Reducing retrieval time modulates the production effect: Empirical evidence and computational accounts

Megan O. Kelly[1]

Tyler M. Ensor[2]

Xinyi Lu[1]

Colin M. MacLeod[1]

Evan F. Risko[1]

[1]*University of Waterloo*

[2]*California State University, Bakersfield*

This is the peer reviewed version of the following article: **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, 104299.***, which has been published in final form at https://doi.org/10.1016/j.jml.2021.104299. ©2022. This manuscript version is made available under the CC-BY-NC-ND 4.0 license https://creativecommons.org/licenses/by-nc-nd/4.0/

For correspondence, please contact Megan O. Kelly at megan.kelly@uwaterloo.ca.

All the data and analyses code are available here: osf.io/un5ca/

The preregistrations of Experiments 1, 2, and 3, are available here, respectively: osf.io/kwzj3, osf.io/xt84h, and osf.io/k8nqt.

Abstract

Memory is reliably better for information read aloud relative to information read silently—*the production effect*. Three preregistered experiments examined whether the production effect arises from a more time-consuming retrieval process operating at test that benefits items that were produced at study. Participants studied items either aloud or silently and then completed a recognition test which required responding within a short deadline, under the assumption that a time-consuming retrieval process would be less able to operate when less time was available. Results generally supported this prediction. Even under speeded responding instructions, however, there was a robust production effect, suggesting that other, more rapid, processes also contribute to the production effect. Based on two extant verbal accounts, a computational model of the production effect using REM is introduced.

Actively engaging with to-be-remembered information has been demonstrated to enhance memory accuracy relative to more passive methods, such as simply silently reading. This has been demonstrated with a variety of manipulations, including generating (Slamecka & Graf, 1978), drawing (Wammes, Meade, & Fernandes, 2018), and enacting (Engelkamp, 1998). Consistent with these findings is an even simpler form of engagement: When individuals vocalize to-be-remembered information, they remember it better than information that they read silently (e.g., MacLeod et al., 2010; Wakeham-Lewis et al, 2021). This memory effect—the *production effect*—is robust, having been replicated using a variety of methods, including spelling, writing, mouthing, whispering, and singing (e.g., Forrin et al., 2012; Quinlan & Taylor, 2013). MacLeod and Bodner (2017) provide a brief review of the burgeoning literature on production.

Determining the mechanism(s) underlying an effect is always challenging, and this is no less true in the case of the production effect (Bodner, Jamieson, Cormack, MacDonald, & Bernstein, 2016; MacLeod et al., 2010). In the current investigation, we tested the account of Forrin et al. (2012) centered on the concept of *relative distinctiveness*, an explanation originally put forward by Conway and Gathercole (1987; and championed more recently by MacLeod et al., 2010). The central claim in this account is that the act of producing an item results in the formation of an item-specific and "distinctive" record within memory. Produced items are distinctive in the sense that they have additional, item-specific, production-associated features encoded with them, thereby differentiating them from items studied silently—and from other items studied aloud. Forrin et al. (2012) suggested that these production-associated details could consist of the motoric (e.g., moving one's mouth to pronounce the item) and/or perceptual (e.g., hearing one's own voice) features involved in producing each item. These details are encoded along with other information about the word (e.g., its meaning). For example, in a recent

computational model of the production effect using MINERVA2, a multitrace model of memory retrieval (see Hintzman, 1984; 1986), Jamieson, Mewhort, and Hockley (2016) captured this aspect of the Forrin et al. (2012) account by adding features to each item (each trace) that was "produced" at encoding/study. In the present investigation, we further develop this account by examining how these production-associated features might influence retrieval during recognition. In particular, we focus on the time available at retrieval. Extant verbal descriptions (Forrin et al., 2012; MacLeod et al., 2010), implementations (Jamieson et al., 2016), and previous work (Ozubko, Gopie, & MacLeod, 2012; but also see Fawcett & Ozubko, 2016) suggest that this might be a critical variable.

One means by which production-associated features might be used at retrieval is if individuals attempt to reinstate the encoding/study context by trying to replay the original encoding/study event. For example, Forrin and colleagues (2012) proposed the following: Any unique production provides a distinctive cue that participants can use at test to help remember studied words. In line with the proceduralist account, this distinct encoding activity is preserved in the original processing record (Kolers, 1973; Kolers & Roediger,

> 1984) and can subsequently be replayed to aid retrieval. (p. 1054)

If the production effect is due, at least in part, to such a replaying process, then it should be sensitive to the amount of time available at retrieval, assuming that engaging in such a process reflects an intentional act over and above more passive retrieval processes. Indeed, in the Jamieson et al. (2016) implementation of the Forrin et al. account, the retrieval mechanism involved multiple iterations with the production-associated features contributing only after the first retrieval attempt. Thus, their implementation also suggested that the time available at retrieval is a critical variable in the influence of production on memory.

More generally, the idea that certain details of a given encoding/study event might become available only later in a given retrieval attempt has some support in the literature (e.g., Gronlund & Ratcliff, 1989; Hintzman & Curran, 1994; McElree, Dolan, & Jacoby, 1999). In understanding the time course of retrieving associative information, Gronlund and Ratcliff (1989) had participants study word pairs and then, at test, identify whether a word pair was originally studied together—*intact*—or separately—*rearranged*. They found evidence that an additional 200 ms was needed to incorporate this associative information from study into recognition decisions. Similarly, Hintzman and Curran (1994) presented participants with a series of nouns, some of which were in plural form (e.g., *apples* rather than *apple*). In a subsequent old-new recognition test, near-target foils took the form of the targets in the opposite plurality (e.g., if *apples* had been studied, then *apple* might be tested). Participants were told to endorse only test items that exactly matched a study item; critically, they accepted more of the near-target foils under shorter signal-response-lag conditions but showed less acceptance of these foils when more time for retrieval was given. Finally, McElree and colleagues (1999) found comparable results when they manipulated study modality (heard vs. read). They asked participants only to judge items as old if they had been heard and found that response sensitivity increased as processing time at test increased. Thus, the information required to distinguish rearranged pairs from intact pairs, near-target foils from actual targets, and undesired targets from desired targets was less available earlier on in the retrieval attempt and became more available with additional time.

In recent computational work, Cox and Shiffrin (2017) modelled the delayed availability of plurality and modality information by having the relevant features contribute to the decision more slowly (after more *time steps*), than other features (also see FESTHER; Lamberts, 2002). Relevant to the present work, Cox and Shiffrin (2017) discussed a couple of potential reasons for

the delay of certain features. One reason might be that "when the relevant modality or context information is not already present in the environment or in the presentation of the test item, participants must generate ("imagine") that information themselves and add it to the memory probe" (p. 823). For example, if participants are presented at test with visual stimuli, the auditory features of the items studied are not included in the test item itself. Thus, if participants are to use the auditory information from the original study episode as a memory cue, they must attempt to do so themselves. This idea mirrors the situation in typical studies of the production effect (in recognition) such that the production-associated features are not present in the visually displayed test item. Thus, if participants are to use the production-associated information as a memory cue for the original study episode, they would have to do so by reinstating the target productionassociated information themselves (i.e., by replaying it).

A second potential reason why certain features might take longer to contribute is if they depend more on recollection. Recollection has long been hypothesized to take longer, on average, than judging whether an item is familiar (Cox & Shiffrin, 2017; Eichenbaum, Yonelinas, & Ranganath, 2007; Yonelinas, 2002). One way that the production effect could, at least in part, rely on recollection-based processing is if participants use recollection of the production to make their recognition decision—an idea first suggested by MacLeod et al. (2010). For example, Dodson and Schacter (2001) suggested that individuals might use the recollection of the study episode of an item as the evidence needed that the item was studied, a strategy they termed the *distinctiveness heuristic*. In fact, there is evidence for an association between the production effect and recollection. Ozubko et al. (2012) found that participants were significantly more likely to report recollection for items produced at study than for silent studied items or new items. If the production effect relies on the recollection of the study episode, and if recollection requires more time on average to benefit memory, then it would be reasonable to expect that the

magnitude of the production effect would be influenced by the amount of time available at retrieval.

Taken together, there exist a number of reasons to expect that the use of productionassociated features will be sensitive to the amount of time that individuals have available at test. If the production effect is the outcome of using of such features, then it should be reduced when responding is speeded. The use of response deadlines and speeded responding in memory research has a long history and has been used to test accounts of memory phenomena in a number of domains (e.g., Gardiner et al., 1999, 2004; Sauvage, Beer, & Eichenbaum, 2009; Toth, 1996; Wammes et al. 2018; Wickelgren & Corbett, 1977). For example, Wammes and colleagues (2018) used speeded responding to disentangle a more time-consuming process from other factors contributing to retrieval accuracy in the context of the drawing effect (i.e., enhanced memory for items drawn at study). They tested recognition memory for written versus drawn items and sped responding for half of their participants (i.e., a deadline of 800 ms). Although the drawing effect persisted in both the speeded and standard (i.e., not speeded) conditions, Wammes et al. (2018) found that the size of the effect was significantly smaller under a response deadline. This finding is consistent with the observed memory advantage for drawn items as emerging, at least partially, through a more time-consuming process at retrieval.

Finally, another popular class of explanation for the production effect is the concept of memory strength (Bodner & Taikh, 2012; Bodner, Taikh, & Fawcett, 2014; Taikh & Bodner, 2016; Fawcett, 2013; Fawcett & Ozubko, 2016). According to strength-based accounts, “reading words aloud strengthens the representations of those items more so than does reading words silently, and therefore words read aloud are easier to recognise and recall than are words read silently” (Ozubko et al., 2014, p. 510). This greater strength results in the produced items being more accessible at the time of retrieval, giving rise to the memory advantage for the produced

items. We will return to the strength-based account during the modelling component of the current work.

Given the account described by Forrin et al., (2012) and the various reasons to expect that the use of production-associated information at retrieval may be sensitive to available retrieval time, we turn to the present investigation examining a manipulation of time available at retrieval together with a production manipulation.

The Present Investigation

In each of three preregistered experiments, we tested recognition memory for items read aloud (i.e., produced) versus items read silently (i.e., not produced) under both standard (i.e., not speeded) retrieval and speeded retrieval conditions. If the production effect relies, at least in part, on a retrieval act sensitive to time available at retrieval (Forrin et al., 2012; Jamieson et al., 2016; Ozubko et al., 2012), then speeding retrieval should reduce the magnitude of the production effect. If the production effect were to remain unaltered under speeded responding (within reasonable limits), then this would suggest that a process which is insensitive to time underlies the production effect. A third possibility is that the production effect will be reliably reduced but still remain under speeded conditions. This finding could support one of two ideas: (1) that the production effect relies on both a more time-sensitive retrieval process (e.g., recollection and/or context reinstatement) and at least one other process that is completed faster at retrieval or (2) that any time-taking process(es) underlying the production effect was (were) still able to occur, at least in part, under speeded conditions.

Experiment 1

Experiment 1 was preregistered at osf.io/kwzj3/. In an otherwise standard production effect experiment, two conditions differed in the deadline for responding on the recognition test. The standard condition had a deadline of 5000 ms to respond to each test item; the speeded

condition had a short deadline of 800 ms per test item.[1] All experiments were preregistered individually, and data collection for each experiment was conducted prior to the preregistrations of subsequent experiments.

Method

Participants. Participants were undergraduate psychology students from the University of Waterloo taking part for course credit. The first 48 usable sets of participant data were analyzed based on an a priori power analysis with a desired power of .80 (with $\alpha = .05$, twotailed) to detect a Cohen's *d* of 0.40. This accounts for a production effect in the speeded condition to be at least a 50% reduction compared with a production effect of approximately $d = 0.80$ in the standard, non-speeded condition. No demographic information was collected from participants in any of the current experiments. All participants across all experiments participated in person before the COVID-19 pandemic and no participants from one experiment participated in another.

Stimuli. The stimulus set consisted of 160 words with lengths ranging from four to ten letters and frequencies ranging from nine to 146371 using FreqCount[2] from *SUBTLEX-UK* (Van Heuven, Mandera, Keuleers, & Brysbaert, 2014) and frequencies ranging from nine to 39491

[1] The preregistered response deadline of 2400 ms for the standard group was due to an oversight in updating the preregistration after the decision was made to apply the 5000-ms response deadline instead. As such, we deviated from the preregistered response deadline only for the standard condition.

[2] Not including *campground* which was not in the database.

using FreqCount[3] from *SUBTLEX-US* (Brysbaert & New, 2009). The 160 items were randomly arranged to form four unique lists containing 40 items each (available at osf.io/un5ca/). Lists were counterbalanced such that each list appeared equally often as the set of items for speeded targets, speeded foils, standard targets, and standard foils. The presentation order of items within each list was also randomized during study/encoding and again during test/retrieval.

Procedure. Participants sat approximately 50 cm in front of a computer monitor and followed instructions given by the monitor and the researcher in the two blocks of the experiment. Each block consisted of a study phase and a recognition test phase. During the study phase, participants were presented with 40 items, half of which randomly appeared in white font (to be read silently) and the other half of which appeared in blue font (to be read aloud), all against a black background. Each item was presented for 2000 ms and was followed by a 400-ms blank interval.

During the recognition test phase, participants were presented with the 40 study items plus 40 foils in a new randomized order, all in gray font against a black background. Participants keyed a response of 'F' for studied items (regardless of whether they were read aloud or silently), and 'J' for non-studied items. In the standard block, participants were given 5000 ms to respond to each presented item. In the speeded block, participants were given 800 ms to respond to each presented item (a value selected based on Wammes et al., 2018). If a participant did not respond before the deadline, a tone sounded to notify them, and no response was collected. Prior to starting each recognition test, participants were reminded that guessing was better than no

[3] Not including *avenue, captain, foundation, harbor, matrix, uncle*, or *valley*, which were not in the database.

response. Half of the participants had the speeded block first and the standard block second; the remaining participants had the two blocks in the reverse order.

Results

Hit rate and response time for hits were each analyzed using a 2 (silent vs. aloud) x 2 (standard vs. speeded) repeated measures analysis of variance (ANOVA). False alarm rate was analyzed using a (standard vs. speeded) repeated measures ANOVA. For each of hit rate and response time for hits, we also conducted an analogous analysis considering block type order (standard first vs. speeded first) as a between-participants factor. Cases in which order interacted with the other factors of interest are noted and presented in Section I of the Supplementary Materials; however, in Experiment 1, no order analyses revealed significant interactions. Note that we report the preregistered analyses for sensitivity in Section II of the Supplementary Materials because the current computation of sensitivity is partially redundant to hit rate—the silent and aloud targets are compared against the same foils. This is the case for all experiments.

Data from 18 participants were replaced as they did not meet the preregistered inclusion criteria: They did not respond in time to at least 80% of recognition trials before the deadline and/or they had hit rates at or below 50% on one of the recognition tests. Trials (4.9% of total trials) wherein participants did not provide responses (i.e., the deadline was reached before a response was provided) were removed before analysis. Given the high proportion of participants needing replacement, we ran parallel analyses including the 18 participants (i.e., $N = 66$) and report these results in square brackets following the analogous main results (the tables and figures accompanying these data are available in Section III of the Supplementary Materials). Results were qualitatively the same as when excluding them. Means of hit rate, false alarm rate,

and correct response time are presented in Table 1 as a function of experimental condition.[4] Data and analyses code for Experiment 1 are available at osf.io/un5ca/.

Hit rate

Mean hit rates as a function of condition are presented in Figure 1. There was a main effect of production on hit rate such that there were higher hit rates for items read aloud at study than for items read silently (aloud: .82 [.81]; silent: .64 [.64]), $F(1, 47) = 118.07$, $p < .001$, $\eta_G^2 = .28$ [$F(1, 65) = 153.45$, $p < 001$, $\eta_G^2 = .24$], indicating a typical production effect, $d = 1.15$ [1.09]. Additionally, there was a main effect of speeding on hit rate such that there were significantly higher hit rates in the standard condition than in the speeded condition (standard: .78 [.78]; speeded: .68 [.67]), $F(1, 47) = 21.52$, $p < .001$, $\eta_G^2 = 0.10$ [$F(1, 65) = 28.31$, $p < 001$, $\eta_G^2 = .09$]. The production effect in the standard condition did not differ from that in the speeded condition (standard: .20 [.20]; speeded: .17 [.15]), $F(1, 47) = 1.01$, $p = .321$, $\eta_G^2 < .01$, [$F(1, 65) = 3.58$, $p = 063$, $\eta_G^2 = .01$].

False alarm rate

The analysis of false alarm rate was not preregistered. As would be expected, however, false alarm rate was significantly lower in the standard condition than in the speeded condition (standard: .12 [.13]; speeded: .20 [.24]), $F(1, 47) = 23.87$, $p < .001$, $\eta_G^2 = .08$ [$F(1, 65) = 29.13$, $p < .001$, $\eta_G^2 = .10$].

[4] Means for incorrect response times by item type, speeding condition, and experiment are available at https://osf.io/un5ca/

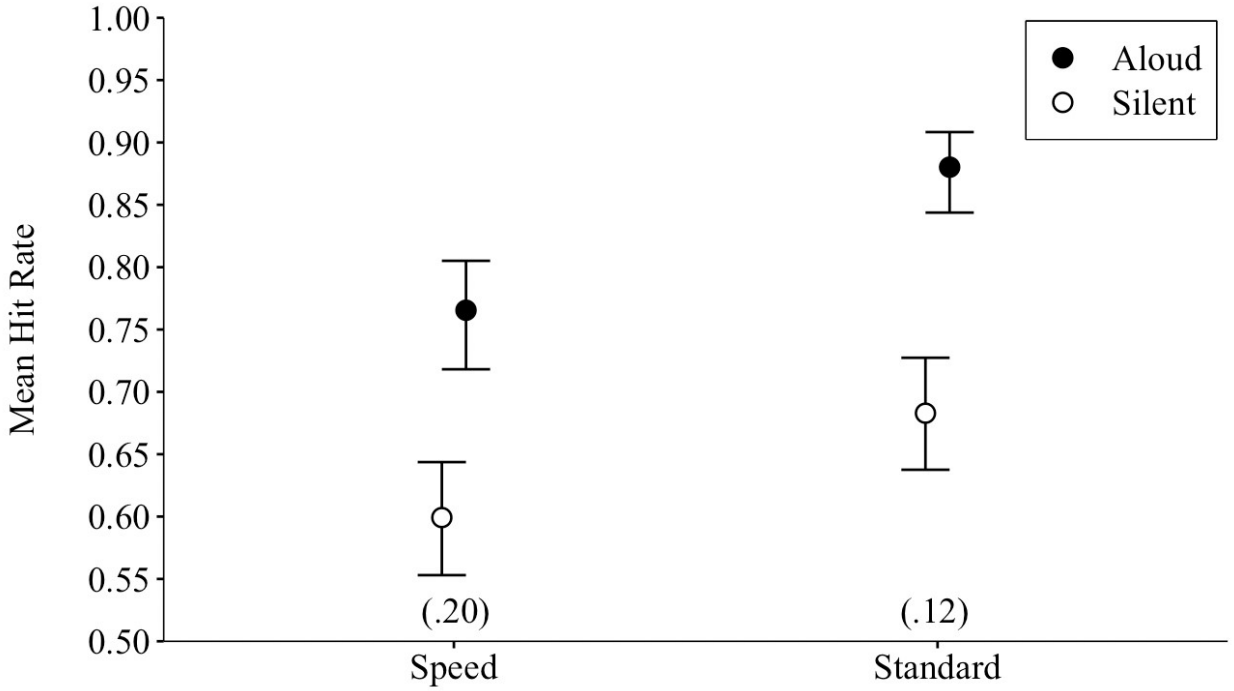


Figure 1. Experiment 1: Mean hit rate by study/encoding condition and speed manipulation. Mean false alarm rates are in parentheses by speed manipulation. Error bars are bias-corrected accelerated bootstrap 95% confidence intervals using 10,000 replications.

Response time for hits

There was a main effect of production on response time for hits such that participants responded significantly faster to items read aloud than to items read silently (aloud: 718 ms [731 ms]; silent: 767 ms [791 ms]), $F(1, 47) = 12.39$, $p < .001$, $\eta_G^2 = .04$ [$F(1, 65) = 19.08$, $p < .001$, $\eta_G^2 = .04$]. As expected, there was a main effect of speeded responding on hits such that response times were significantly slower in the standard condition than in the speeded condition (standard: 906 ms [934 ms]; speeded: 579 ms [588 ms]), $F(1, 47) = 184.84$, $p < .001$, $\eta_G^2 = .70$ [$F(1, 65) = 135.44$, $p < .001$, $\eta_G^2 = .59$]. The interaction between production and deadline was significant such that the production effect in response time for hits was smaller in the speeded condition than in the standard condition (standard: -81 ms [-119 ms]; speeded: -2 ms [-0.5 ms]), $F(1, 47) =$

8.97, $p = .004$, $\eta_G^2 = .04$ [$F(1, 65) = 17.05$, $p = .001$, $\eta_G^2 = .04$]. There was a significant production effect in response time for hits in the standard condition, $t(47) = 3.34$, $p = .001$, $d = 0.48$ [$t(65) = 4.34$, $p < .001$, $d = 0.53$], but not in the speeded condition, $t(47) = 0.34$, $p = .733$, $d = 0.05$ [$t(65) = 0.08$, $p = .934$, $d = 0.01$].

Table 1

Experiment 1: Mean Hit Rate, False Alarm Rate, and Response time for Hits and Correct Rejections for Each Item Type by Condition (CI_{95} in Parentheses)

		Silent	Aloud	New
Standard	Hits and false alarms	.68 [.64, .73]	.88 [.84, .91]	.12 [.09, .16]
	Response time (ms) for hits and correct rejections	922 [868, 988]	841 [803, 885]	918 [860, 992]
	Timeout proportion	.001 [0, .003]	.001 [0, .003]	.001 [0, .003]
Speeded				
	Hits and false alarms	.60 [.55, .64]	.77 [.72, .80]	.20 [.16, .25]
	Response time (ms) for hits and correct rejections	590 [576, 602]	588 [575, 599]	580 [568, 592]
	Timeout proportion	.11 [.09, .13]	.09 [.07, .10]	.10 [.08, .12]

Note. Confidence intervals are bias-corrected accelerated bootstrap 95% confidence intervals using 10,000 replications.

Exploratory

We also analyzed the proportion of timeouts (i.e., the proportion of trials wherein participants did not respond before the deadline). Given that the proportion of timeouts was negligible in the standard condition (see Table 1), we focused on analyzing the speeded condition. The proportion of timeouts did not differ among silent, aloud, and foil items (silent: .11 [.14]; aloud: .09 [.13]; foil: .10 [.14]), $F(2, 94) = 1.85$, $p = .163$, $\eta_G^2 = .02$ [$F(2, 130) = 0.87$, $p = .421$, $\eta_G^2 < .01$].

Discussion

We found consistent effects of production on hit rate and hit rate response time, as well as consistent effects of speeded responding on hit rate, false alarm rate, and correct response time. The results of Experiment 1 did not provide clear evidence of an effect of speeding recognition responses on the size of the production effect in hit rate, which is inconsistent with the prediction we derived from the account described by Forrin and colleagues. We did, however, observe an interaction between production and speeded responding for response time. Taken together, the results of Experiment 1 are mixed. The primary dependent variable in studies of the production effect—hit rate—did not show the interaction predicted based on the account described by Forrin et al. (2012). That said, the means were in the predicted direction and the interaction was marginal when all of the data (i.e., including those excluded in the main data set) was included (the interaction was also significant in the pre-registered sensitivity analysis reported in Section II of the Supplementary Materials). Thus, Experiment 2 further examined the interaction between production and speeding.

Experiment 2

In Experiment 2, we sought to provide a second test of the interaction between production and speeded retrieval by reducing the response deadline in the speeded condition. Specifically, rather than 800 ms, the speeded retrieval condition now featured a 750 ms deadline with the intent of reducing the time available at retrieval but with caution about potentially losing responses due to too little time to respond. Experiment 2 was preregistered at osf.io/xt84h/.

Method

The method of Experiment 2 was identical with that of Experiment 1 with three exceptions. First, the response deadline of the recognition task was set to 750 ms (i.e., 50 ms earlier; although this was incorrectly preregistered as using an identical deadline to that in Experiment 1). We also changed the minimum response rate to 60% (i.e., from 80% in Experiment 1) and we increased power by collecting 64 sets of usable participant data based on the determined *N* of Experiment 1 (i.e., 48), and the addition of two full cycles of counterbalances (i.e., 2 x 8 = 16). As in Experiment 1, participants were undergraduate students from the University of Waterloo taking part for course credit, and none had taken part in Experiment 1. All other aspects of *Participants, Stimuli,* and *Procedure* for Experiment 2 were identical to those of Experiment 1.

Results

As in Experiment 1, hit rate and response times were each analyzed using a 2 (silent vs. aloud) x 2 (standard vs. speeded) repeated measures ANOVA; false alarm rate was analyzed with a (standard vs. speeded) repeated measures ANOVA. For each of hit rate, false alarm rate, and response time (for hits and correct rejections), we also conducted an analogous analysis considering block type order (standard first vs. speeded first) as a between-participants factor. Only the cases where order interacted with the other factors of interest are noted and are presented in Section I of the Supplementary Materials. These few interactions demonstrated that when the speeded block occurred first, this occasionally led to a smaller effect of speed compared to when the speeded block occurred second. However, in general, these instances did not qualitatively affect the results reported here. Note that the preregistration for Experiment 2

specified analyzing response time for false alarms specifically but we have deviated from that to, instead, report response time for correct rejections.

We replaced data from 25 participants because their response rates were lower than the 60% minimum criterion and/or their recognition accuracy was at or below 50% on one of the recognition tests. Trials wherein participants did not provide responses (i.e., the deadline was reached before a response was provided) were removed before analysis (6% of total trials). Given the high proportion of participants needing replacement, we ran parallel analyses including the 25 excluded participants (i.e., $N = 89$). Results were consistent with those when excluding them (these results, which include the preregistered exclusions, are reported in square brackets following each of the analogous main results; see Section III of the Supplementary Materials for the tables and figures accompanying these data), except for two minor deviations from the analyses with $N = 64$ (see the sections titled *Response time for hits* and *Exploratory*). Means of hit rate, false alarm rate, and correct response time are presented in Table 2 as a function of experimental condition. Data and analyses code for Experiment 2 are available at osf.io/un5ca/.

Hit rate

The main effect of production was significant such that hit rate was significantly higher for items read aloud than for those read silently (aloud: .79 [.77]; silent: .65 [.63]), $F(1, 63) = 100.90$, $p < .001$, $\eta_G^2 = .16$ [$F(1, 88) = 126.48$, $p < .001$, $\eta_G^2 = .14$], a typical production effect, $d = 0.93$ [0.82]. There was also a main effect of speeding such that hit rate was significantly higher in the standard condition than in the speeded condition (standard: .78 [.77]; speeded: .66 [.62]), $F(1, 63) = 47.20$, $p < .001$, $\eta_G^2 = .12$ [$F(1, 88) = 65.87$, $p < .001$, $\eta_G^2 = .16$]. Critically—and unlike in Experiment 1—the interaction between production and speeding was significant, such that the production effect was larger in the standard condition than in the speeded condition

(standard: .17 [.19]; speeded: .11 [.09]), $F(1, 63) = 5.82$, $p = .019$, $\eta_G^2 = .01$ [$F(1, 88) = 19.57$, $p < .001$, $\eta_G^2 = .02$]. The production effect was, however, significant in both conditions: standard, $t(63) = 9.14$, $p < .001$, $d = 1.14$ [$t(88) = 11.43$, $p < .001$, $d = 1.21$], and speeded, $t(63) = 5.98$, $p < .001$, $d = 0.75$ [$t(88) = 4.94$, $p < .001$, $d = 0.52$]. Mean hit rates by production and speed manipulation are presented in Figure 2.

False alarm rate

False alarm rate was significantly lower in the standard condition than in the speeded condition (standard: .16 [.16]; speeded: .23 [.29]), $F(1, 63) = 25.75$, $p < .001$, $\eta_G^2 = .08$ [$F(1, 88) = 45.70$, $p < .001$, $\eta_G^2 = .14$], a typical mirror effect.

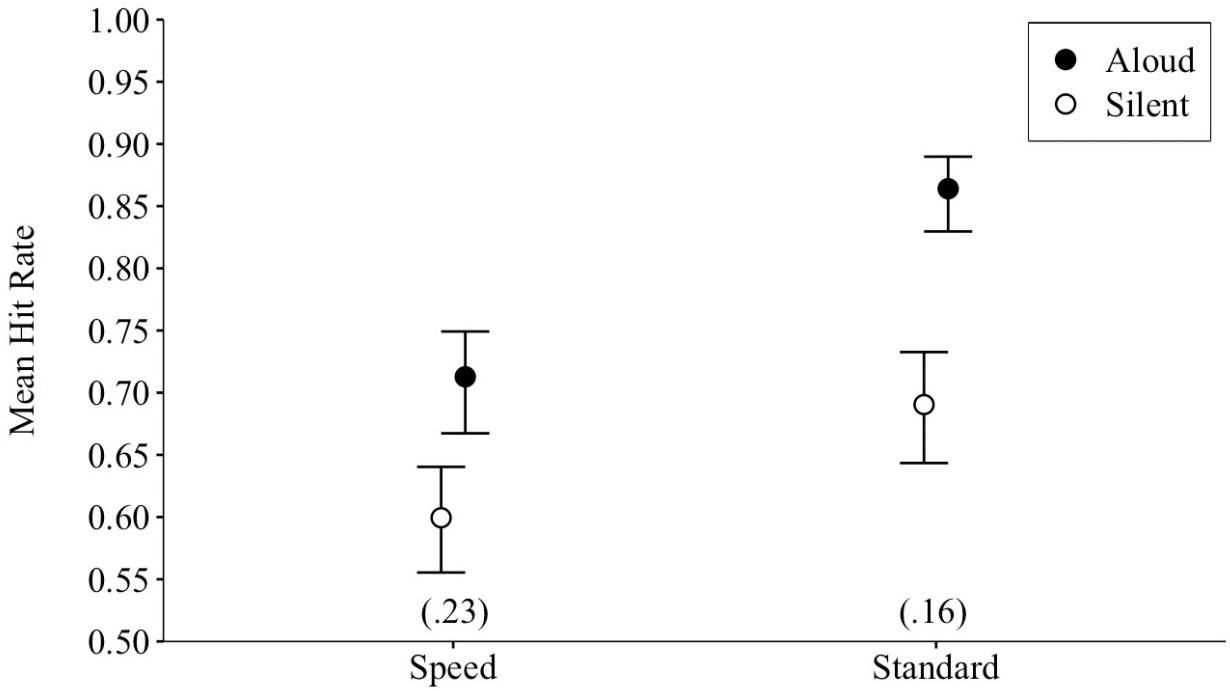


Figure 2. Experiment 2: Mean hit rate by item condition and by speed manipulation. Mean false alarm rates are in parentheses by speed manipulation. Error bars are bias-corrected accelerated bootstrap 95% confidence intervals using 10,000 replications.

Response time for hits

There was a significant main effect of production on response time for hits, such that response time was significantly faster for aloud items than for silent items (aloud: 712 ms [728 ms]; silent: 736 ms [760 ms]), $F(1, 63) = 6.64$, $p = .012$, $\eta_G^2 = .01$ [$F(1, 88) = 16.09$, $p < .001$, $\eta_G^2 = .01$]. Not surprisingly, there was also a significant main effect of speeding on response time for hits such that response time was significantly slower in the standard condition compared with the speeded condition (standard: 882 ms [926 ms]; speeded: 566 ms [562 ms]), $F(1, 63) = 206.63$, $p < .001$, $\eta_G^2 = .70$ [$F(1, 88) = 246.09$, $p < .001$, $\eta_G^2 = .54$]. Although the pattern was similar to that in Experiment 1, in Experiment 2 the production effect in response time for hits did not differ significantly between the standard condition and the speeded condition (standard: 40 ms; speeded: -6 ms), $F(1, 63) = 2.65$, $p = .108$, $\eta_G^2 = .01$. [When not excluding any participants ($N = 89$), the production effect in response time for hits was significantly different between the standard condition and the speeded condition (standard: -59 ms; speeded: -3 ms), $F(1, 88) = 9.00$, $p = .004$, $\eta_G^2 = .01$.]

Response time for correct rejections

For correct rejections, as expected, a standard versus speeded repeated measures ANOVA revealed that response time was significantly faster in the speeded condition than in the standard condition (standard: 945 ms [999 ms]; speeded: 550 ms [541 ms]), $F(1, 63) = 133.58$, $p < .001$, $\eta_G^2 = .50$ [$F(1, 87) = 187.36$, $p < .001$, $\eta_G^2 = .51$].[5] No other effects were significant.

Table 2

Experiment 2: Mean Hit Rate, False Alarm Rate, and Response time for Hits and Correct Rejections for Each Item by Condition (CI_{95} in Parentheses)

[5] In the analysis with all participants, one participant had to be removed as they never made a correct rejection.

	Silent	Aloud	New
Standard			
Hit and false alarm rate	.69 [.64, .73]	.86 [.83, .89]	.16 [.14, .19]
Response time (ms) for hits and correct rejections	902 [855, 959]	862 [821, 911]	945 [888, 1025]
Timeout proportion	.003 [.001, .006]	.001 [0, .002]	0 [0, 0]
Speeded			
Hit and false alarm rate	.60 [.55, .64]	.71 [.67, .75]	.23 [.20, .27]
Response time (ms) for hits and correct rejections	569 [556, 581]	563 [549, 574]	550 [537, 561]
Timeout proportion	.14 [.12, .16]	.10 [.08, .11]	.11 [.09, .13]

Note. All confidence intervals are bias-corrected accelerated bootstrap 95% confidence intervals using 10,000 replications.

Exploratory

As in Experiment 1, we analyzed the proportion of timeouts in the speeded condition (this analysis was not preregistered). Unlike in Experiment 1, the proportion of timeouts significantly differed across item types (silent: .14 [.16]; aloud: .10 [.13]; foil: .11 [.15]), $F(2, 126) = 6.06$, $p = .003$, $\eta_G^2 = .04$ [$F(2, 176) = 4.18$, $p = .017$, $\eta_G^2 = .01$]. Paired comparisons using t-tests revealed a significant difference between silent and aloud items, $t(63) = 3.50$, $p = .001$, $d = 0.44$ [$t(88) = 3.03$, $p = .003$, $d = 0.32$], and between silent and foil items, $t(63) = 2.15$, $p = .035$, $d = 0.27$, but no difference between aloud and foil items, $t(63) = 1.29$, $p = .201$, $d = 0.16$ [$t(88) = 1.84$, $p = .069$, $d = 0.20$]. [In the analysis that did not exclude any poorly performing participants ($N = 89$), there was no significant difference between silent and foil items, $t(88) = 0.99$, $p = .324$, $d = 0.11$.]

Discussion

As seen in Experiment 1, there were robust effects of production and of response speeding. Unlike in Experiment 1, however, Experiment 2 found that speeded responding reduced the observed production effect as predicted. That is, when responding was speeded, the production effect was significantly smaller in hit rate (the pattern in RTs was similar but not significant). This is consistent with the idea that when responding was speeded, there was reduced opportunity for a time-consuming process to unfold and to support the emergence of a production effect. Interestingly, there nevertheless remained a reliable production effect for hit rate in the speeded condition. As noted, this effect might reflect the slower process running to completion even in the speeded condition, just less frequently, or it might reflect a contribution to the production effect that is relatively insensitive to the manipulation of retrieval time.

Experiment 3

The goal of Experiment 3 was to replicate and extend Experiment 2. We again used a speeded response deadline of 750 ms to investigate the effect of speeded responding on the production effect. In addition, we added a measure of guessing. The increased difficulty associated with the shorter deadline in the speeded condition likely leads individuals to "guess" more often in that condition especially given that, throughout our experiments, participants are told that guessing is better than not responding. This could cloud the interpretation of the interaction observed between speeded responding and the production effect because guesses would not be expected to yield a production effect. That is, the production effect might be smaller under speeded retrieval because speeded retrieval increases the propensity to guess (and guessing would not yield a production effect), rather than because speeding reduces the opportunity for a time intensive process to unfold.

By attempting to identify guesses, we can begin to address this alternative account by focusing on responses reported not as guesses. If speeded responses still lead to a reduction of the production effect when guesses are excluded, then guessing is unlikely to be a complete explanation of the reduced production effect observed in the speeded condition. Experiment 3 was preregistered at osf.io/k8nqt/.

Method

As in Experiment 2, data were analyzed from 64 undergraduate students from University of Waterloo taking part for course credit, with none having taken part in Experiments 1 or 2. *Stimuli* for Experiment 3 were identical to those used in Experiments 1 and 2. The *Procedure* for Experiment 3 was identical to that of Experiment 2 with one key exception: During the recognition tests, after each individual response as to whether an item was old ("f") or new ("j"), participants were prompted to answer yes (*"y"*) or no ("n") to whether they felt that their previous response was a guess. Responding to subsequent guess prompts was not speeded but was to be completed within 3 s (considered ample time, e.g., Wammes et al., 2018). Due to the task-switching nature of this modification, we included a 500-ms interstimulus interval (ISI) between responding old/new to the presented item and the onset of the guess prompt, as well as a 1500-ms ISI between responding yes/no to the guess prompt and the onset of the subsequent trial. For consistency, participants were still urged to favor responding (even if it meant guessing) over not responding.

Results

As in previous experiments, hit rate and response time (for hits and correct rejections) were each analyzed using a 2 (silent vs. aloud) x 2 (standard vs. speeded) repeated measures ANOVA; false alarm rate was analyzed using a (standard vs. speeded) repeated measures ANOVA. We also conducted an analogous analysis considering block type order (standard first

vs. speeded first) as a between-participants factor for each of hit rate, false alarm rate, and response time. Only cases where order interacted with the other factors of interest are noted in Section I of the Supplementary Materials. As in Experiment 2, when the speeded block occurred first, this occasionally led to a smaller effect of speed compared to when the speeded block occurred second. However, in general, these instances did not qualitatively affect the results.

Data from 12 participants were replaced because either their response rates were lower than the 60% preregistered criterion (for Experiment 3, this criterion was specified at the condition level: standard-silent, standard-aloud, speeded-silent, speeded-aloud) or their recognition accuracy was at or below 50% on the recognition test. We updated the 60% minimum response criterion to apply at the condition level to prevent overly unbalanced observations within each condition, which could potentially occur with inclusion of individuals scoring perfectly on standard and/or aloud trials while having an inordinate number of timed out responses for silent and/or speeded trials. To remain consistent with Experiments 1 and 2, we ran parallel analyses which included the 12 participants (i.e., $N = 76$; available at osf.io/un5ca/) and results were qualitatively the same (these results which include the preregistered exclusions are reported in square brackets following the analogous main results; see Section III of the Supplementary Materials for the tables and figures accompanying these data).

Trials on which participants did not provide guess responses to the guess prompt (i.e., the deadline for the guess prompt was reached before responding) were removed before analyzing (7% of total trials). The majority of these removed trials (85%) were also trials wherein participants did not provide responses to the presented item (i.e., the deadline was reached before a response was provided). Although very infrequent, trials (.26% of total trials with responses to guess prompts) wherein participants did not provide responses were also removed before analyzing. Note that, like in Experiment 2, the preregistration for Experiment 3 specified

analyzing response time for false alarms but, again, we have deviated from that to report response time for correct rejections.

For this experiment, we preregistered three sets of analyses analogous to those reported in Experiments 1 and 2. The first set of analyses include all data (aside from the exclusions outlined previously); these are described in tables/figures as "with guesses." The second set of analyses are a subset of the first wherein only responses to targets explicitly reported by participants as not being guesses are included. Participants were removed from this subset if they had fewer than eight responses remaining in each condition (i.e., standard-silent, standard-aloud, speededsilent, speeded-aloud) after removing guesses. From here on, such participants are referred to as participants with high guessing rates. This second set of data is described in tables/figures as "excluding guesses." The final set of analyses are a subset of the first data set (i.e., guesses included), but with the 17 participants with high guessing rates also removed (23 of the 76 participants when no participants were excluded). The results of these final analyses were generally qualitatively similar to the results of the first data set and, as such, are not reported (though they are available at osf.io/un5ca/). A Greenhouse-Geisser correction was applied in the cases of a sphericity violation. Means of hit rate, false alarm rate, and response time are presented in Table 3 as a function of experimental condition. Data and analyses code for Experiment 3 are available at osf.io/un5ca/.

Hit rate

When guesses were included, there was a significant main effect of production such that hit rate was significantly higher for aloud items than for silent items (aloud: .78 [.75]; silent: .63 [.61]), $F(1, 63) = 92.44$, $p < .001$, $\eta_G^2 = .16$ [$F(1, 75) = 107.65$, $p < .001$, $\eta_G^2 = .14$], a typical production effect, $d = 0.81$ [0.81]. There was also a main effect of speeding such that hit rate was

significantly higher in the standard condition than in the speeded condition (standard: .74 [.73]; speeded: .66 [.63]), $F(1, 63) = 18.66$, $p < .001$, $\eta_G^2 = .05$ [$F(1, 75) = 31.98$, $p < .001$, $\eta_G^2 = .07$]. Consistent with Experiment 2, the interaction between production and speeding was significant, $F(1, 63) = 28.34$, $p < .001$, $\eta_G^2 = .05$ [$F(1, 75) = 35.66$, $p < .001$, $\eta_G^2 = .04$], such that the production effect was significantly larger in the standard condition than in the speeded condition (standard: .22 [.22]; speeded: .06 [.07]). Still, there was a significant production effect in both conditions: standard, $t(63) = 11.02$, , $p < .001$, $d = 1.38$ [$t(75) = 12.24$, $p < .001$, $d = 1.40$], and speeded, $t(63) = 3.42$, $p = .001$, $d = 0.43$ [$t(75) = 3.63$, $p = .001$, $d = 0.42$].

When excluding guesses and participants with high guessing rates, there remained a main effect of production such that hit rate was still significantly higher for items read aloud than for those read silently (aloud: .86 [.85]; silent: .72 [.71]), $F(1, 46) = 53.15$, $p < .001$, $\eta_G^2 = .15$ [$F(1, 52) = 66.64$, $p < .001$, $\eta_G^2 = .14$]. However, there was no main effect of speeding as hit rate did not differ significantly between the standard and speeded conditions (standard: .80 [.79]; speeded: .79 [.77]), $F(1, 46) = 0.66$, $p = .422$, $\eta_G^2 < .01$ [$F(1, 52) = 1.76$, $p = .19$, $\eta_G^2 = .01$]. The interaction between production and response speeding remained significant, $F(1, 46) = 9.39$, $p = .004$, $\eta_G^2 = .02$ [$F(1, 52) = 10.81$, $p = .002$, $\eta_G^2 = .02$], such that the production effect was smaller when responses were speeded (standard: .19 [.19]; speeded: .09 [.09]). As in the previous analyses, there were significant production effects in both conditions: standard, $t(46) = 7.13$, $p < .001$, $d = 1.04$ [$t(52) = 7.97$, $p < .001$, $d = 1.09$], and speeded, $t(46) = 2.55$, $p = .014$, $d = 0.37$ [$t(52) = 3.11$, $p = .003$, $d = 0.43$]. Mean hit rates by production, speed manipulation, and whether guesses were included are presented in Figure 3.

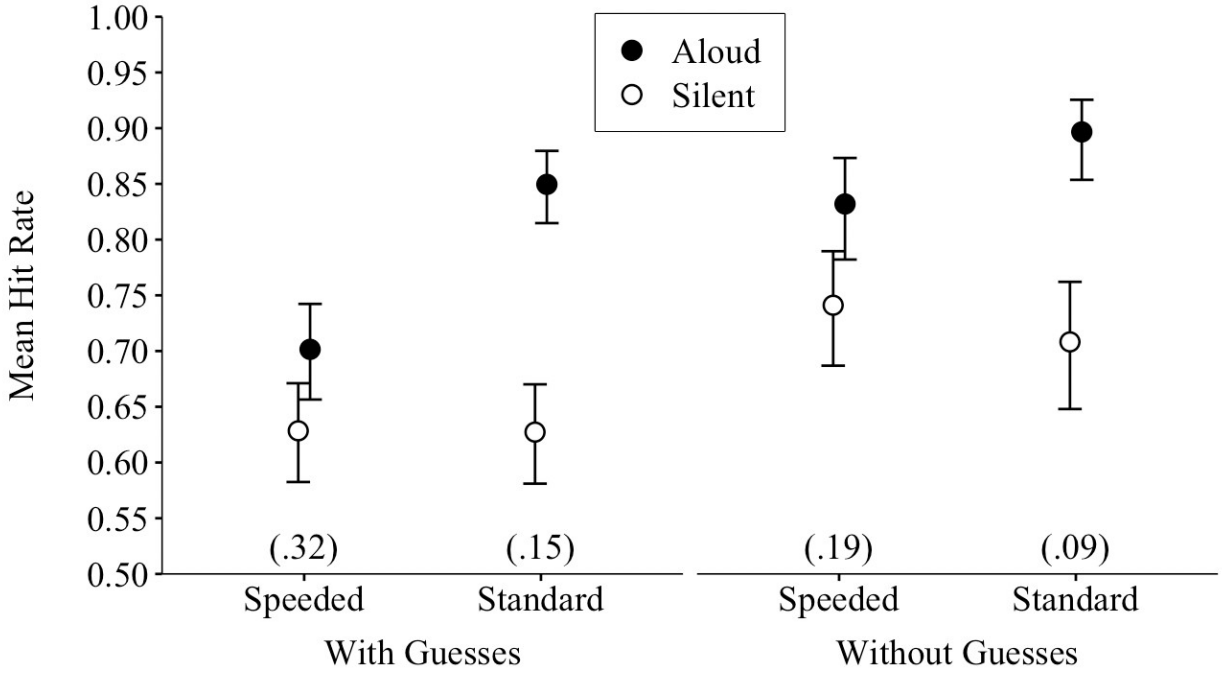


Figure 3. Experiment 3: Mean hit rate by item condition and by speed manipulation with mean false alarm rates in parentheses by speed manipulation, all by whether guesses were included or excluded. Error bars are bias-corrected accelerated bootstrap 95% confidence intervals using 10,000 replications.

False alarm rate

When guesses were included, the false alarm rate was significantly lower in the standard condition than in the speeded condition (standard: .15 [.16]; speeded: .32 [.32]), $F(1, 63) = 61.26$, $p < .001$, $\eta_G^2 = .24$ [$F(1, 75) = 78.03$, $p < .001$, $\eta_G^2 = .24$]. When guesses were excluded (as were participants with high guessing rates), the false alarm rate was still significantly lower in the standard condition than in the speeded condition (standard: .09 [.09]; speeded: .19 [.21]), $F(1, 46) = 18.08$, $p < .001$, $\eta_G^2 = .10$ [$F(1, 52) = 25.24$, $p < .001$, $\eta_G^2 = .12$].

Response time for hits

When guesses were included, there was a main effect of production on response times for hits such that response times for hits were significantly faster for aloud items than for silent items

(aloud: 823 ms [830 ms]; silent: 855 ms [866 ms]), $F(1, 63) = 10.18$, $p = .002$, $\eta_G^2 < .01$ [$F(1, 75) = 14.70$, $p < .001$, $\eta_G^2 < .01$]. Not surprisingly, there was also a main effect of speeding such that participants were also faster in the speeded condition than in the standard condition (standard: 1136 ms [1155 ms]; speeded: 543 ms [541 ms]), $F(1, 63) = 211.94$), $p < .001$, $\eta_G^2 = .61$ [$F(1, 75) = 230.30$), $p < .001$, $\eta_G^2 = .59$]. These two effects interacted significantly such that the production effect in response time for hits was significantly smaller in the speeded condition than in the standard condition (standard: -69 ms [-74 ms]; speeded: 4 ms [3 ms]), $F(1, 63) = 11.22$, $p = .001$, $\eta_G^2 = .01$ [$F(1, 75) = 16.75$, $p < .001$, $\eta_G^2 = .01$]. The production effect in response time for hits was significant in the standard condition, $t(63) = 3.39$, $p < .001$, $d = 0.42$ [$t(75) = 4.23$, $p < .001$, $d = 0.49$], but not in the speeded condition, $t(63) = 0.75$, $p = .454$, $d = 0.09$ [$t(75) = 0.49$, $p = .625$, $d = 0.06$].

When guesses and high guessing rate participants were excluded, findings were consistent with the pattern observed when they were included. That is, there was a main effect of production on response times for hits such that response times were significantly faster for aloud items than for silent items (aloud: 759 ms [757 ms]; silent: 788 ms [788 ms]), $F(1, 46) = 6.36$, $p = .015$, $\eta_G^2 = .01$ [$F(1, 52) = 7.57$, $p = .001$, $\eta_G^2 < .01$]. There was also a main effect of speeding such that response times were longer in the standard condition than in the speeded condition (standard: 997 ms [1155 ms]; speeded: 550 ms [541 ms]), $F(1, 46) = 206.68$, $p < .001$, $\eta_G^2 = .66$ [$F(1, 52) = 224.43$, $p < .001$, $\eta_G^2 = .64$]. Also consistent with the previous findings, these effects significantly interacted, $F(1, 46) = 5.18$, $p = .028$, $\eta_G^2 = .01$ [$F(1, 53) = 7.04$, $p = .010$, $\eta_G^2 = .01$], such that the production effect in response time was smaller in the speeded condition (standard: 56 ms [-72 ms]; speeded: 2 ms [2 ms]). The production effect in response time was significant only in the standard condition, $t(46) = 2.48$, $p = .017$, $d = 0.36$ [$t(52) = 2.86$, $p = .006$, $d = 0.392$], and not in the speeded condition, $t(46) = 0.29$, $p = .777$, $d = 0.04$ [$t(52) = 0.10$, $p = .918$, $d =$

0.01].

Response time for correct rejections

When guesses were included, as expected, the response times for correct rejections were significantly faster when responses were speeded (standard: 1148 ms [1151 ms]; speeded: 542 ms [540 ms]), $F(1, 63) = 194.90$, $p < .001$, $\eta_G^2 = .59$ [$F(1, 75) = 223.65$, $p < .001$, $\eta_G^2 = .58$]. When guesses and high guessing rate participants were excluded, results were similar to the previous analysis: Response times for correct rejections were significantly faster when responding was speeded than when it was not (standard: 1028 ms [1018 ms]; speeded: 552 ms [545 ms]), $F(1, 46) = 140.19$, $p < .001$, $\eta_G^2 = .57$ [$F(1, 52) = 154.52$, $p < .001$, $\eta_G^2 = .56$].

Exploratory

As in the previous experiments, we analyzed the proportion of timeouts in the speeded condition. Consistent with Experiment 1, the proportion of timeouts did not differ among silent, aloud, and foil items (silent: .14 [.16]; aloud: .14 [.15]; foil: .15 [.17]), $F(2, 126) = 0.54$, $p = .583$, $\eta_G^2 < .01$ [$F(1.71, 127.94) = 0.62$, $p = .514$, $\eta_G^2 < .01$], when including guesses. When guesses and high guess-rate participants were excluded, the proportion of timeouts did not differ among silent, aloud, and foil items (silent: .02 [.02]; aloud: .01 [.02]; foil: .02 [.01]), $F(1.62, 102.03) = 1.14$, $p = .316$, $\eta_G^2 = .01$ [$F(1.74, 128.46) = 0.28$, $p = .722$, $\eta_G^2 < .01$], similar to when guesses were included, though these timeouts were very infrequent in comparison.

We also analyzed the proportion of guesses using a 3 (silent vs. aloud vs. foil) x 2 (standard vs. speeded) repeated measures ANOVA. There was a main effect of item type (silent: .33 [.33]; aloud: .24 [.25]; foil: .32 [.32]), $F(1.52, 95.68) = 14.66$, $p < .001$, $\eta_G^2 = .04$ [$F(1.49, 111.69) = 13.63$, $p < .001$, $\eta_G^2 = .03$]. Paired comparisons using t-tests revealed significant differences between silent and aloud items (silent: .33 [.33]; aloud: .24 [.25]), $t(127) = 6.30$, $p < .001$, $d = 0.56$ [$t(151) = 6.16$, $p < .001$, $d = 0.50$], and between aloud and foil items (aloud: .24

[.25]; foil: .32 [.32]), $t(127) = 4.22$, $p < .001$, $d = 0.37$ [$t(151) = 3.95$, $p < .001$, $d = 0.32$], but not between silent and foil items (silent: .33 [.33]; foil: .32 [.32]), $t(127) = 0.86$, $p = .394$, $d = 0.08$ [$t(151) = 0.97$, $p = .333$, $d = 0.08$]. There was also a main effect of speeding such that the proportion of guesses was significantly lower in the standard condition than in the speeded condition (standard: .27 [.26]; speed: .33 [.34]), $F(1, 63) = 16.30$, $p < .001$, $\eta_G^2 = .03$ [$F(1, 75) = 24.40$, $p < .001$, $\eta_G^2 = .03$]. There was no significant interaction between production and speed, $F(2, 126) = 1.90$, $p = .154$, $\eta_G^2 < .01$ [$F(2, 150) = 2.62$, $p = .076$, $\eta_G^2 < .01$].

Table 3

Experiment 3: Mean Hit Rate, False Alarm Rate, and Response time for Hits and Correct Rejections for Each Item Type by Condition for Data With and Without Guesses (CI_{95} in Parentheses)

	Guesses	Silent	Aloud	New
Standard	*with*			
Hit and False alarm rate		.63 [.58, .67]	.85 [.81, .88]	.15 [.13, .18]
	without	.71 [.65, .76]	.90 [.85, .93]	.09 [.07, .12]
Response time (ms) for hits and correct rejections	*with*	1170 [1092, 1254]	1101 [1028, 1186]	1148 [1066, 1236]
	without	1025 [967, 1085]	969 [910, 1041]	1028 [954, 1117]
	with			
Timeout proportion		.01 [.01, .02]	.01 [.01, .02]	.01 [.01, .02]
	without	0 [0, -]	0 [0, -]	0 [0, -]
	with			
Guess proportion		.31 [.26, .35]	.20 [.16, .24]	.29 [.25, .34]
Speeded	*with*			
Hit and False alarm rate		.62 [.58, .67]	.70 [.65, .74]	.32 [.27, .36]
	without	.74 [.69, .79]	.83 [.78, .87]	.19 [.14, .26]

Response time (ms) for hits and correct rejections	*with*	541 [521, 558]	545 [528, 560]	542 [524, 557]
	without	551 [528, 568]	549 [533, 563]	552 [532, 568]
	with			
Timeout proportion		.14 [.12, .17]	.14 [.12, .16]	.15 [.13, .18]
	without	.02 [.01, .05]	.01 [.003, .03]	.02 [.01, .03]
	with			
Guess proportion		.36 [.31, .41]	.29 [.24, .34]	.34 [.30, .39]

Note. All confidence intervals are bias-corrected accelerated bootstrap 95% confidence intervals using 10,000 replications.

Discussion

The results of Experiment 3 are consistent with those of Experiment 2, both when including responses reported as guesses and when excluding such responses. When including guesses, we observed a robust production effect. That is, as in Experiment 2, speeded responding significantly reduced the production effect. Moreover, in the speeded responding condition, the production effect remained for hit rate. In Experiment 3, we sought to reduce the potential influence of a higher guess rate in the speeded condition. Critically, however, when excluding responses reported as guesses, results were similar to those including them. Indeed, across hit rates, false alarm rates, and RT analyses, the results were qualitatively identical. Thus, there is little reason to suspect that the reduction of the production effect observed under speeded conditions is driven by a dilution due to increased guessing. While it is possible that participants are unable to accurately monitor and report their own guessing behaviour, the data show that guessing rates were indeed higher in the speeded condition and that the removal of guess trials generally increased performance, as would be expected. Finally, it is worth noting that the interactions between the production and retrieval speed manipulations are characterized as "removable" because they can be eliminated using a monotonic transformation (Bogartz & Wackwitz, 1970; Loftus, 1978; Wagenmakers et al., 2012). We discuss this issue further in *Section IV* of the *Supplementary materials*.

Modelling the Production Effect

Empirical work on the production effect has accumulated rapidly, however theoretical accounts have remained largely verbal in nature (the exception being Jamieson et al., 2016; also see MacLeod, Pottruff, Forrin, & Masson, 2012 for other modelling approaches). This is disappointing given how robust the production effect has proven to be (as is the case for

potentially related phenomena, e.g., drawing effect, enactment effect, generation effect; MacLeod et al., 2010; Wammes, Meade, & Fernandes, 2018). This robustness suggests that the phenomenon is capturing an attribute of memory well worth attempting to integrate into existing computational approaches to memory. To this end, we next attempt to implement a version of the Forrin et al. (2012) account into one such computational framework. Following this, and for comparison, we also implement a version of another popular account of the production effect based on "strength."

The Current Modelling Framework

The current modelling is situated in the Retrieving Effectively from Memory framework (REM; Shiffrin & Steyvers, 1997), a computational model of recognition memory. REM was originally developed (at least partially) in response to the inability of prevailing global memory models to account for the memory advantage of low-frequency words (Bowers & Davis, 2012) and borrowed elements from various extant models including SAM (e.g., Raaijmakers & Shiffrin, 1981) and MINERVA2 (e.g., Hintzman, 1984; 1986). REM has been successful in accounting for a large assortment of memory phenomena including, but not limited to, the wordfrequency effect (Malmberg & Murnane, 2002), feature frequency effects (Malmberg, Steyvers,

Stevens, & Shiffrin, 2002), effects of repetition and similarity (Malmberg, Holden, & Shiffrin, 2004), list-strength and spacing effects (Ensor, Surprenant, & Neath, 2020; Malmberg & Shiffrin, 2005; Osth et al., 2018), intentional forgetting (Lehman & Malmberg, 2011), retrievalinduced forgetting (Verde, 2013), source memory (Osth et al., 2018), output interference (Annis, Malmberg, Criss, & Shiffrin, 2013) and judgments of frequency (Annis & Malmberg, 2013).

To model the account of Forrin et al., we adapted REM.1, the version of REM most often implemented to model recognition, in addition to being its most simple instantiation (Ensor et al., 2020, Shiffrin & Steyvers, 1997). In REM.1, memory is treated as a collection of *traces* (or, traditionally, *images*) each represented as a vector of feature values, such that knowledge about a feature is denoted with a positive integer while absence of knowledge about a feature is represented by a value of zero. In general, the number of features, w, is set to 20. The specific feature values are drawn from a geometric distribution with parameter g. At study, features from studied items are stored into memory with probability u. If the feature information (i.e., its value) is stored, then it is accurately copied with probability c, otherwise, a different value is drawn from the geometric distribution for that feature. If no information is copied, then the assigned value of that feature is zero.

Modelling retrieval during a recognition test involves "presenting" a probe (either a target or a foil) to all traces stored in "memory." All of the feature values of the probe and each trace are aligned to compare their similarities and differences. A likelihood ratio, λ, incorporating the Bayes rule, is computed for each stored trace by dividing the probability of observing the similarities and differences between the probe and trace if the probe was a target by the probability of observing the similarities and differences if the probe was a foil. For each trace stored in memory, M is the set of feature positions in the vector wherein the nonzero feature values match that of the probe (i.e., the features within the trace and probe are compared based on aligning their analogous feature positions) and Q is the set of feature positions wherein the nonzero feature values mismatch the probe. $P_{km}(i)$ is the probability that a feature value i would occur in feature position k if the trace derived from the probe and the feature value was stored accurately. $P_{\text{kq}}(i)$ is the probability that the feature value i would be stored in feature position k if

the trace derived from the probe and the feature value was stored inaccurately. Finally, $P_{kd}(i)$ is the probability that feature value i would occur in position k provided that the trace did not derive from the probe. Then, the odds calculation for any trace stored in memory is the following:

$$\lambda = \frac{\$\ \%P_{!"}(i)^{*}\ \$\ +P_{!\&}(i),}{!\in\%\ P_{!\#}(i)\ !\in'\ P_{!\#}(i)}$$

Based on the overall average of the likelihood ratios (λs) computed from each probe-trace comparison, REM computes the average odds that the probe is "old" over "new." Typically, a decision of "old" is reached if the average odds that the probe is a target exceed 1 (see Shiffrin & Steyvers 1997 for expanded formulas). The following equation is the calculation for odds, Φ, such that λ_j is the likelihood ratio computed for each stored trace, j, and wherein n is the total number of stored traces in memory compared to the probe:

$$\Phi = \left(\frac{1}{n}\sum_{j=0}^{n}\lambda_j\right)^{(*+}$$

Adapting REM.1 to the Forrin et al. (2012) account

Our adaptation to REM.1 is based on what we consider the central idea in the Forrin et al. (2012) account—that engaging in production leads to the storage (at study/encoding) and use (at test/retrieval) of production-associated features (see also Jamieson et al., 2016). These production-associated features might, for example, consist of the motoric or perceptual information generated through the act of production (Forrin et al., 2012, Jamieson et al., 2016). In the adapted model, we represent production-associated information with an additional set of features—10 features beyond the 20 base features—for to-be-remembered items produced at

study, with these production-associated features containing values (i.e., they are filled in); items not produced at study do not contain such features. Thus, on average, production at study/encoding leads to the storage of additional information.

The account by Forrin et al. (2012) also suggests that these production-associated features are used at test (e.g., "*Do I remember saying this aloud?*"). In the current model, we represent the use of production-associated features at test by including production-associated features in the probe vectors that are presented to memory at test. Thus, memory is probed with both the base features (e.g., lexical and semantic features) and the features representing the information related to the production of the item. Note that all probe vectors, regardless of condition (produced target, silent target, foil), could include production-associated features. This captures the assumption that participants do not know whether the item was studied prior to this information being used.

The presence and use of production-associated features should yield a production effect (i.e., proportion of "yes" responses for produced targets > silent targets). This is because items produced at study have a greater number of opportunities for a feature match at test (i.e., there are more feature positions with non-zero values on average). As a result, assuming that features are encoded accurately, the computed likelihood that a probe is a target will tend to be higher for probes representing items that were produced at study. It is important to note that the feature locations reserved for production-associated information do not apply to the memory traces representing items *not produced* at study. Specifically, the vectors across item type are virtually/functionally equal, except that the vectors representing traces of items *not* produced at study did *not* contain production-associated features (i.e., to increase the model processing speed, these feature positions were not included in these vectors); this is functionally equivalent to

adding ten additional 0s to vectors representing silent traces. Figure 4 below demonstrates this idea visually.

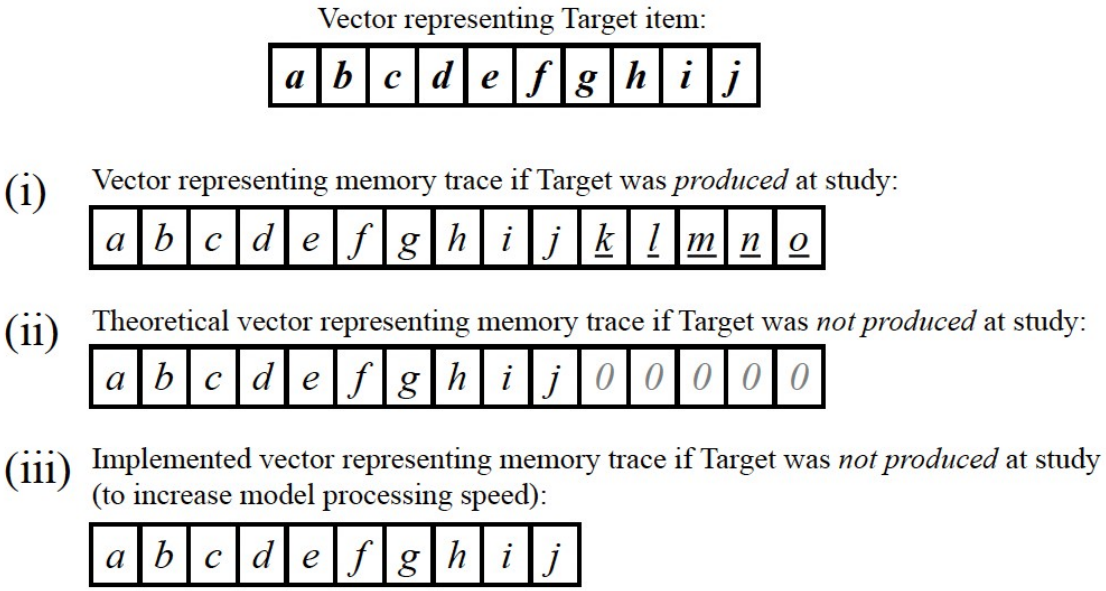


Figure 4. Visual depiction of how a given Target item vector, wherein each letter represents the corresponding feature position within the vector, is represented in "memory" depending on if it was produced at study, case (i), or not, cases (ii) and (iii). Production-associated features are underlined in the relevant case of (ii). Note the lack of functional difference between (ii) and (iii) as explained in text.

In REM.1, for the traces of items not produced at study, this leads to their production-associated features contributing nothing (no increase, nor decrease) to the computed likelihood that a probe is a target or foil.

Finally, we attempted to capture the idea that the use of production-associated features at retrieval takes additional time (and may well be subject to other contextual influences). Although REM.1 does not include a mechanism to model the passage of time at retrieval, we approximated the effect of restricting retrieval time by varying the number of probe features available. This

approach is similar to earlier work modelling the time course of retrieval and can be conceptualized as responses being required prior to the complete processing of the test stimulus under speeded retrieval conditions (such as in the current procedure or in signal-response methods; Brockdorff & Lamberts, 2000; Cox & Shiffrin, 2017; Lamberts, 2002). In REM.1, having fewer features in the stimulus probe will, on average, leave fewer opportunities for matching features between the probe and the trace. Consequently, the overall probability that the probe is a target will be reduced compared to what would transpire if all of the features in the probe were available at retrieval, all else being equal. Critically, to capture the idea that some features would be less likely to contribute under limited retrieval time, we classified all features as either "fast" or "slow" and, importantly, all production associated features as "slow." Thus, in simulating the influence of restricting retrieval time, we first made "slow" features unavailable followed by making "fast" features unavailable in the probe. This is clearly a simplification (e.g., it need not be all production-associated features), but as noted in the Introduction and as the empirical data presented here support, this assumption appears to rest on solid ground.

Simulations

Where possible, we set parameters to values considered conventional based on previous investigations using REM. We set each item to have 20 base features, 10 of which we considered "fast" and 10 of which we considered "slow". Because we had no principled basis for selecting a given number of production-associated features for produced items short of approximating the qualitative patterns in empirical data, we set the number of production-associated features to 10.[6]

[6] Note that although we have opted to set the number of production-associated features to 10, the same qualitative pattern emerges when varying the number of features per feature type (e.g., 5) and the proportion of features available in the probe. Consistently, as the number of productionassociated features increases, so does the resulting production effect, all else being equal.

For each type of feature (faster base, slower base, and production-associated), we set *c* to .70, *g* to .40, and *u* to .28. The odds criterion was set to 1, the number of produced targets and silent targets to 20 each, and the number of foils to 40 (these numbers correspond to those used in the present empirical work).

To model the probability that the "slow" features and production-associated features are used at test, we varied the availability of these features, incrementally reducing their availability by one feature per simulation set. That is, to simulate results under "full" retrieval time, all features were available. To simulate decreasing time available at retrieval, in each step, we removed one additional feature. The first features to be removed were the "slow" base features and production-associated features, followed by the "fast" base features. Whenever an odd number of features was made unavailable, half of the simulations had one more "slow" feature removed than production-associated features and the other half of the simulations had the reverse. We ran a total of 1000 simulations at each of the 30 increments of feature availability, for a total of 30,000 simulations. Simulations were coded in Python 2. For reproducibility, all simulations (including those for the strength-based model to follow) were conducted with the *numpy* random seed set to 1 (seed did not influence any qualitative findings throughout). The code for the model is available at osf.io/un5ca/.

Simulation results & discussion

Mean proportions of items classed as "old" as a function of item type (produced target vs. silent target vs. foil) and the production effect in hit rate are presented in Figure 5, all as a function of the availability of features in the probe. As is clear in Figure 5, when productionassociated features are available in the probe, there is a production effect such that the

hit rate for items produced at study is higher than that for items not produced at study. In addition, as feature availability increases, the production effect in hit rate increases. If we consider the present empirical manipulation of speeding retrieval as a manipulation of the availability of features contributing to the recognition decision (such that "slow" and production-associated features take longer), then this generally follows the pattern of results in the reported experiments. Similarly, under speeded recognition instructions, the hit rate for items not produced at study also reduces, while the false alarm rate increases. As seen in Figure 5, these patterns qualitatively match those in the observed participant data.

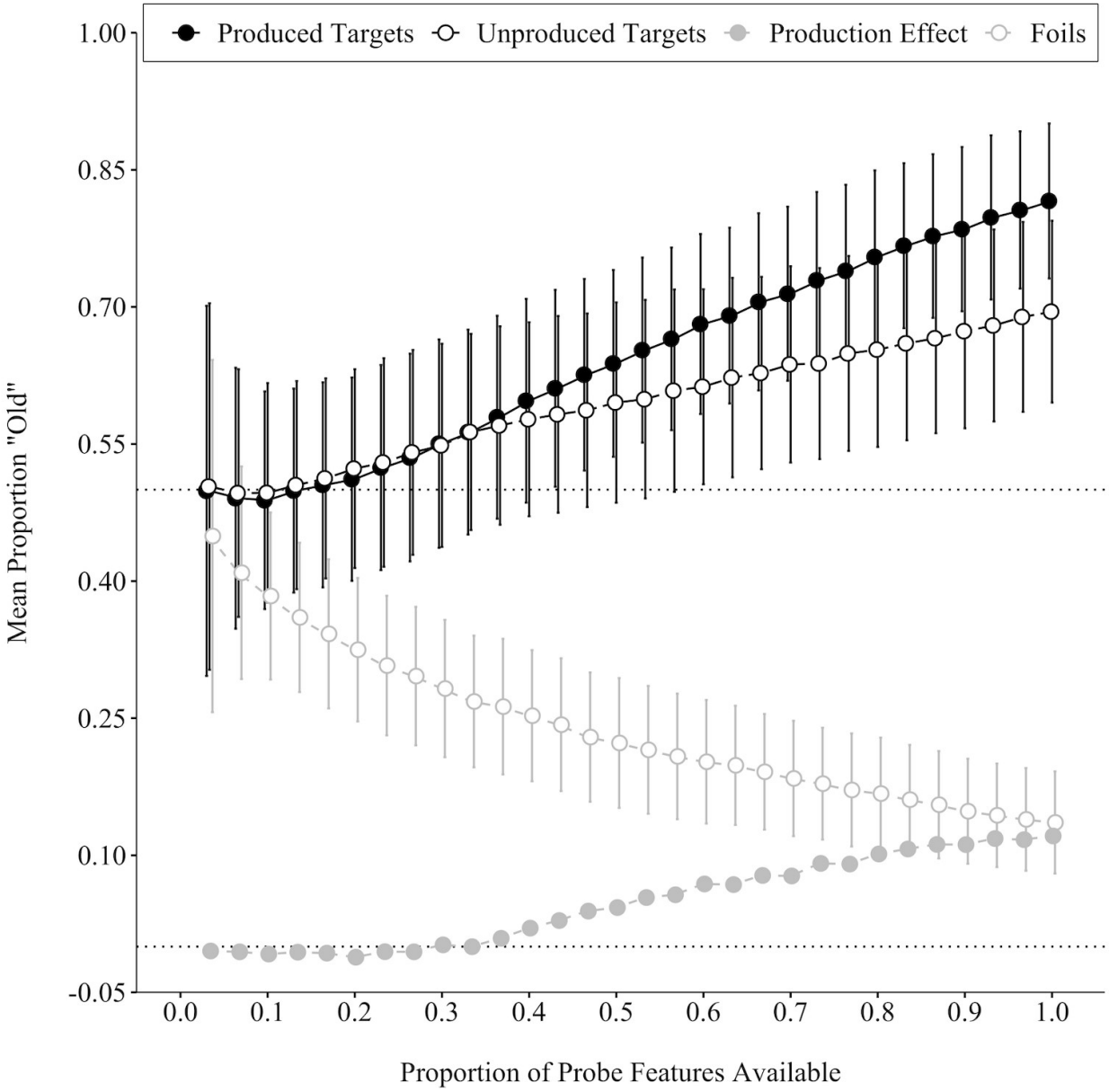


Figure 5. Production-associated features account in REM: Hit rate as a function of production with false alarm rates and the production effect in hit rate, all as a function of the proportion of probe features available. Error bars are ± the standard deviation.

The results of the present simulation suggest that an instantiation of the Forrin et al. (2012) account in REM can capture at least the basic production effect and, with additional assumptions (i.e., that the probability of using “slow” features increases when more time is

available and that production-associated features are such features), the experimental results reported here. It is important to note that the simulation demonstrates, as expected based on the Forrin et al. account, that if production-associated features are not used, then there would be no production effect (i.e., when only "fast" base features are available). This should not be interpreted as the model predicting no production effect under the retrieval speed manipulation used in the present experiments. The latter would require the additional assumption that our deadline (800 ms in Experiment 1; 750 ms in Experiments 2 and 3) provided *no opportunity* for the use of production-associated features across all of the participants. This would be a strong assumption. What is more important, in our view, is that the pattern in the simulations qualitatively matches the empirical pattern. That is, the less the retrieval context permits use of the production-associated features, the smaller the production effect should be and, we suggest, speeding retrieval represents once such manipulation that would have such an effect on the retrieval context. Critically, speeding retrieval time, as implemented in the present computational account, shows that the reduction to the predicted effect is largely due to reducing performance of produced items. This is consistent with the empirical findings. We discuss the model further in the General Discussion. But first, we examine an alternative strength-based account.

Adapting REM.1 to a strength-based account

One way to model a strength-based explanation of the production effect in REM is by manipulating the likelihood that a feature is stored into the trace, *u*. For example, Shiffrin and Steyvers (1997) suggested that *u* be set to .28 to represent "weak" encoding and to .40 for "strong" encoding. Similarly, Jamieson at al. (2016) implemented a strength-based mechanism in their MINERVA2 model of the production effect using encoding quality—the probability that a feature is copied accurately from the study item to the memory trace. In REM.1, having a higher value of the strength parameter, *u*, for items produced at study leads to a higher proportion of

features that are accurately copied into memory for each to-be-remembered item. This leads to produced items having more matching features than items not produced, on average, upon comparing the probes and memory traces, and ultimately this contributes to a higher computed likelihood that a test probe is 'old' when it is indeed 'old.' Of course, there might exist alternative means of implementing a strength-based account.

Simulation

To investigate a strength-based account in the present model, we introduced two changes to the model implemented earlier: (1) There were no production-associated features stored into memory and hence no use of them at retrieval, and (2) as suggested by Shiffrin and Steyvers (1997), we varied u based on whether an item was in the produced condition (u = .40) or the silent condition (u = .28). Again, we removed one feature per increment as an approximation of the time available at retrieval (note that in the model of the strength-based account, "fast" and "slow" feature classifications have no material impact). We ran 1000 simulations across the 20 increments for 20,000 simulations in total.

Simulation Results & Discussion

Mean proportion of items determined as "old" as a function of item type (produced target vs. silent target vs. foil) and the production effect in hit rate are presented in Figure 6, all as a function of the proportion of features available in the probe at retrieval. As is clear in Figure 6, a strength-based account in REM.1 also results in a production effect. Interestingly, again if we consider the present empirical manipulation of speeding retrieval as a manipulation of the availability of features contributing to the recognition decision, this model also captures the reported pattern such that as feature availability increases, the production effect in hit rate increases. While the overall pattern is the same, the form is quite different from that produced in

the model using our implementation of the Forrin et al. (2012) mechanism. The strength-based model captures this pattern because the "weaker" silent items hit the floor before the "stronger" produced items and as the latter items approach the floor the production effect decreases. This difference in how the production effect is influenced by feature availability across the two models likely reflects the fact that, in the strength model, the *u* or "strength" parameter in REM provides proportional increases in the likelihood that a feature is stored. As such, its manipulation contributes a relatively more constant influence as the number of features contributing to the recognition decision decreases (i.e., as feature availability decreases) compared to having production-associated features in the probe drive the production effect.

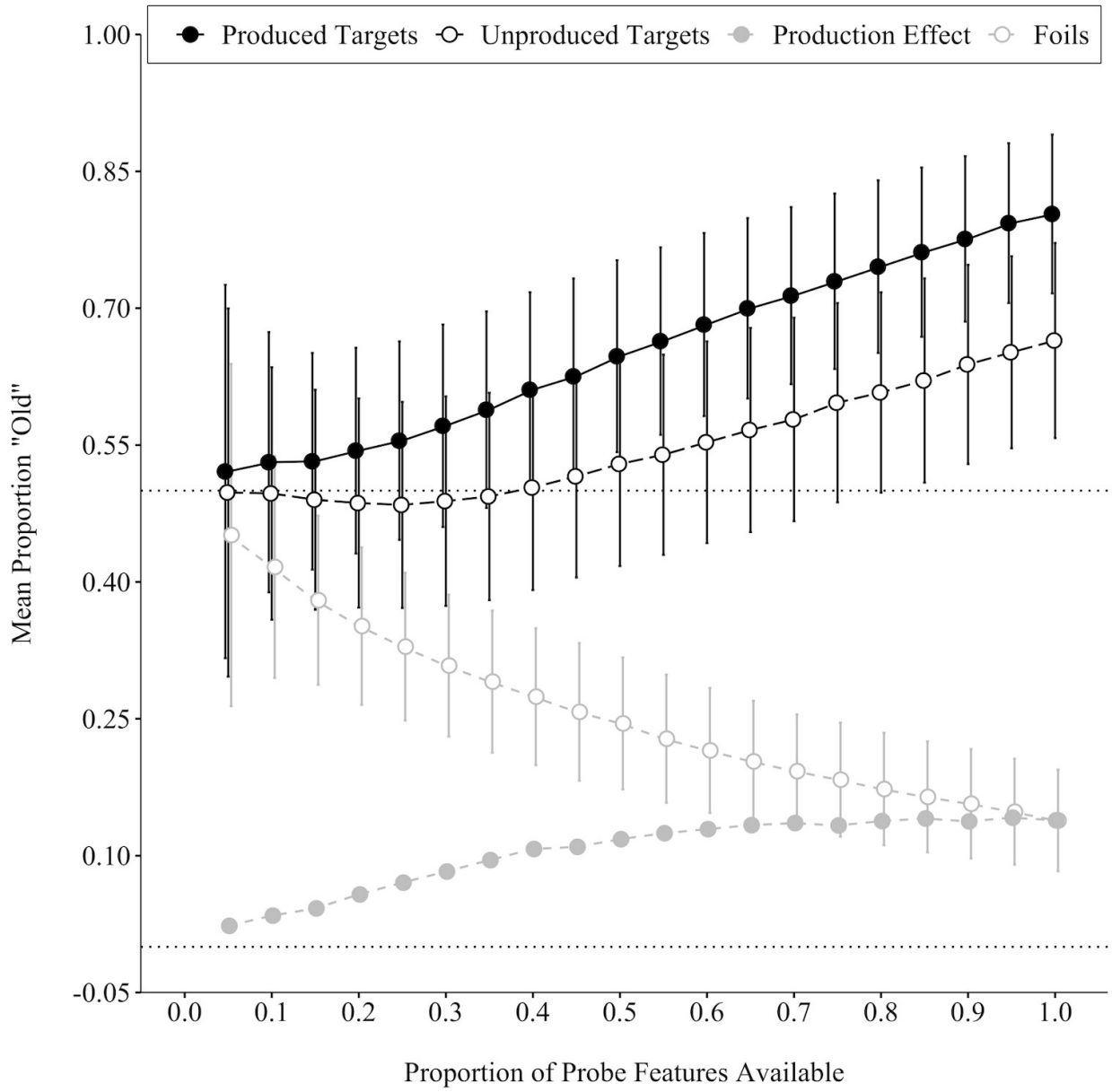


Figure 6. Strength-based account in REM: Hit rate as a function of production with false alarm rates and the production effect in hit rate, all as a function of the proportion of probe features available. Error bars are ± the standard deviation.

As noted above, both the model inspired by Forrin et al. (2012) and a basic strength account generate a production effect and capture the general pattern relating reducing time at retrieval and the production effect (if we assume that speeding influences the number of probe

features available). While there may appear to be some qualitative differences between the results of each model type (e.g., in the strength model simulations, it is not until unproduced performance is near the floor that there are substantive reductions in the magnitude of the production effect), taken at face value, the modelling of the present work demonstrates both accounts remain plausible and there is clearly more to be done in terms modelling the production effect. Future work may further distinguish between the proposed accounts perhaps with use of quantitative model fits and comparisons.

General Discussion

In the current investigation, we examined a prediction of the account of the production effect described by Forrin et al. (2012). Specifically, if the production effect arises from the use at the time of test of additional production-associated features created at the time of study, we reasoned that such a process would take time and, therefore, under a condition of limited time to respond, the production effect should be reduced. To test this prediction, we examined the effect of a response deadline on the magnitude of the production effect. Across three preregistered experiments, speeded responding reduced the size of the production effect (though not significantly in Experiment 1). In addition, in Experiment 3, we found that speeded responding reduced the magnitude of the production effect even when controlling for self-reported guessing. This reduction to the production effect was, as anticipated, due to a larger effect of speeding on the performance of produced items relative to unproduced items.

In all three experiments, we observed robust production effects in hit rate even in the speeded condition. There are (at least) two viable accounts of this effect: (1) the time-taking process occurred often enough to generate a production effect, even under speeded conditions or (2) this effect reflects a contribution to the production effect due to a process that is not slower

(e.g., strength). Of course, both (1) and (2) could be true (i.e., they are not mutually exclusive). Regardless of how this effect is to be explained, the observation (i.e., a robust production effect even under speeded retrieval) provides an important new constraint on accounts of the production effect.

Modelling

In addition to testing a prediction based on an extant account of the production effect and providing a new empirical constraint on theories of it, we also developed versions of both the Forrin et al. (2012) account and a strength account of the production effect using the REM computational framework. To represent restricting retrieval time in the models, we manipulated the number of probe features available such that *less* retrieval time is approximated by *fewer* features available in the probe. Both models produced the basic production effect, such that items produced at study resulted in higher hit rates than did items not produced at study. In addition, our manipulation of retrieval time produced the predicted pattern (i.e., decreased hit rates and increased false alarm rate). In implementing Forrin et al.'s account in REM, the model produced results in line with the qualitative results of the current experiments. That is, the production effect was smaller when our representation of retrieval time was more restricted. This success reflects the idea that the "additional" features associated with production are the types of features that take longer to contribute to the recognition decision, thus a manipulation of retrieval time has a direct effect on the magnitude of the production effect. In the introduction we detailed various reasons why this might be the case. Research exploring this idea further would be valuable for understanding the production effect and related phenomenon (e.g., drawing effect; Wammes et al., 2018) and the dynamics of recognition decisions more broadly (e.g., Cox & Shiffrin, 2017).

The strength-based model also yielded a production effect. This production effect was also modulated by the availability of features in the probe, particularly once most probe features were unavailable. Thus, this model does seem to be able to capture key elements of the findings here—and does so in a different way than in the model inspired by the Forrin et al. (2012) account, as discussed above. The latter differences might provide one means of adjudicating between the two accounts.

As previously noted, the present computational modelling effort is not the first attempt to model the production effect. Jamieson et al. (2016) also implemented a version of the Forrin et al. (2012) account, in their case in MINERVA2, so it is instructive to contrast these efforts. In comparing our model to theirs, both models represent the encoding of additional productionassociated features at study as additional (on average) filled-in features of the stored traces of items produced at study. We have reported simulations featuring 10 additional features whereas they focused on simulations with five additional features. However, Jamieson et al. also reported a simulation varying the number of production-associated features as a means of simulating the differences found across different production types. Specifically, they compared the production effect when 3 additional production-associated features were used versus when 5 additional production-associated features were used to capture, respectively, the difference between whispering and saying produced items aloud. Importantly, in both models, the greater the number of production-associated features, the greater the magnitude of the production effect, at least within the values used.

Although the implementation of the encoding of production-associated features is similar across our REM.1 variant model and Jamieson et al.'s (2016) MINERVA2 model, how these features are used differs. In our simulation, probes presented at test (sometimes) included

production-associated features, regardless of whether the probe was a produced target or otherwise. That is, the production-associated features were used to probe memory (in addition to the base features). Jamieson et al. (2016) used a different approach. Specifically, they made use of the *deblurring* process in MINERVA2, wherein an iterative retrieval process takes place. In their model, the initial probing of memory does not include production-associated features. However, memory is probed multiple times, such that the probes used at each iteration derive from the retrieval results of the previous comparison between probe and trace. This results in production-associated information being included in the probe only when memory has already been probed at least once already. Thus, in the Jamieson et al. (2016) implementation, the production-associated features emerge from memory to contribute to the recognition decision rather than being placed there by the modeller, as in the current implementation. While the theoretical insight is similar, in this respect, the former model appears preferable.

In an attempt to develop our own simulations in the Jamieson et al. model (2016; and in our earlier work), we initially tried to use the number of iterations of retrieval to represent the time spent in the act of retrieval. This was motivated by the idea that speeding retrieval would reduce the time to sample memory repeatedly (i.e., to engage in the iterations required for iterative retrieval—or *deblurring*). This simulation captured the overall pattern such that the magnitude of the production effect increased as the number of iterations/ "time" increased. However, it also produced behavior inconsistent with the observed data. For example, the model predicted that false alarm rate would increase as more "time" became available for retrieval (i.e., as number of iterations increased) but this was not the case in our data. Moreover, it predicted that the increase in the production effect as number of iterations/ "time" increased would be driven by increased confusion between silent items and foils, with the produced hit rate

remaining stable. This, too, was not the case in our data; to the contrary, it was hit rate for the produced items that was differentially influenced. That being said, these issues were specific to our particular adaptation of the model by Jamieson et al. (2016), not their original model.[7] There likely are other ways to implement the time available at retrieval within MINERVA2, including by having the number of available probe features depend on retrieval condition (i.e., speeded vs. standard), similar to what we have presented in the current instantiation of REM.1. Future computational work comparing these two different approaches to implementing the Forrin et al. (2012) account would be valuable.

Response Time

While not the focus of the present investigation, in the standard condition of each experiment, response times for hits were always faster for items studied aloud than for items studied silently. This effect of production on response times has been reported previously (MacLeod, 2011) but has received little scrutiny. The repeated demonstration of this effect here and the current demonstration that the production effect in recognition performance appears sensitive to the amount of time available at retrieval both suggest that greater attention paid to this observation might benefit theory. Indeed, in an exploratory analysis of the standard condition across all the experiments here, we found that the increase in hit rate when items were produced was positively correlated with the decrease in response time when items were produced, r = -.23, p = .003. Thus, the two phenomena appear moderately correlated.

[7] We thank Randall Jamieson for noting these limitations in the computations in our original adaptation of the model by Jamieson et al. (2016).

While current theoretical accounts of the production effect in recognition performance (including the ones considered here) do not articulate in enough detail how production might influence response time in recognition, we suspect that such consideration would be valuable in advancing our understanding of how the act of production influences memory. Indeed, it might be worth considering the influence of production on response time to reflect a benchmark phenomenon for accounts of the production effect to explain. One particularly fruitful avenue in this direction would be to develop a computational model that jointly simulates both recognition accuracy and response time. Recent efforts that have adopted this approach with other phenomena have yielded important new insights (e.g., Cox & Shiffrin, 2017; Rae et al., 2014).

Conclusion

In the present investigation, we have provided evidence from three preregistered experiments which generally support the idea that part, but possibly not all, of the benefit of production is sensitive to the time available at retrieval. We also examined two computational accounts of the production effect and its interaction with the time available at retrieval—one based on the storage, during study/encoding, and use, during testing, of production associated features and another based on encoding strength. Both models could capture the basic empirical patterns. Future empirical and computational work distinguishing between and further developing these accounts promises a deeper understanding of how more elaborative forms of encoding, such as production, result in improvement in memory.

Acknowledgements

This work was supported by a Discovery Grant (#04091) from the Natural Sciences and

Engineering Research Council of Canada (NSERC), an Early Researcher Award from the Province of Ontario (#ER14-10-258), funding from the Canada Foundation for Innovation and Ontario Research Fund (#37872) from the Canada Research Chairs (#950-232147) program, and Alexander Graham Bell Canada Graduate Scholarships from NSERC.

References

Annis, J., & Malmberg, K. J. (2013). A model of positive sequential dependencies in judgments of frequency. *Journal of Mathematical Psychology, 57*(5), 225-236.

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

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.

Bodner, G. E., Taikh, A., & Fawcett, J. M. (2014). Assessing the costs and benefits of production in recognition. *Psychonomic Bulletin & Review, 21*(1), 149–154.

Bogartz, R. S., & Wackwitz, J. H. (1970). Transforming response measures to remove interactions or other sources of variance. *Psychonomic Science, 19,* 87-89.

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin, 138*(3), 389-414.

Brockdorff, N., & Lamberts, K. (2000). A feature-sampling account of the time course of old-new recognition judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26, 77–102.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved frequency measure for American English. *Behavior Research Methods*, 41, 977-990.

Conway, M. A., & Gathercole, S. E. (1987). Modality and long-term memory. *Journal of*

Memory and Language, 26(3), 341-361.

Cox, G. E., & Shiffrin, R. M. (2017). A dynamic approach to recognition memory. *Psychological Review, 124*(6), 795-860.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning & Verbal Behaviour, 11,* 671-684.

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.

Engelkamp, J. (1998). *Memory for actions*. Hove, UK: Psychology press Ltd.

Eichenbaum, H., Yonelinas, A. P., Ranganath, C. (2007). The medial temporal lobe and recognition memory. *Annual Review of Neuroscience, 30*, 123-152.

Ensor, T. M., Surprenant, A. M., & Neath, I. (2020). Modeling list-strength and spacing effects using version 3 of the retrieving effectively from memory (REM.3) model and its superimposition-of-similar-images assumption. *Behavior Research Methods, 53,* 4-21.

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

Fawcett, J. M., & Ozubko, J. D. (2016). Familiarity, but not recollection, supports the betweensubject production effect in recognition memory. *Canadian Journal of Experimental Psychology, 70*(2), 99-115.

Forrin, N. D., MacLeod, C. M., & Ozubko, J. D. (2012). Widening the boundaries of the production effect. *Memory & Cognition, 40*, 1046-1055.

Gardiner, J. M., Konstantinou, I., Karayianni, I., & Gregg, V. H. (2004). Memory awareness following speeded compared with unspeeded picture recognition. *Experimental Psychology, 52,* 140-149.

Gardiner, J. M., Ramponi, C., & Richardson-Klavehn, A. (1999). Response deadline and subjective awareness in recognition memory. *Consciousness and Cognition, 8*(4), 484–496.

Gronlund, S. D., & Ratcliff, R. (1989). Time course of item and associative information: Implications for global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15, 846–858.

Hintzman, D. L. (1984). MINERVA2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers, 16*, 96-101.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93*, 411-428.

Hintzman, D. L., & Curran, T. (1994). Retrieval dynamics of recognition and frequency judgments: Evidence for separate processes of familiarity and recall. *Journal of Memory and Language*, 33, 1–18.

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.

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

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

Lamberts, K. (2002). Feature sampling in categorization and recognition of objects. *Quarterly*

Journal of Experimental Psychology, 55A(1), 141-154.

Lehman, M., & Malmberg, K. J. (2011) Overcoming the effects of intentional forgetting. *Memory & Cognition, 39,* 335-347.

Loftus, G. R. (1978). On interpretation of interactions. *Memory & Cognition, 6*, 312-319.

MacLeod, C. M., & Bodner, G. E. (2017). The production effect in memory. *Current Directions in Psychological Science, 26,* 390-395.

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

MacLeod, C.M., Pottruff, M. M., Forrin, N.D., & Masson, M. E. J. (2012). The next generation: The value of reminding. *Memory & Cognition, 40,* 693-702.

Malmberg, K. J., Holden, J. E., & Shiffrin, R. M. (2004). Modeling the effects of repetition, similarity, and normative word frequency on old-new recognition and judgments of frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(2), 319-331.

Malmberg, K. J., & Murnane, K. (2002). List composition and the wordfrequency effect for recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28, 616–630.

Malmberg, K. J., Steyvers, M., Stephens, J. D., & Shiffrin, R. M. (2002). Feature frequency effects in recognition memory. *Memory & Cognition*, 30, 607–613.

McElree, B., Dolan, P. O., & Jacoby, L. L. (1999). Isolating the contributions of familiarity and source information to item recognition: A time course analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25, 563–582.

Osth, A. F., Fox, J., McKague, M., Heathcote, A., & Dennis, S. (2018). The list strength effect in source memory: Data and a global matching model. *Journal of Memory and Language*, 103, 91–113

Ozubko, J. D., Gopie, N., & MacLeod, C. M. (2012). Production benefits both recollection and familiarity. *Memory & Cognition, 40,* 326-338.

Ozubko, J. D., Major, J., & MacLeod, C. M. (2014). Remembered study mode: Support for the distinctiveness account of the production effect. *Memory, 22*(5), 509-524.

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

Quinlan, C. K., & Taylor, T. L. (2013). Enhancing the production effect in memory. *Memory, 21*(8), 904-915.

Raaijmakers, J. G., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88*(2), 93–134.

Rae, B., Heathcote, A., Donkin, C., Averell, L., & Brown, S. (2014). The hare and the tortoise: Emphasizing speed can change the evidence used to make decisions. *Journal of Experimental Psychology: Learning, Memory, and, Cognition, 40*(5), 1226–1243.

Sauvage, M. M., Beer, Z., & Eichenbaum, H. (2010). Recognition memory: Adding a response deadline eliminates recollection but spares familiarity. *Learning & Memory, 17*(2), 104–108.

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

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4*(6), 592-604.

Taikh, A., & Bodner, G. E. (2016). Evaluating the basis of the between-group production effect in recognition. *Canadian Journal of Experimental Psychology, 70*(2), 186–194.

Toth, J. P. (1996). Conceptual automaticity in recognition memory: Levels-of-processing effects on familiarity. *Canadian Journal of Experimental Psychology, 50*(1), 123–138.

Van Heuven, W.J.B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology, 67*, 1176-1190.

Verde, M. F. (2013). Retrieval-induced forgetting in recall: Competitor interference revisited. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39, 1433–1448.

Wagenmakers, E.-J., Krypotos, A.-M., Criss, A. H., & Iverson, G. (2012). On the interpretation of removable interactions: A survey of the field 33 years after Loftus. *Memory & Cognition, 40,* 145-160.

Wakeham-Lewis, R. M., Ozubko, J., & Fawcett, J. M. (2021). Characterizing production: The production effect is eliminated for unusual voices unless they are frequent at study.

Wammes, J. D., Meade, M. E., & Fernandes, M. A. (2018). Creating a recollection-based memory through drawing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*(5), 734-751.

Wickelgren, W. A., & Corbett, A. T. (1977). Associative interference and retrieval dynamics in yes-no recall and recognition. *Journal of Experimental Psychology: Human Learning and*

Memory, 3(2), 189–202.

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