Journal of Memory and Language 140 (2025) 104584

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

# Production increases both true and false recognition

Xinyi Lu [a,**,1], Jianqin Wang [b,*,2], Colin M. MacLeod [a,3]

[a] Department of Psychology, University of Waterloo, Waterloo, Ontario, Canada
[b] Department of Psychology, Fudan University, Shanghai, China

ARTICLE INFO

*Keywords:*
Production effect
Recognition
False memory
Fuzzy trace theory
Relational processing
Context

ABSTRACT

The production effect is the finding that reading information aloud enhances memory relative to reading information silently. In five experiments, we examined the influence of production on true and false memory in the DRM paradigm. In Experiments 1a, 1b, 3a, and 3b, reading aloud was compared to reading silently. In Experiment 2, reading aloud was compared to reading silently while hearing the words spoken by another voice. In all experiments, reading aloud consistently resulted in better recognition of studied words, but it also consistently resulted in more false alarms to unstudied lures that were semantically related to the studied words. We advance an argument based on current theoretical accounts of false memory wherein reading aloud selectively enhances relational or gist processing—the encoding of shared features across items—rather than item or verbatim processing—the encoding of specific details of individual items. This selective enhancement could be for the shared semantic network (gist), for the shared context of reading aloud (misattributed source memory), or for both. Thus, the benefit of production is best captured by the combination of adding new features (contextual information) together with enriching existing features (semantic information).

## Introduction

Words that have been read aloud are remembered better than words that have been read silently (MacLeod et al., 2010). This robust phenomenon—the *production effect*—also extends beyond speaking to production via writing, mouthing, whispering, and even singing (e.g., Forrin et al., 2012; Quinlan & Taylor, 2013, 2019), and beyond individual words to longer texts (Ozubko et al., 2012b; Roberts et al., 2024). MacLeod and Bodner (2017) provide a brief review of this burgeoning literature.

The production effect has primarily been explained as due to the distinctive processing applied to the produced items (Conway & Gathercole, 1987; Forrin et al., 2012; MacLeod et al., 2010; MacLeod & Bodner, 2017). According to the *relative distinctiveness* account, the act of producing an item during encoding results in the formation of an item-specific, distinctive record that is then useful during retrieval. Produced items are distinctive in the sense that they have additional, item-specific, production-associated features encoded in memory—including auditory and articulatory features (Forrin & MacLeod, 2018)—thereby differentiating them from items studied silently. Computational modeling of the production effect has captured this by storing in memory additional features for the produced items (Cyr et al., 2022; Jamieson et al., 2016; Kelly et al., 2022). Participants may also benefit from this distinctive record through use of a *distinctiveness heuristic* at test (Dodson & Schacter, 2001), a strategy that involves trying to retrieve whether a word on the test was produced at study. Remembering having spoken a word during the study phase is an additional way to verify that it was indeed studied.

An alternative explanation of the production effect posits that the produced items are more strongly encoded than the silent items (Bodner & Taikh, 2012; Fawcett, 2013; Fawcett & Ozubko, 2016). Under this *strength* account, the act of producing an item simply results in a stronger memory representation being formed during encoding. Fawcett and Ozubko (2016) provided evidence for influences of both distinctiveness and strength (see also Ozubko et al., 2012a). They used two approaches—Remember/Know judgments (see Gardiner, 1988) and receiver

* Corresponding author at: 200 University Avenue West, Waterloo, ON N2L 3G1, Canada
** Corresponding author at: 220 Handan Rd, Yangpu District, Shanghai 200437, China.
*E-mail addresses:* xinyilu@stanford.edu (X. Lu), wjq@fudan.edu.cn (J. Wang).
[1] https://orcid.org/0000-0001-6898-979X.
[2] https://orcid.org/0000-0003-4542-6518.
[3] https://orcid.org/0000-0002-8350-7362.

https://doi.org/10.1016/j.jml.2024.104584
Received 30 July 2023; Received in revised form 3 November 2024; Accepted 5 November 2024
Available online 16 November 2024

operating characteristic (ROC) confidence ratings (Yonelinas, 1994, 1997)—to dissociate the effect of production on recollection versus familiarity. Computational modeling of the production effect has usually modeled strength by increasing the likelihood that a feature is accurately copied into memory for the produced items (Jamieson et al., 2016; Kelly et al., 2022).

In the production effect, distinctiveness is viewed as operating primarily via recollective processes because the produced features provide a basis for a recollection boost during retrieval. In contrast, memory strength could reflect familiarity as well as recollection. Fawcett and Ozubko (2016; see also MacLeod & Bodner, 2017) have argued that familiarity alone underlies the small advantage for production seen in between-subjects, pure-list manipulations—a strength effect. In contrast, both familiarity *and* recollection underlie the considerably larger effect seen in within-subject, mixed-list manipulations—a strength effect plus a larger distinctiveness effect.

### *The phenomenon of false memory*

In the present investigation, we set out to explore the influence of production in the Deese-Roediger-McDermott (DRM) paradigm (Deese, 1959; Roediger & McDermott, 1995), the widely used procedure for inducing false memories for words. In this paradigm, participants study a list of words (e.g., *bed, rest, awake…*), all of which are semantically related to a non-presented word (e.g., *sleep*), known as the critical lure. Although it is never presented during study, participants nonetheless show substantial false recognition of the critical lure; indeed, this is even seen when attention to the studied words is very limited (Dodd & MacLeod, 2004). Moreover, false memories experienced in the DRM paradigm are often subjectively compelling. For instance, when participants are asked to distinguish whether they *Remember* or *Know* that a word was studied (Tulving, 1985; Gardiner, 1988), false recognition of the critical lure is often experienced in the "Remember" sense (e.g., Payne et al., 1996; Roediger & McDermott, 1995): The critical lure is not just familiar but actually seems to be recollected as having been studied.

One of the major theoretical accounts of false memory, activation monitoring theory (Roediger et al., 2001), posits that false recognition is a result of the interplay of automatic spreading activation and a more controlled source-monitoring process. As each word is presented, activation spreads to associated words that are connected in a semantic network (Anderson, 1983; Collins & Loftus, 1975). Because the critical lure in the DRM paradigm is, by design, strongly associated with all of the presented words, it is repeatedly activated during encoding of them. Source monitoring involves participants making attributions about the source of the aforementioned activation (i.e., distinguishing between words that were actually presented versus those that were merely activated but not presented). Thus, false recognition of the critical lure occurs when that word is activated and the participant mistakenly attributes the source of this activation to that word having been presented during study.

A second major theoretical account, fuzzy-trace theory (Reyna & Brainerd, 1995), proposes that false recognition is a result of two types of memory traces that are created in parallel during encoding: gist traces and verbatim traces. Gist traces contain the general meaning or theme of the information whereas verbatim traces contain item-specific, surface information (Brainerd et al., 2001). In the DRM paradigm, despite never having been presented during study, the critical lure is nonetheless repeatedly semantically cued by the presented words, each of which was selected for its association with the critical lure. Of course, the critical lure also is highly consistent with the general theme of the list. Thus, fuzzy-trace theory considers false recognition in the DRM paradigm to be based predominantly on gist traces, with true recognition based on both gist and verbatim traces. Under some circumstances, it is also possible for verbatim traces to suppress false recognition in a process known as *recollection rejection* (Brainerd et al., 2003): New yet gist-consistent items presented at test, notably the critical lure, can be rejected as 'new' on the basis of lacking verbatim detail associated with retrieval of the true 'old' items.

In activation monitoring theory, false recognition is assumed to arise from strong semantic activation in conjunction with failures of source monitoring. In fuzzy-trace theory, false recognition is assumed to arise from strong gist traces in the absence of verbatim recall. Consequently, both of these frameworks assume that false recognition is primarily a result of the retrieval of semantic information shared by the critical lure and the presented items (i.e., the common activation in a semantic network, or the gist trace).

There is also a third kind of theoretical framework: global-matching models (Arndt, 2010; Hintzman, 1988; Shiffrin & Steyvers, 1997). These accounts propose that false recognition arises because critical lures are highly similar to the memory representations of the presented items. This similarity can be due to shared semantics but it can also be due to shared context. This account can explain why, when responding to the critical lure, participants often retrieve specific visual, auditory, and contextual details shared with actually studied items (Lampinen et al., 2005; Payne et al., 1996).

### *Production and false memory*

In the series of experiments reported here, our aim was to determine how production would influence false recognition in the DRM paradigm. As has been repeatedly shown (see MacLeod & Bodner, 2017), production should benefit memory for the actually studied words, but how would it affect the likelihood of falsely recognizing the critical (unstudied) lures?

Within a relative distinctiveness account, items read aloud are rendered more distinctive compared to items read silently. This suggests that it should be easier to reject the associated critical lures of words studied aloud via recollection rejection, given that they would lack these distinctive produced features in memory. Thus, this account would predict a primarily recollection-based increase in true recognition for words spoken aloud, but a decrease in false recognition, relative to words read silently.

A strength account would appear to make a different prediction. If strength primarily reflects familiarity (as argued by Fawcett & Ozubko, 2016), then for true recognition the words spoken aloud should be better recognized than the words read silently. But now, reading words aloud should also lead to greater false recognition than reading words silently. That is, a strength-based account of production would predict primarily familiarity-based increases in both true and false recognition.

Only two prior studies have investigated the influence of reading aloud on DRM false recognition rates. Dodson and Schacter (2001, Experiment 2) presented all lists visually and had participants read half of the lists aloud and hear a recorded voice speak the other half of the lists. They reported no difference in recognition of critical lures as a function of whether words had been read aloud or read silently plus heard. Culbreth and Putnam (2019) obtained the same result—no difference in false alarm rate to critical lures—when they used the more usual production procedure of comparing reading words aloud to reading them silently. Dodson and Schacter (2001, p. 158) compared the null result of reading aloud on false recognition in their within-subject experiment to their earlier finding of a lower false alarm rate for aloud vs. heard words when using a between-subjects manipulation (Dodson & Schacter, 2001, Experiment 1). They argued that in a between-subjects design, participants successfully used a distinctiveness heuristic (*I remember saying it aloud, so it must be old*), but in a within-subject design "there is no longer a particular kind of information that is solely diagnostic of a test item's oldness or newness" (p. 158).

Other studies also have examined encoding manipulations in the DRM procedure. Particularly relevant, Huff and Bodner (2013) showed less false recognition after item-specific encoding than after relational encoding. Huff et al. (2015) reviewed the influence of distinctive encoding on correct and false recognition, in particular summarizing

studies that showed less false recognition after generation of the list items than after reading the items, a finding confirmed by Huff et al. (2021). In these cases, it appears that emphasizing encoding of individual items in a DRM list resulted in less false recognition of critical lures than did emphasizing connections among items in the list. This different influence of encoding on false recognition than that observed by Dodson and Schacter (2001) and by Culbreth and Putnam (2019) was one of the motivations for us undertaking the present investigation.

### *Overview of the current investigation*

In the experiments reported here, we manipulated production within subject because that is where the larger effect of production is seen. Thus, a random half of the DRM lists were read aloud during study, with the other half read silently. We report a series of pre-registered experiments (Experiment 1a: https://osf.io/j3a6g; Experiment 1b: https://osf.io/whbax; Experiment 2: https://osf.io/swkaj; Experiment 3a: https://osf.io/xnzcv; Experiment 3b: https://osf.io/eg84a) that examined the effect of reading aloud versus silently on true and false memory. All experiments were approved by the University of Waterloo Research Ethics Board (protocol #41398).

In each experiment, participants were presented with ten 12-word DRM lists. In Experiments 1a and 1b, participants were to read all of the words in five of these DRM lists aloud and all of the words in the other five lists silently. In Experiment 2, participants were to read all of the words in five of the lists aloud and all of the words in the other five lists silently while listening to those words spoken by another voice. In these first three experiments, participants performed a recognition test incorporating a tripartite decision: Remember vs. Know vs. New (see Gardiner, 1988). Under this procedure, Remember judgments are taken to index recollection whereas Know judgments are taken to index familiarity (see Yonelinas, 2002, for a review). Experiments 3a and 3b were similar to Experiments 1a and 1b except that the recognition test involved a binary response (Old vs. New), collapsing Remember/Know into Old. We did this in part because it is how the Dodson and Schacter (2001) and the Culbreth and Putnam (2019) studies were conducted and in part to increase the generalizability of our study.

## Experiments 1a and 1b

### *Method*

Experiment 1a was conducted in person in the laboratory; Experiment 1b was conducted online and was remotely supervised by an experimenter to ensure that participants followed the reading instructions. The general procedure for the two experiments was intentionally very similar so they are described together.

**Participants.** 80 students from the University of Waterloo participated for course credit in each experiment. This pre-registered sample size was determined via a power analysis (Faul et al., 2007) using an effect size of $d = 0.34$ estimated from pilot testing, for the critical *t*-test of interest—the effect of production on false recognition. No participants took part in both experiments, and none had taken part in a previous production effect or DRM experiment from our laboratory.

**Stimuli.** The stimuli in the study phase consisted of ten DRM lists that had been used in previous studies (e.g., Roediger & McDermott, 1995; see Stadler et al., 1999, for norms for DRM materials). Each DRM list consisted of twelve words such as *bed, rest, awake,* and *tired*, all related to a non-presented critical lure such as *sleep*. For each participant, five lists were presented in a color (red/blue) indicating "read aloud" and the other five lists were presented in the other color indicating "read silently." The assignment of both list and color to reading condition, as well as the order of list presentation, was randomized for each participant. Each word was presented in lowercase 30-pt Arial font against a white background.

The recognition test list contained 80 words: 30 studied words (three per studied list, drawn from positions 1, 6, and 10), 10 critical lure words (one per studied list), and 40 words that had not been presented. Of the 40 non-presented words, 30 were taken from ten unstudied DRM lists (again, three words for each given list were drawn from positions 1, 6, and 10), and the remaining 10 words were the critical lures for the ten unstudied lists. This ensured that the distractors were comparable to the targets. As in the study phase, on the recognition test each word was presented in lowercase 30-pt Arial font against a white background, except that now all test words were presented in black. The order of test words was randomized for each participant. The Appendix contains all of the stimuli.

**Procedure.** Participants sat in front of computer monitors and followed on-screen and/or oral instructions for the duration of the experiment. At the beginning of the experiment, all participants were given the following instructions:

> "You will be seeing a series of common words presented one at a time at the center of the screen. If the word is blue [red], please read the word out loud in a normal speaking voice. If the word is red [blue], please read the word silently in your mind, without moving your lips."

Once they indicated that they understood the instructions, participants proceeded to the practice study task, where they read (aloud/silently) five blue words and five red words to become familiar with the procedure. They then proceeded to the study task proper, consisting of the ten DRM lists of 12 words each. Each DRM list was presented in its entirety before the next list began, and the order of the 10 lists was randomized for each participant. All 120 words were presented consecutively as a continuous stream; there were no breaks between the DRM lists. Words were presented one at a time for 2000 ms each, with a 500-ms blank between successive words. Within each DRM list, words were presented in a single consistent color and in a fixed order (decreasing backward associative strength), as is typical of DRM studies (e.g., Lu et al., 2020; Roediger & McDermott, 1995; Schacter et al., 2001). Upon completion of the study phase, participants played on-screen Tetris (Experiment 1a, for five minutes; Experiment 1b, for two minutes) before proceeding to the test phase.

Prior to the test phase, participants were given the following instructions:

> "In the next task, words will again be shown on the screen one at a time. Some of these words were presented in the previous phase and some were not. Your task is to decide whether each word is **'Old'** (i.e., it was presented in the previous phase) or **'New'** (i.e., it was not presented before). If you think the word was **not** presented before, click the **'New'** button. If you think the word was presented before, there are two possible **'Old'** responses that you can make. When you clearly remember that the word was presented (i.e., you can recall specific information about it, like that it was early in the list), click the **'Old: Remember'** button. When you feel that the item was presented before but you cannot recall specific details, click the **'Old: Know'** button."

Once they indicated having understood these instructions, participants proceeded to the practice test consisting of four words from the practice study phase, where they familiarized themselves with the choice buttons. Then they proceeded to the 80-word recognition test proper. Test words were presented with three clickable button options—Old: Remember, Old: Know, and New—below the word. Each word remained on the screen until the participant responded, and there was a 500-ms blank screen between the response to one word and the presentation of the next word. After completing the recognition test, participants were thanked for their participation and debriefed.

**Data Analysis**. All analyses were conducted in R (R Core Team, 2022). ANOVAs were conducted using the *afex* package (Singmann et al., 2023). We report Cohen's $d_z$ as a measure of within-subject effect size (using the standard deviation of the difference scores as the

denominator; Cohen, 1988, p. 48), calculated using the *effectsize* package (Ben-Shachar et al., 2020). Data and analysis code for all experiments are available at https://osf.io/tjm2n/files.

### Results

Experiment 1a was pre-registered at https://osf.io/j3a6g and Experiment 1b at https://osf.io/whbax. The final sample for each experiment included data from 80 participants. Six participants in Experiment 1a and three in Experiment 1b had to be removed and replaced due either to not following instructions, technical malfunctions, or scoring lower than the pre-registered criteria. Those participants who participated after the pre-registered sample size stopping rule of 80 (one in Experiment 1a; nine in Experiment 1b) also were excluded from analysis.

#### True and false recognition rates

The recognition test was comprised of four stimulus types: 30 studied words drawn from ten DRM lists, 10 critical lures associated with each of the ten studied lists, and 40 unstudied words (30 unstudied list words and their 10 unstudied critical lures). The overall false alarm rates for the unstudied words were .18 ($SE$ = .02) in Experiment 1a and .23 ($SE$ = .02) in Experiment 1b. Breaking these down further, false alarm rates for the unstudied critical lures and for the unstudied list words (see Appendix) were, respectively, .23 and .16 in Experiment 1a, and .28 and .21 in Experiment 1b. In each experiment, there was a small and consistent false memory effect at test for the unstudied DRM lists: The unstudied critical lures were incorrectly recognized as old more often than were the unstudied list words, Experiment 1a: $F(1,79) = 11.31$, $MSE = .01$, $p = .001$, $\eta^2_G = .034$; Experiment 1b: $F(1,79) = 14.42$, $MSE = .01$, $p < .001$, $\eta^2_G = .028$.

Turning to the studied words, we begin simply, collapsing Old: Remember and Old: Know responses into 'old' responses. These proportions are shown in Fig. 1 for actually studied DRM list words (hits) and for their critical lures (critical false alarms) in both experiments. The principal analysis was a 2 (Reading Condition: Aloud vs Silent) x 2 (Word Type: Studied vs Critical Lure) repeated-measures ANOVA, followed by separate paired $t$-tests for studied words and their critical lures.

*Experiment 1a.* There was a significant main effect of Reading Condition, $F(1,79) = 75.09$, MSE = .04, $p < .001$, $\eta^2_G = .111$, with more old responses to words read aloud ($M = .76$, $SE = .02$) than to words read silently ($M = .60$, $SE = .03$), the difference of .16 being a quite typical production effect. There was also a significant main effect of Word Type, $F(1,79) = 22.46$, MSE = 0.04, $p < .001$, $\eta^2_G = .046$, with more old responses to studied words ($M = .73$, $SE = .02$) than to critical lures ($M = .63$, $SE = .03$), suggesting some discrimination. The interaction was not significant, however, $F(1,79) = 1.64$, MSE = .03, $p = .204$, $\eta^2_G = .003$, indicating roughly equivalent production effects for studied words and for their critical lures.

We then conducted pre-registered paired $t$-tests separately for studied words and for their critical lures. For studied words, there was a significant production effect, $t(79) = 8.72$, $p < .001$, $dz = 0.97$ [95 % CI: 0.71, 1.24]: Reading aloud resulted in a higher true recognition rate than did reading silently. For critical lures, there also was a significant production effect, $t(79) = 4.57$, $p < .001$, $dz = 0.51$ [95 % CI: 0.28, 0.74]: Having read DRM list words aloud resulted in a higher false recognition rate for their critical lures than did having read list words silently.

*Experiment 1b.* There was again a significant main effect of Reading Condition, $F(1,79) = 51.51$, MSE = .04, $p < .001$, $\eta^2_G = .105$, with more old responses to words read aloud ($M = .72$, $SE = .03$) than to words read silently ($M = .56$, $SE = .03$), again a .16 production effect. There was also a significant main effect of Word Type, $F(1,79) = 14.43$, MSE = .04, $p = .003$, $\eta^2_G = .031$, with more old responses to studied words ($M = .68$, $SE = .02$) than to their critical lures ($M = .60$, $SE = .03$), again suggesting evidence of discrimination. This time, however, the interaction was significant, $F(1,79) = 4.03$, MSE = .02, $p = .048$, $\eta^2_G = .005$, indicating that the production effect was somewhat larger for true recognition than for false recognition.

We again conducted pre-registered paired $t$-tests separately for studied words and for their critical lures, and the results were identical to those of Experiment 1a. There was a significant production effect for studied words, $t(79) = 9.40$, $p < .001$, $dz = 1.05$ [95 % CI: 0.73, 1.32]: Reading aloud resulted in a higher true recognition rate than did reading silently. There also was a significant production effect for their critical lures, $t(79) = 3.79$, $p < .001$, $dz = 0.42$ [95 % CI: 0.19, 0.65]: Having read DRM list words aloud resulted in a higher false recognition rate than did having read them silently.

#### Familiarity and recollection with Remember/Know responses

Based on the Remember/Know responses, we calculated measures of recollection and familiarity using the independent remember/know method (Mangels et al., 2001; Ochsner, 2000; Ozubko et al., 2012a; Yonelinas & Jacoby, 1995). It is important to keep in mind that unadjusted "know" responses underestimate the influence of familiarity. In the remember/know procedure, participants are instructed to respond "know" only if an item is familiar *and* they cannot recollect any details surrounding it. In essence, then, as the proportion of "remember" responses increases, the proportion of "know" responses must necessarily decrease. Therefore, if the raw proportion of "know" responses was taken as a measure of familiarity, this would make it appear that the influence of familiarity was decreasing, when in fact only recollection was increasing. Consequently, because they are limited by the proportion of "remember" responses, "know" responses themselves do not provide ideal measures of familiarity. In the independent remember/know method, Recollection (R) is measured as the proportion of "remember" responses; familiarity (F) is measured as the proportion of "know" responses divided by the proportion of non–"remember" responses.[4] Using this correction, researchers have demonstrated that estimates of recollection and familiarity, as measured in the remember/know procedure, correspond with estimates of recollection and familiarity as measured by other techniques (Yonelinas, 2002; Yonelinas & Jacoby, 1995; Yonelinas, 1997).

These estimates were calculated separately for each participant, both for studied words (aloud or silently studied) and for their critical lures (associated with either aloud or silently studied words). We then analyzed these recollection and familiarity estimates using 2 X 2 X 2 repeated-measures ANOVAs, the factors being Reading Condition (Aloud vs Silent), Word Type (Studied vs Critical Lure), and Response Type (Recollection vs Familiarity). These analyses were followed by two-way ANOVAs—2 (Reading Condition: Aloud vs Silent) X 2 (Response Type: Recollection vs Familiarity)— done separately for studied words and for their critical lures. Table 1 presents the proportions of Recollection/Familiarity responses as a function of Reading Condition and Word Type for Experiments 1a and 1b.

*Experiment 1a.* Despite the three-way interaction not being significant, $F(1,79) = 1.30$, MSE = .04, $p = .258$, $\eta^2_G = .001$, as pre-registered, we conducted separate two-way ANOVAs for studied words and for critical lures, the factors being Reading Condition (Aloud vs Silent) and Response Type (Recollection vs Familiarity). For studied words, the main effect of Reading Condition was significant, $F(1,79) = 38.23$, MSE = .04, $p < .001$, $\eta^2_G = .059$, with reading aloud ($M = .53$, $SE = .03$) resulting in more Recollection/Familiarity responses than reading silently ($M = .40$, $SE = .02$). The main effect of Response Type was not significant, $F(1,79) = 0.33$, MSE = .13, $p = .566$, $\eta^2_G = .002$, indicating that the proportions of Recollection responses ($M = .47$, $SE = .03$) and of Familiarity responses ($M = .45$, $SE = .03$) did not differ. However, the interaction of Response Type with Reading Condition was significant,

[4] Cases where participants had zero know responses and zero non-remember responses in a condition were treated as Familiarity = 0.

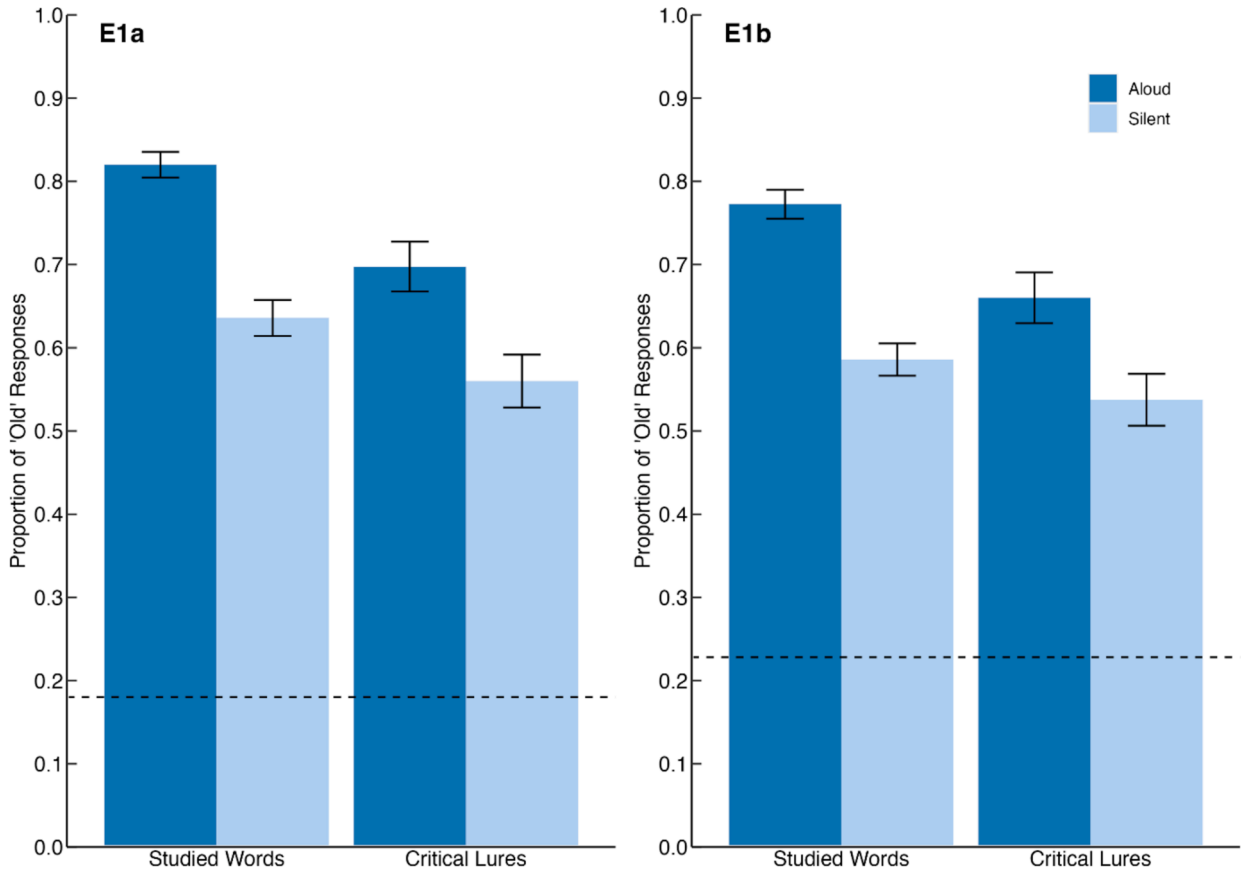


**Fig. 1.** Experiments 1a (left panel) and 1b (right panel): Mean proportions of 'old' responses to studied words (hits) and to critical lures (critical false alarms) in the aloud and silent study conditions. Dashed lines are mean false alarm rates for the unstudied items; error bars are standard errors.

**Table 1**
Experiments 1a and 1b: Mean (SE) proportions of Recollection/Familiarity responses for studied words and for their critical lures as a function of reading condition.

| | Experiment 1a | | Experiment 1b | |
|---|---|---|---|---|
| | Recollection | Familiarity | Recollection | Familiarity |
| **Studied Words** | | | | |
| Aloud | .57 (.03) | .49 (.04) | .41 (.03) | .58 (.03) |
| Silent | .38 (.02) | .42 (.02) | .27 (.02) | .42 (.02) |
| **Critical Lures** | | | | |
| Aloud | .43 (.03) | .47 (.04) | .31 (.03) | .49 (.04) |
| Silent | .29 (.03) | .38 (.03) | .19 (.03) | .42 (.03) |

$F(1,79) = 7.00$, MSE = .04, $p = .010$, $\eta^2_G = .014$. We followed up by examining the effect of reading aloud separately on the proportions of Recollection responses and of Familiarity responses. Relative to reading silently, reading aloud resulted in a greater proportion of Recollection responses, $t(79) = 7.35$, $p < .001$, $dz = 0.77$ [95 % CI: 0.54, 1.01], but not of Familiarity responses, $t(79) = 1.92$, $p = .058$, $dz = 0.24$ [95 % CI: −0.01, 0.50].

Turning to the critical lures, the main effect of Reading Condition was significant, $F(1,79) = 22.13$, MSE = .05, $p < .001$, $\eta^2_G = .034$, again indicating that reading aloud ($M = .45$, $SE = .04$) resulted in a higher proportion of Recollection/Familiarity responses than did reading silently ($M = .34$, $SE = .03$). The main effect of Response Type was not significant, $F(1,79) = 3.30$, MSE = .11, $p = .073$, $\eta^2_G = .013$, indicating that the proportions of Recollection responses ($M = .36$, $SE = .04$) and of Familiarity responses ($M = .43$, $SE = .03$) did not differ. The interaction of Response Type with Reading Condition also was not significant, $F(1,79) = 0.70$, MSE = .07, $p = .404$, $\eta^2_G = .002$.

*Experiment 1b.* The three-way interaction again was not significant, $F(1,79) = 1.04$, MSE = .05, $p = .311$, $\eta^2_G = .001$. As in Experiment 1a, and in keeping with our pre-registration, we carried out separate two-way ANOVAs for studied words and their critical lures. The results were somewhat different from Experiment 1a. For studied words, the main effect of Reading Condition was significant, $F(1,79) = 79.44$, MSE = .02, $p < .001$, $\eta^2_G = .098$, with reading aloud ($M = .50$, $SE = .03$) resulting in more Recollection/Familiarity responses than reading silently ($M = .35$, $SE = .04$). The main effect of Response Type was also significant, $F(1,79) = 22.45$, MSE = .10, $p < .001$, $\eta^2_G = .114$, indicating a higher proportion of Familiarity responses ($M = .50$, $SE = .03$) than of Recollection responses ($M = .34$, $SE = .03$). The interaction of Response Type with Reading Condition was not significant, $F(1,79) = 0.16$, MSE = .03, $p = .690$, $\eta^2_G < .001$.

Turning to the critical lures, the main effect of Reading Condition was significant, $F(1,79) = 17.05$, MSE = .04, $p < .001$, $\eta^2_G = .023$, indicating that reading aloud ($M = .40$, $SE = .04$) resulted in a higher proportion of Recollection/Familiarity responses than did reading silently ($M = .31$, $SE = .03$). The main effect of Response Type was also significant, $F(1,79) = 26.71$, MSE = .12, $p < .001$, $\eta^2_G = .104$, indicating a higher proportion of Familiarity responses ($M = .45$, $SE = .04$) than of Recollection responses ($M = .25$, $SE = .03$). The interaction was not significant, $F(1,79) = 0.74$, MSE = .08, $p = .392$, $\eta^2_G = .002$.

### *Discussion*

In two experiments, one conducted in the laboratory and one online, we consistently found a robust production effect: Reading aloud resulted in a higher recognition rate than did reading silently. We also obtained a robust DRM effect: The critical lures for studied lists were falsely recognized as having been presented at quite high rates (see Fig. 1).

Additionally, we observed a small but significant DRM effect for the unstudied items: The critical lures (e.g., *doctor*) from the unstudied lists (e.g., *nurse, health, patient*) were falsely recognized at a higher rate than were the other words from those lists that were presented only at test. The critical lures of the unstudied lists may have been repeatedly cued and activated as participants encountered the unstudied list words during the recognition test (see also Coane & McBride, 2006). Overall, these results clearly demonstrate that both our reading aloud and false memory manipulations were successful.

Critically, we found that, relative to reading silently, reading aloud increased not only true recognition rates but also false recognition rates. That is, relative to reading silently, reading aloud resulted in a higher proportion of 'old' responses both to words from the studied lists and to critical lures from those same lists. However, this effect did not operate differentially on Recollection vs. Familiarity responses (except in Experiment 1a, where reading aloud was associated with a higher proportion of Remember responses for studied words). We also did not replicate the finding of production leading to a greater increase in recollection than in familiarity (Fawcett & Ozubko, 2016; Ozubko et al., 2012a): Instead, production increased both recollection and familiarity responses to both studied words and their critical lures. It is possible, of course, that the semantically related structure of DRM lists is responsible for this difference in pattern.

Clearly, our results are inconsistent with those of the two prior related studies. Both Dodson and Schacter (2001, Experiment 2, $N = 27$) and Culbreth and Putnam (2019, $N = 73$) manipulated production within-subject and reported that reading the DRM list words aloud made no difference in false recognition rates for their critical lures compared either to having seen and heard the list words (Dodson & Schacter) or having read the list words silently (Culbreth & Putnam) at study. Given this difference in findings, and because of the different comparison conditions in the prior studies, we conducted Experiment 2 in which we compared reading words aloud to reading words silently while also hearing them spoken by someone else—the comparison condition used by Dodson and Schacter. Previous literature suggests that auditory presentation of words is associated with higher rates of false memory than is visual presentation (Smith & Hunt, 1998). Hence, one of the aims of Experiment 2 was to examine whether reading aloud would result in substantially different effects on false memory when compared to a control condition where words were presented auditorily as well as visually.

## Experiment 2

In Experiment 2, the only change was that the silent reading condition in Experiments 1a and 1b was replaced with a condition involving silent reading and simultaneously hearing the words spoken by a recorded voice. Experiment 2 was conducted online and was remotely supervised.

### *Method*

**Participants.** We chose a sample size of 80 to be consistent with Experiments 1a and 1b. Participants were students from the University of Waterloo participating online for course credit; none had taken part in a previous production effect or DRM experiment from our laboratory. None had to be replaced.

**Stimuli.** The DRM lists were those used in Experiments 1a and 1b. Voice files were generated using the male-gendered "Alex" voice on a Mac computer.

**Procedure.** Reading aloud was compared to reading silently while hearing the words spoken by a male voice. This was the only procedural difference compared to Experiment 1b: Both experiments were conducted online and were remotely supervised by an experimenter via web camera.

### *Results*

Experiment 2 was pre-registered at https://osf.io/swkaj.

#### *True and false recognition rates*

The overall false alarm rate for the unstudied items was .25 ($SE$ = .02). As in Experiments 1a and 1b, the false alarm rate was lower for unstudied list words (.24) than for unstudied critical lures (.30), $F(1,79) = 20.41$, MSE = .01, $p < .001$, $\eta^2_G = .027$, again showing a DRM effect at test for unstudied materials.

As in Experiments 1a and 1b, we begin simply, collapsing Old: Remember and Old: Know responses into 'old' responses. Fig. 2 shows the proportions of 'old' responses to studied words and to their critical lures as a function of reading condition. There was a significant main effect of Reading Condition, $F(1,79) = 35.29$, MSE = .04, $p < .001$, $\eta^2_G = .088$, with reading aloud ($M = .76$, $SE = .02$) resulting in greater recognition than reading silently while hearing the printed words read aloud ($M = .62$, $SE = .03$). The .14 production advantage is quite typical. The main effect of Word Type was also significant, $F(1,79) = 14.99$, MSE = .03, $p < .001$, $\eta^2_G = .033$, with participants responding old more often to studied words ($M = .73$, $SE = .02$) than to their critical lures ($M = .65$, $SE = .03$), demonstrating some discrimination. There also was a significant interaction, $F(1,79) = 8.79$, MSE = .02, $p = .004$, $\eta^2_G = .011$, indicating that the effect of reading aloud was larger for true recognition than for false recognition. Paired $t$-tests revealed significant production effects both for true recognition of studied words, $t(79) = 9.18$, $p < .001$, $dz = 1.12$ [95 % CI: 0.81, 1.43], and for false recognition of their critical lures, $t(79) = 2.70$, $p = .008$, $dz = 0.34$ [95 % CI: 0.08, 0.59].

#### *Familiarity and recollection with Remember/Know responses*

As in Experiments 1a and 1b, we calculated Recollection and Familiarity estimates separately for each participant, for studied words (aloud or silently studied) and for their critical lures (associated with either aloud or silently studied words). Table 2 presents the proportions of Recollection/Familiarity responses as a function of Reading Condition and Word Type.

Although the three-way interaction was not significant, $F(1,79) = 2.65$, MSE = .04, $p = .107$, $\eta^2_G = .002$, in keeping with our pre-registration, we followed up with separate planned two-way ANOVAs for studied words and their critical lures, the factors being Reading Condition (Aloud vs Silent) and Response Type (Remember vs Know). For studied words, the main effect of Reading Condition was significant, $F(1,79) = 85.82$, MSE = .03, $p < .001$, $\eta^2_G = .133$: Reading aloud ($M = .57$, $SE = .03$) resulted in a higher proportion of Recollection/Familiarity responses than did reading silently and listening ($M = .39$, $SE = .02$). The main effect of Response Type also was significant, $F(1,79) = 9.49$, MSE = .09, $p = .003$, $\eta^2_G = .049$, indicating a higher proportion of Familiarity responses ($M = .53$, $SE = .03$) than of Recollection responses ($M = .43$, $SE = .03$). The interaction was not significant, $F(1,79) = 0.03$, MSE = .02, $p = .855$, $\eta^2_G < .001$.

For critical lures, the main effect of Reading Condition was marginally significant, $F(1,79) = 3.77$, MSE = .06, $p = .056$, $\eta^2_G = .008$, consistent with reading aloud ($M = .43$, $SE = .04$) resulting in a higher proportion of Recollection/Familiarity responses than did reading silently while listening ($M = .37$, $SE = .03$). The main effect of Response Type was significant, $F(1,79) = 14.58$, MSE = .12, $p < .001$, $\eta^2_G = .060$, indicating that there was a higher proportion of Familiarity responses ($M = .48$, $SE = .04$) than of Recollection responses ($M = .33$, $SE = .03$). The interaction was not significant, $F(1,79) = 2.68$, MSE = .07, $p = .106$, $\eta^2_G = .006$.

#### *Exploratory Analysis: Combined analysis for Familiarity and recollection*

We conducted an exploratory combined analysis across Experiments 1a, 1b, and 2 because the numeric trends in Table 2 suggested that production might increase recollection more than familiarity. Although we did not find a significant interaction with reading condition in any

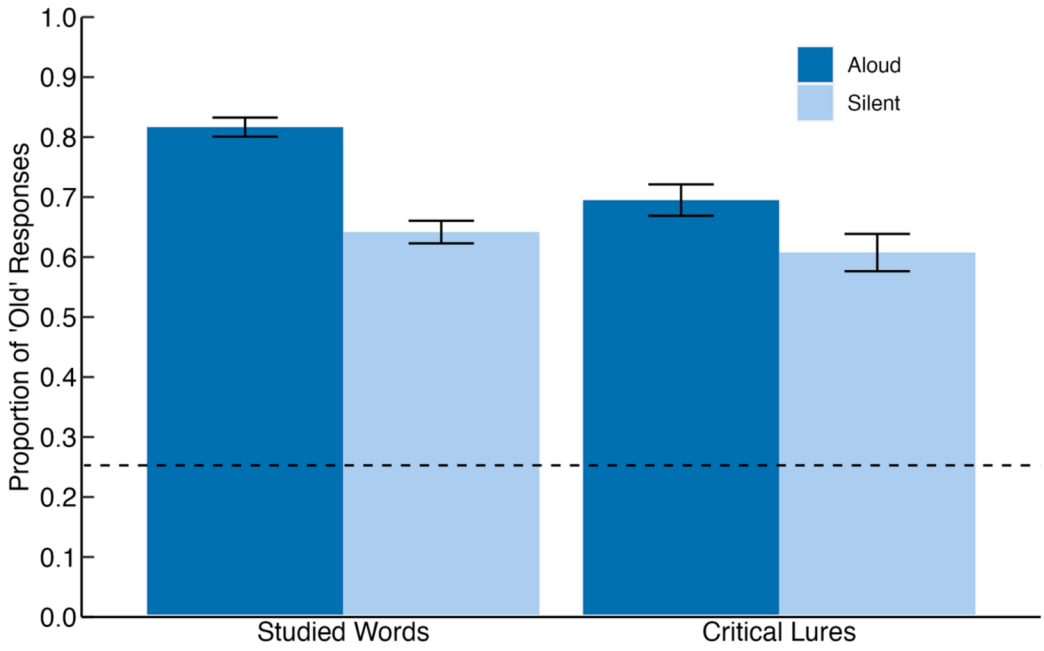


**Fig. 2.** Experiment 2: Mean proportions of 'old' responses to studied words (hits) and to critical lures (critical false alarms) in the aloud and silent/heard study conditions. The dashed line represents mean false alarm rate for the unstudied items; error bars are standard errors.

**Table 2**
Experiment 2: Mean (SE) proportions of Recollection/Familiarity responses for studied words and for their critical lures as a function of reading condition.

| | Recollection | Familiarity |
|---|---|---|
| **Studied Words** | | |
| Aloud | .52 (.03) | .63 (.03) |
| Silent/Heard | .34 (.02) | .44 (.02) |
| **Critical Lures** | | |
| Aloud | .38 (.03) | .48 (.04) |
| Silent/Heard | .28 (.03) | .47 (.04) |

experiment, this combined analysis would provide more power to detect an interaction. As in our previous analyses, we found no significant three-way interaction, $F(1,239) = 0.74$, $MSE = .04$, $p = .389$, $\eta^2_G < .001$. Below we report separate planned two-way ANOVAs for studied items and their critical lures, the factors being Reading Condition (Aloud vs Silent) and Response Type (Remember vs Know). Table 3 presents the proportions of Recollection/Familiarity responses in the combined data as a function of Reading Condition and Word Type.

For studied words, the main effect of Reading Condition was significant, $F(1,239) = 190.99$, $MSE = .03$, $p < .001$, $\eta^2_G = .089$, indicating that reading aloud resulted in a higher proportion of Recollection/Familiarity responses than did reading silently or reading silently and listening. The main effect of Response Type also was significant, $F(1,239) = 14.35$, $MSE = .11$, $p < .001$, $\eta^2_G = .027$, indicating a higher proportion of Familiarity responses than of Recollection responses. The interaction remained non-significant, $F(1,239) = 1.90$, $MSE = .03$, $p = .170$, $\eta^2_G = .001$.

For critical lures, the main effect of Reading Condition was significant, $F(1,239) = 36.48$, $MSE = .05$, $p < .001$, $\eta^2_G = .020$, consistent with reading aloud resulting in a higher proportion of Recollection/Familiarity responses than did reading silently or reading silently and listening. The main effect of Response Type was significant, $F(1,239) = 38.65$, $MSE = .12$, $p < .001$, $\eta^2_G = .051$, indicating that there was a higher proportion of Familiarity responses than of Recollection responses. The interaction was marginally significant, $F(1,239) = 3.67$, $MSE = .07$, $p = .057$, $\eta^2_G = .003$. Follow-up analyses suggested that production had a larger effect on recollection, $F(1,239) = 43.72$, $MSE = .04$, $p < .001$, $\eta^2_G = .048$, than on familiarity, $F(1,239) = 4.34$, $MSE = .08$, $p = .038$, $\eta^2_G = .006$.

**Table 3**
Combining Experiments 1a, 1b, and 2: Mean (SE) proportions of Recollection/Familiarity responses for studied words and for their critical lures as a function of reading condition.

| | Recollection | Familiarity |
|---|---|---|
| **Studied Words** | | |
| Aloud | .50 (.02) | .57 (.02) |
| Silent/Heard | .33 (.01) | .43 (.01) |
| **Critical Lures** | | |
| Aloud | .37 (.02) | .48 (.02) |
| Silent/Heard | .25 (.02) | .43 (.02) |

*Discussion*

The change in the comparison condition—from silent reading in Experiments 1a and 1b to hearing the words spoken by a recorded voice while reading silently in Experiment 2—did not alter the pattern of results. Again, false recognition of the critical lures was greater following reading aloud than following the comparison condition, just as was the case for true recognition of the studied words. We also replicated the finding from Experiments 1a and 1b that production was associated with increases in both recollection-based and familiarity-based responses.

## Experiment 3a

All three of our earlier reported experiments demonstrated that both true recognition of the studied words and false recognition of the critical lures were greater following reading aloud than following the comparison condition. As noted, this finding contradicts that of previous investigations that found no effect of reading aloud on false recognition (Dodson & Schacter, 2001; Culbreth & Putnam, 2019).

Another potential point of difference was our use of the tripartite Remember/Know/New procedure, in contrast to the two previous investigations that used a binary Old/New test. One potential concern was that the one-step Remember/Know/New procedure tends to be associated with a more liberal response criterion (Hicks & Marsh, 1999),

leading to increased hit and false alarm rates, though we saw no reason to anticipate that this would affect the aloud and silent conditions differently in a within-subject design. Nevertheless, because our results are the first demonstration of production increasing both hit and false alarm rates, we saw it as important to establish that this effect generalized to the more common Old/New recognition test format. Thus, in the following two experiments, we sought to replicate the basic pattern of results with a binary recognition test.

*Method*

Experiment 3a constituted a replication of Experiment 1a with a binary Old/New recognition test instead of a tripartite Remember/Know/New recognition test.

**Participants.** We chose a sample size of 80, the same as earlier experiments. Participants were students from the University of Waterloo participating for course credit; none had taken part in a previous production effect or DRM experiment from our laboratory.

**Stimuli and Procedure.** The DRM lists were the same as those used in the previous experiments. The main procedural difference compared to Experiment 1a was that the recognition test required participants to respond with "Old" or "New" instead of with "Old: Remember," "Old: Know," or "New." Participants were also asked to self-report their age and gender at the end of the experiment.

*Results*

Experiment 3a was pre-registered at https://osf.io/xnzcv. The final sample included data from 80 participants (64 women, 15 men, 1 undisclosed, $M = 19.09$ years, $SD = 1.60$). Seven participants had to be removed and replaced due either to not following instructions, technical malfunctions, or scoring lower than the pre-registered criteria. Finally, five participants who participated after the pre-registered sample size stopping rule of 80 were excluded from analysis.

*True and false recognition rates*

The overall false alarm rate for the unstudied words was .19 ($SE = .04$). As in previous experiments, the false alarm rate was lower for the unstudied words (.18) than for their unstudied critical lures (.24), $F(1,79) = 14.85$, $MSE = .01$, $p < .001$, $\eta^2_G = .032$, again showing a DRM effect at test for unstudied materials.

Fig. 3 (left panel) shows the proportions of 'old' responses to presented words and to critical lures as a function of reading condition. There was a significant main effect of Reading Condition, $F(1,79) = 28.95$, $MSE = .04$, $p < .001$, $\eta^2_G = .077$, with reading aloud ($M = .76$, $SE = .02$) resulting in greater recognition than reading silently ($M = .63$, $SE = .02$), a quite typical .13 production effect. The main effect of Word Type was also significant, $F(1,79) = 24.40$, $MSE = .03$, $p < .001$, $\eta^2_G = .055$, with participants responding old more often to studied words ($M = .75$, $SE = .01$) than to their critical lures ($M = .64$, $SE = .02$), demonstrating some discrimination. These effects were qualified by a significant interaction, $F(1,79) = 18.21$, $MSE = .02$, $p < .001$, $\eta^2_G = .030$,

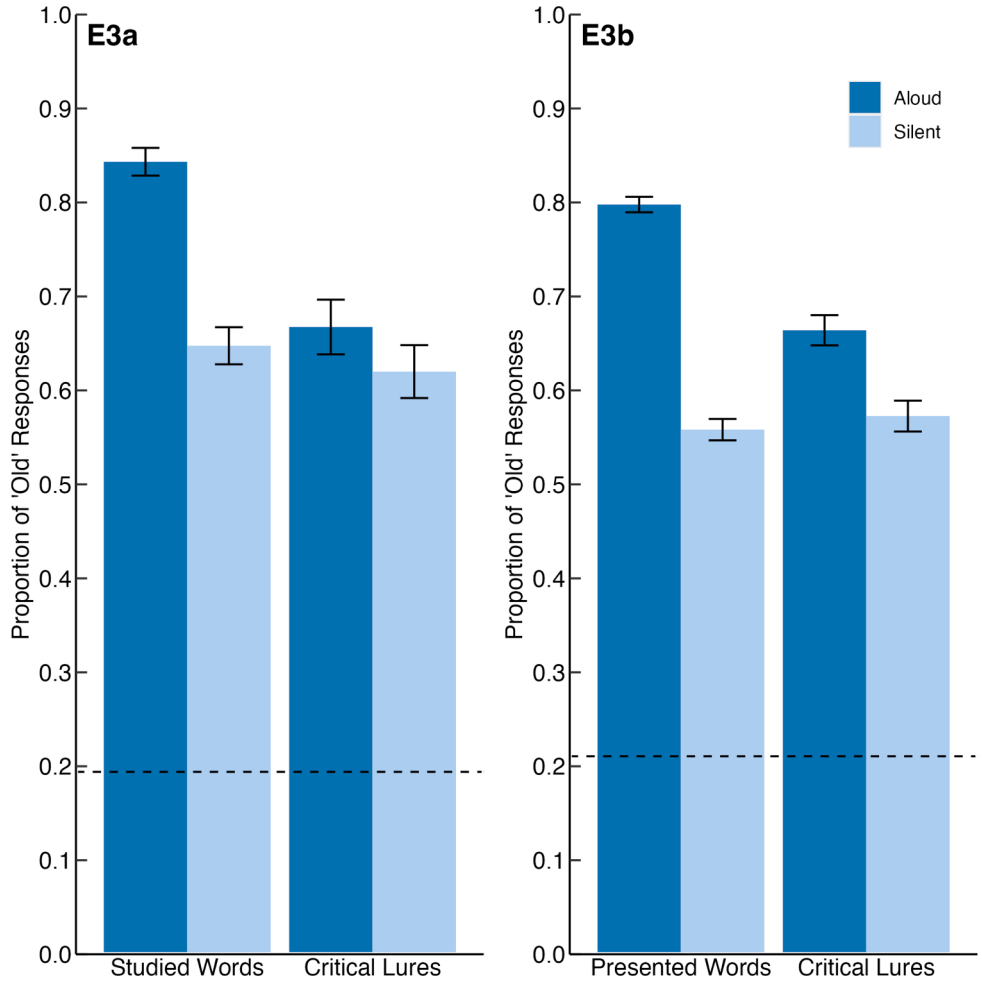


**Fig. 3.** Experiments 3a (left panel) and 3b (right panel): Mean proportions of 'old' responses to studied words (hits) and to their critical lures (critical false alarms) in the aloud and silent study conditions. Dashed lines are mean false alarm rates for the unstudied items; error bars are standard errors.

suggesting that the effect of reading aloud was larger for true recognition than for false recognition. Paired *t*-tests revealed a significant production effect for true recognition of studied words, $t(79) = 8.90$, $p < .001$, $dz = 0.99$ [95 % CI: 0.72, 1.26], but not for false recognition of their critical lures, $t(79) = 1.40$, $p = .164$, $dz = 0.16$ [95 % CI: −0.06, 0.38].

### *Discussion*

In Experiment 3a, as in the previous three experiments, we found a main effect of production and an interaction with memory type, such that we observed higher true recognition rates for the studied words that had been read aloud. Unlike the previous experiments, however, reading aloud did not result in a significant effect on false recognition of the critical lures. One possibility is that the true influence of production on false recognition is a null effect when a binary recognition test is used. Another possibility is that our experiment was underpowered so that we failed to detect the true effect of production. Thus, we conducted Experiment 3b as a follow-up replication experiment with a larger sample size that was determined after conducting a power analysis using the estimates obtained from Experiment 3a.

## Experiment 3b

Experiment 3b was a replication of Experiment 3a with a substantially larger sample size.

### *Method*

**Participants.** We pre-registered a sample size of 300; Superpower's *ANOVA_exact* (Lakens & Caldwell, 2021) deemed this *N* to be sufficient for .80 power to detect a production effect in the critical lures ($M_{aloud} = 67$, $M_{silent} = .62$, $SD = .26$, $r = .30$; estimates from Experiment 3a). Participants were recruited from Prolific (https://www.prolific.com/) and paid GB £1.88; none had taken part in a previous production effect or DRM experiment from our laboratory.

**Stimuli and Procedure.** The DRM lists were the same as those used in the previous experiments. Experiment 3b was conducted remotely on Prolific; participants were audio-recorded during the study phase so that we could check for adherence to the reading instructions.

### *Results*

Experiment 3b was pre-registered at https://osf.io/eg84a. The final sample included data from 300 participants (169 women, 127 men, 3 gender nonconforming, 1 undisclosed, $M = 40.70$ years, $SD = 12.98$). Forty-one other participants had to be removed and replaced due either to not following instructions, technical malfunctions, or scoring lower than the pre-registered criteria. Finally, three participants who participated after the pre-registered sample size stopping rule of 300 were excluded from analysis.

#### *True and false recognition rates*

The overall false alarm rate for the unstudied words was .21 ($SE = .01$). As in previous experiments, the false alarm rate was lower for unstudied words (.20) than for their unstudied critical lures (.25), $F(1,299) = 39.17$, $MSE = .01$, $p < .001$, $\eta^2_G = .016$, again showing a DRM effect at test for unstudied materials.

Fig. 3 (right panel) shows the proportions of 'old' responses to studied words and to their critical lures as a function of reading condition. There was a significant main effect of Reading Condition, $F(1,299) = 170.44$, $MSE = .05$, $p < .001$, $\eta^2_G = .112$, with reading aloud ($M = .73$, $SE = .01$) resulting in greater recognition than reading silently ($M = .57$, $SE = .01$), a quite typical .16 production effect. The main effect of Word Type was also significant, $F(1,299) = 26.18$, $MSE = .04$, $p < .001$, $\eta^2_G = .016$, with participants responding old more often to studied words ($M = .68$, $SE = .01$) than to their critical lures ($M = .62$, $SE = .01$). These effects were qualified by a significant interaction, $F(1,299) = 70.97$, $MSE = .02$, $p < .001$, $\eta^2_G = .025$, indicating that the effect of reading aloud was larger for true recognition than for false recognition. Paired *t*-tests revealed significant production effects both for true recognition of studied words, $t(299) = 19.04$, $p < .001$, $dz = 1.10$ [95 % CI: 0.96, 1.24], and for false recognition of their critical lures, $t(299) = 5.13$, $p < .001$, $dz = 0.30$ [95 % CI: 0.18, 0.41].

**Table 4**
Combined data for all of the experiments: Mean (SE) proportions of 'old' responses for studied words and their critical lures as a function of reading condition and test type.

| | Studied Words | | Critical Lures | |
|---|---|---|---|---|
| | Aloud | Silent | Aloud | Silent |
| E1a/b and E2: Tripartite (R/K) | .80 (.01) | .62 (.01) | .68 (.02) | .57 (.02) |
| E3a/b: Binary (Old) | .81 (.01) | .58 (.01) | .67 (.01) | .58 (.01) |

### *Discussion*

In Experiment 3b, we again found a main effect of production and an interaction with memory type: Reading aloud resulted in higher true recognition rates for studied words and in higher false recognition rates for their associated critical lures, with the latter effect being smaller than the former. A combined analysis (see Supplementary Materials) of the data from Experiments 3a and 3b indicated that the effect of production on false recognition was small but robust. Table 4 presents the mean proportions of 'Old' responses combining the data of Experiments 3a and 3b, along with the comparable proportions from Experiments 1a, 1b, and 2 (collapsing Old:Remember and Old:Know responses). Clearly, the tripartite test and the binary test procedures led to virtually identical outcomes, indicating that the tripartite test did not lead to inflated hit and false alarm rates.

## General Discussion

Across multiple experiments (total $N = 620$), we have seen a thoroughly consistent pattern: False recognition of critical unstudied lures was amplified by having studied the list words aloud rather than silently—or silently and aurally. This was the same amplification as was consistently found for true recognition—for words that were actually read aloud during study— although the increase in false recognition was consistently found to be significantly smaller than that for true recognition.

As noted in the introduction, neither the pure distinctiveness account nor the pure strength account would appear to fully explain the current results. In the relative distinctiveness account (e.g., Conway & Gathercole, 1987; MacLeod & Bodner, 2017), the additional features associated with the act of production (relative to the silent condition) provide a basis for a recollection boost during retrieval. Computational modeling tends to capture this by storing in memory additional features for produced items, which can be thought of as perceptual/motoric features associated with the act of producing an item (Cyr et al., 2022; Jamieson et al., 2016; Kelly et al., 2022). Given that the critical lures

were not encoded with distinctive produced features, the distinctiveness account would expect it to be easier to reject critical lures. However, we instead found an increase in false alarm rates for the critical lures associated with the produced words. This is at odds with a pure distinctiveness account that would predict a primarily recollection-based increase in true recognition for the words spoken aloud, but a decrease in false recognition. The pure memory strength account (e.g., Bodner & Taikh, 2012; Fawcett & Ozubko, 2016) is also insufficient, as we did not find the effect of production on false recognition to be primarily due to familiarity. Instead, in Experiments 1a, 1b, and 2, production increased recollection and familiarity responses for both true and false recognition (if anything, our results suggested potentially a greater boost for recollection over familiarity).

The upshot is that each of these theoretical accounts of production appears to be incomplete: In Experiments 1a, 1b, and 2, reading aloud did not selectively boost familiarity-based false memory nor did it selectively increase memory for the distinctive produced features in a way that made recollection rejection easier. Instead, we observed increased rates of "phantom recollection" (Brainerd et al., 2001; Brainerd et al., 2003) for the critical lures associated with the words studied aloud, although this increase due to production was less than that observed for the studied words themselves.

How do our findings fit with the theoretical accounts of false recognition? In activation monitoring theory (e.g., Roediger et al., 2001), increased false recognition arises from a high level of activation of the critical lure, which is predominantly based on the semantic information that is repeatedly cued by the list words during study. Furthermore, if a critical lure becomes strongly activated during study, it can also become associated with the study context (Roediger et al., 2001; Roediger et al., 2004). In global matching models, false recognition occurs due to high similarity between the critical lure and a multitude of encoded memory traces (Arndt, 2010), and similarity can be based on item/semantic information as well as contextual information. In fuzzy trace theory (e.g., Reyna & Brainerd, 1995) false recognition largely depends on gist (semantic and contextual) traces rather than on verbatim (item-specific) traces. Although they differ in some assumptions, all three frameworks account for false memory phenomena via a semantic activation-based process (i.e., spreading activation, global matching, gist trace). In the following sections, we will emphasize fuzzy-trace theory because it provides the most articulate distinction between non-semantic item information, semantic information, and context information (Brainerd et al., 2014).

Within fuzzy trace theory, false memories are thought to stem primarily from the retrieval of gist (semantic) traces in memory, whereas true memories can be based on the retrieval of either (or both) of gist traces and verbatim traces. Retrieval of verbatim traces should increase true memory and potentially suppress (via recollection rejection) false memories; in sharp contrast, retrieval of gist traces should increase false memories. Our principal result—that production led to an increase in false memory (albeit to a lesser extent than was evident for true memory)—suggests increases in both verbatim-based and gist-based retrieval. That we observed enhancement, not suppression, of false memories additionally suggests that the influence on false memory of the gist traces outweighed that of the verbatim traces.

A gist-based account may appear at first blush to contradict our observation in Experiments 1a, 1b, and 2 of a production-associated increase in *both* recollection and familiarity responses across both true and false memory. However, the distinction between gist and verbatim traces does not map directly onto recollection and familiarity. Brainerd et al. (2014) argue that there are two types of recollection: recollection of the target items themselves and recollection of their contextual details. This distinction between item information and context information goes back to McGeoch (1932), and can also be seen in computational models of memory (e.g., Polyn et al., 2009; also the item vs. associative information distinction in TODAM2; Murdock, 1993). Contextual information includes such features as when or where an item was presented and the conditions (internal/external) present when the item was encoded.

In the fuzzy trace account, false recognition is thought to be suppressed by target recollection but increased by contextual recollection (Brainerd et al., 2014). Context recollection can be based on retrieval of verbatim and/or gist traces (because both are episodically tagged), but target item recollection should only be based on the retrieval of verbatim traces (i.e., retrieval of item-specific information). In other words, gist traces can be expected to support contextual recollection (which increases false memory), but not target recollection (which suppresses false memory). Familiarity, on the other hand, occurs when gist traces are retrieved without stored contextual details coming to mind (Brainerd et al., 2014). Therefore, a production-associated increase in gist would be expected to enhance context recollection as well as familiarity but not to influence target recollection.

### *Production, gist, and context*

The current pattern of results, where production increased recollection and familiarity responses in both true and false recognition, is consistent with production-related enhancement of semantic and/or contextual gist. Zhou and MacLeod (2022) have previously argued that production constitutes an additional kind of contextual information (i.e., a kind of condition present for some items during encoding). The distinctive produced features establish a global contextual cue that is associated only with the produced items (from the participant's point of view, the context can be considered to be "items that were studied aloud"). The current results fit well with the Zhou and MacLeod contextual account of production whereby the distinctive produced features establish a global contextual cue that becomes associated with the produced items and consequently, in the present case, with their associated critical lures.

Another possibility (that does not preclude the first) is that the effect of production as observed on false recognition is based on increased semantic activation. Fawcett et al. (2022, Experiment 2) have provided converging evidence that production enhances semantic processing. Following a standard production paradigm study phase, Fawcett et al. administered a two-alternative forced-choice recognition test. When the lure was a synonym of a target word, rather than being unrelated, the production effect was reliably reduced. Coupled with their Experiment 1 finding that homophone lures did not reduce the effect, they suggested (p. 2261) "that items read aloud are distinctive not only due to the inclusion of sensorimotor elements, but also because the act of production encourages broader conceptual encoding." Greater false recognition for the better-remembered aloud lists than for the silent lists likely is the consequence of greater priming of the critical lures. Such priming would necessarily be semantic, given the associative structure of DRM lists and their critical lures.

We note that these two possibilities—semantic retrieval and context retrieval—are likely to co-occur and have been documented in theoretical accounts of false memory. In activation-monitoring theory, the critical lure becomes associated with the same context as the studied words because of strong semantic-based activation during the original encoding context. Thus, production may have increased semantic processing, such that critical lures associated with the produced items may also have become associated with a produced context, leading to more false recognition. These processes presumably outweighed any effect that production may have had on target item (non-contextual verbatim information) recollection given that our participants were less successful at rejecting the critical lures associated with the produced items.

### *Comparison of current results with previous literature*

Our results—that reading aloud was associated with greater false alarms to critical lures—contradict those reported by Dodson and Schacter (2001) and by Culbreth and Putnam (2019). Both of those

studies reported no effect on false recognition of whether the list words had been studied aloud, doing so with respect to different comparison conditions—silent reading with simultaneous hearing of recorded speaking (Dodson & Schacter) or simply silent reading (Culbreth & Putnam).[5] In Experiments 1a, 1b, 3a, and 3b, we used the silent reading comparison condition typical of production effect studies; in Experiment 2, we used the silent reading with concurrent hearing comparison condition. Overall, we had considerably increased power than the prior studies and we demonstrated very consistent replicability of our patterns. Given the power and replication in our study, we see our experiments as providing the most complete picture with respect to false recognition resulting from production. Additionally, our use of the Remember/Know/New procedure in Experiments 1a, 1b, and 2 also allowed us to dissociate production's effects on recollection and familiarity in both true and false recognition.

### *Theoretical and practical significance*

Although both the production effect and the DRM effect are well established in the literature, our experiments provide a considerably clearer picture of their relation than has been available to date. The obtained pattern of results, that reading aloud was associated with greater false memory rates as well as with greater true memory rates, clarifies the mechanisms underlying how reading aloud influences memory. Instead of a distinctiveness/global strength dichotomy, we posit that fuzzy trace theory in particular may be a promising framework for understanding the effects of production on memory for semantic gist, contextual gist, and verbatim detail. Based on the current results, we have suggested that reading aloud enhances memory for semantic gist and/or context in particular. Future research might explore this hypothesis further by examining how reading aloud can influence memory for context (e.g., source memory).

In terms of broader practical significance, our results may have implications in such domains as the legal field—for example, asking eyewitnesses to orally retrieve their memories may potentially increase the risk of unwanted false recall and hence of erroneous testimony (see also Loftus, 1996). While practical settings such as these clearly differ from the experimentally controlled, within-subject presentation of words explored here, this is certainly an empirical question that warrants further research.

### *Concluding Remarks*

The emphasis in prior studies of the production effect has been on the benefit of production—the enhanced recognition of actually studied material. This makes sense given that the lists in prior studies consisted of unrelated words. Presumably, saying those words aloud enhanced verbatim traces. We saw this benefit again here for the studied words. But we have shown that production also has a "dark side": It simultaneously increases the likelihood of false recognition of semantically related information, presumably because that information matches the gist trace of what was actually studied. Reading aloud leads to enhanced memory for meaning or for context, or both, which in turn provides the basis both for accurately remembering the items presented in a list and for falsely remembering unstudied associated items as having been studied.

## Data availability

Data and analysis code for all experiments are available at https://osf.io/tjm2n/files.

## CRediT authorship contribution statement

**Xinyi Lu:** Writing – review & editing, Writing – original draft, Visualization, Software, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Jianqin Wang:** Writing – review & editing, Methodology, Investigation, Conceptualization. **Colin M. MacLeod:** Writing – review & editing, Supervision, Project administration, Methodology, Investigation, Conceptualization.

## Funding

This research was supported by an Alexander Graham Bell Canada Graduate Scholarship from the Natural Sciences and Engineering Research Council (NSERC) of Canada to XL, by NSERC Discovery Grant [A7459] to CMM, and by a grant from the Ministry of Education in China (MOE) Project of Humanities and Social Sciences [24YJA190016] to JW.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

We thank Jocelyn Cheng, Batul Karimjee, and Kristen Sullivan for their assistance with data collection. Correspondence may be directed to xinyilu@stanford.edu (X. Lu) or wjq@fudan.edu.cn (J. Wang).

[5] There were, of course, small procedural differences between their studies and ours, including different presentation durations (3 s for D & S; 2 s for C & P; 2.5 s for us), different numbers of studied lists (16 for Dodson & Schacter, 2001; 12 for Culbreth & Putnam, 2019; 10 for us), different numbers of words per list (15 for D & S; 10 for C & P; 12 for us), and different numbers of test items (96 for D & S; 72 for C & P; 80 for us). The D & S study likely suffered from the noise inherent in small samples, whereas the C & P study may have been less sensitive than ours due to using shorter lists and fewer test items; of course, it may simply reflect a Type II error.

## Appendix

*Study lists (critical lures marked with *; items appearing on the recognition test in bold).*
*Backward associative strength of each list in parentheses (M = 0.26).*

| | | | | |
|---|---|---|---|---|
| **bed** | **truck** | **table** | **apple** | **blouse** |
| rest | bus | sit | vegetable | sleeves |
| awake | automobile | seat | orange | pants |
| tired | vehicle | couch | kiwi | tie |
| **dream** | **drive** | **desk** | **citrus** | **button** |
| wake | jeep | recliner | ripe | shorts |
| snooze | ford | sofa | pear | polo |
| blanket | keys | cushion | banana | collar |
| doze | garage | swivel | berry | pocket |
| **snore** | **highway** | **stool** | **cherry** | **belt** |
| nap | van | rocking | basket | linen |
| drowsy | taxi | bench | cocktail | cuffs |
| **sleep* (0.52)** | **car* (0.38)** | **chair* (0.37)** | **fruit* (0.25)** | **shirt* (0.21)** |
| **shoe** | **sour** | **hard** | **white** | **boy** |
| hand | candy | light | dark | dolls |
| toe | sugar | pillow | charred | female |
| kick | bitter | plush | night | young |
| **sandals** | **taste** | **cotton** | **funeral** | **dress** |
| walk | nice | fur | colour | pretty |
| ankle | honey | touch | grief | niece |
| arm | soda | fluffy | death | beautiful |
| boot | chocolate | feather | ink | cute |
| **inch** | **cake** | **downy** | **coal** | **date** |
| sock | tart | kitten | brown | daughter |
| knee | pie | tender | grey | sister |
| **foot* (0.19)** | **sweet* (0.22)** | **soft* (0.18)** | **black* (0.16)** | **girl* (0.12)** |

*Unrelated distractor items (critical lures marked with *).*

| | | | | |
|---|---|---|---|---|
| **nurse** | **hill** | **thread** | **fast** | **door** |
| **health** | **top** | **sharp** | **snail** | **ledge** |
| **patient** | **goat** | **thorn** | **hesitant** | **frame** |
| **doctor*** | **mountain*** | **needle*** | **slow*** | **window*** |
| **nose** | **garbage** | **sky** | **want** | **house** |
| **hear** | **sewage** | **low** | **think** | **cabin** |
| **reek** | **scraps** | **airplane** | **true** | **roses** |
| **smell*** | **trash*** | **high*** | **wish*** | **cottage*** |

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jml.2024.104584.

## References


Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior, 22*(3), 261–295. https://doi.org/10.1016/S0022-5371(83)90201-3

Arndt, J. (2010). The role of memory activation in creating false memories of encoding context. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(1), 66–79. https://doi.org/10.1037/a0017394

Ben-Shachar, M., Lüdecke, D., & Makowski, D. (2020). effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software, 5*(56), 2815. https://doi.org/10.21105/joss.02815

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. https://doi.org/10.1037/a0028466

Brainerd, C. J., Gomes, C. F. A., & Moran, R. (2014). The two recollections. *Psychological Review, 121*(4), 563–599. https://doi.org/10.1037/a0037668

Brainerd, C. J., Reyna, V. F., Wright, R., & Mojardin, A. H. (2003). Recollection rejection: False-memory editing in children and adults. *Psychological Review, 110*(4), 762–784. https://doi.org/10.1037/0033-295X.110.4.762

Brainerd, C. J., Wright, R., Reyna, V. F., & Mojardin, A. H. (2001). Conjoint recognition and phantom recollection. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(2), 307–327. https://doi.org/10.1037/0278-7393.27.2.307

Coane, J. H., & McBride, D. M. (2006). The role of test structure in creating false memories. *Memory & Cognition, 34*(5), 1026–1036. https://doi.org/10.3758/BF03193249

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge Academic.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82*(6), 407–428. https://doi.org/10.1037/0033-295X.82.6.407

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

Culbreth, J. L., & Putnam, A. L. (2019). November). *Speaking of memory: The effect of production on the DRM illusion. Poster presented at the annual meeting of the Psychonomic Society*.

Cyr, V., Poirier, M., Yearsley, J. M., Guitard, D., Harrigan, I., & Saint-Aubin, J. (2022). The production effect over the long term: Modeling distinctiveness using serial positions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 48*(12), 1797–1820. https://doi.org/10.1037/xlm0001093

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology, 58*(1), 17–22. https://doi.org/10.1037/h0046671

Dodd, M. D., & MacLeod, C. M. (2004). False recognition without intentional learning. *Psychonomic Bulletin & Review, 11*(1), 137–142. https://doi.org/10.3758/BF03206473

Dodson, C. S., & Schacter, D. L. (2001). "If I had said it I would have remembered it": Reducing false memories with a distinctiveness heuristic. *Psychonomic Bulletin & Review, 8*(1), 155–161. https://doi.org/10.3758/BF03196152

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/BF03193146

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

Fawcett, J. M., & Ozubko, J. D. (2016). Familiarity, but not recollection, supports the between-subject production effect in recognition memory. *Canadian Journal of Experimental Psychology, 70*(2), 99–115. https://doi.org/10.1037/cep0000089

Fawcett, J. M., Bodner, G. E., Paulewicz, B., Rose, J., & Wakeham-Lewis, R. (2022). Production can enhance semantic encoding: Evidence from forced-choice recognition with homophone versus synonym lures. *Psychonomic Bulletin & Review, 29*(6), 2256–2263. https://doi.org/10.3758/s13423-022-02140-x

Forrin, N. D., & MacLeod, C. M. (2018). This time it's personal: The memory benefit of hearing oneself. *Memory, 26*(4), 574–579. https://doi.org/10.1080/09658211.2017.1383434

Forrin, N. D., MacLeod, C. M., & Ozubko, J. D. (2012). Widening the boundaries of the production effect. *Memory & Cognition, 40*(7), 1046–1055. https://doi.org/10.3758/s13421-012-0210-8

Gardiner, J. M. (1988). Functional aspects of recollective experience. *Memory & Cognition, 16*(4), 309–313. https://doi.org/10.3758/BF03197041

Hicks, J. L., & Marsh, R. L. (1999). Remember-know judgments can depend on how memory is tested. *Psychonomic Bulletin & Review, 6*(1), 117–122. https://doi.org/10.3758/BF03210818

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95*(4), 528. https://doi.org/10.1037/0033-295X.95.4.528

Huff, M. J., & Bodner, G. E. (2013). When does memory monitoring succeed versus fail? Comparing item-specific and relational encoding in the DRM paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(4), 1246–1256. https://doi.org/10.1037/a0031338

Huff, M. J., Bodner, G. E., & Fawcett, J. M. (2015). Effects of distinctive encoding on correct and false memory: A meta-analytic review of costs and benefits and their origins in the DRM paradigm. *Psychonomic Bulletin & Review, 22*(2), 349–365. https://doi.org/10.3758/s13423-014-0648-8

Huff, M. J., Bodner, G. E., & Gretz, M. R. (2021). Distinctive encoding of a subset of DRM lists yields not only benefits, but also costs and spillovers. *Psychological Research, 85*(1), 280–290. https://doi.org/10.1007/s00426-019-01241-y

Jamieson, R. K., Mewhort, D. J. K., & Hockley, W. E. (2016). A computational account of the production effect: Still playing twenty questions with nature. *Canadian Journal of Experimental Psychology, 70*(2), 154–164. https://doi.org/10.1037/cep0000081

Kelly, M. O., Ensor, T. M., Lu, X., MacLeod, C. M., & Risko, E. F. (2022). Reducing retrieval time modulates the production effect: Empirical evidence and computational accounts. *Journal of Memory and Language, 123*, 1–14. https://doi.org/10.1016/j.jml.2021.104299

Lakens, D., & Caldwell, A. R. (2021). Simulation-based power analysis for factorial analysis of variance designs. In *Advances in Methods and Practices in Psychological Science*. https://doi.org/10.1177/2515245920951503

Lampinen, J. M., Meier, C. R., Arnal, J. D., & Leding, J. K. (2005). Compelling untruths: Content borrowing and vivid false memories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(5), 954–963. https://doi.org/10.1037/0278-7393.31.5.954

Loftus, E. F. (1996). *Eyewitness Testimony*. Cambridge, MA: Harvard University Press.

Lu, X., Kelly, M. O., & Risko, E. F. (2020). Offloading information to an external store increases false recall. *Cognition, 205*, Article 104428. https://doi.org/10.1016/j.cognition.2020.104428

MacLeod, C. M., & Bodner, G. E. (2017). The production effect in memory. *Current Directions in Psychological Science, 26*(4), 390–395. https://doi.org/10.1177/0963721417691356

MacLeod, C. M., Gopie, N., Hourihan, K. L., Neary, K. R., & Ozubko, J. D. (2010). The production effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(3), 671–685. https://doi.org/10.1037/a0018785

Mangels, J. A., Picton, T. W., & Craik, F. I. M. (2001). Attention and successful episodic encoding: An event-related potential study. *Cognitive Brain Research, 11*(1), 77–95. https://doi.org/10.1016/S0926-6410(00)00066-5

McGeoch, J. A. (1932). Forgetting and the law of disuse. *Psychological Review, 39*(4), 352–370. https://doi.org/10.1037/h0069819

Murdock, B. B. (1993). TODAM2: A model for the storage and retrieval of item, associative, and serial-order information. *Psychological Review, 100*(2), 183–203. https://doi.org/10.1037/0033-295X.100.2.183

Ochsner, K. N. (2000). Are affective events richly recollected or simply familiar? The experience and process of recognizing feelings past. *Journal of Experimental Psychology: General, 129*(2), 242–261. https://doi.org/10.1037/0096-3445.129.2.242

Ozubko, J. D., Gopie, N., & MacLeod, C. M. (2012a). Production benefits both recollection and familiarity. *Memory & Cognition, 40*(3), 326–338. https://doi.org/10.3758/s13421-011-0165-1

Ozubko, J. D., Hourihan, K. L., & MacLeod, C. M. (2012b). Production benefits learning: The production effect endures and improves memory for text. *Memory, 20*(7), 717–727. https://doi.org/10.1080/09658211.2012.699070

Payne, D. G., Elie, C. J., Blackwell, J. M., & Neuschatz, J. S. (1996). Memory illusions: Recalling, recognizing, and recollecting events that never occurred. *Journal of Memory and Language, 35*(2), 261–285. https://doi.org/10.1006/jmla.1996.0015

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review, 116*(1), 129–156. https://doi.org/10.1037/a0014420

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

Quinlan, C. K., & Taylor, T. L. (2019). Mechanisms underlying the production effect for singing. *Canadian Journal of Experimental Psychology, 73*(4), 254–264. https://doi.org/10.1037/cep0000179

R Core Team. (2022). *R: A language and environment for statistical computing*. [Computer software].

Reyna, V. F., & Brainerd, C. J. (1995). Fuzzy-trace theory: An interim synthesis. *Learning and Individual Differences, 7*(1), 1–75. https://doi.org/10.1016/1041-6080(95)90031-4

Roberts, B. R. T., Hu, Z. S., Curtis, E., Bodner, G. E., McLean, D., & MacLeod, C. M. (2024). Reading text aloud benefits memory but not comprehension. *Memory & Cognition, 52*(1), 57–72. https://doi.org/10.3758/s13421-023-01442-2

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(4), 803–814. https://doi.org/10.1037/0278-7393.21.4.803

Roediger, H. L., McDermott, K. B., Pisoni, D., & Gallo, D. (2004). Illusory recollection of voices. *Memory, 12*(5), 586–602. https://doi.org/10.1080/09658210344000125

Roediger, H. L., Watson, J. M., McDermott, K. B., & Gallo, D. A. (2001). Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin & Review, 8*(3), 385–407. https://doi.org/10.3758/BF03196177

Schacter, D. L., Cendan, D. L., Dodson, C. S., & Clifford, E. R. (2001). Retrieval conditions and false recognition: Testing the distinctiveness heuristic. *Psychonomic Bulletin & Review, 8*(4), 827–833. https://doi.org/10.3758/BF03196224

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review, 4*, 145–166. https://doi.org/10.3758/BF03209391

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2023). *afex: Analysis of factorial experiments* (Version 1.3-0) [Computer software]. https://CRAN.R-project.org/package=afex.

Smith, R. E., & Hunt, R. R. (1998). Presentation modality affects false memory. *Psychonomic Bulletin & Review, 5*, 710–715. https://doi.org/10.3758/BF03208850

Stadler, M. A., Roediger, H. L., & McDermott, K. B. (1999). Norms for word lists that create false memories. *Memory & Cognition, 27*(3), 494–500. https://doi.org/10.3758/BF03211543

Tulving, E. (1985). *Memory and consciousness. Canadian Psychology/Psychologie canadienne, 26*(1), 1–12. https://doi.org/10.1037/h0080017

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(6), 1341–1354. https://doi.org/10.1037/0278-7393.20.6.1341

Yonelinas, A. P. (1997). Recognition memory ROCs for item and associative information: The contribution of recollection and familiarity. *Memory & Cognition, 25*(6), 747–763. https://doi.org/10.3758/BF03211318

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. https://doi.org/10.1006/jmla.2002.2864

Yonelinas, A. P., & Jacoby, L. L. (1995). The relation between remembering and knowing as bases for recognition: Effects of size congruency. *Journal of Memory and Language, 34*(5), 622–643. https://doi.org/10.1006/jmla.1995.1028

Zhou, Y., & MacLeod, C. M. (2022). Production as a distinctive contextual cue for retrieving intentionally forgotten information. *Canadian Journal of Experimental Psychology, 76*(3), 226–233. https://doi.org/10.1037/cep0000284