent words.

Further supporting evidence in Experiment 2 showed that the mean hit rate for the red words when they were read aloud was almost the same as the mean hit rate for the to-be-remembered blue aloud words, indicating that participants actually processed and remembered these red words effectively. Because participants in the red-aloud condition were actually remembering a greater number of aloud words compared to silent words, the aloud words were less statistically distinct, so the observed decrease in memory for the aloud words would be expected according to this extension of the distinctiveness account. The main finding was replicated in Experiment 3: The magnitude of the production effect was reduced in the red-aloud condition because memory for the to-be-remembered blue aloud items declined relative to our baseline at the same time as there was no corresponding change in memory for the silent words.

Our findings do not, however, support the statistical distinctiveness account in the opposite direction: The account calls for a greater production effect when the aloud words are more statistically distinct than the silent words; that is, when there are fewer aloud words than silent words at study. In the red-silent conditions, participants also showed good memory for the red words. Because participants were now effectively studying more silent words than aloud words, under a statistical distinctiveness framework we would have expected an increase in the size of the production effect, where memory for the aloud words should have increased relative to our baseline. Instead, we found no difference in performance between the red-silent conditions relative to our baseline.

This difference between the red-silent and the red-aloud conditions cannot be attributed to more attention having been devoted to the red-aloud items such that they became associated as one group with the mixed-list aloud items, whereas the red-silent items were treated as a "to-be-ignored" third group and thus did not affect statistical distinctiveness. In the red-silent condition, participants showed memory for the red items comparable to that for the white-silent items (see MacDonald & MacLeod, 1998, for a related result in Experiment 3 of their study: Where participants could truly ignore the non-spoken items, they showed no recognition memory for these items). In addition, although the numerical differences in the magnitudes of the production effect across our three experiments showed a trend in the direction of a greater production effect in the red-silent conditions compared to the baseline (.279 and .236 versus .178), there is one main difference between our results and the results of Icht et al. (2014): They showed a memory boost for the aloud words in the condition with fewer aloud items (aloud 20%) relative to their other conditions, whereas the numerical difference in our results derived mainly from reduced memory for the silent items. Moreover, this difference was not significant across experiments and Icht et al. (2014), using a recognition test, also found no difference in memory for silent items.[7] We therefore think that the change predicted by the statistical distinctiveness account may not be as robust in the case where there are more silent words relative to aloud words, a possibility worthy of further study in the future.

An interesting connection to the directed forgetting literature (see MacLeod, 1998, for a review) is that, although we instructed participants that they did not need to remember the irrelevant red words, memory for the red words was similar to that for the to-be-remembered blue aloud and white silent words respectively in the red-aloud and red-silent conditions in Experiment 2. Our method resembled the item-method directed forgetting paradigm where a directed forgetting effect has been consistently found when using recognition tests. Accordingly, we would have expected to at least find worse memory for the red words compared to the to-be-remembered words in Experiment 2, even if participants were able to remember them above chance level. Why was this not the case? Indicating at the time of test whether an item had been a "remember" or "forget" item at study should have no effect on the magnitude of directed forgetting (Taylor et al., 2018), so it appears that the absence of a directed forgetting effect was due to the nature of the task affecting the encoding phase.

Selective rehearsal of the to-be-remembered items, but not of the to-be-forgotten items, is the most supported explanation of the item-method directed forgetting effect (see MacLeod, 1998; Tan et al., 2020). When the irrelevant red words were studied aloud, it would have been impossible not to rehearse these words at all, although we might have expected the red words still to have been rehearsed less than the blue aloud words. Compounding the mystery is the fact that no directed forgetting was seen when the red words were studied silently despite it being possible not to rehearse these words – or perhaps even not to study them at all. Moreover, we would have expected the red words to add to the memory load, reducing overall performance relative to Experiment 1: In directed forgetting, the presence of to-be-forgotten words in a list has been shown to reduce memory for the to-be-remembered words relative to a list containing only to-be-remembered words (e.g., Muther, 1965). However, memory for the to-be-remembered white silent words in Experiments 2 and 3 was similar to our Experiment 1 baseline. These curious violations of the selective rehearsal account warrant further examination in the future.

One might argue that the greater-than-typical production effect for the pure-list items reflects not usage of the distinctiveness heuristic but rather an effect of incidental versus intentional learning on the silent items. In the generation effect literature it has been suggested that, compared to intentional learning, incidental learning may result in a greater generation effect (e.g., Watkins & Sechler, 1988). The idea is that encoding of the non-generated items is better under intentional than incidental learning whereas memory for the generated items remains similar in the two procedures. In our experiment, consequently, it is possible that participants showed poorer memory for the silent pure-list items than they would have had those items been intentionally encoded, while there was no effect of distinctiveness on the aloud items.

Unfortunately, there currently are no studies directly comparing intentional versus incidental learning in the production effect. But in an experiment embedding the production effect within the item-method directed forgetting paradigm, Hourihan and MacLeod (2008) showed a greater production effect for the items that participants were told to forget relative to those they were instructed to remember, as a result of reduced memory for the silent "forget" items. However, the two main differences between the Hourihan and MacLeod (2008) study and ours are that (1) they used an entirely mixed-list paradigm whereas we also aimed to examine pure-list effects, and (2)

we did not observe directed forgetting in our Experiment 2. Further work is needed to specifically investigate the intentional versus incidental learning hypothesis.

Our experimental design was set up to distinguish the pure-list items from the mixed-list items. Given the similarity in the pure-list and mixed-list data in Experiment 2, however, we cannot rule out that participants may have integrated the pure-list and mixed-list items into a longer mixed list, rather than processing the pure lists as separate entities. We do believe that initially testing memory for the pure-list items in Experiment 2 immediately reminded participants of their status as a separate group of words. Future studies should continue to devise alternative ways of introducing the distinctiveness heuristic to participants without doing so explicitly.

There is one further consideration. In Experiment 1, inserting the same design of a row of red Xs between successive to-be-studied items may have resulted in participants growing accustomed to this stimulus, making it easier to ignore and thus providing participants some extra rehearsal time compared to Experiments 2 and 3. A future study could try using differing numbers of red Xs (and perhaps asking participants to indicate the number of Xs shown) to mitigate this potential issue.[8]

In summary, these experiments demonstrate that the pure list, between-subjects production effect can be enhanced if the distinctiveness of the pure aloud items is made more apparent by a concurrent mixed list, within-subject manipulation. These experiments also provide evidence partially in support of the statistical distinctiveness account, showing that even the addition of words that participants do not have to remember influences the relative proportions of aloud and silent words, thereby altering the magnitude of the production effect. Taken together, these findings offer further support for a distinctiveness explanation of the production effect.

Notes

1. Studying non-visually presented items may, however, lead to a different pattern of results. For example, Mama and Icht (2016) reported that words presented auditorily were better remembered when written down relative to spoken aloud. In general, the form and the magnitude of the production effect may be related to the number of encoding modalities during learning (see also Forrin & MacLeod, 2016, 2018).
2. Estimates of Hedges' *g* were computed according to the formulae provided in Cumming (2012). For confidence intervals of the effect sizes, we first computed the boundaries for the noncentrality parameter using the MBESS package in R Version 3.6.3 (Kelley, 2007), then (except for paired *t*-tests) computed the corresponding confidence intervals for the effect sizes according to the formulae provided in Smithson (2003), chapters 4 and 5. For paired comparisons, confidence intervals for the effect sizes were computed according to Algina and Keselman (2003).
3. Effect size confidence intervals for *t*-tests were calculated based on *d* (see Cumming, 2012, chapter 11).
4. For effect sizes represented by the partial eta squared statistic, 90% confidence intervals were computed instead of 95% confidence intervals since partial eta squared cannot be negative (see Steiger, 2004).
5. We only examined the aloud items here because in referencing the results from Icht et al. (2014), we did not expect to find differences across groups for memory of the silent items.
6. A supporting Bayesian repeated measures ANOVA, conducted in JASP using uniform priors, revealed anecdotal evidence for the null hypothesis (Lee & Wagenmakers, 2013), BF_{01}=2.273. This may be a consequence of the blue aloud and white silent items having been tested second in Experiment 2, yielding somewhat lower performance numerically compared to Experiment 3. Because there is agreement between the individual analyses in Experiments 2 and 3, we see this as a reasonable basis to conclude that the results of Experiments 2 and 3 were similar. (Note that the error percentage from this analysis was 6.886%, indicating that the Bayes Factor will change slightly when the analysis is repeated; however, this is within the acceptable error percentage limit: See van Doorn et al., 2020.)
7. A supporting Bayesian one-way ANOVA examining just the white silent items across all three experiments, conducted using uniform priors in JASP Version 0.14 (JASP Team, 2020), revealed a Bayes factor indicating moderate evidence for the null hypothesis (Lee & Wagenmakers, 2013), BF_{01}=3.069 (with error percentage < 0.001%).
8. We thank Reviewer 1 for suggesting this idea.

Acknowledgment

We thank David McLean, Junwen Liu, Zoey Hu, and Zelin Chen for their assistance with data collection.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was supported by the Natural Sciences and Engineering Research Council of Canada Discovery [grant number A7459].

Data availability

The datasets and programme code for this study are available from the authors upon request.

References

Algina, J., & Keselman, H. J. (2003). Approximate confidence intervals for effect sizes. *Educational and Psychological Measurement*, *63*(4), 537–553. https://doi.org/10.1177/0013164403256358

Bodner, G. E., Jamieson, R. K., Cormack, D. T., McDonald, D.-L., & Bernstein, D. M. (2016). The production effect in recognition memory: Weakening strength can strengthen distinctiveness. *Canadian Journal of Experimental Psychology*, *70*(2), 93–98. https://doi.org/10.1037/cep0000082

Bodner, G. E., & MacLeod, C. M. (2016). The benefits of studying by production … and of studying production: Introduction to the special issue on the production effect in memory. *Canadian Journal of Experimental Psychology*, *70*(2), 89–92. https://doi.org/10.1037/cep0000094

Conway, M. A., & Gathercole, S. E. (1987). Modality and long-term memory. *Journal of Memory and Language*, *26*(3), 341–361. https://doi.org/10.1016/0749-596X(87)90118-5

Cumming, G. (2012). *Understanding the New Statistics*. Routledge.
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*(1), 155–161. https://doi.org/10.3758/BF03196152
Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146
Fawcett, J. M. (2013). The production effect benefits performance in between-subject designs: A meta-analysis. *Acta Psychologica*, *142* (1), 1–5. https://doi.org/10.1016/j.actpsy.2012.10.001
Fawcett, J. M., & Ozubko, J. D. (2016). Familiarity, but not recollection, supports the between-subject production effect in recognition memory. *Canadian Journal of Experimental Psychology*, *70*(2), 99–115. https://doi.org/10.1037/cep0000089
Forrin, N. D., Jonker, T. R., & MacLeod, C. M. (2014). Production improves memory equivalently following elaborative vs non-elaborative processing. *Memory (Hove, England)*, *22*(5), 470–480. https://doi.org/10.1080/09658211.2013.798417
Forrin, N. D., & MacLeod, C. M. (2016). Auditory presentation at test does not diminish the production effect in recognition. *Canadian Journal of Experimental Psychology*, *70*(2), 116–124. https://doi.org/10.1037/cep0000092
Forrin, N. D., & MacLeod, C. M. (2018). Cross-modality translations improve recognition by reducing false alarms. *Memory (Hove, England)*, *26*(1), 53–58. https://doi.org/10.1080/09658211.2017.1321129
Forrin, N. D., MacLeod, C. M., & Ozubko, J. D. (2012). Widening the boundaries of the production effect. *Memory & Cognition*, *40*(7), 1046–1055. https://doi.org/10.3758/s13421-012-0210-8
Gathercole, S. E., & Conway, M. A. (1988). Exploring long-term modality effects: Vocalization leads to best retention. *Memory & Cognition*, *16*(2), 110–119. https://doi.org/10.3758/BF03213478
Hopkins, R. H., & Edwards, R. E. (1972). Pronunciation effects in recognition memory. *Journal of Verbal Learning and Verbal Behavior*, *11* (4), 534–537. https://doi.org/10.1016/S0022-5371(72)80036-7
Hourihan, K. L., & MacLeod, C. M. (2008). Directed forgetting meets the production effect: Distinctive processing is resistant to intentional forgetting. *Canadian Journal of Experimental Psychology*, *62*(4), 242–246. https://doi.org/10.1037/1196-1961.62.4.242
Hunt, R. R. (2013). Precision in memory through distinctive processing. *Current Directions in Psychological Science*, *22*(1), 10–15. https://doi.org/10.1177/0963721412463228
Icht, M., & Mama, Y. (2019). The effect of vocal production on vocabulary learning in a second language. *Language Teaching Research*. Advance online publication. https://doi.org/10.1177/1362168819883894
Icht, M., Mama, Y., & Algom, D. (2014). The production effect in memory: Multiple species of distinctiveness. *Frontiers in Psychology*, *5*, 886. https://doi.org/10.3389/fpsyg.2014.00886
Jamieson, R. K., & Spear, J. (2014). The offline production effect. *Canadian Journal of Experimental Psychology*, *68*(1), 20–28. https://doi.org/10.1037/cep0000009
JASP Team. (2020). JASP (Version 0.14)[Computer software].
Kelley, K. (2007). Methods for the behavioral, educational, and social sciences: An R package. *Behavior Research Methods*, *39*(4), 979–984. https://doi.org/10.3758/BF03192993
Kurtz, K. H., & Hovland, C. I. (1953). The effect of verbalization during observation of stimulus objects upon accuracy of recognition and recall. *Journal of Experimental Psychology*, *45*(3), 157–164. https://doi.org/10.1037/h0061022
Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian Cognitive Modeling: A Practical course*. Cambridge University Press.
Lin, O. Y., & MacLeod, C. M. (2012). Aging and the production effect: A test of the distinctiveness account. *Canadian Journal of Experimental Psychology*, *66*(3), 212–216. https://doi.org/10.1037/a0028309
MacDonald, P. A., & MacLeod, C. M. (1998). The influence of attention at encoding on direct and indirect remembering. *Acta Psychologica*, *98*(2-3), 291–310. https://doi.org/10.1016/S0001-6918(97)00047-4
MacLeod, C. M. (1998). Directed forgetting. In J. M. Golding, & C. M. MacLeod (Eds.), *Intentional forgetting: Interdisciplinary approaches* (pp. 1–57). Lawrence Erlbaum Associates.
MacLeod, C. M. (2011). I said, you said: The production effect gets personal. *Psychonomic Bulletin & Review*, *18*(6), 1197–1202. https://doi.org/10.3758/s13423-011-0168-8
MacLeod, C. M., & Bodner, G. E. (2017). The production effect in memory. *Current Directions in Psychological Science*, *26*(4), 390–395. https://doi.org/10.1177/0963721417691356
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*(3), 671–685. https://doi.org/10.1037/a0018785
Mama, Y., & Icht, M. (2016). Auditioning the distinctiveness account: Expanding the production effect to the auditory modality reveals the superiority of writing over vocalising. *Memory (Hove, England)*, *24*(1), 98–113. https://doi.org/10.1080/09658211.2014.986135
Muther, W. S. (1965). Erasure or partitioning in short-term memory. *Psychonomic Science*, *3*(1-12), 429–430. https://doi.org/10.3758/BF03343215
Ozubko, J. D., Gopie, N., & MacLeod, C. M. (2012). Production benefits both recollection and familiarity. *Memory & Cognition*, *40*(3), 326–338. https://doi.org/10.3758/s13421-011-0165-1
Ozubko, J. D., Hourihan, K. L., & MacLeod, C. M. (2012). Production benefits learning: The production effect endures and improves memory for text. *Memory (Hove, England)*, *20*(7), 717–727. https://doi.org/10.1080/09658211.2012.699070
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*(6), 1543–1547. https://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 (Hove, England)*, *22*(5), 509–524. https://doi.org/10.1080/09658211.2013.800554
Quinlan, C. K., & Taylor, T. L. (2013). Enhancing the production effect in memory. *Memory (Hove, England)*, *21*(8), 904–915. https://doi.org/10.1080/09658211.2013.766754
Smithson, M. (2003). *Quantitative applications in the social sciences: Confidence intervals*. SAGE Publications, Inc.
Steiger, J. H. (2004). Beyond the F test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, *9*(2), 164–182. https://doi.org/10.1037/1082-989X.9.2.164
Tan, P., Ensor, T. M., Hockley, W. E., Harrison, G. W., & Wilson, D. E. (2020). In support of selective rehearsal: Double-item presentation in item-method directed forgetting. *Psychonomic Bulletin & Review*, *27*(3), 529–535. https://doi.org/10.3758/s13423-020-01723-w
Taylor, T. L., Cutmore, L., & Pries, L. (2018). Item-method directed forgetting: Effects at retrieval? *Acta Psychologica*, *183*, 116–123. https://doi.org/10.1016/j.actpsy.2017.12.004
van Doorn, J., van den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., Etz, A., Evans, N. J., Gronau, Q. F., Haaf, J. M., Hinne, M., Kucharský, Š., Ly, A., Marsman, M., Matzke, D., Gupta, A. R. K. N., Sarafoglou, A., Stefan, A., Voelkel, J. G., & Wagenmakers, E.-J. (2020). The JASP guidelines for conducting and reporting a Bayesian analysis. Advance online publication. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-020-01798-5
Watkins, M. J., & Sechler, E. S. (1988). Generation effect with an incidental memorization procedure. *Journal of Memory and Language*, *27*(5), 537–544. doi:10.1016/0749-596X(88)90024-1

Appendix A

List of words used in the present studies. The words assigned to each condition (aloud, silent, distractor) were always selected randomly from the entire set of words for each participant in every experiment.

Experiment 1					
account	address	afternoon	amount	answer	arrow
attention	attitude	author	avenue	basket	battery
beauty	border	branch	building	campaign	capital
captain	castle	century	clothes	daughter	debate
department	dinner	direction	distance	education	election
engine	entrance	envelope	evening	factory	fashion
forest	foundation	friend	furniture	garden	glass
gravity	guardian	handle	harbour	history	holiday
industry	invention	invitation	island	journey	judge
justice	kettle	kingdom	kitchen	knock	ladder
language	laugh	leather	lesson	machine	market
meadow	merchant	message	minute	neighbour	nephew
ocean	office	orchard	package	painting	partner
peace	pebble	plate	pocket	porch	powder
quarrel	quarter	queen	record	resort	reward
river	sailor	school	shadow	shoulder	speech
station	steam	stream	summer	teacher	theatre
thread	ticket	traffic	travel	treasure	trousers
turnip	uncle	uniform	vacation	valley	victory
village	wagon	wheat	wheel	whisper	winter
Additional words used in Experiment 2					
article	automobile	battle	block	blossom	bottle
breakfast	bridge	business	character	circle	citizen
coast	cotton	creature	crowd	curtain	desire
doctor	dream	empire	expense	family	feather
fence	figure	fortune	frame	frost	governor
guest	harvest	health	interest	knight	knowledge
league	library	match	material	member	moment
mountain	nation	notice	object	occasion	opinion
penny	permit	plain	prince	property	province
purse	region	result	review	ribbon	robin
saddle	season	section	sense	service	shape
shell	shelter	shore	society	space	stairs
stamp	street	string	surface	temple	vessel
window	witness				
Additional words used in Experiment 2 and removed in Experiment 3					
banner	barrel	breath	breeze	bucket	channel
charity	colony	column	compass	committee	continent
council	discovery	display	division	flight	information
instruction	method	monarch	monument	motion	passenger
pattern	policy	population	portion	position	principle
prospect	receipt	relation	reserve	slipper	source
sunshine	territory	vision	voyage		