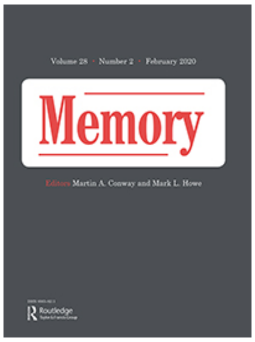

Memory

ISSN: 0965-8211 (Print) 1464-0686 (Online) Journal homepage: https://www.tandfonline.com/loi/pmem20

Distinctive encodings and the production effect: failure to retrieve distinctive encodings decreases recollection of silent items

Jason D. Ozubko, Luke D. Bamburoski, Kayla Carlin & Jonathan M. Fawcett

To cite this article: Jason D. Ozubko, Luke D. Bamburoski, Kayla Carlin & Jonathan M. Fawcett (2020) Distinctive encodings and the production effect: failure to retrieve distinctive encodings decreases recollection of silent items, Memory, 28:2, 237-260, DOI: 10.1080/09658211.2019.1711128

To link to this article: https://doi.org/10.1080/09658211.2019.1711128

Published online: 20 Jan 2020.

MEMORY
2020, VOL. 28, NO. 2, 237–260
https://doi.org/10.1080/09658211.2019.1711128

Routledge
Taylor & Francis Group

Distinctive encodings and the production effect: failure to retrieve distinctive encodings decreases recollection of silent items

Jason D. Ozubko[a], Luke D. Bamburoski[a], Kayla Carlin[a] and Jonathan M. Fawcett[b]

[a]Department of Psychology, SUNY Geneseo, Geneseo, NY, USA; [b]Department of Psychology, Memorial University, Saint John's, Canada

ABSTRACT

Studies have shown that when aloud and silent items are studied together, silent items are remembered more poorly than when they are studied independently. We hypothese that this cost to silent items emerges because, at test, participants search for memories of having said items aloud and when those memory searches fail, participants become uncertain about whether silent items were studied. This effect should be exaggerated if other unique distinctive encoding conditions are also included at study (e.g., mumbling, writing, typing, etc.). To test this prediction, we examined the impact of introducing mumbled, "important" (i.e., words that participants are told are the most important to remember), and mouthed words to a study list of aloud and silent words. Introducing mumbled and "important" words further impaired the recollection of silent items. Introducing mouthed items did not further impair the memorability of silent items because mouthing and speaking aloud are so similar and hence, are not fully unique from each other. The memorability of aloud items was unaffected in all conditions. These results suggest that participants search for distinctive encoding information at test, and only for items that fail those searches (i.e., silent items) do they lose confidence.

ARTICLE HISTORY
Received 23 July 2019
Accepted 24 December 2019

KEYWORDS
The production effect; recollection; familiarity; costs; benefits

Words that are produced (i.e., read aloud) are better remembered than words read silently. This simple effect was first reported in experiments examining modality (Gathercole & Conway, 1988; Hopkins, Boylan, & Lincoln, 1972) and verbal frequency (Ekstrand, Wallace, & Underwood, 1966; Hopkins & Edwards, 1972). Despite theoretical interest at the time of its initial discovery, this phenomenon was largely forgotten over the years and it was not until it was revived and rebranded as the *production effect* by MacLeod and colleagues that memory researchers began to take notice (MacDonald & MacLeod, 1998; MacLeod, Gopie, Hourihan, Neary, & Ozubko, 2010). Since its revival, the production effect has been demonstrated to be a robust phenomenon, occurring in a wide range of conditions, such as in recognition (Dodson & Schacter, 2001; Fawcett & Ozubko, 2016; Ozubko, Gopie, & MacLeod, 2012), recall (Castel, Rhodes, & Friedman, 2013; Jones & Pyc, 2014; Jonker, Levene, & MacLeod, 2014; Mama & Icht, 2019), and fill-in-the-blank tests (Lin & MacLeod, 2012; Ozubko, Hourihan, & MacLeod, 2012).

In their initial investigation, MacLeod et al. (2010) were particularly interested in determining whether production led to a more consistent or reliable benefit than other simple encoding techniques. MacLeod et al. tested two alternative kinds of "productions": a rote response production, such as pressing a button (i.e., the SPACEBAR) or saying "yes" to some words, and mouthing (i.e., moving one's lips to pronounce a word but without actually producing vocal sounds). MacLeod et al. found that simply pressing SPACEBAR in response to some words or verbally saying "yes" to some words did not lead to a production effect. However, mouthing words did lead to a production effect. From these findings MacLeod et al. suggested that the production effect was not necessarily driven by the fact that a response was made to produced words, but that a *unique* response was made. The mouthing results in particular suggested that vocalisation and auditory information were not necessary components of the memory benefit, so long as the response made to each word at study was relatively distinct.

The results of MacLeod et al. (2010) shed light on the factors behind the production effect, but as one of the major reasons production is so interesting is because it is so effective, and because both mouthing and speaking aloud were found to result in a production effect, a natural follow-up question emerges: which would be *more* effective at enhancing memory, speaking aloud or mouthing? Mouthing differs from speaking aloud both in that it lacks vocalisation and auditory information but also in that it is a less automatic response (i.e., reading aloud is more practiced in our day-to-day lives than mouthing). Perhaps by lacking vocalisation and auditory information mouthing is inherently less effective an encoding technique than speaking aloud. Alternatively, perhaps by

CONTACT Jason D. Ozubko ozubko@geneseo.edu Department of Psychology, SUNY Geneseo, 1 College Circle, Geneseo, NY 14454, USA

being less practiced mouthing leads to a more distinctive and memorable encoding, thereby being more effective than speaking aloud (see Bjork, 1994; Bjork & Bjork, 2011). Though MacLeod et al. (2010) did not address this question directly, subsequent researchers have broached this question using a 3-condition study phase variant of the production effect paradigm (Forrin, MacLeod, & Ozubko, 2012).

Production has commonly been studied in 2-condition mixed-list experiments where participants are presented with a series of words, often in one of two colours, and are asked to read words presented in one of those colours aloud and words in the other colour silently. In 3-condition study phases, words are shown in one of three colours, and participants are asked to read one colour of words silently, another aloud, and asked to perform another action with the third colour of words (such as mouthing, writing, or whispering). Using this technique, Forrin et al. (2012) directly compared words that were read silently, aloud, or mouthed at study. The researchers found that mouthing led to an intermediate effect between aloud and silent words. Hence, reading aloud persevered as the most effective encoding technique. In fact, Forrin et al. also examined writing and whispering. Although each of these encoding methods were found to be more memorable than reading silently, reading aloud was always found to be the most memorable condition.

The cost of production

Though Forrin et al. (2012) used the 3-condition study phase to emphasise the memory advantage that aloud words had over other production-like manipulations (writing, mouthing, and whispering), an interesting pattern exists in the memorability of aloud and silent items across their experiments. In Forrin et al.'s 2-condition study phase experiments, when silent words were studied in the presence of one production-like condition (i.e., spell, write, or type), mean hit rates for silent items were approximately .66 across the conditions. However, mean hit rates for silent items in the 3-condition study phase, when aloud and silent items were studied along with written, mouthed, or whispered words, were only .57. This decline in memorability was never emphasised or statistically examined, but it suggests that the introduction of aloud words in the 3-condition study phase may have impaired the memorability of silent items.

The idea that the memorability of silent items can be impaired by the presence of aloud items in mixed-list production effect experiments is not new. Bodner and colleagues have conducted several experiments showing that the memorability (i.e., hit rates) of silent items is often worse when they are studied with aloud items (i.e., in 2-condition study phases) than when they are studied alone (i.e., in pure lists) (Bodner & Taikh, 2012; Bodner, Taikh, & Fawcett, 2014). Forrin et al.'s (2012) findings agree that silent items incur a cost in hit rates when aloud items are present in mixed-lists, but critically, Forrin et al. never conducted a 2-condition experiment with aloud and silent words. Hence, their results are ambiguous on one important point: Is it the case that silent items suffer a cost to memorability when aloud words are present, regardless of what other encoding conditions exist? Or do silent items suffer an increasing cost to memorability as multiple distinctive encoding conditions (e.g., aloud, mouthed, written, spelled, etc.) are introduced? Furthermore, if silent items are incurring costs when multiple distinctive conditions are included at study, would these distinctive conditions incur costs on one another as well? That is, could aloud items be less memorable in the presence of a condition like mouthing, spelling, or writing, than in their absence? To address these questions, let us appeal to the distinctiveness model of production.

The distinctiveness model of production

The basic distinctiveness framework suggests that production requires additional processing (e.g., activation of phonological and motoric representations) not required of reading silent; these elements are integrated into the representation of the memory, and at test these elements can be relied upon to recognise that an item was studied (Dodson & Schacter, 2001; Fawcett, 2013; Fawcett & Ozubko, 2016; Ozubko & MacLeod, 2010; Ozubko, Major, & MacLeod, 2014). Thus, when presented with a test probe, participants can try to recall if they said that word aloud. If they recall having heard it or said it or moved their lips to pronounce it, then they can rest assured that they probably studied it and identify it as "old". However, if they cannot recall any of that information then they are left with two possibilities, either the word was read silently at study, or it was never studied. That is, the inability to recall if a word was said aloud is not diagnostic, and cannot help participants identify the word as studied.

An extension of this distinctiveness model is that a failure to find "aloudness" information in memory for a given test probe may make that test probe seem less likely to have been studied than if "aloudness" information were never sought out. Thus, one way to explain why the memorability of silent items is better when they are not studied or tested with aloud items is that when only silent items are studied participants simply do not bother searching memory for distinctive information about having said the word aloud (because no words were studied aloud), and focus on other attributes of memory instead. Supporting this idea, some studies have shown that the production effect in mixed-list production effect designs often has a notable recollective component (Fawcett & Ozubko, 2016), which is often related to conscious recall of perceptual experiences, in pure-list designs where only aloud or silent items are studied and tested by themselves, the effect lacks this recollective component (see Yonelinas, 2002 for a review of recollection and familiarity).[1]

From a distinctiveness framework then, silent items may incur a cost in mixed-list designs because participants are searching for "aloudness" memories for each individual item at test, disadvantaging retrieval of items studied silently. Participants lose confidence that the silent items were indeed studied when "aloudness" information cannot be found, and are more likely to reject them as a result. As well, this effect should be predominantly a recollective effect. If this account is correct, the results of Forrin et al. (2012) may indicate the reason hit rates for silent items declined in the 3-condition experiments was not because aloud items were being introduced necessarily, but because another distinctive encoding that was being introduced, and that distinctive encoding would be searched for in memory at test, further reducing confidence in silent items. Hence, if there were two distinctive encoding conditions at study, such as reading aloud and mumbling, participants would thus search for "aloudness" and "mumbleness" information for a given test probe, and failing to find either kind of information they may be even *less* likely to now believe that the test probe were studied than if mumble items had not been present.

Turning back to Forrin et al.'s (2012) finding that hit rates for silent items decreased when aloud items were introduced at study, our new theory would suggest that the hit rates for silent items declined not because aloud words were introduced specifically, but because aloud items are distinctive and a distinctive encoding condition was introduced. If we had instead began with aloud and silent words, introducing a third condition like mumbling or writing would have had the same effect on the hit rates for silent words. Hence, aloud words are not "special" by this account, and any distinctive encoding condition should incur further costs to the memorability of silent items.

Beyond explaining the cost to silent items that can be noted in Forrin et al.'s (2012) 3-condition mixed-list experiments, there are two other novel predictions of the theoretical account that we are proposing. First, is that the second distinctive encoding condition should only affect the memorability of silent items. In other words, aloud words should not impair the memory of the other distinctive condition and nor should the other distinctive condition impair memory for aloud words. The reason for this is that when presented with a test probe, if participants search for evidence of "aloudness" and find it, then it does not matter what the other potential kind of evidence that could have been searched for was. It does not matter if the other distinctive condition was a mumbling condition or a singing condition, "aloudness" information was found and so that test probe can be recognised as old. Hence, one prediction that our account makes is that the second distinctive encoding condition will incur a cost only on silent items, and the memorability of aloud items will remain the same as in a 2-condition production effect experiment.

The second novel prediction of our account is that to the degree that the second distinctive encoding condition is similar to production, it should incur less of a cost to silent items. To understand this prediction, consider an experiment where participants study aloud, mouthed, and silent items at study. A test probe is later shown and a participant searches their memory for evidence of having said that word aloud and fails. Because the experience of mouthing is almost identical to that of reading aloud, and the participant has failed to find evidence that the test probe was said aloud in their memory, the participant may consider it unsurprising if they also have no evidence of having mouthed the word. Hence, the participant failed to find evidence of either distinctive encoding condition, but because the mouthing information and aloud information are so similar, failing to find evidence of both is not as surprising as it would be if the two distinctive encoding conditions had been more dissimilar from one another. Therefore, we would predict in this case that the cost to silent items would be smaller than it would be if the two distinctive encoding conditions were more unique from each other.

Present experiments

To begin our investigation we will first replicate a basic production effect using a 3-condition study phase (Forrin et al., 2012) with two aloud and one silent condition. This first experiment will serve as a baseline for subsequent experiments. Unlike Forrin et al., we will thus begin with a standard production effect of aloud and silent words, even though we will have three encoding conditions (Forrin et al. reported spell, write, or type conditions along with silent words, but had no true 2-condition production effect experiment for comparisons). This first experiment will allow us to examine the memorability of aloud and silent items when only aloud and silent items are encoded but also when 3 encoding conditions are present. Furthermore, rather than focusing solely on old/new recognition which is most typical in production effect research, we will collect remember and know ratings to estimate recollection and familiarity separately (see Yonelinas, 2002 for review). Because all of the mnemonic costs are believed to be the result of consciously searching memory for contextual encoding experiences, we believe that costs should be primarily limited to more explicit forms of memory. We expect to see costs to the recollection of silent items specifically, since recollection is often characterised as the ability to retrieve episodic and contextual details associated with a previous experience. The familiarity of silent words should be less affected by the introduction of a third encoding condition, whatever its nature. Our analysis of costs will therefore be a more subtle analysis than previous studies by Bodner et al. (2014); Bodner and Taikh (2012) or Forrin et al. (2012). This is by design as if silent items are already suffering costs due to the presence of aloud items, further costs may be difficult to observe in overall hit rates, hence our emphasis on conscious recollection.

From there, we will investigate two manipulations aimed at the distinctiveness of the third set of study items: mumbling and "important" words. For mumbling, participants will be instructed to mumble some words aloud at study. For "important" words, participants will be told that a third condition of words is especially important to remember, and more important than either aloud or silent words, and so should be emphasised above all others during the study phase. Both mumbling and "important" words then will be very different attempts to introduce a second distinctive encoding condition to study. As per our theory, even though the mumbling and "important" conditions are so different from one another, they should both similarly impair the recollective memory of silent items above and beyond what is seen in a 2-condition production effect experiment (i.e., our baseline experiment; Experiment 1).

In our fourth experiment we will examine a 3-condition study phase with aloud, mouthed, and silent words. Mouthing was selected because it was examined in a 2-condition study phase (mouthed vs. silent) by MacLeod et al. (2010) and a 3-condition study phase (aloud vs. mouthed vs. silent) by Forrin et al. (2012). Hence, this experiment will directly connect the present investigation to two seminal works. More importantly however, as described above, our account specifically predicts that mouthing may be less effective at incurring further costs to silent items because mouthing and reading aloud are so similar.

Lastly, two final experiments will seek to replicate the results of Experiments 2 and 3 to both verify our findings and resolve any lingering methodological issues from the across experimental comparisons we will be making in Experiments 1–4. Because Experiments 2–3 will rely on an across experimental comparison with Experiment 1 to test our hypotheses, Experiments 5 and 6 serve as a within-subjects replication of Experiments 2 and 3 which each also include Experiment 1.

Experiment 1

To begin our investigation we will first replicate a basic production effect using a 3-condition study phase (Forrin et al., 2012). Rather than focusing solely on old/new recognition though, we will be gathering estimates of recollection and familiarity (see Yonelinas, 2002 for review). Past studies have shown that the recollection, rather than familiarity, best differentiates aloud and silent items in mixed-list designs (Fawcett & Ozubko, 2016; Ozubko, Gopie, et al., 2012), and we hypothesise that this effect should be exaggerated when a second distinctive encoding condition is introduced (i.e., the recollection of silent items should be depressed further). To evaluate this hypothesis, we will need measures of recollection and familiarity in all experiments and hence, we will gather remember/know estimates at test as estimates of recollection and familiarity (see Tulving, 1985).

In this investigation then, measuring recollection and familiarity using remember-know ratings (Fawcett & Ozubko, 2016; Ozubko, Gopie, et al., 2012; Tulving, 1985), we will examine whether, in mixed-lists, the recollective benefit of production can be perturbed by introducing a third encoding condition. Because all experiments in the present paper will use the 3-condition study phase and gather remember-know ratings to estimate recollection and familiarity, Experiment 1 will serve as a baseline in which to gather estimates of recollection and familiarity in a 3-condition study phase where no unusual condition beyond aloud and silent encoding conditions exist. That is, in Experiment 1 participants were presented with a series of words and were asked to either read each silently or aloud, with the exception that two separate cues were associated with reading aloud and only one with reading silently.[2] Following the study phase, participants engaged in a remember-know recognition test, to gather estimates of recollection and familiarity. Thus, Experiment 1 featured 3 encoding conditions, but nonetheless featured only aloud and silent conditions. Experiment 1 allowed us to gather baseline data regarding the proportion of recollection and familiarity-based responses in a mixed-list 3-condition production effect design. Furthermore, Experiments 2 and 3 will be replacing one of the aloud conditions of Experiment 1 with either mumbled or "important" words. The aloud-cost interpretation of the results of Forrin et al. (2012) is that the cost to silent items is purely a function of the presence or amount of aloud words. If there are more aloud words in Experiment 1 compared to any subsequent experiment, this basic interpretation would predict that hit rates for silent words should increase in subsequent experiments. Experiment 1 thus positions us to best test several competing accounts with our subsequent experiments.

Method

Participants

A total of 45 students from the State University of New York at Geneseo received extra credit in exchange for taking part in Experiment 1. One subject was excluded from analyses for failing to note the distinction between recollection and familiarity.

Stimuli and apparatus

The stimulus pool consisted of 348 words drawn from the MRC (Medical Research Council) Linguistic Database (www.psy.uwa.edu.au/MRCDataBase/uwa_mrc.htm). The words were nouns from 5 to 10 letters long, with Kučera and Francis (1967) frequencies between 30 and 847 and a mean frequency of 114.71 ($SD = 120.59$). For each participant, words were selected randomly from this pool and randomly assigned to each condition. The experiment was programmed in Psychopy version 1.84.2 (www.psychopy.org) and was carried out using a 17-in. colour monitor and a Hewlett-Packard computer running Windows 8. Words were presented in white ink on a black background in 14-point font.

Procedure

After being welcomed into the lab and signing a consent form, participants were told that they would be presented with a series of words to study, and that their memory for these words would later be tested.

Study Phase. In total, participants studied a list of 90 words, of which 30 were read silently and 60 were read aloud. Before each word an image was presented on the computer screen to instruct participants whether the upcoming word was to be read silently or read aloud. In Experiment 1, an image of an eye was used to indicate the upcoming word should be read silently, and two mouth images were used to indicate that the upcoming word should be read aloud. That is, of the 60 words read aloud, 30 words were cued by an image of an open mouth and 30 were cued by an image of a closed mouth (see Figure 1). In all cases, the image was presented for 1000 ms followed by a 250 ms blank screen before the stimulus appeared. Words were displayed in white Arial font for 2000 ms, followed by a 500 ms inter-stimulus interval.

Test Phase. An old/new recognition test containing all 90 studied ("old") words as well as 90 new words followed the study phase. Old and new words were randomly intermixed and presented individually during this test. For each word, participants were asked to provide a confidence rating on a scale from 1 to 6. Participants pressed 1, 2, or 3 if they believed that an item was new: 1 if they were *very* sure, 2 if they were *somewhat* sure, and 3 if they were *less* sure. Participants pressed 4, 5, or 6 if they believed that an item was old: 6 if they were *very* sure, 5 if they were *somewhat* sure, and 4 if they were *less* sure. After participants provided a confidence rating, recollection and familiarity were measured by having participants provide remember/know ratings for each item (Tulving, 1985). Participants pressed the "R" key if they could recollect the word, the "F" key if the word was familiar (i.e., a "know" response), and the "N" key if the word could not be recollected and was not familiar. To ensure that participants were not conflating the concepts of confidence with remember/know ratings, detailed instructions were provided to participants that emphasised the difference between confidence and recollection and familiarity.[3] Detailed instructions such as these have been shown to produce remember/know ratings that converge with other independent measures of recollection and familiarity (Rotello, Macmillan, Hicks, & Hautus, 2006; Rotello, Macmillan, Reeder, & Wong, 2005; Yonelinas, Dobbins, Szymanski, Dhaliwal, & King, 1996). After providing a remember/know rating, the test trial ended and there was a 500 ms inter-stimulus interval before the next test word was shown.

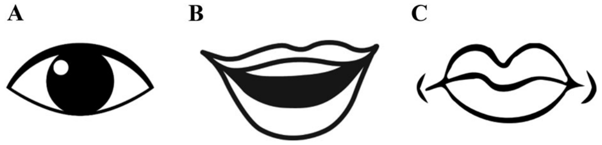

Figure 1. The three images used to cue production at study. Image (A) was used to cue the silent condition in all experiments. Image (B) was used to cue half of the aloud items in Experiment 1 and all of the aloud items in every other experiment. Image (C) was used to cue the other half of the aloud items in Experiment 1. In Experiments 2–4 image (B) was used to cue the aloud items and image (C) was used to cue the remaining condition (i.e., mumbling in Experiment 2 and 5, the important condition in Experiment 3 and 6, and mouthing in Experiment 4).

Results

All tests of significance used an alpha level of .05. Effect sizes are reported as partial-eta squared (η_p^2) for ANOVA results and Cohen's *d* for t-test results. As there were two aloud conditions in this experiment, they will be referred to as Aloud 1 and Aloud 2 respectively. Aloud 1 indicates words that were cued with the open mouth (Figure 1(B)) whereas Aloud 2 indicates words that were cued with the closed mouth (Figure 1(C)). Beyond this difference in cues, there was no experimental difference between these two aloud conditions.

Hit Rates. To analyse our results, we first examined overall hit rates before examining recollection and familiarity responses separately. To compute hit rates, we collapsed "4", "5", and "6" confidence responses into "old" responses and calculated hit rates for each condition according. We also calculated false alarm rates for comparison purposes. Hit and false alarm rates can be seen in Figure 2(A), with false alarm rates noted in the caption. A one-way analysis of variance (ANOVA) comparing the probability of an *old* response rates across the three studied conditions (Silent, Aloud 1, and Aloud 2) showed a significant effect, $F(2, 86) = 14.70$, $MSE = .008$, $p < .01$, $\eta_p^2 = .26$. Follow-up comparisons revealed, unsurprisingly, that there was no significant difference between the two aloud conditions, $t(43) = 1.42$, $p = 0.16$, $d = 0.43$. However, there were significantly more hits in both the Aloud 1 and Aloud 2 conditions compared to the silent condition, $t(43) = 4.84$, $p < .01$, $d = 1.48$ and $t(43) = 3.92$, $p < .01$, $d =$ 1.20, respectively. Averaging the two aloud conditions together to create a single aloud condition also yielded more hits than the silent condition, $t(43) = 4.87$, $p < .01$, d $= 1.49$. Hence, a production effect of similar magnitude was observed across both aloud conditions.

Recollection. Recollection responses were measured as the raw proportion of "R" responses at test and can be seen in Figure 3(A). A one-way ANOVA comparing recollection scores for the three studied conditions showed a significant effect, $F(2,86) = 9.60$, $MSE = .006$, $p < .01$, $\eta_p^2 = 0.18$. Follow-up comparisons revealed a similar

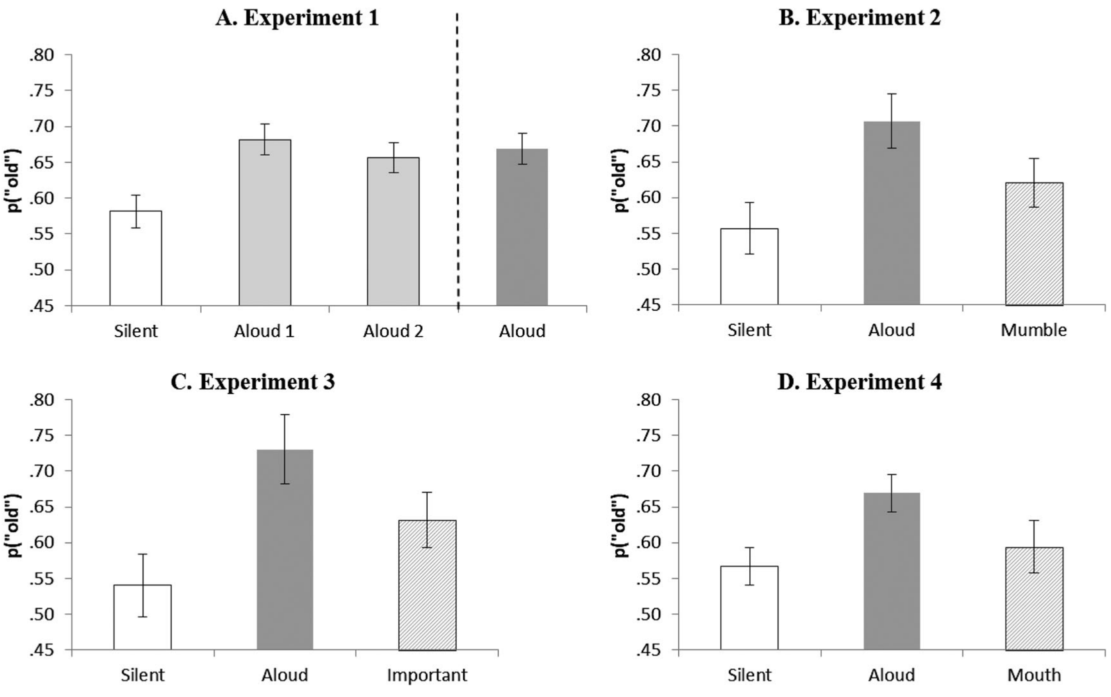

Figure 2. Mean hit for Experiments 1 through 4. For Experiment 1, Aloud 1 and 2 results were combined into a single Aloud condition. Error bars represent standard errors of the mean. Mean false alarm rates were .33 (*SE* = .02), .36 (*SE* = .03), .32 (*SE* = .04), and .33 (*SE* = .03) for Experiments 1 through 4 respectively.

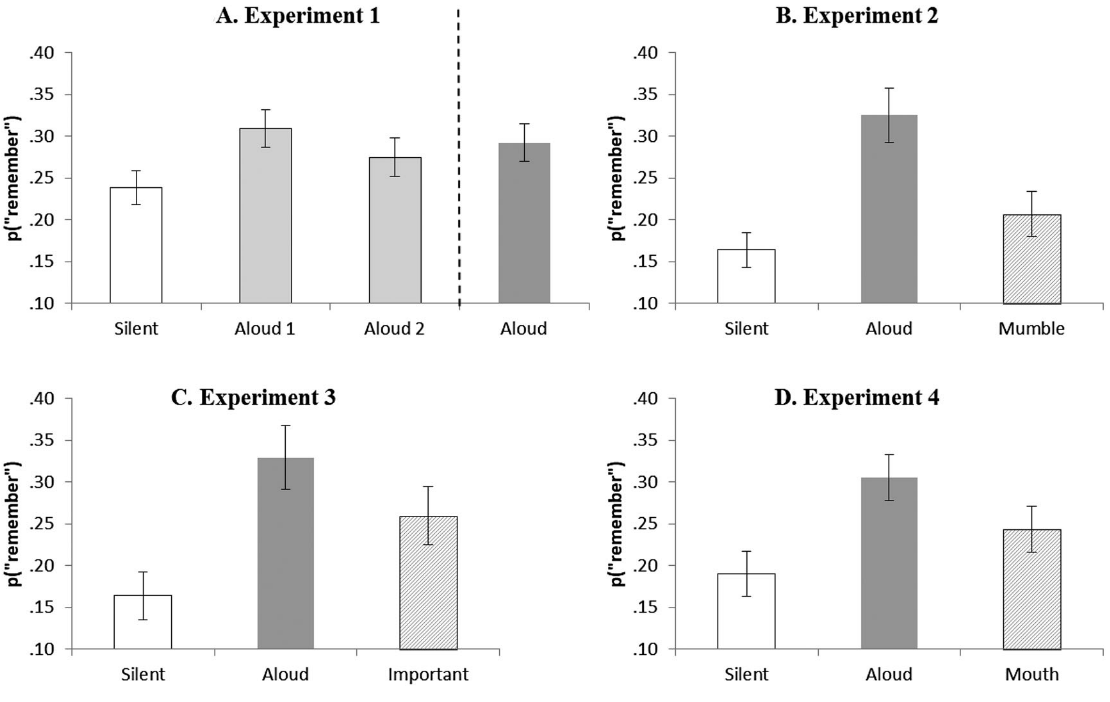

Figure 3. Mean proportion of "remember" response for items in Experiments 1 through 4. For Experiment 1, Aloud 1 and 2 results were combined into a single Aloud condition. Error bars represent standard errors of the mean. Mean false "remember" rates were .05 (*SE* = .01), .05 (*SE* = .01), .06 (*SE* = .02), and .06 (*SE* = .01) for Experiments 1 through 4 respectively.

pattern as overall hit rates in that there were significantly more recollective responses in both the Aloud 1 and Aloud 2 conditions compared to the silent condition, $t(43) = 4.00$, $p < .01$, $d = 1.22$ and $t(43) = 2.11$, $p < .05$, $d = 0.64$ respectively. Averaging the two aloud conditions together to create a single aloud condition also yielded more recollective responses than the silent condition, $t(43) = 3.27$, $p < .01$, $d = 1.00$. Unexpectedly though, there were significantly more recollective responses in the Aloud 1 condition than the Aloud 2 condition, $t(43) = 2.63$, $p < 0.05$, $d = 0.80$. In the end, a production effect was observed across both aloud conditions for recollective responses, although it was statistically larger for the Aloud 1 than the Aloud 2 condition.

Familiarity. To compute familiarity scores we began by calculating the proportion of "know" responses defined as the proportion of responses given a "familiar" rating and a confidence rating of 4, 5, or 6. Rather than using the raw proportion of "know" ratings to examine familiarity, familiarity was estimated using the *independent remember-know procedure*. The independent remember-know procedure involves adjusting the raw proportion of "know" responses at test by the opportunity to produce a "know" rating (see Rotello et al., 2005; Yonelinas et al., 1996). The independent remember-know procedure was used to adjust the proportion of "know" ratings because it has been shown to more directly converge with estimates of familiarity using independent procedures. Familiarity (F) was thus measured by the proportion of "know" responses divided by the proportion of non-recollected responses: $F = p(\text{"know"})/[1-p(\text{"recollection"})]$.

Familiarity scores are shown in Figure 4(A). A one-way ANOVA comparing familiarity scores for the three studied conditions showed a significant effect, $F(2,86) = 8.93$, $MSE = .01$, $p < .01$, $\eta_p^2 = 0.17$. Follow-up comparisons revealed a similar pattern as overall hit rates and recollective rates in that there were significantly greater familiarity scores in both the Aloud 1 and Aloud 2 conditions compared to the silent condition, $t(43) = 3.57$, $p < .01$, $d = 1.09$ and $t(43) = 3.69$, $p < .01$, $d = 1.13$ respectively. Averaging the two aloud conditions together to create a single aloud condition also yielded greater familiarity scores than in the silent condition, $t(43) = 4.20$, $p < .01$, $d = 1.28$, and familiarity scores in the Aloud 1 and Aloud 2 condition did not significantly differ from one another, $t(43) = 0.43$, $p = .67$, $d = 0.13$. Once again, a production effect of similar magnitude was observed across both aloud conditions, this time for familiarity scores.

Discussion

The results of Experiment 1 show a clear production effect emerges between aloud and silent items when there are two aloud conditions at test. Furthermore, the production effect was found at the level of hit rates, recollection rates, and familiarity scores, consistent with past studies (Fawcett & Ozubko, 2016; Ozubko, Gopie, et al., 2012) and therefore indicating that it is a robust effect and that the 3-condition paradigm is a valid method for observing the production effect. Interestingly, one of the aloud conditions exhibited greater recollection rates than the other. While it is always possible that there was some systematic bias that led to this result, such as the cueing icon for one of the aloud conditions being slightly more effective than the other, considering that the order of stimulus presentation was completely randomised and inter-mixed (i.e., the Aloud 1 condition was not studied before the Aloud 2 condition) and the instructions for the

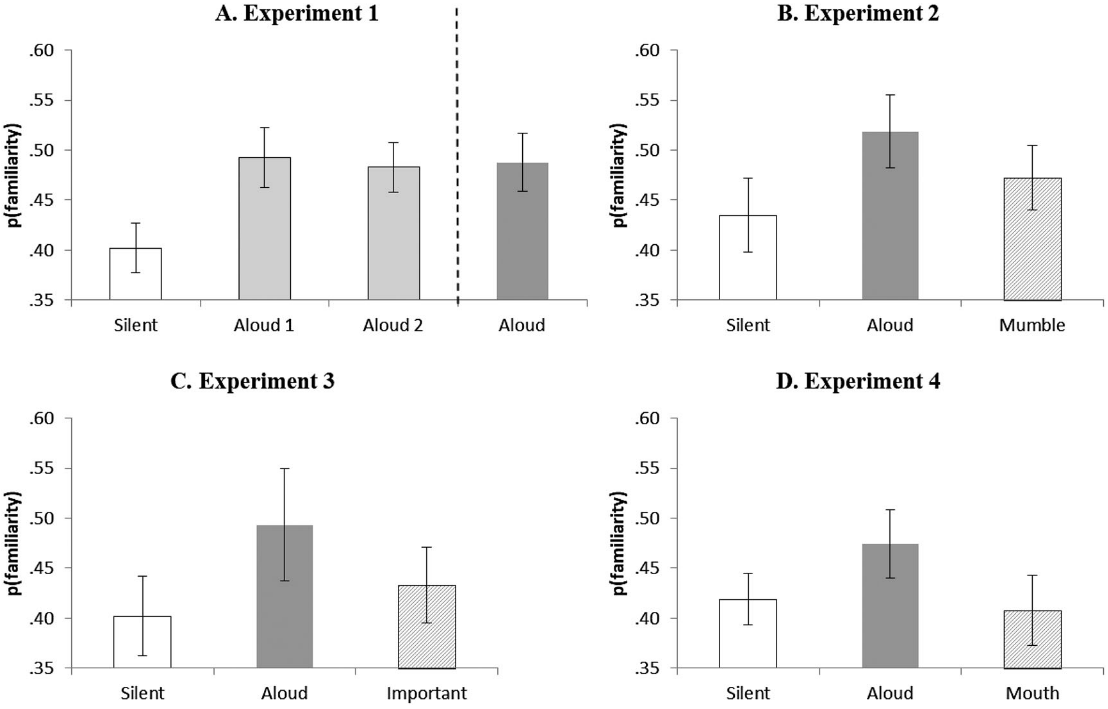

Figure 4. Mean familiarity score for items in Experiments 1 through 4. For Experiment 1, Aloud 1 and 2 results were combined into a single Aloud condition. Error bars represent standard errors of the mean. Mean false familiarity rates were .26 ($SE = .02$), .28 ($SE = .02$), .23 ($SE = .02$), and .24 ($SE = .02$) for Experiments 1 through 4 respectively.

two aloud conditions were identical, we consider it more likely that this difference was simply due to chance. Past studies that have examined the standard aloud/silent production in 2-condition within-subjects paradigms have observed production effects ranging in size from a hit rate advantage of .07 up to .27 (Dodson & Schacter, 2001; Fawcett & Ozubko, 2016; Hopkins & Edwards, 1972; Icht, Bergerzon-Biton, & Mama, 2019; Lin & MacLeod, 2012; MacLeod et al., 2010; Ozubko, Gopie, et al., 2012). Given that some variability exists within the exact size of the production effect in past research, for our purposes it is sufficient that we observed the effect for both of our aloud conditions in Experiment 1. We will therefore take the average of the two aloud conditions in Experiment 1 as our reference point for aloud conditions in the upcoming experiments.[4]

Experiment 2: An effortful and vocal encoding (mumbled words)

Having established the baseline production effect in Experiment 1, the goal of Experiment 2 is to introduce a second distinctive condition at study: mumbling. Mumbling involves reading a word aloud but altering its pronunciation so as to make it more difficult to say and hear. Mumbling is unpracticed, unusual, and less acoustically perceptible than speaking. Such an act, we argue, is therefore distinct from speaking (though not completely unrelated).

In their investigation of speaking and mumbling, Forrin et al. (2012) did observe that the hit rates for silent items were lower when aloud and mumbled words were studied with silent words than when silent words were studied with spelled, written, or typed words. However, as previously stated, no aloud/silent condition was reported by Forrin et al., nor was a mumbled/silent condition reported. Hence, it is not clear if aloud items by themselves were impairing the memorability of silent items, or whether silent items were incurring more cost when aloud and mumbling items were both studied with silent items, as opposed to if only aloud and silent items had been studied. In essence, do aloud items themselves impair the memorability of silent items, or are silent items impaired as a function of the number of distinct encoding conditions at study?

If the cost to silent items emerges only because of aloud items (aloud-cost), then the memorability of silent items in this experiment should actually increase compared to Experiment 1, as in Experiment 1 there were two sets of aloud items at study and in this experiment there will be only one. On the other hand, if the cost to silent items is emerging, as we propose, because participants are searching memory for distinctive encodings (i.e., "aloudness" and "mumbleness"), then we expect the introduction of mumbled items at study to impair memory of silent items compared to Experiment 1. We also expect this impairment to be in the recollection, but not the familiarity, of silent items. Furthermore, the recollection of aloud items should not be affected by the introduction of mumbled items, further demonstrating that this effect selectively occurs only for the items at study that lack distinctive encoding dimensions (i.e., silent items).

Methods

Participants

A total of 25 students from the State University of New York at Geneseo received extra credit in exchange for taking part in Experiment 2. One subject was excluded from analyses for failing to note the distinction between recollection and familiarity.

Stimuli and apparatus

The stimuli and apparatus were the same as those used in Experiment 1.

Procedure

In Experiment 2, participants were instructed to mumble 30 words at study, cued by the closed mouth cue (Figure 1(C)). Mumbling was defined as vocalising the word while minimising movement of the lips. As in Experiments 1 and 2, 30 words were read silently, by the eye cue (Figure 1(A)), and 30 words were read aloud, preceded by the open mouth cue (Figure 1(B)). The timing and presentation details of all words were identical to those of Experiments 1.

Following the study phase, participants engaged in the same recognition test as Experiment 1. Confidence ratings and remember/know ratings were gathered in Experiment 2 the same as in Experiments 1.

Results

Hit Rates. Hit and false alarm data is shown in Figure 2(B). A one-way ANOVA among *old* response rates for the three studied conditions (Silent, Aloud, and Mumbled) revealed a significant overall effect, $F(2,46) = 20.2$, $MSE = .007$, $p < .01$, $\eta_p^2 = .47$. Follow-up analyses revealed more hits for words read aloud than for words read silently, $t(23) = 6.24$, $p < .01$, $d = 0.53$. Furthermore, the hit rates for silent and aloud words in Experiment 2 did not differ from those in Experiment 1, $t(66) = 0.59$, $p = .56$, $d = 0.15$ and $t(66) = 1.08$, $p = .28$, $d = 0.27$ respectively. False alarm rates between Experiment 1 and 2 were not significantly different, $t(66) = 0.83$, $p = .41$, $d = 0.20$, meaning that hit rates could be directly compared and interpreted between these two experiments. Hence, the overall hit rate pattern for aloud and silent words in Experiment 2 replicated that of Experiment 1.

Mumbled words produced significantly more hits than silent words, $t(23) = 2.35$, $p < .05$, $d = 0.98$, though significantly fewer hits than aloud words, $t(23) = 4.35$, $p < .01$, $d = 1.81$. Nonetheless, this pattern indicates that a production effect was observed for mumbled words, albeit a smaller production effect than was observed for aloud words.

Recollection. Recollection data is shown in Figure 3(B). A one-way ANOVA analysing recollection scores among studied word conditions showed an overall significant effect, $F(2, 46) = 22.96$, $MSE = .007$, $p < .01$, $\eta_p^2 = .50$. Follow-up comparisons revealed that words read aloud were more recollectable than words read silently, $t(23) = 6.15$, $p < .01$, $d = 2.56$. However, compared to Experiment 1, there were significantly fewer recollective responses to silent words in Experiment 2, $t(66) = 2.29$, $p < .05$, $d = 0.56$. Recollection rates for aloud words did not significantly differ between Experiment 1 and Experiment 2, $t(66) = 0.86$, $p = .40$, $d = 0.21$. False recollection rates between Experiment 1 and 2 were not significantly different, $t(66) = 0.17$, $p = .86$, $d = 0.04$, meaning that recollection rates could be directly compared and interpreted between these two experiments. Hence, a production effect was observed in the recollection rates of Experiment 2, however the recollection rates of silent items were selectively impaired in Experiment 2 compared to Experiment 1.

Mumbled words produced significantly more recollective responses than silent words, $t(23) = 2.11$, $p < .05$, $d = 0.88$, though significantly fewer recollection responses than aloud words, $t(23) = 4.52$, $p < .01$, $d = 1.88$. Like with hit rates then, a production effect was observed for recollective responses to mumbled words, albeit a smaller production effect than was observed for aloud words.

Familiarity. Adjusted familiarity score data is shown in Figure 4(B). A one-way ANOVA comparing the adjusted familiarity scores for the three studied conditions showed an overall significant effect, $F(2, 46) = 4.52$, $MSE = .009$, $p < .05$, $\eta_p^2 = .16$. Follow-up comparisons revealed a similar pattern as overall hit rates and recollection rates in that there were significantly greater familiarity scores in the aloud condition compared to the silent condition, $t(23) = 3.21$, $p < .01$, $d = 1.34$. Familiarity scores for aloud and silent words in Experiment 2 also did not significantly differ from those in Experiment 1, $t(66) = 0.69$, $p = .49$, $d = 0.17$ and $t(66) = 0.76$, $p = .45$, $d = 0.19$ respectively. False familiarity scores between Experiment 1 and 2 were not significantly different, $t(66) = 0.83$, $p = .41$, $d = 0.20$, meaning that familiarity scores could be directly compared and interpreted between these two experiments.

Familiarity scores for mumbled words did not significantly differ from either the silent or the aloud conditions, $t(23) = 1.28$, $p = .21$, $d = 0.53$ and $t(23) = 1.66$, $p = .11$, $d = 0.69$ respectively, indicating that mumbling may have produced some intermediate level of familiarity. A production effect was therefore observed in familiarity scores between aloud and silent words but not for mumbled words.

Discussion

As predicted by our account, the introduction of mumbled words at study impaired the recollection, but not the familiarity, for silent words. Silent words showed lower recollection scores in Experiment 2 compared to Experiment 1, but familiarity was unaffected. By demonstrating that the memorability of silent items can be impaired by the replacement of an aloud condition with a mumble condition, we have shown that the cost to silent items is not solely the function of silent items being studied along with aloud items per se (i.e., the aloud-cost proposal), but instead arises as a function of the number of distinct encoding conditions at study.

Indeed, our interpretation is strengthened by the fact that mumbled items are less memorable than aloud items. That is, remember also that in Experiment 1 there were two sets of aloud items and in Experiment 2 one of those sets of aloud items was replaced with mumbled items. Mumbled items were ultimately *less* memorable than aloud items but *increased* costs for silent items compared to when they were not present (i.e., Experiment 1). If one were to presume that costs to silent items depend on the effectiveness of the other distinctive encoding conditions, then one would expect that mumbling, by virtue of being less memorable than speaking aloud, should impose less costs on silent items. In fact, just the opposite occurred. This seeming contradiction is easily explained however, by our distinctiveness account, which suggests that costs to silent items are a function solely of the number of distinct encoding conditions at study, and not their relative effectiveness. Furthermore, it is important to emphasise that the introduction of mumbled words also did not universally impair recollection, as recollection for aloud words did not significantly differ in Experiment 2 compared to Experiment 1. If mumbled items had impaired recollection for aloud and silent items, one could have proposed that mumbling was a difficult encoding condition and took attention away from other trials at study, even if mumbling ended up being a poor encoding condition compared to speaking aloud. Considering that mumbling selectively affected only silent items however, this explanation seems unlikely.

Experiment 2 therefore shows that silent items are selectively impaired in the presence of distinctive encoding conditions. These data are consistent with our predictions, and supports the notion that at test, participants may have been searching their memory for "aloudness" and "mumbleness", and in instances where they failed to find either, they were less confident that the probe in question had been studied (as opposed to if the probe simply didn't have "aloudness" information like in Experiment 1).

Experiment 3: An effortful and covert encoding ("important" words)

Experiment 2 showed that including a second distinctive condition at study selectively impaired the recollection of silent words. The second distinctive encoding condition in Experiment 2 was mumbling. Although mumbling is somewhat distinct from reading aloud, and we argue, is distinct enough to act as a separate encoding condition from speaking, it nonetheless shares many characteristics with speaking. Hence, in Experiment 3 we sought to include a second distinctive encoding condition that was as unlike speaking aloud as possible. In Experiment 3 then, participants once again engaged in a 3-condition production effect experiment. At study, some items were read silently, others aloud, and others were read silently but participants were told that words in this third condition were more important than all other words, and participants therefore needed to focus extra attention on trying to memorise those items. The encoding condition selected was therefore "important" words, and this "important" manipulation was selected specifically to address two key issues.

First, the "important" condition was selected to be as different as possible from speaking. Whereas speaking involves overt behavioural actions, "important" involves covert cognitive actions; and whereas speaking is a consistent action taken on all words, "important" has less clearly defined parameters and may involve variable encoding strategies. Remember however, that our distinctiveness model suggests that the specifics of the second distinctive encoding condition should not matter, so long as participants can search for the encoding technique at test. Controlling precisely what participants chose to do for "important" items should not matter as long as whatever participants do, they can search for those encoding experiences at test. What then did participants do to encode "important" words? From informal post-experiment comments, the important instruction led to more attention and effort for the "important" items. Sometimes this resulted in increased repetition or rehearsal of items, but other times resulted in semantic or imagery-based strategies. Past research has shown that participants can differentially recall items that were repeated silently once, twice, or said aloud at study (Ozubko et al., 2014), suggesting that repetition is a distinctive trait that can be recollected at test. Studies which have examined the recollection of words that are paired with pictures, sounds, or faces at study have shown that the sensory areas of brain, associated with initial perception at study, are re-activated at test for recollected items (Khader, Burke, Bien, Ranganath, & Rösler, 2005; Nyberg, Habib, Mcintosh, & Tulving, 2000; Vaidya, Zhao, Desmond, & Gabrieli, 2002; Waldhauser, Braun, & Hanslmayr, 2016; Wheeler, Peterson, & Buckner, 2000; see Skinner, Grady, & Fernandes, 2010 for review), suggesting that semantic elaboration and visual imagery are cognitive experiences which can be recollected or re-experienced at test. Hence, whether engaging in rote rehearsal or more elaborative rehearsal, the "important" items should carry distinctive encoding experiences which can be selectively recollected at test. From this fact, our model would thus predict that "important" items should impose a cost on the recollection of silent items, but not aloud items.

The second major consideration that led us to select the "important" condition is that this condition allows us to rule out a few alternative accounts of our results. First, like Experiment 2, Experiment 3 allows us to evaluate the aloud-cost idea that costs to silent items are driven by the presence of aloud items. Because in Experiment 3 we are replacing one of the aloud conditions from Experiment 1 with an "important" condition, the hit rates for silent items should increase in Experiment 3 compared to Experiment 1 if the costs for silent items are purely driven by the presence of aloud items. Second, however, the "important" condition allows us to evaluate simple encoding explanations of our results, namely the lazy-reading account (cf. Begg & Snider, 1987; see MacLeod et al., 2010). In Experiment 1, there were two sets of aloud items and one set of silent items. A lazy-reading account could suggest that aloud items in Experiment 1 were being interpreted as more important than silent items and thus, stealing rehearsals from silent items, resulting in silent items being less memorable in our Experiment 1 than they would be in a pure-list of only silent items, as has been observed in the past (Bodner et al., 2014; Bodner & Taikh, 2012). This explanation suggests that for silent items to be further impaired in Experiment 3 compared to Experiment 1, more rehearsals would have to be stolen from silent items in Experiment 3 compared to Experiment 1. Hence, would need to predict that "important" words in Experiment 3 are considered more important than aloud words in general.

If "important" items were considered by participants to be more important than the aloud items they replaced were, then "important" items should steal *more* rehearsals than the aloud items did. This would arrive at the prediction that the recollection for silent items should decline in Experiment 3 compared to Experiment 1. However, because "important" items are considered more important than both aloud and silent items in this account, they should steal rehearsals not just from silent items but also from the remaining aloud items. As a result, this account would actually predict that recollection for both aloud *and* silent items should be impaired in Experiment 3 compared to Experiment 1. We should acknowledge that it may be possible to entertain an even more sophisticated lazy-reading hypothesis that could selectively predict a decline in recollection for silent items but not aloud items in Experiment 3, but we will save consideration of this issue for the General Discussion.

In sum, Experiment 3 is similar to Experiment 1 except that one of the aloud conditions is being replaced with the "important" word condition. If the costs to silent words are driven solely by the presence of aloud items (aloud-cost) then the memorability of silent items should improve in Experiment 3 compared to Experiment 1. If the costs to silent words are being driven by lazy reading of silent items, because aloud items are perceived as more important by participants, then replacing aloud items with "important" words should either have no effect on the memorability of silent words or could further impair the memory of silent items but should also then impair the memory of aloud items. Finally, if neither of these alternative accounts is correct and instead, as we have argued, silent items incur costs because of failed memory searches at test for distinctive encoding information, then recollection but not familiarity for silent items should be impaired in Experiment 3 compared to Experiment 1, and the memory of aloud items should be equivalent across the experiments.

Method

Participants

A total of 26 students from the State University of New York at Geneseo received extra credit in exchange for taking part in Experiment 3. Two subjects were excluded from analyses for failing to follow instructions.

Stimuli and apparatus

The stimuli were the same as those used in Experiment 1.

Procedure

At study, participants were instructed to read 30 words silently, preceded by the eye cue (Figure 1(A)) and 30 words aloud, preceded by the open mouth cue (Figure 1(B)). The remaining 30 words were preceded by the closed mouth cue (Figure 1(C)); these words were to be read silently, but participants were told that these words would be especially important for the later recognition test. The timing and presentation details of all words in Experiment 3 were identical to those of Experiment 1 and 2.

Following the study phase, participants engaged in the same recognition test as Experiments 1 and 2. Confidence ratings and remember/know ratings were gathered in Experiment 3 the same as in Experiments 1 and 2.

Results

Hit Rates. Hit and false alarm data can be found in Figure 2 (C). A one-way ANOVA among *old* response rates for studied conditions (Silent, Aloud, and Important) indicated a significant overall effect, $F(2, 46) = 15.41$, $MSE = .01$, $p < .01$, $\eta_p^2 = .40$. Follow-up analyses showed that words in the aloud condition had significantly greater hit rates than words in the silent condition, $t(23) = 5.28$, $p < .01$, $d = 2.20$. Furthermore, the hit rates for silent and aloud words in Experiment 3 did not differ from those in Experiment 1, $t(66) = 0.98$, $p = .33$, $d = 0.24$ and $t(66) = 1.65$, $p = .10$, $d = 0.41$ respectively. False alarm rates between Experiment 1 and 3 were not significantly different, $t(66) = 0.20$, $p = .84$, $d = 0.05$, meaning that hit rates could be directly compared and interpreted between these two experiments. Hence, the overall hit rate pattern for aloud and silent words in Experiment 3 replicated that of Experiment 1.

Important words produced significantly more hits than silent words, $t(23) = 2.50$, $p < .05$, $d = 1.04$, though significantly fewer hits than aloud words, $t(23) = 3.34$, $p < .01$, $d = 1.39$. This pattern indicates that a memory benefit was observed for important words, albeit a smaller benefit than the production effect observed for aloud words.

Recollection. Recollection rates can be found in Figure 3 (C). A one-way ANOVA analysing recollection scores among studied word conditions showed an overall significant effect, $F(2, 46) = 14.23$, $MSE = 0.01$, $p < .01$, $\eta_p^2 = .38$. Follow-up analyses revealed that words read aloud were more recollectable than words read silently, $t(23) = 6.20$, $p < .01$, $d = 2.59$. However, compared to Experiment 1, there were significantly fewer recollective responses to silent words in Experiment 3, $t(66) = 2.07$, $p < .05$, $d = 0.51$. Recollection rates for aloud words did not significantly differ between Experiment 1 and Experiment 3, $t(66) = 0.88$, $p = .38$, $d = 0.22$. False recollection rates between Experiment 1 and 3 were not significantly different, $t(66) = 0.64$, $p = .52$, $d = 0.16$, meaning that recollection rates could be directly compared and interpreted between these two experiments. Hence, a production effect was observed in the recollection rates of Experiment 3, however the recollection rates of silent items were selectively impaired in Experiment 3 compared to Experiment 1.

Important words produced significantly more recollective responses than silent words, $t(23) = 2.98$, $p < .01$, $d = 1.24$. Important words did not produce significantly fewer recollective responses than aloud words, though this non-significant result was marginal, $t(23) = 2.03$, $p = .05$, $d = 0.85$. A production effect was therefore observed in recollection for important words. Conservatively, the effect was of equivalent magnitude as that observed for aloud words, though it is possible that the effect is marginally smaller. In any even the important take away is that a significant production effect was observed.

Familiarity. Familiarity score data is shown in Figure 4(C). A one-way ANOVA comparing familiarity scores across the three studied conditions did not show an overall significant effect, $F(2, 46) = 2.49$, $MSE = .02$, $p = .09$, $\eta_p^2 = 0.10$. Follow-up analyses found that aloud words produced marginally greater familiarity scores than silent words, $t(23) = 2.05$, $p = .05$, $d = 0.85$, indicating a production effect may have been present for aloud words, albeit less robustly than observed in both recollection and overall hits. However, familiarity scores for aloud and silent words in Experiment 3 did not significantly differ from those in Experiment 1, $t(66) = 0.08$, $p = .94$, $d = 0.02$ and $t(66) = 0.01$, $p = .99$, $d = 0.002$ respectively. False familiarity scores between Experiment 1 and 3 were not significantly different, $t(66) = 1.02$, $p = .31$, $d = 0.25$, meaning that familiarity scores could be directly compared and interpreted between these two experiments. Within Experiment 3, important words did not significantly differ from silent words or aloud words, $t(23) = 0.93$, $p = .37$, $d = 0.39$ and $t(23) = 1.32$, $p = .20$, $d = 0.55$ respectively.

Discussion

Consistent with the predictions of our distinctiveness account, the introduction of "important" items selectively impaired the recollection of silent items. The memorability of aloud items was unaffected in Experiment 3, compared to Experiment 1. Interestingly, aloud items were remembered better than "important" items, showing that merely telling participants that "important" items should be remembered at the expense of all other items was not sufficient to overcome the production effect. Nonetheless, the results of Experiment 3 therefore replicate those of Experiment 2 and once again show that the recollection, but not the familiarity, of silent items was impaired with the introduction of a second distinctive encoding condition, consistent with our account.

It is worth emphasising one last time how different mumbling and "importantness" are. Whereas mumbling is a vocal condition, similar to but distinct from, speaking, "important" is a much less clearly defined encoding manipulation likely involving subject-selected attention, mental imagery, elaboration, or other rehearsal processes. Indeed, we contend that this extreme difference between the mumbling and "important" conditions coupled with the fact that they had the same impact on the memorability of silent items serves as good support for our distinctiveness account. We believe that mumbling and "importantness" had the same effect because the nature of distinctiveness does not matter so much as the ability for participants to search for distinctiveness at test. If participants search for "aloudness" and another kind of encoding information ("mumbleness" or "importantness") and fail, that probe will seem less likely to have been studied than if they only search for "aloudness" and failed. Hence, the recollection of silent items should and did decline compared to Experiment 1.

The results of the present experiment therefore support the idea that the effects of Experiments 2 and 3, and especially the costs to silent items, were not occurring just because of the presence of aloud items (aloud-cost), or as a result of selective rehearsal to more important items (lazy-reading). Indeed, participants were explicitly told that "important" items were the most important items to remember and yet they remembered them less well than aloud items. This shows us that the production effect generally may be a retrieval phenomenon, wherein the memory benefit arises from the attempt to retrieve "aloudness" information at test. If production were simply an encoding effect, wherein aloud items garnered increased attention and encoding efforts by virtue of seeming important, then "important" items should have been even better remembered than aloud items. They were not.

Experiment 4: An encoding similar to speaking (mouthed words)

Based on our distinctiveness model, we have argued thus far that introducing a second distinctive encoding condition should impair the recollection, but not the familiarity, of silent items. Experiments 2 and 3 both show this pattern, consistent with our predictions. However, a presently untested aspect of our model is that the second distinctive encoding condition should only be effective to the degree that it is distinct from production. In other words, if aloud and silent items are studied along with some other encoding condition which is itself very similar to speaking aloud, then the costs to silent items should be reduced or eliminated. In Experiment 2 we compared speaking aloud with mumbling and found those conditions sufficiently distinct from one another to impose a cost on the memorability of silent items. What then could be more similar to speaking aloud than mumbling? In Experiment 4 we selected mouthing.

Mouthing involves reading a word aloud without actually speaking. Hence, just like speaking, one must plan on how to move ones lips, perform the action, and often hear in one's own mind the word being said. The only missing component is the actual verbalisation of the word, which requires one to consciously control oneself to prevent actual verbalisation as reading is often considered to be a relatively automatic process (Macleod, 1991). One could therefore argue that mouthing is verymuch like production except that it requires a bit more conscious attention and control, to avoid accidentally and automatically reading the word aloud. Alternatively, the effort needed to not speak aloud may be minimal, and mouthing does lack the acoustic dimension of speaking aloud, hence one could consider mouthing to be slightly less distinct and involved than speaking aloud. Both perspectives should agree however, that mouthing is very close to speaking aloud, and possibly as close as one can get while still being obviously different encoding conditions.

In the context of our previous experiment, Experiment 4 serves an important purpose, as one could argue that the "important" words in Experiment 3 were not being encoded as well as aloud words because the "important" manipulation could only lead to increased attention or rehearsal whereas the aloud items were explicitly acted on. Explicit behavioural actions, like speaking aloud, may be a more effective encoding technique than just treating items as "important", and hence, one could have predicted that aloud items would still encoded better than "important" items in Experiment 3. If this argument is correct, mouthing may actually turn out to be a more effective encoding condition than the "important" condition, because mouthing is an overt and explicit action like speaking aloud, and might even require more conscious control and effort than speaking aloud (i.e., mouthing is not an automatic process and so one must be sure to not accidentally say the word aloud; Macleod, 1991). If the costs to silent items observed in Experiment 3 arose from the introduction of "important" items, which were not encoded as well as aloud items (by virtue of not being acted on with explicit behavioural actions), then the introduction of mouthed items in Experiment 4 should result in costs to silent items that are as large in magnitude as Experiment 3, or perhaps even larger.

In contrast, our own interpretation of Experiment 3 is that of a retrieval phenomenon. The reason "important" items led to a cost to silent items is because at test participants were searching for "aloudness" and "importantness", rather than just "aloudness" (as in Experiment 1). Hence, we do not believe that the results of Experiment 3 are the result of aloud items being encoded better than "important" items. If our account is correct then two highly similar encoding conditions should have less of an impact on silent items than two clearly different encoding conditions. For example, if a participant fails to retrieve "aloudness" information for a test probe, then it would almost be expected that the test probe would also not elicit any "mouthness" information from memory, since "aloudness" almost completely encompasses "mouthness". Subjectively then, the test probe might be considered to have only really failed to show evidence for one type of encoding, even though two were searched for. In contrast if a test probe fails to elicit "aloudness" information and "important" information then it may seem, subjectively, like a more significant failure on the part of the test probe because "aloudness" and "important" information as so distinct from one another and non-overlapping in nature.

Hence, having mouthed items at study should cause less of an impairment to the recollection of silent words than having mumbled or "important" words. Furthermore, as we have seen in all experiments so far, the familiarity of silent items and the recollection and familiarity of aloud items should be unaffected by mouthed items. Should all of these patterns emerge they would provide good evidence for the predictions of our account.

Method

Participants

A total of 24 students from the State University of New York at Geneseo received extra credit in exchange for taking part in Experiment 4.

Stimuli and apparatus

Experiment 4 used the same pool of words and general apparatus as Experiment 1.

Procedure

Experiment 4 had a nearly identical procedure to Experiment 1 except that instead of two aloud conditions at study there was one aloud condition and one "mouthed" condition. Mouthing was defined as moving lips as if you were to say the word, but without making any vocalisation. Thus, 30 words were read silently, 30 words were read aloud, and 30 words were mouthed. Silent words were preceded by the eye cue (Figure 1(A)), aloud words were preceded by the open mouth cue (Figure 1(B)), and mouthed words were preceded by the closed mouth cue (Figure 1 (C)). The timing and presentation details of all words in Experiment 4 were identical to those of Experiment 1.

Following the study phase, participants engaged in the same recognition test as Experiment 1. Confidence ratings and remember/know ratings were gathered in Experiment 4 the same as in Experiment 1.

Results

Hit Rates. Hit and false alarm rate data for Experiment 4 can be found in Figure 2(D). A one-way ANOVA comparing *old* response rates across the three studied conditions (Silent, Aloud, and Mouthed) revealed an overall significant effect, $F(2, 46) = 7.51$, $MSE = .009$, $p < .01$, $\eta_p^2 = .25$. Follow-up analysis revealed a production effect between words read aloud and words read silently, $t(23) = 3.79$, $p < .01$, $d = 1.58$. Furthermore, the hit rates for silent and aloud words in Experiment 4 did not differ from those in Experiment 1, $t(66) = 0.39$, $p = .70$, $d = 0.10$ and $t(66) = 0.01$, $p = .99$, $d = 0.002$ respectively. False alarm rates between Experiment 1 and 4 were not significantly different, $t(66) = 0.14$, $p = .89$, $d = 0.03$, meaning that hit rates could be directly compared and interpreted between these two experiments. Hence, the overall hit rate pattern for aloud and silent words in Experiment 4 replicated that of Experiment 1.

Mouthed words did not lead to significantly more hits than silent words, $t(23) = 1.05$, $p = .31$, $d = 0.44$, and produced significantly fewer hits than aloud words, $t(23) = 2.64$, $p < .05$, $d = 1.10$. Hence, a production effect was not observed for mouthed words in overall hit rates.

Recollection. Recollection rates are shown in Figure 3(D). A one-way ANOVA analysing recollection scores among studied conditions showed an overall significant effect, $F(2, 46) = 12.12$, $MSE = .07$, $p < .01$, $\eta_p^2 = .35$. Follow-up analyses revealed that recollection rates were higher among words read aloud than words read silently, $t(23) = 4.21$, $p < .01$, $d = 1.76$. As with overall hit rates, the recollection rates for silent and aloud words in Experiment 4 did not differ from those in Experiment 1, $t(66) = 1.35$, $p = .18$, $d = 0.33$ and $t(66) = 0.38$, $p = .70$, $d = 0.09$ respectively. False recollection rates between Experiment 1 and 4 were not significantly different, $t(66) = 0.84$, $p = .40$, $d = 0.21$, meaning that recollection rates could be directly compared and interpreted between these two experiments. Hence, the recollective pattern for aloud and silent words in Experiment 4 replicated that of Experiment 1.

Mouthed words produced significantly more recollective responses than silent words, $t(23) = 2.16$, $p < .05$, $d = 0.91$, though significantly fewer recollective responses than aloud words, $t(23) = 3.61$, $p < .01$, $d = 1.51$. Nonetheless, this pattern means that a production effect was observed for mouthed words in recollective responses, albeit a smaller production effect than was observed for aloud words.

Familiarity. Familiarity score data for Experiment 4 is shown in Figure 4(D). A one-way ANOVA comparing familiarity scores among the three studied word conditions found no significant effect, $F(2,46) = 2.61$, $MSE = .01$, p

= .09, $\eta_p^2 = 0.10$. Follow-up analyses confirmed that there was no significant difference between familiarity scores for aloud and silent words, $t(23) = 1.80$, $p = .09$, $d = 0.75$, though familiarity scores for aloud and silent words in Experiment 4 did not significantly differ from those in Experiment 1, $t(66) = 0.42$, $p = .67$, $d = 0.10$ and $t(66) = 0.31$, $p = .76$, $d = 0.08$ respectively. False familiarity scores between Experiment 1 and 4 were also not significantly different, $t(66) = 0.42$, $p = .67$, $d = 0.10$, meaning that familiarity scores could be directly compared and interpreted between these two experiments.

Mouthed words were not significantly different from silent or aloud words, $t(23) = 0.41$, $p = .67$, $d = 0.17$ and $t(23) = 1.92$, $p = .07$, $d = 0.80$ respectively. Overall then, there was no production effect observed in familiarity scores in Experiment 4.

Discussion

Experiments 2 and 3 demonstrated that the recollection of silent words was impaired by the introduction of a third encoding condition that was distinct from speaking. Experiment 4 demonstrated that this effect was statistically eliminated by the introduction of a third encoding condition that was similar to speaking – namely, mouthing. Although mouthing was a distinctive encoding manipulation, leading to more recollection than for silent items, it did not affect the recollection of silent items significantly. It is worth noting that numerically it does appear as if the recollected rates for silent items may be on a downward trajectory in Experiment 4 ($M = .19$) compared to Experiment 1 ($M = .24$). We do not wish to over-emphasise this nonsignificant difference, but it is interesting to note that there may be a subtle effect here, albeit one that would likely be very difficult to demonstrate significantly given its small size.[5] Importantly though, the competing encoding interpretation of Experiment 3 predicted that the cost to silent items in Experiment 4 should be as large or larger in magnitude. The fact that, if anything, the costs to silent items appears to be smaller in magnitude in Experiment 4 speaks strongly against this competing account. As predicted then, Experiment 4 demonstrates that unlike mumbling and "important" words, mouthing has no significant effect on the recollection of silent words.

Experiment 5: A within-subjects replication of mumbling

Before proceeding to our General Discussion and interpretation of our findings, it is worth acknowledging that given the novelty of our results, an attempt at replication is prudent. Indeed, the finding that the recollection of silent words was significantly impaired in the presence of mumbled words (Experiment 2) and "important" words (Experiment 3) relies on a cross-experiment comparison with Experiment 1, and across experiment comparisons need to be treated with some caution. In Experiments 5 and 6 then, we sought to investigate if Experiments 1–3 would replicate within-subjects. Namely, beyond performing a within-experiment comparison between Experiments 1 and 2 or Experiments 1 and 3, we sought to further reduce variability between conditions by eliminating the difference between the groups of subjects. Thus, if the same subjects had participated in Experiment 1 and Experiments 2, would the results we observed replicate? Similarly, if the same subjects had participated in Experiments 1 and 3, would the results we observed replicate? Experiments 5 and 6 then aim to replicate Experiments 1 with Experiments 2 and 3 using a within-subjects variant of our design.

In Experiment 5, participants took part in a two back-to-back study-test sessions. In our Baseline condition participants studied two sets of aloud items and one set of silent items (replicating Experiment 1), whereas in the Mumble condition participants studied aloud, mumbled, and silent items (replicating Experiment 2). The order of these sessions was randomly determined for each participant. The test phase was a recognition test identical to that of previous experiments. Experiment 5 therefore replicates Experiments 1 and 2 in a within-subjects design. Experiment 6 will follow this same procedure but use aloud, "important", and silent words in the experimental condition, thus replicating Experiments 1 and 3 in a within-subjects design.

Methods

Participants

A total of 27 students from the State University of New York at Geneseo received extra credit in exchange for taking part in Experiment 5. One subject was excluded from analyses for failing to note the distinction between recollection and familiarity.

Stimuli and apparatus

The stimuli and apparatus were the same as those used in Experiments 1 and 2.

Procedure

Experiment 5 used the same procedures as Experiments 1 and 2 combined. In the Baseline condition participants studied aloud and silent words and so this condition served to replicate Experiment 1. In the Mumble condition participants studied aloud, mumbled, and silent words and so this condition served to replicate Experiment 2. In both conditions the study phase was immediately followed with a test phase, as in prior experiments. Confidence ratings and remember/know ratings were gathered during this test, the same as in Experiments 1 and 2.

Regarding the Baseline and Mumble conditions, all participants took part in both conditions, one followed by the other. In Experiment 5 then, participants essentially participated in both Experiments 1 and 2 back-to-back. The order of the Baseline and Mumble conditions was randomly determined for each participant, with 17 participants receiving the Baseline-Mumble order and 9 receiving the Mumble-Baseline order. No significant order effects were observed in the data and so data for Baseline and Mumble conditions were collapsed together respectively, irrespective of condition order. Words randomly selected for the Baseline condition were not repeated in the Mumble condition, and so each condition received its own unique set of words at study and test.

Finally, unlike Experiment 1 where the two sets of aloud items were cued with two separate icons (see Figure 1), in the Baseline condition of Experiment 5 all aloud items were cued with the mouth icon (Figure 1(B)). From the subjects' perspectives then, there were simply aloud and silent items at study in the Baseline condition.

Results

The results of Experiment 5 can be seen in Figure 5. Hit rates, recollection rates, and familiarity scores are plotted separately for the Baseline and Mumble experimental conditions, and for the aloud, silent, and mumbled conditions from each.

Hit Rates. Hit and false alarm rates are shown in top panels of Figure 5. Considering first the Baseline condition, a one-way ANOVA among *old* response rates for the three studied item conditions (Silent, Aloud 1, and Aloud 2) revealed a significant overall effect, $F(2,50) = 15.26$, $MSE = .01$, $p < .01$, $\eta_p^2 = .38$. Follow-up analyses revealed more hits for Aloud 1 and Aloud 2 items compared to the silent items, $t(25) = 4.59$, $p < .01$, $d = 1.84$ and $t(25) = 3.66$,

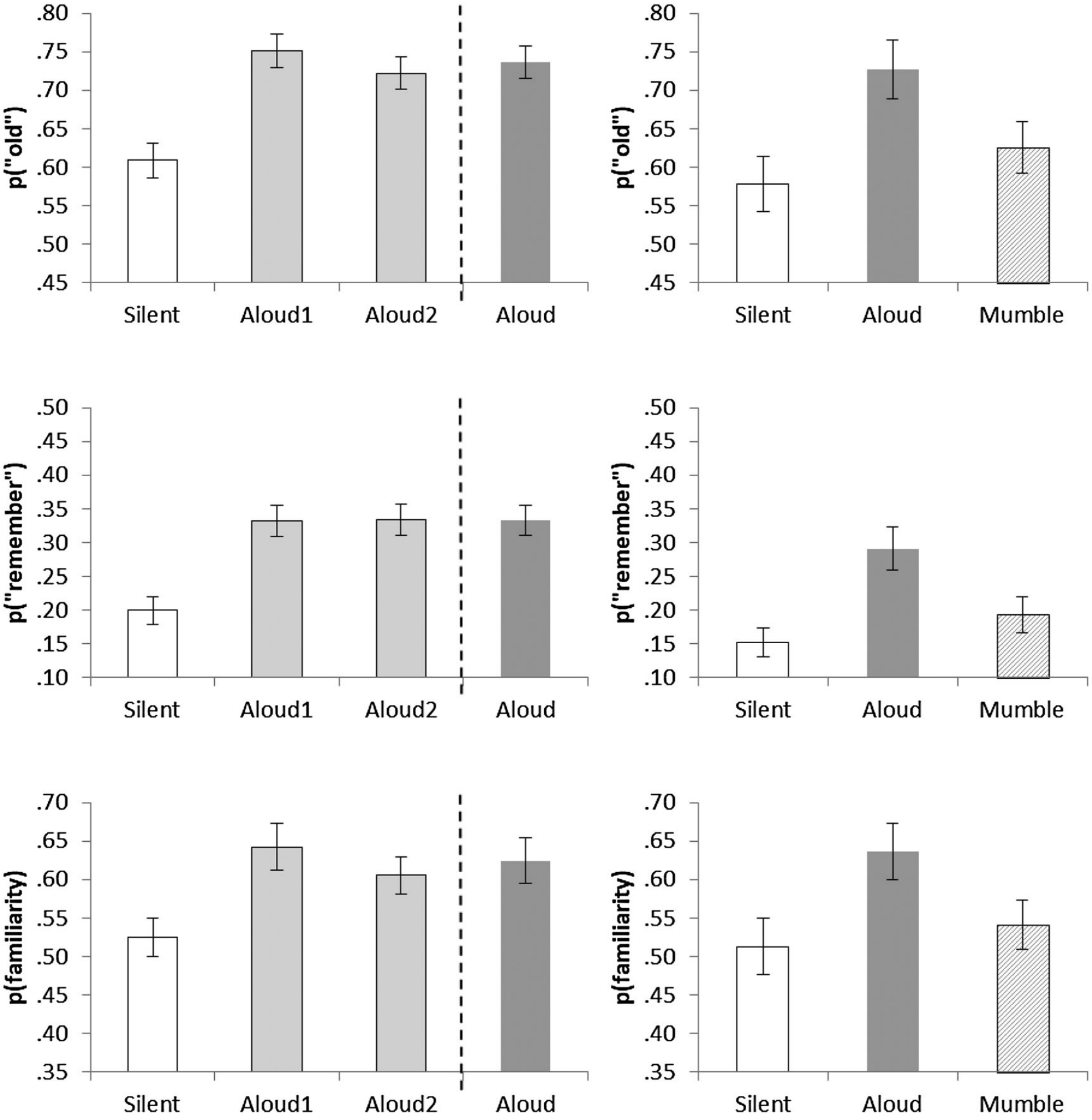

Figure 5. Mean hit and false alarm rates, recollection rates, and familiarity scores for the baseline and experimental condition of Experiment 5. For the baseline condition, Aloud 1 and Aloud 2 results were combined into a single Aloud condition. Error bars represent standard errors of the mean. For the Baseline condition mean false alarm rates were .35 (*SE* = .03), false recollection rates were .04 (*SE* = .01), and false familiarity rates were .32 (*SE* = .03). For the Mumble condition mean false alarm rates were .32 (*SE* = .04), false recollection rates were .03 (*SE* = .01), and false familiarity rates were .31 (*SE* = .03).

$p < .01$, $d = 1.46$ respectively. Hit rates did not differ between Aloud 1 and Aloud 2, $t(25) = 1.69$, $p = .10$, $d =$ 0.68. Hence, a production effect was observed in both the Aloud 1 and Aloud 2 conditions and was of equivalent magnitude. Hit rates for these two conditions would be combined into a single hit rate for subsequent comparison to the Mumble condition.

Considering the Mumble condition, a one-way ANOVA among *old* response rates for the three studied item conditions (Silent, Mumbled, and Aloud) revealed a significant overall effect, $F(2,50) = 11.80$, $MSE = .01$, $p < .01$, $\eta_p^2 = .32$. Follow-up analyses revealed more hits for aloud items compared to silent items, $t(25) = 4.38$, $p < .01$, $d = 1.75$, hence a production effect was observed. Mumbled items had lower hit rates than aloud items, $t(25) = 3.17$, $p < .01$, $d = 1.27$, and did not have higher hit rates than silent items, $t(25) = 1.76$, $p = .09$, $d = 0.70$. Hence, mumbled items did not show a significant production effect.

Comparing the Mumble to the Baseline condition, hit rates for aloud and silent items did not significantly decline between these conditions, both t's < 1.02, p's > .32. False alarm rates between these two conditions were not significantly different, $t(25) = 0.94$, $p = .36$, $d =$ 0.38, meaning that hit rates could be directly compared and interpreted between the Baseline and Mumble condition. Hence, the presence of mumble items in the Mumble condition did not affect the overall hit rates of silent items.

Recollection. Recollection rates are shown in the middle panels of Figure 5. Considering first the Baseline condition, a one-way ANOVA for recollection scores among studied word conditions (Silent, Aloud 1, and Aloud 2) showed an overall significant effect, $F(2,50) = 14.71$, $MSE = .01$, $p < .01$, $\eta_p^2 = .37$. Follow-up comparisons revealed that words read aloud in the Aloud 1 and Aloud 2 conditions were more recollectable than words read silently, $t(25) = 5.16$, $p < .01$, $d = 2.06$ and $t(25) = 4.10$, $p < .01$, $d = 1.64$ respectively. The rate of recollection responses did not differ between Aloud 1 and Aloud 2 items, $t(25) = 0$, $p = 1.00$, d $= 0$. Hence, a production effect was observed in both the Aloud 1 and Aloud 2 conditions and was of equivalent magnitude. Recollection rates for these two conditions would be combined into a single recollection rate for subsequent comparison to the Mumble condition.

Considering the Mumble condition, a one-way ANOVA analysing recollection scores among studied word conditions (Silent, Mumbled, and Aloud) showed an overall significant effect, $F(2,50) = 17.55$, $MSE = .01$, $p < .01$, $\eta_p^2 = .41$. Follow-up analyses revealed higher recollection rates for aloud items compared to silent items, $t(25) = 5.26$, $p < .01$, $d = 2.10$, hence a production effect was observed. Mumbled items had lower recollection rates than aloud items, $t(25) = 3.86$, $p < .01$, $d = 1.54$, but did have higher hit rates than silent items, $t(25) = 2.07$, $p < .05$, $d = 0.83$. Hence, mumbled items did show a significant production effect in recollection responses, albeit a smaller production effect than aloud items.

Comparing the Mumble to the Baseline condition, recollection rates for aloud items did not significantly decline between these conditions, $t(25) = 1.82$, $p = .08$, $d = 0.73$, however recollection rates for silent items did decline, t $(25) = 2.49$, $p < .05$, $d = 1.00$. False recollection rates between these two conditions were not significantly different, $t(25) = 0.83$, $p = .42$, $d = 0.33$, meaning that recollection rates could be directly compared and interpreted between the Baseline and Mumble condition. The presence of mumbled items thus impaired the recollection of silent items but not aloud items.

Familiarity. Familiarity scores are shown in the bottom panels of Figure 5. Considering first the Baseline condition, a one-way ANOVA for familiarity scores among studied word conditions (Silent, Aloud 1, and Aloud 2) showed an overall significant effect, $F(2,50) = 7.68$, $MSE = .01$, $p < .01$, $\eta_p^2 = .24$. Follow-up comparisons revealed that words read aloud in the Aloud 1 and Aloud 2 conditions had higher familiarity scores than words read silently, $t(25) = 3.40$, $p < .01$, d $= 1.36$ and $t(25) = 2.57$, $p < .05$, $d = 1.03$ respectively. Familiarity scores did not differ between Aloud 1 and Aloud 2 items, $t(25) = 1.43$, $p = .16$, $d = 0.57$. Hence, a production effect was observed in both the Aloud 1 and Aloud 2 conditions for familiarity scores and was of equivalent magnitude. Familiarity scores for these two conditions would be combined into a single familiarity score for subsequent comparison to the Mumble condition.

Considering the Mumble condition, a one-way ANOVA analysing familiarity scores among studied word conditions (Silent, Mumbled, and Aloud) showed an overall significant effect, $F(2,50) = 6.58$, $MSE = .02$, $p < .01$, $\eta_p^2 = .21$. Follow-up analyses revealed higher familiarity scores for aloud items compared to silent items, $t(25) = 3.26$, $p < .01$, $d = 1.30$, hence a production effect was observed. Mumbled items had lower familiarity scores than aloud items, $t(25) = 2.47$, $p < .05$, $d = 0.99$, and did not have higher familiarity scores than silent items, $t(25) = 0.93$, p $= .36$, $d = 0.37$. Hence, mumbled items did not show a significant production effect in familiarity scores.

Comparing the Mumble to the Baseline condition, familiarity scores for aloud and silent items did not significantly decline between these conditions, both t's < 0.50, p's > .62. False familiarity scores between these two conditions were not significantly different, $t(25) = 0.71$, $p = .48$, $d = 0.28$, meaning that familiarity scores could be directly compared and interpreted between the Baseline and Mumble condition. The presence of mumbled items thus did not impact the familiarity scores of aloud or silent items.

Discussion

The results of Experiment 5 replicate our critical findings from Experiments 1 and 2 insomuch as the recollection rates of silent items declined in the Mumble condition compared to the Baseline condition, whereas the recollection rates of aloud items did not change between these conditions. This finding once again demonstrates that recollection for silent items is impaired in the presence of additional distinctive encoding conditions.

Interestingly a mnemonic benefit for mumbled words was not observed in the overall hit rates like in Experiment 2, but rather only in the recollection rates. In Experiment 2 the mean hit rate for mumbled words was .62 (*SE* = .03) and for silent words was .56 (*SE* = .02). In Experiment 5 the mean hit rate for mumbled words was .62 (*SE* = .04) and for silent words was .58 (*SE* = .04). Numerically then, the benefit produced by mumbled words in Experiment 2 and 5 was very similar. The failure to find a significant mnemonic benefit for mumbled words here in overall hit rates may be the result of a power issue. That said, the mnemonic benefit of mumbling is clearly smaller than that of a true production effect, and this was already noted in Experiment 2. Overall, the cost to silent items manifests more clearly as a cost to recollection.

Hit rates aside, Experiment 5 does replicate our key finding related to recollective costs for silent items when mumbled items are present. More so, by virtue of being a within-subjects replication of Experiments 1 and 2, Experiment 5 also addresses any concerns we might have had that the results of our analysis of Experiments 1 and 2 relied on a cross-experiment comparison. Experiment 5 addressed another methodological issue as well, namely the number of cues use in the Baseline condition. In Experiment 1 two separate cues were used to cue the Aloud 1 and Aloud 2 condition. Experiment 5 used only a single cue for aloud items, addressing the minor (but valid) concern that utilising two separate cues for aloud items in Experiment 1 may have been perceived as unusual by participants and led them to behave in an unusual manner. As the results of the Baseline condition closely mirror those of Experiment 1, it seems unlikely that the two cues in Experiment 1 made a notable difference.

Experiment 6: A within-subjects replication of importance

As stated previously, Experiment 6 seeks to replicate Experiments 1 and 3 in a within-subjects paradigm, with a Baseline and Important condition respectively. It therefore uses the same design and logic of Experiment 5; in the Baseline condition participants study aloud and silent items before being tested and in the Important condition participants study aloud, "important", and silent items before being tested.

Method

Participants

A total of 25 students from the State University of New York at Geneseo received extra credit in exchange for taking part in Experiment 6.

Stimuli and apparatus

The stimuli were the same as those used in Experiments 1 and 3.

Procedure

The procedure of Experiment 6 was identical to that of Experiment 5 except that instead of a Baseline and Mumble condition, Experiment 6 consisted of a Baseline and Important condition. In the Important condition participants studied aloud, "important", and silent words at study.

The order of the Baseline and Important conditions was randomly counterbalanced, just like in Experiment 5. This time, 12 participants were in the Baseline-Important ordering and 14 were in the Important-Baseline ordering. The order of the conditions had no significant impact on the results and data was therefore collapsed across participants into Baseline and Important conditions respectively.

Results

The results of Experiment 6 can be seen in Figure 6. Hit rates, recollection rates, and familiarity scores are plotted separately for the Baseline and Important experimental conditions, and for the aloud, silent, and important conditions from each.

Hit Rates. Hit and false alarm rates are shown in top panels of Figure 6. Considering first the Baseline condition, a one-way ANOVA among *old* response rates for the three studied item conditions (Silent, Aloud 1, and Aloud 2) revealed a significant overall effect, $F(2,50) = 14.29$, $MSE = .01$, $p < .01$, $\eta_p^2 = .36$. Follow-up analyses revealed more hits for Aloud 1 and Aloud 2 items compared to the silent items, $t(25) = 4.95$, $p < .01$, $d = 1.98$ and $t(25) = 4.53$, $p < .01$, $d = 1.81$ respectively. Hit rates did not differ between Aloud 1 and Aloud 2, $t(25) = 1.20$, $p = .24$, $d =$ 0.48. Hence, a production effect was observed in both the Aloud 1 and Aloud 2 conditions and was of equivalent magnitude. Hit rates for these two conditions would be combined into a single hit rate for subsequent comparison to the Important condition.

Considering the Important condition, a one-way ANOVA among *old* response rates for the three studied item conditions (Silent, Important, and Aloud) revealed a significant overall effect, $F(2,50) = 16.25$, $MSE = .01$, $p < .01$, $\eta_p^2 = .39$. Follow-up analyses revealed more hits for aloud items compared to silent items, $t(25) = 7.04$, $p < .01$, $d = 2.82$, hence a production effect was observed. Important items also had a higher hit rate than silent items, $t(25) = 4.31$, $p < .01$, $d = 1.72$. Hit rates did not significantly differ between important and aloud items, $t(25) = 0.56$, $p = .58$, $d = 0.22$. Hence, important items showed a significant production effect and one of equivalent magnitude to that of aloud items.

Comparing the Important to the Baseline condition, hit rates for aloud items did not significantly decline between these conditions, $t(25) = 1.00$, $p = .33$, $d = 0.4$, however, hit rates for silent items did significantly decline, $t(25) = 2.65$, $p < .05$, $d = 1.06$. False alarm rates between these two conditions were not significantly different, $t(25) = 1.25$, $p = .22$,

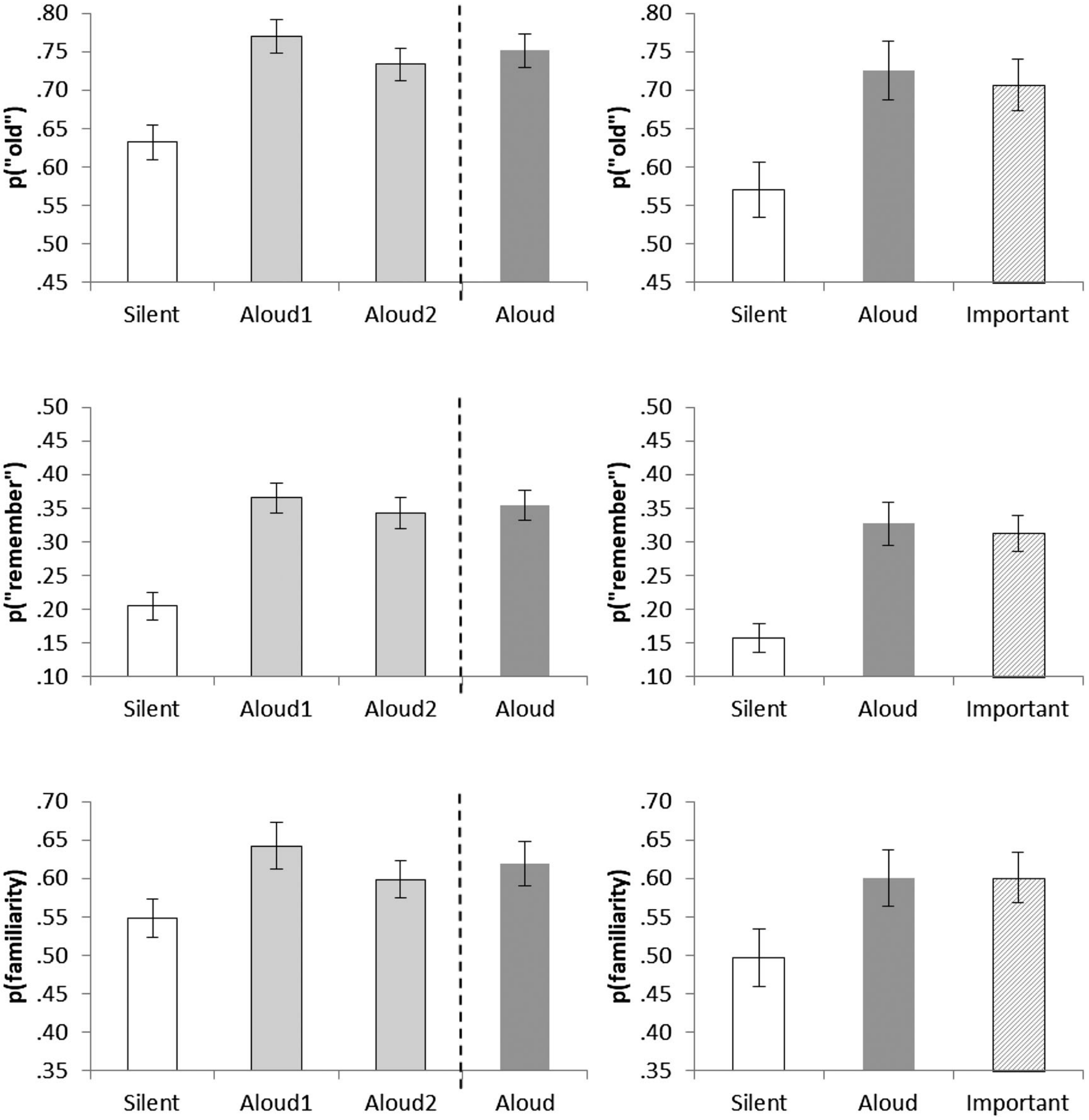

Figure 6. Mean hit and false alarm rates, recollection rates, and familiarity scores for the baseline and experimental condition of Experiment 6. For the baseline condition, Aloud 1 and Aloud 2 results were combined into a single Aloud condition. Error bars represent standard errors of the mean. For the Baseline condition mean false alarm rates were .33 (*SE* = .03), false recollection rates were .04 (*SE* = .02), and false familiarity rates were .30 (*SE* = .02). For the Important condition mean false alarm rates were .36 (*SE* = .03), false recollection rates were .04 (*SE* = .02), and false familiarity rates were .33 (*SE* = .03).

$d = 0.5$, meaning that hit rates could be directly compared and interpreted between the Baseline and Important condition. Hence, the presence of important items impaired the memorability of silent items in the Important condition.

Recollection. Recollection rates are shown in the middle panels of Figure 6. Considering first the Baseline condition, a one-way ANOVA for recollection scores among studied word conditions (Silent, Aloud 1, and Aloud 2) showed an overall significant effect, $F(2,50) = 19.66$, $MSE = .01$, $p < .01$, $\eta_p^2 = .44$. Follow-up comparisons revealed that words read aloud in the Aloud 1 and Aloud 2 conditions were more recollectable than words read silently, $t(25) = 5.42$, $p < .01$, $d = 2.17$ and $t(25) = 5.05$, $p < .01$, $d = 2.02$ respectively. The rate of recollection responses did not differ between Aloud 1 and Aloud 2 items, $t(25) = 0.90$, $p = .38$, $d = 0.36$. Hence, a production effect was observed in both the Aloud 1 and Aloud 2 conditions and was of equivalent magnitude. Recollection rates for these two conditions would be combined into a single recollection rate for subsequent comparison to the Important condition.

Considering the Important condition, a one-way ANOVA analysing recollection scores among studied word conditions (Silent, Important, and Aloud) showed an overall significant effect, $F(2,50) = 13.44$, $MSE = .01$, $p < .01$, $\eta_p^2 = .35$. Follow-up analyses revealed higher recollection rates for aloud items compared to silent items, $t(25) = 6.55$, $p < .01$, $d = 2.62$, hence a production effect was observed. Important items also had higher recollection rates than silent items, $t(25) = 4.58$, $p < .01$, $d = 1.83$, and did not significantly differ from aloud items, $t(25) = 0.33$,

$p = .75$, $d = 0.13$. Hence, important items showed a significant production effect in recollection responses, and one of equivalent magnitude to that of aloud items.

Comparing the Important to the Baseline condition, recollection rates for aloud and did not significantly decline between these conditions, $t(25) = 0.88$, $p = .39$, $d = 0.35$, however recollection rates for silent items did decline, $t(25) = 2.31$, $p < .05$, $d = 0.92$. False recollection rates between these two conditions were not significantly different, $t(25) = 0.46$, $p = .65$, $d = 0.18$, meaning that recollection rates could be directly compared and interpreted between the Baseline and Important condition. The presence of important items thus impaired the recollection of silent items but not aloud items.

Familiarity. Familiarity scores are shown in the bottom panels of Figure 6. Considering first the Baseline condition, a one-way ANOVA for familiarity scores among studied word conditions (Silent, Aloud 1, and Aloud 2) showed an overall significant effect, $F(2,50) = 3.90$, $MSE = .02$, $p < .05$, $\eta_p^2 = .14$. Follow-up comparisons revealed that words read aloud in the Aloud 1 condition had higher familiarity scores than words read silently, $t(25) = 2.95$, $p < .01$, $d = 1.18$, however, words in the Aloud 2 condition did not have higher familiarity scores than words read silently, $t(25) = 1.87$, $p = .07$, $d = 0.75$. The familiarity scores for the Aloud 1 and Aloud 2 conditions did not significantly differ from each other however, $t(25) = 1.06$, $p = .30$, $d = 0.42$. Given that this nonsignificant effect was marginal and that the Aloud 1 and Aloud 2 conditions were arbitrary (i.e., there was no functional distinction between them from the subject's perspective and words were randomly assigned to them by the researcher), we considered this unexpected finding to be a likely Type II error. Consistent with this interpretation, collapsing Aloud 1 and Aloud 2 into a single aloud condition showed higher familiarity scores than the silent condition, $t(25) = 3.17$, $p < .01$, $d = 1.27$. Hence, overall a production effect was observed in for the aloud items. Familiarity scores for these two conditions would be combined into a single familiarity score for subsequent comparison to the Important condition.

Considering the Important condition, a one-way ANOVA analysing familiarity scores among studied word conditions (Silent, Important, and Aloud) showed an overall significant effect, $F(2,50) = 6.24$, $MSE = .02$, $p < .01$, $\eta_p^2 = .20$. Follow-up analyses revealed higher familiarity scores for aloud items compared to silent items, $t(25) = 3.35$, $p < .01$, $d = 1.34$, hence a production effect was observed. Important items also showed higher familiarity scores than silent items, $t(25) = 2.94$, $p < .05$, $d = 1.18$. Familiarity scores for aloud and important items did not significantly differ, $t(25) = 0.02$, $p = .98$, $d = 0.01$. Hence, important items showed a significant production effect in familiarity scores, and one of equivalent magnitude to that of aloud items.

Comparing the Important to the Baseline condition, familiarity scores for aloud items did not significantly decline between these conditions, $t(25) = 0.57$, $p = .58$, $d = 0.23$. Familiarity scores for silent items marginally declined between these conditions, $t(25) = 1.96$, $p = .06$, $d = 0.78$. False familiarity scores between these two conditions were not significantly different, $t(25) = 1.19$, $p = .24$, $d = 0.48$, meaning that familiarity scores could be directly compared and interpreted between the Baseline and Important condition. The presence of important items thus did not impact the familiarity scores of aloud items but may have impaired the familiarity scores of silent items.

Discussion

The results of Experiment 6 replicate the key findings of Experiments 1 and 3. Specifically, the presence of "important" words at study impaired both the overall hit rates and the recollection of silent words (compared to baseline). Hit rates and recollection rates for aloud words were unaffected. This finding replicates our prior experiments and by doing so in a within-subjects design affirms that the conclusions drawn from our cross-experimental comparison of Experiments 1 and 3 were indeed valid.

Interestingly, Experiment 6 found that the familiarity scores of silent items may have declined in the presence of "important" items. Considering that out of all of our experiments Experiment 6 is the only experiment to demonstrate this finding, it must be taken with some caution. Nonetheless, perhaps the familiarity scores of silent items can indeed be impacted by the presence of other distinctive encoding conditions. The reason this effect may not have been observed thus far is that it could be a more subtle effect than the recollection costs we have been able to consistently observe. Pursuing this possibility in future work may be worthwhile, however, for our purposes, both because our theoretical account focuses on the recollective costs to silent items and because recollective costs have been more consistently observed across all experiments, we will not interpret this marginal effect on familiarity further.

General discussion

Past studies have shown that silent items are less memorable when studied in the presence of aloud items than when studied alone (Bodner et al., 2014; Bodner & Taikh, 2012). This effect appeared to extend in Forrin et al. (2012), in that the memory for silent items appears to be further impaired in the presence of aloud items. However, Forrin et al. never presented a 2-condition aloud/silent experiment, making this inference rather indirect. From their data, two possibilities existed: the aloud-cost idea that aloud items may by themselves impair the memorability of silent items (with other distinctive manipulations like mumbling, written, spelling, etc. having no effect), or the memorability of silent items may decline as a function of the number of distinctive encoding conditions. Furthermore, it was unclear if the memorability for aloud items could itself be impaired by the presence of another distinctive encoding condition or whether only silent items were vulnerable to this memorability cost.

Across 6 experiments, we investigated these issues. Though past studies focused on the costs that emerged in overall hit rates (Bodner et al., 2014; Bodner & Taikh, 2012), our theoretical model suggested that we consider recollection rates specifically. We found that although aloud items may introduce a general cost to the memorability of silent items, replacing some aloud items with other distinctive encoding conditions at study further impairs the memorability of silent items (Experiments 2, 3, 5, and 6); this impairment primarily affects recollections over familiarity (Experiments 2, 3, 5, and 6); aloud items are not susceptible to a memorability impairment by the introduction of another distinctive encoding condition (Experiments 2–6); an encoding condition that is very similar to speaking aloud (i.e., mouthing) does not appear to affect the memorability of silent items, or does so to a less consistent degree (Experiment 4); and finally, the other distinctive encoding conditions that replace some aloud words can be *less* memorable than the aloud words they replace and yet cause a *greater* cost to the memorability of silent words (Experiments 2, 3, 5 and 6). Our findings have broader implications for models of the production effect as well as our understanding of the effectiveness of production. We will consider these two issues in turn but first we will consider the specifics of our results in more detail.

Interpreting the cost to recollection vs. the cost to hits

An interesting element of our results that we have yet to emphasise is the fact that the recollection of silent items was regularly impaired by the introduction of a second distinctive encoding condition, yet the overall hit rates for silent items was less often observed. This data pattern suggests that the cost incurred by silent items is specifically related to the experience of conscious recollection, or at least there is a smaller cost to hits than to recollection. Indeed, as our model emphasises, when presented with a test probe we believe that participants explicitly search memory for evidence of having produced that item at study in one of the distinctive encoding conditions. Failure to find such evidence would be akin to failure to recollect the probe, resulting in participants feeling uncertain about that item. Of course, at this stage participants are free to base their final recognition decision on any other lingering factor, such as feelings of familiarity. When participants fail to recollect the encoding conditions of an item then, they may default to a familiarity-based response. This is indeed what researchers believe typically happens in the remember-know procedure and hence why the use of independent remember-know scores is advocated when analysing remember-know data (see Yonelinas, 2002 for a review).

So when a test probe's encoding condition cannot be recollected, participants may be defaulting to their sense of relative familiarity for the probe when making the final recognition decision. If the probability of recognising an item based on familiarity is greater than zero, then at least some of the items for which recollection has failed should still be recognised. Hence, a recognition failure will not always lead to a miss. Hit rates, by virtue of being based on recollection or familiarity, would thus be expected to decline more slowly than recollection rates. Interpreting our findings then, the presence of another distinctive encoding condition reduced participants' ability to recollect information about silent items (or mades them more skeptical of any noncriterion recollection that might have occurred for silent items), and yet had no effect on familiarity scores (i.e., the probability of a familiarity-based recognition). In cases where recollection thusly fails, every silent item would have some probability of still being recognised based on familiarity. As a result, when silent items suffered a cost to explicit memory, recollection rates significantly declines whereas hit rates declined at a slower, and as it turns out, mostly significant rate.

In sum, when looking at the costs for silent items in production effect designs, it is prudent to measure both recollection and familiarity, not just overall hit rates, as these metrics can provide a more sensitive and detailed analysis of how memory is changing. Had we not measured recollection and familiarity in our experiments, and focused solely on hit rates, we would not have had the capacity to detect the cost that silent items were incurring in the presence of other distinctive encoding conditions. Certainly costs occur for silent items in overall hit rates (see Bodner et al., 2014; Bodner & Taikh, 2012), but sometimes costs are more subtle or focused specifically on conscious, recollected memory, as we report in our experiments.

Consequences for models of production

The data across all of our experiments are consistent with a distinctiveness model of production that we have proposed. When presented with a test probe in a recognition test, participants search memory for evidence of any distinctive encoding condition from study (e.g., "aloudness", "mumbleness", "importantness", etc.). Failing to find evidence of more than one distinctive encoding condition results in less certainty in the test probe than failing to find evidence of only "aloudness". In mixed-lists with two distinctive encoding conditions participants may be less likely to accept silent test probes, or exhibit less explicit memory for silent test probes than in mixed-lists with only one distinctive encoding condition (i.e., aloud items). And to the degree that the two distinctive encoding conditions are not distinct from one another (i.e., aloud vs. mouthed), the cost to silent items is not exaggerated beyond that observed when only aloud and silent items are studied together.

The major alternative accounts we considered were an aloud-cost model and the lazy-reading hypothesis (cf. Begg & Snider, 1987; see MacLeod et al., 2010). Regarding the aloud-cost idea, we never observed the memorability of silent items improving when aloud items were removed. In fact, as we stated above, one of the more noteworthy empirical findings we report is that the costs to silent items actually increase when less memorable items replace some aloud items at study. This finding speaks strongly against the idea that aloud items are specially responsible for the cost to silent items and instead speaks to the idea that dynamics of retrieval must be at play instead.

Regarding the lazy-reading hypothesis, this account suggests that aloud items are perceived as more important than silent items, and hence, steal rehearsals from silent items at encoding, resulting in costs for silent items. In Experiment 3, we proposed two versions of this account. One version considered aloud items to be more important, whereas the other considered "important" items to be more important (both considered silent items to be least important). From these proposals, a simple interpretation of the lazy-reading hypothesis suggested that either silent items should incur no further costs to recollection in Experiment 3 compared to Experiment 1, or both aloud and silent items should incur costs to recollection in Experiment 3 compared to Experiment 1. Neither of these predictions came to pass (nor did they occur in the replication; Experiment 6). At the time we interpreted the results of Experiment 3 as opposing the lazy-reading hypothesis, but is it possible to reconcile these results with a more sophisticated account?

First and foremost we should acknowledge that some evidence consistent with a lazy-reading hypothesis was observed in Experiment 6. Namely, familiarity scores for silent items marginally declined in the Important condition compared to the Baseline condition. This suggests that the presence of "important" items may have degraded the encoding of silent items, while having no impact on the encoding of aloud items. Given that this finding was marginal and only observed in Experiment 6 however, another way to reconcile a lazy-reading-like account with the results of Experiment 3 (and 6) would be to appeal to both the overt behavioural action of production and the implied importance of production and the "important" words. That is, perhaps aloud and "important" words are both perceived as more important than silent words and so given explicit attention at encoding, but in addition the behavioural action of speaking aloud may provide independent distinctive information at encoding. Thus, both attention and action act to encode aloud items, allowing aloud items to remain relatively memorable in Experiment 3, even if "important" items stealing rehearsals from both aloud and silent items, resulting in costs to silent items becoming exaggerated in Experiment 3 compared to Experiment 1. Admittedly this account is a bit vague and can only speculate as to why aloud items suffered no discernable cost in memorability despite having rehearsals stolen, but it at least overs an avenue of consideration for lazy-reading type accounts. Namely, perhaps it is the interaction between attentional efforts and action that makes production so effective and could somehow protect it from the effect of other attentionally demanding encoding conditions at study. One possibility is that aloud items regularly steal attention from other items at study but that attention provides no further mnemonic benefit over the act of production.

Though it may be possible to come up with further versions of the lazy-reading hypothesis to encompass our results, at present these endeavours are somewhat speculative and not necessarily consistent with our experiments as a set. Instead, we argue that our distinctiveness account provides a good framework in which to both predict and understand the results of our experiments. In a broader sense, the distinctiveness framework that we have proposed connects to the wider memory literature in that it shares ideas with the source monitoring framework (Johnson, Hashtroudi, & Lindsay, 1993). Source monitoring is a more established and general framework of memory, which suggests that rather than memories being conceived of as abstract tag or labels, individuals strategically combine remembered aspects of experience (thoughts, images, feelings) with their knowledge, expectations, and schemas in order to dynamically construct memories at retrieval. From the source monitoring perspective, during a recognition memory test, participants can strategically search for "aloudness" information for any given test probe by trying to recall experiential elements of having said the word aloud (i.e., what it was like to move one's lips or hear that word spoken in one's own voice). Recalling that a test probe was indeed said aloud recently should be regarded as good evidence that the word may have been studied. Failing to recall that a test probe was said aloud would cause participants to strategically evaluate the test probe, and may cause them to weigh it as less likely to have been studied than if the "aloudness" information could be retrieved. In essence, our model fits precisely within the source-monitoring framework and offers a similar characterisation as to how memory retrieval occurs. Given that our model was able to correctly predict the impact of introducing another distinctive encoding condition at study, it may be fruitful for future researchers to more explicitly use the source-monitoring framework when considering the production effect.

On the effectiveness of production

Taken as a set, the results of Experiments 2–6 were predicted well by our distinctiveness model. The present study and our model therefore expands our current knowledge of the production effect by further elaborating why a cost must exist for silent items. Because silent items have no uniquely distinctive encoding information to search memory for, they will always incur costs when encoded and tested along with distinctive encoding conditions. Interestingly though, the present research also demonstrates how distinctive encoding conditions can be immune to those very costs, and how for instance, aloud and mumble items can be encoded together without impairing one another. Because the costs to silent items occur at retrieval, the relative distinctiveness of the distinctive encoding conditions do not affect one another (i.e., mumbled, "important", and mouthed never impaired the recollection or familiarity of aloud items).

The results of Experiment 3 and 6 are also especially interesting in the context of the production effect more generally, as participants were explicitly told that "important" items are the most important items to remember. Although "important" items were remembered better than silent items, there were not remembered better than aloud items. This finding not only rules out a simple lazy reading account (and as we saw, paired with Experiment 4, it also rules out a more sophisticated lazy reading account), but it suggests that when participants are given free reign to encode items in any way they chose, they do worse than simply reading aloud. In essence, better than trying to consciously and effortfully remember a word, Experiments 3 and 6 suggest it is simpler and more effective to just say that word aloud.

The findings of Experiment 3 and 6 illustrate just how effective production is. Indeed, in the Introduction we noted that the robustness and reliability of the production is one of its most fascinating properties. Simply speaking aloud has often been shown to be a highly effective mnemonic. Indeed, in their examination, Forrin et al. (2012) examined whispering, spelling, writing, typing, and yet none of these manipulations was found to be more effective than simply reading aloud. Similarly, across our experiments examining mumbling, "important", and mouthing, we did not find any manipulation that was more effective than speaking aloud. One might wonder then, the costs to silent items may turn out differently if a condition that is more memorable than speaking aloud were introduced as the second distinctive encoding condition at study.

Finding manipulations that are comparable but more effective than production is a surprisingly difficult endeavour. As already mentioned, Forrin et al. (2012) compared typing, writing, and spelling to production and found them all to be less effective than production. In the original design of our experiments own, we had hoped that mumbling or "important" words might prove to be more memorable than speaking aloud. They did not. Recent work has found evidence that singing may be more memorable than reading aloud (Hassall, Quinlan, Turk, Taylor, & Krigolson, 2016; Quinlan & Taylor, 2013), however, despite piloting several production effect designs that include singing as a second encoding condition, we have yet to replicate this finding. From our pilot experiments so far, singing has not to produced more memorable traces than simply reading aloud in 3-condition study phases, but interestingly, singing may impair the recollection *and* familiarity response of both aloud *and* silent words. We are currently pursuing a line of research to attempt to clarify these preliminary findings, but they suggest that in some circumstances it may be possible to affect the memorability of aloud items. Whether this is an encoding or retrieval effect however, remains to be seen.

Another avenue to pursue in search of an encoding condition more effective than production might have been the generation effect. The generation effect is a phenomenon whereby participants are required to generate items from cues (Slamecka & Graf, 1978; see Bertsch, Pesta, Wiscott, & McDaniel, 2007 for a review). For example, giving participants the word HOT and asking them to generate an antonym that begins with C often leads participants to generate COLD. Participants will demonstrate better memory for COLD if it is generated in this manner than if the word COLD is simply provided to them. The generation effect is believed to enhance memory because it requires participants to actively process the cue, and engage in mental effort to search semantic memory for an appropriate response. These cognitive acts are effortful, and ultimately memorable, leading to superior memory for items that must be generated as opposed to those that are just provided. However, despite the complexity of generation, it regularly leads to a 9% increase in hit rates (see Bertsch et al., 2007). Compared to generation, reading words aloud is considered to be a simple and commonplace activity, yet it reliably enhances memory, frequently leading to an improvement of approximately 15%[6] in recognition memory accuracy. In other words, the mnemonic benefit of generation is at best comparable to that of production, yet it is a more sophisticated cognitive act than production.

Suffice to say that coming up with manipulations that are simple yet more effect than production is difficult. If we did find a manipulation that reliably led to better memory than speaking aloud, our distinctiveness account would predict that the memorability of aloud items would not be impaired by being studied and tested along with this new, highly effective mnemonic. Specifically, because our account proposes that the recollection of silent items is impaired because silent test probes fail to produce memories for having said the word aloud or engaged in the other distinctive encoding, the memorability for aloud words should only be affected by the ease of retrieving "aloudness" information for aloud test probes. Thus, items studied with a distinctive encoding technique at study should be relatively immune from the contents of the other encoding conditions, so long as the other encoding conditions do not interfere with one another. This final point is interesting because Forrin et al. (2012) do report a 3-condition mixed-list design where aloud and silent items are studied along with whispered items. This condition sees the hit rates for aloud items decline from an average score of .74 (when aloud and silent items were studied with written or mouthed words) down to .66. In this one condition, the ability to remember "aloudness" and "whisperedness" might have interfered with one another, and hence, the only real observable case of a decline in hit rates for aloud items occurred. It is also equally possible that the decline in the magnitude of the production effect was simply a measurement error, as without enough production experiments, eventually the magnitude will change, purely by chance. Nonetheless, there may be a fruitful future avenue of research in exploring the confusability of "aloudness" information with other vocalised responses to see the memorability of aloud items decline as a result.

Conclusion

Across 6 experiments we have shown that while production is a reliable and effective way to enhance memory, recollection for silent items is impaired in the presence of distinctive encoding conditions. The number of other encoding conditions and the degree to which the other encoding conditions are distinct from one another impacts the amount of impairment observed. These effects appear to arise due to participants searching their memories at test for evidence that a probe was studied in a distinctive condition. The distinctiveness model provides a reliable framework for understanding how and why the production effect occurs, why it incurs costs on silent items, and when those costs may be mitigated.

Notes

1. Admittedly, the issue of whether participants are searching for distinctive information at test in pure-lists is a bit murky. Participants do not self-report relying on distinctiveness strategies at test more often in mixed-list experiments than in pure-list experiments (Fawcett & Ozubko, 2016), which seems to create a mystery as to why the production effect is smaller in pure-lists than mixed-lists. However, regarding the memorability of silent items specifically, in pure-lists there simply is no "aloud" information to search for. That is, participants are clearly aware of the fact that all items at study were silent, and hence searching for "aloud" information at test does not make sense. It thus, seems improbable that participants would be engaging in this kind of distinctiveness search for silent items in pure-lists, and therefore fits with the account that we are proposing that the costs for silent items in mixed-lists are associated with searches at test for "aloudness" and other distinctive types of encoding information.
2. It should be noted that we intentionally wanted to include two aloud conditions and one silent condition because in Experiments 2, 3, and 4, we would be including one aloud condition and one other distinctive condition (mouthed words, mumbled words, or "important" words). These other distinctive conditions are more active than simply reading a word silently, and so to ensure Experiment 1 was as equivalent to later experiments as possible in terms of complexity, it makes sense to have two active encoding conditions rather than two silent conditions and one active encoding condition. This would allow us more assurance that the remember-know ratings from Experiment 1 were derived under similarly complex encoding conditions in Experiment 1 as in the other experiments.
3. See Appendix A of Ozubko and Seli (2016) for a detailed description of our remember/know instructions. Note that the labels "re-experience" and "familiar" were used to explain recollection and familiarity to participants, rather than "remember" and "know". The "re-experience" and "familiar" labels were used because they have been found to be more effective at intuitively conveying the distinction between recollection and familiarity than the labels "remember" and "know" (Ozubko & Seli, 2016)
4. It should be further noted that no reported patterns change when reporting comparisons between aloud conditions of a later experiment vs. either Aloud 1 or Aloud 2 of Experiment 1. Thus, using the average of the two aloud conditions simplifies later analyses.
5. Indeed, power analyses for this effect indicate that we would need more than 140 participants per experiment in order to have a power level of .80 between Experiments 1 and 4.
6. Our estimate of the typical production effect size comes from a convenience sample of studies reporting a standard, 2-condition (aloud/silent), within-subjects production effect design that utilized a recognition memory test. This dataset included 1 experiment from Dodson and Schacter (2001), 2 experiments from Fawcett and Ozubko (2016), 2 experiments from Hopkins and Edwards (1972), 1 experiment from Icht et al. (2019), 2 experiments from Lin and MacLeod (2012), 1 experiment from MacLeod et al. (2010), and 2 experiments from Ozubko, Gopie, et al. (2012). The mean difference in hit rates between aloud and silent words was .15 (*SD* = .06). Mean differences ranged from .07 to .27 across all experiments.

Acknowledgements

We wish to thank Sarah Simon, Justin Leonti, Sophia Phillips, Ezekiah Bacchus, Lucia Verrelli, and Elsie Martin for their assistance in collecting data in the reported experiments.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Jason D. Ozubko http://orcid.org/0000-0003-0351-0957

References

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

Bertsch, S., Pesta, B., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition*, *35*(2), 201–210.

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

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). New York, NY: Worth Publishers.

Bodner, G. E., & Taikh, A. (2012). Reassessing the basis of the production effect in memory. *Journal of Experimental Psychology:*

Learning, Memory, and Cognition, *38*(6), 1711–1719. doi: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.

Castel, A. D., Rhodes, M. G., & Friedman, M. C. (2013). Predicting memory benefits in the production effect: The use and misuse of self-generated distinctive cues when making judgments of learning. *Memory & Cognition*, *41*(1), 28–35. doi:10.3758/s13421-012-0249-6

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

Ekstrand, B. R., Wallace, W. P., & Underwood, B. J. (1966). A frequency theory of verbal-discrimination learning. *Psychological Review*, *73* (6), 566–578. doi:10.1037/h0023876

Fawcett, J. M. (2013). The production effect benefits performance in between-subject designs: A meta-analysis. *Acta Psychologica*, *142* (1), 1–5. doi: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), doi:10.1037/cep0000089

Forrin, N. D., MacLeod, C. M., & Ozubko, J. D. (2012). Widening the boundaries of the production effect. *Memory and Cognition*, *40*(7), doi: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*, 110–119.

Hassall, C. D., Quinlan, C. K., Turk, D. J., Taylor, T. L., & Krigolson, O. E. (2016). A preliminary investigation into the neural basis of the production effect. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *70*(2), 139–146. doi:10.1037/cep0000093

Hopkins, R. H., Boylan, R. J., & Lincoln, G. L. (1972). Pronunciation and apparent frequency. *Journal of Verbal Learning and Verbal Behavior*, *11*(1), 105–113. doi:10.1016/S0022-5371(72)80066-5

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

Icht, M., Bergerzon-Biton, O., & Mama, Y. (2019). The production effect in adults with dysarthria: Improving long-term verbal memory by vocal production. *Neuropsychological Rehabilitation*, *29*(1), 131–143. doi:10.1080/09602011.2016.1272466

Johnson, M. K., Hashtroudi, S., & Lindsay, S. D. (1993). Source monitoring. *Psychological Bulletin*, *114*(1), 3–28.

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*(1), 300–305.

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.

Khader, P., Burke, M., Bien, S., Ranganath, C., & Rösler, F. (2005). Content-specific activation during associative long-term memory retrieval. *Neuroimage*, *27*(4), 805–816.

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

Lin, O., & MacLeod, C. M. (2012). Aging and the production effect: A test of the distinctiveness account. *Canadian Journal of Experimental Psychology*, *66*(3), 212–216.

MacDonald, P. A., & MacLeod, C. M. (1998). The influence of attention at encoding on direct and indirect remembering. *Acta Psychologica*, *98*, 291–310. doi:10.1016/s0001-6918(97)00047-4

Macleod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, *109*(2), 163–203. doi:10.1037//0033-2909.109.2.163

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. doi:10.1037/a0018785

Mama, Y., & Icht, M. (2019). Production effect in adults with ADHD with and without methylphenidate (MPH): Vocalization improves verbal learning. *Journal of the International Neuropsychological Society*, *25* (2), 230–235. doi:10.1017/S1355617718001017

Nyberg, L., Habib, R., Mcintosh, A. R., & Tulving, E. (2000). Reactivation of encoding-related brain activity during memory retrieval. *Proceedings of the National Academy of Sciences*, *97*(20), 11120–11124.

Ozubko, J. D., Gopie, N., & MacLeod, C. M. (2012). Production benefits both recollection and familiarity. *Memory & Cognition*, *40*(3), 326–338. doi: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. doi: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. doi: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*(5), doi:10.1080/09658211.2013.800554

Ozubko, J. D., & Seli, P. (2016). Forget all that nonsense: The role of meaning during the forgetting of recollective and familiarity-based memories. *Neuropsychologia*, *90*, 136–147. doi:10.1016/j.neuropsychologia.2016.06.026

Quinlan, C. K., & Taylor, T. L. (2013). Enhancing the production effect in memory. *Memory*, *21*(8), 904–915. doi:10.1080/09658211.2013.766754

Rotello, C. M., Macmillan, N. A., Hicks, J. L., & Hautus, M. J. (2006). Interpreting the effects of response bias on remember-know judgments using signal detection and threshold models. *Memory & Cognition*, *34*(8), 1598–1614. doi:10.3758/BF03195923

Rotello, C. M., Macmillan, N. A., Reeder, J. A., & Wong, M. (2005). The remember response: Subject to bias, graded, and not a process-pure indicator of recollection. *Psychonomic Bulletin & Review*, *12*(5), 865–873. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16524003

Skinner, E. I., Grady, C. L., & Fernandes, M. A. (2010). Reactivation of context-specific brain regions during retrieval. *Neuropsychologia*, *48*(1), 156–164. doi:10.1016/j.neuropsychologia.2009.08.023

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

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology*, *21* (1), 1–12.

Vaidya, C. J., Zhao, M., Desmond, J. E., & Gabrieli, J. D. E. (2002). Evidence for cortical encoding specificity in episodic memory: Memory-induced re-activation of picture processing areas. *Neuropsychologia*, *40*(40), 2136–2143.

Waldhauser, G. T., Braun, V., & Hanslmayr, S. (2016). Episodic memory retrieval functionally relies on very rapid reactivation of sensory information. *The Journal of Neuroscience*, *36*(1), 251–260. doi:10.1523/JNEUROSCI.2101-15.2016

Wheeler, M. E., Peterson, S. E., & Buckner, R. L. (2000). Memory's echo: Vivid remembering reactivates sensory-specific cortex. *Proceedings of the National Academy of Sciences*, *97*(20), 11125–11129.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*(3), 441–517. doi:10.1006/jmla.2002.2864

Yonelinas, A. P., Dobbins, I. G., Szymanski, M. D., Dhaliwal, H. S., & King, L. (1996). Signal-detection, threshold, and dual-process models of recognition memory: ROCs and conscious recollection. *Consciousness and Cognition*, *5*(4), 418–441. doi:10.1006/ccog.1996.0026