Psychonomic Bulletin & Review (2022) 29:2256–2263
https://doi.org/10.3758/s13423-022-02140-x

BRIEF REPORT

Production can enhance semantic encoding: Evidence from forced-choice recognition with homophone versus synonym lures

Jonathan M. Fawcett[1] · **Glen E. Bodner**[2] · **Borys Paulewicz**[3] · **Julia Rose**[1] · **Rachelle Wakeham-Lewis**[1]

Accepted: 15 June 2022 / Published online: 11 July 2022

Abstract

The production effect—better memory for words read aloud rather than silently—has been attributed to responses at test being guided by memory for the act of production. In Experiment 1, we evaluated this distinctiveness account by comparing production effects in forced-choice recognition when lures were either homophones of the targets (*toad* or *towed?*) or unrelated words (*toad* or *seam?*). If the production effect at test was driven solely by memory for the productive act (e.g., articulation, auditory processing), then the effect should be reduced with homophone lures. Contrary to that prediction, the production effect did not differ credibly between homophone-lure and unrelated-lure groups. Experiment 1 led us to hypothesize that production may also boost semantic encoding, and that participants use memory of semantic encoding to guide their forced-choice responses. Consistent with these hypotheses, using synonym lures to interfere with semantic-based decisions (*poison* or *venom?*) reduced the production effect relative to using unrelated lures (*poison* or *ethics?*) in Experiment 2. Our findings suggest that enhanced conceptual encoding may be another useful product of production.

Keywords Production effect · Forced-choice recognition · Distinctiveness

Memories play a critical role in our lives, guiding our decisions (Pillemer, 2003), contributing to our sense of identity (Wilson & Ross, 2003), and allowing us to master complex tasks across multiple encoding episodes (Herzfeld et al., 2014). For this reason, much cognitive research has focused on identifying effective encoding strategies. One simple yet effective strategy that has grown in popularity over the past decade is reading aloud. This strategy has been advocated informally by historical figures as far back as Abraham Lincoln (Herndon & Weik, 1896) and has been studied sporadically over several decades (e.g., Conway & Gathercole, 1987; Hopkins & Edwards, 1972). The memory advantage favouring items read aloud over those read silently has since been dubbed the *production effect* (MacLeod et al., 2010), and has also been obtained using other forms of production such as writing (Forrin et al., 2012), drawing (Wammes et al., 2018), and singing (Quinlan & Taylor, 2013).

Regardless of the modality, the production effect has most often been attributed to the distinctive encoding processes performed on "produced" items at study facilitating their later retrieval (MacLeod et al., 2010; MacLeod & Bodner, 2017). In the case of reading aloud, the *production trace* would incorporate additional elements that the silent items lack (e.g., articulatory/motor and auditory processes; Fawcett & Ozubko, 2016; Fawcett et al., 2012; Forrin et al., 2012). On a recognition test, participants may also base their decisions on whether they can recollect these additional elements (i.e., *If I can remember reading this item aloud, I must have studied it*; Ozubko & Macleod, 2010)—a strategy known as the *distinctiveness heuristic* (Dodson & Schacter, 2001). Although this heuristic is often deemed the basis of production effects (e.g., MacLeod et al., 2010), computational models have questioned whether use of a production trace need be strategic or even conscious (Jamieson et al., 2016).

Regardless of whether access to the production trace is used strategically or via intrinsic retrieval dynamics, distinctiveness has been deemed the "active ingredient" in the production effect (MacLeod et al., 2010, p. 681). This claim

✉ Jonathan M. Fawcett
jfawcett@mun.ca

[1] Department of Psychology, Memorial University of Newfoundland, St. John's, NL, Canada

[2] College of Education, Psychology and Social Work, Flinders University, Adelaide, SA, Australia

[3] Psychology Institute, Jagiellonian University, Krakow, Poland

has been backed by a sizable body of evidence, including the fact that (a) the effect emerges for specific responses (e.g., reading "dog" aloud) but not for nonspecific responses (e.g., pressing the spacebar for each word; MacLeod et al., 2010); (b) the production effect is eliminated when participants have previously read aloud the foil items used at test (thus eliminating the utility of the production trace; Ozubko & MacLeod, 2010); and (c) the production effect is weaker in between-subject designs than in within-subject designs (e.g., Fawcett, 2013; MacLeod et al., 2010).

However, there have also been indications that distinctiveness is not the sole basis of the production effect. For example, early theorists argued that the effect was dependent on the *relative* distinctiveness between aloud and silent items, meaning that it should be observed only when aloud items are studied against a "backdrop" of silent items. Although early studies did not detect the between-subject effects (e.g., Dodson & Schacter, 2001; Hopkins & Edwards, 1972; MacLeod et al., 2010), the existence of a smaller between-subject production effect has since been confirmed experimentally (e.g., Fawcett & Ozubko, 2016; Forrin et al., 2016; Taikh & Bodner, 2016) and in meta-analyses (Bodner et al., 2014; Fawcett, 2013; Fawcett et al., 2022).

The reduction in the magnitude of the production effect between subjects has been explained by some with reference to a dual-process account. Ozubko et al. (2012) demonstrated that the within-subject production effect is driven by both familiarity-based and recollective processes. However, Fawcett and Ozubko (2016) revealed that the between-subject production effect is driven by familiarity alone. Thus, the within-subject effect is larger because both familiarity and recollection contribute, whereas only familiarity contributes to the between-subject effect. Indeed, Fawcett and Ozubko speculated that the production effect might be driven by multiple processes, including attentional or motivational factors that might produce a stronger, better integrated, or more elaborate memory trace (see also Fawcett, 2013; Ozubko et al., 2012). At the very least, these findings suggest that distinctiveness cannot explain all aspects of the data. Findings such as a reverse-production effect (Icht et al., 2014), and a cost to memory for silent items in a within-subject condition relative to a pure-silent list (e.g., Bodner et al., 2014) fall outside the umbrella of the distinctiveness account (MacLeod & Bodner, 2017).

Our study examined what happens to the production effect when the probative value of the production trace is undermined by the presence of related lures in a two-alternative forced-choice (2AFC) recognition task. Experiment 1 accomplished this using a standard within-subject production manipulation at study, and then testing recognition in a 2AFC task using either homophone lures (e.g., *bare* or *bear?*) or unrelated lures (e.g., *bare* or *merry*?). The standard distinctiveness-based account of the production effect predicts a reduction of the production effect in the homophone-lure condition where the diagnostic value of the production trace is undermined. The results of Experiment 1 led us to evaluate the influence of synonym lures (e.g., *error* or *mistake*?) on the production effect in Experiment 2, to establish whether production enhances semantic processing.

Experiment 1: Homophone lures

Experiment 1 evaluated the distinctiveness account's claim that memory of the production trace (i.e., of having said studied words aloud) underlies the production effect. To this end, we compared the production effect in a forced-choice recognition task across groups that either received homophone lures or unrelated lures. According to the distinctiveness account, homophone lures should reduce or eliminate the production effect relative to the unrelated lure group, because memory of saying the target aloud will not advantage selection of the target over the lure. However, if production also strengthens memory for other aspects of encoding, such as the item's meaning, then the production effect might not be affected by this manipulation. Participants studied a mixed list of words: half were read aloud and half were read silently. They then performed a 2AFC recognition task requiring a confidence judgment with respect to their decision about which word in each pair had been studied.

Method

Participants

University of Calgary undergraduates participated for course credit and were randomly assigned to receive either homophone lures or unrelated lures (48 per group).

Materials

The critical stimuli were 80 word-triplets, each comprising a homophone target, a homophone lure, and an unrelated homophone lure (e.g., *bare-bear-merry*), selected from online sources (see Tables S1 and S3 of the Online Supplement). Which homophone served as the target within a given triplet was randomized for each participant. Another six triplets served as practice items. Word length and frequency were similar for the three item types across triplets (see the supplementary information for details). Half of each set were assigned to the silent condition, and half to the aloud condition, determined randomly for each participant.

Procedure

Participants were tested individually in a lab room. The experiment was run on a Mac computer using PsychoPy (Version 1.90.3; Peirce et al., 2019). Stimuli and instructions were presented on a 24-inch monitor in white 32-pt Arial font against a grey background.

Participants were told that they would read a list of words, half silently and half aloud, for an unspecified memory test. Words to be read silently were preceded by an eye icon, and words to be read aloud were preceded by a mouth icon. The six practice and 80 critical study trials followed; each set randomly ordered for each participant. Each study trial comprised a fixation stimulus ("+") for 500 ms, the eye/mouth icon for 1,500 ms, the lowercase word for 3,000 ms, and an intertrial interval of 1,000 ms. The experimenter ensured compliance with the silent/aloud cues during the practice trials.

The test phase immediately followed. Participants were told that on each trial, two words would be shown side by side. Their task was to judge which word they had studied using a 6-point rating scale, provided at the bottom of the screen (1 = *very sure left*, 6 = *very sure right*). Participants entered their responses using the number keys. The test consisted of 6 practice trials (based on the practice study trials; not analyzed) and 80 critical trials, each set randomly ordered for each participant. The experimenter ensured the task was clear to participants after the practice trials. Each trial consisted of a fixation stimulus ("+") for 500 ms, followed by the word pair with the confidence scale underneath. The studied target appeared on the left for half the trials, and on the right for half the trials, determined randomly.

Statistical approach

Following previous work (see Fawcett et al., 2016; Fawcett & Ozubko, 2016), a fully Bayesian analytic approach was used, based on fitting a series of multilevel models implemented in Version 2.9.0 of the *brms* package (Bürkner, 2017, 2018) within R (Version 3.5.2; R Core Team, 2018). The models used uninformative, mildly regularizing priors for all parameters. We fit our models using four separate chains with random starting points for each parameter and at least 6,000 iterations per chain (half of which were used as a warm-up period). This resulted in a minimum total post-warm-up sample of 12,000 iterations for each model. Model convergence metrics indicated that our models had converged (R-hat ~ 1 and $N_{Effective} > 2000$ across all parameters; Gelman et al., 2014; Gelman & Hill, 2007).

The data were analyzed two ways. First, we dichotomized the confidence ratings and estimated d' for each condition using probit regression, as described by Fawcett and Ozubko (2016). We then fit a multilevel ordinal regression model making full use of the confidence ratings to estimate d' (Paulewicz & Blaut, 2020). Both models supported the same conclusions (with nearly indistinguishable condition estimates), therefore the more complex ordinal model is reported in the Online Supplement. Estimates of d' were divided by the square-root of 2, placing them on a scale comparable to typical d' values. Either modelling approach included random effects (intercepts and slopes, as appropriate) for both participant and item.

Table 1 Experiments 1 and 2: Empirical mean (*SE*) accuracy (%) and d' by production condition and group

	Accuracy		d'	
	Aloud	Silent	Aloud	Silent
Experiment 1				
Homophone-lure Group	81.9 (1.4)	75.2 (1.5)	1.44 (0.09)	1.06 (0.08)
Unrelated-lure Group	86.1 (1.1)	79.2 (1.5)	1.73 (0.09)	1.29 (0.08)
Experiment 2				
Synonym-lure Group	82.3 (1.3)	78.6 (1.3)	1.49 (0.09)	1.22 (0.08)
Unrelated-lure Group	85.7 (1.2)	77.6 (1.2)	1.69 (0.09)	1.15 (0.07)

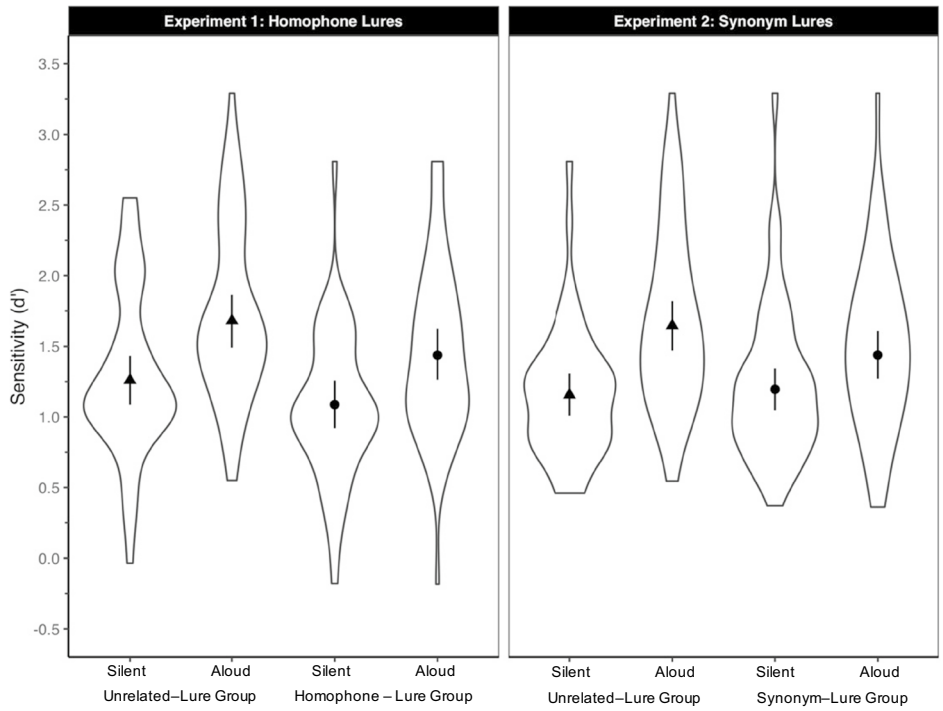
Results

Table 1 provides mean accuracy and d' for each condition based on the raw data. As depicted in Fig. 1, our probit model showed a production effect in the homophone-lure group, $PE = 0.35$, 95% CI [0.21, 0.50], and in the unrelated-lure group, $PE = 0.42$, 95% CI [0.27, 0.58], that were of similar magnitude, difference = 0.07, 95% CI [−0.13, 0.27].[1] In short, eliminating the utility of memory for having said the target words aloud at study had surprisingly little impact on the magnitude of the production effect.

Discussion

The production effect was of similar magnitude whether the lures sounded like the targets or not. Clearly, the production effect in the homophone-lure group was not driven

[1] An analysis of the d' values reported in Table 1 using a frequentist ANOVA produces a similar outcome, with a main effect of production condition, $F(1, 94) = 54.66$, $p < .001$, and a main effect of group, $F(1, 94) = 5.70$, $p = .019$, but no interaction, $F(1, 94) = 0.27$, $p = .605$. Although we favor parameter estimation over model comparison, a Bayes factor (BF) was likewise calculated for the critical interaction using the *hypothesis* function from the *brms* package. Supporting preceding conclusions, there was substantial evidence supporting the null ($BF_{01} = 5.5$).



Note. Error bars represent 95% confidence intervals. Violin plots reflect the distribution of the *d'* values summarized in Table 1.

Fig. 1 Experiments 1 and 2: Sensitivity (d′) Estimated from the probit regression model as a function of production condition and group

by reliance on memories of having said the studied words aloud. Had that been the case, the production effect should have been weak or absent in this group. Although our findings trended toward a smaller effect in the homophone-lure group relative to the unrelated-lure group, this difference was not credible. The similarity of the production effect across lure types is surprising, and indicates that at least under some circumstances, production enhances memory through a means other than retrieval of the production trace.

In their classic study of transfer-appropriate processing, Morris et al. (1977) argued that recognition decisions are typically based on consideration of the meanings rather than on the sound of the test items. Consistent with that possibility, their participants were more likely to recognize a target like TRAIN if it had been studied in a semantic task (e.g., judging whether TRAIN makes sense in the phrase "The ____ had a silver engine") than if they had studied it in a rhyme-based task (e.g., judging whether TRAIN rhymes with BRAIN). In Experiment 1, reliance on meaning (rather than sound) for guiding 2AFC recognition decisions would enable participants to successfully choose the targets regardless of whether the lures sounded like the targets. Furthermore, if production strengthens semantic encoding, this could explain why the production effect was similar across groups. Experiment 2 tested the hypotheses that (1) our 2AFC recognition judgments relied on semantic encoding, and (2) production boosts semantic encoding. Synonym lures should impair participants' ability to use memory of having processed the meaning of items during study to guide their recognition judgments. If so, then the production effect on 2AFC recognition should be *smaller* in the synonym-lure group than in the unrelated lure group. Alternatively, the synonym-lure group might simply shift their focus from the meaning to the sound of the two alternatives, in line with the distinctiveness account of the production effect. If so, then by this account both groups will rely on memory for the production trace and thus should show similar production effects.

Experiment 2: Synonym lures

Experiment 1 tested two nested premises: (1) production improves the semantic encoding of items and (2) participants rely on semantics when making standard 2AFC recognition decisions (cf. in a rhyme-based recognition task; Morris et al., 1977). If production improves semantic encoding, and if participants rely on semantics at test, then making it more difficult for participants to use meaning to make their decisions should reduce the production effect. To test this possibility, Experiment 2 compared the size of the production effect across a synonym-lure group, in which both alternatives overlapped in meaning (e.g., *error* or *mistake?*), and an unrelated-lure group, in which meaning overlap was minimal (e.g., *error* or *purchase?*). The

synonym-lure group should have a harder time using meaning to pick out the studied targets than the unrelated-lure group, and therefore should show a smaller production effect. In contrast, a distinctiveness account based on reliance on memory for the act of production predicts a similar production effect in both groups.

Method

Participants

Participants from the same pool as Experiment 1 were randomly assigned to either the synonym-lure group or the unrelated-lure group (48 per group).

Materials

A new set of critical stimuli were created as per Experiment 1, except the six practice and 80 critical triplets each consisted of a target, a synonym lure, and an unrelated lure (e.g., *poison-venom-ethics*; see Tables S2 and S4 in the Online Supplement).

Procedure

The procedure was identical to Experiment 1.

Results

Table 1 provides mean accuracy and d' for each condition based on the raw data. As depicted in Fig. 1, we observed a production effect in the synonym-lure group, $PE = 0.24$, 95% CI [0.09, 0.39], and in the unrelated-lure group, $PE = 0.48$, 95% CI [0.34, 0.64].[2] Critically, the production effect was credibly smaller in the synonym-lure group than the unrelated-lure group, difference = 0.25, 95% CI [0.04, 0.45]. Thus, participants exhibited a smaller production effect when forced to choose between semantically related target–lure pairs than between unrelated target–lure pairs.

Discussion

Building on Morris et al.'s (1977) classic demonstration, Experiment 2 suggests that participants use memory of prior semantic processing to guide their recognition responses: Reducing the utility of memory of prior semantic encoding mitigated the production effect. In turn, this finding indicates that production may enhance semantic encoding. Thus, the traditional distinctiveness account, which holds that production works by encoding sensorimotor features associated with production (e.g., articulation and/or auditory processing), is too limited. By that account, the production effect should not have been reduced in the synonym-lure group relative to the unrelated-lure group in Experiment 2 (contrary to what we found)—whereas the production effect should have been reduced in the homophone-lure group relative to the unrelated-lure group in Experiment 1 (contrary to what we found). Taken together, Experiments 1 and 2 suggest that production may facilitate conceptual encoding.

General discussion

Our study interrogated the form of the production trace in the production effect. In Experiment 1, the production effect on the 2AFC recognition test was similar whether target–lure pairs were matched homophones (*bare* or *bear?*) or unmatched homophones (*bare* or *merry?*). The production effect survived even when the diagnostic utility of the sensorimotor components comprising the production trace was eliminated. However, in Experiment 2, requiring participants to choose between matched synonyms (*error* or *mistake?*) roughly halved the magnitude of the production effect relative to selecting between unmatched words (*error* or *purchase?*). This pattern suggests that the production effect may also reflect enhanced conceptual encoding, rather than relying solely on sensorimotor representations. Thus, our findings invite reconsideration of the loci of production effects and consideration of additional mechanisms.

One plausible explanation for our findings is that reading aloud facilitates activation of semantic information and the binding of that information to the encoding episode. This claim is compatible with a recent functional magnetic resonance imaging study (Bailey et al., 2021), which observed greater activation in regions of the anterior temporal lobe associated with semantic processing on aloud trials than

[2] An analysis of the d' values reported in Table 1 using a frequentist ANOVA produces a similar outcome, with a significant effect of production condition, $F(1, 94) = 43.65$, $p < .001$, a nonsignificant effect of group, $F(1, 94) = 0.42$, $p = .517$, but a significant interaction, $F(1, 94) = 4.99$, $p = .028$. As before, a supplementary BF was calculated for the critical interaction. Here, support was less clear, with only anecdotal evidence supporting the alternative hypothesis ($BF_{10} = 2.9$). Further, because our conclusions might imply a three-way interaction between production, matching, and experiment, an additional BF was calculated evaluating whether the matching-related decrease in the production effect differed between Experiments 1 and 2: Here, the idea that phonology plays a special role in the production effect is represented by a directional hypothesis predicting a greater reduction for Experiment 1 than Experiment 2. Although a comparison against the null revealed weak evidence favoring no difference ($BF_{01} = 2.4$), the directional test found substantial evidence favoring a larger reduction, difference = 0.24, 95% CI [−0.15, 0.63], in the magnitude of the production effect in Experiment 2, rather than Experiment 1 ($BF = 7.7$). This outcome is opposite what would be expected if phonological representations were central to the production effect in this task. Importantly, our experiments were not powered or designed for between-study comparisons. Additional data are needed to resolve this comparison fully.

on silent trials (see also Fawcett, 2013; Fawcett & Ozubko, 2016; Ozubko et al., 2012). However, this claim is challenged by studies reporting that requiring deep processing of both aloud and silent items (e.g., Forrin et al., 2014; MacLeod et al., 2010; Zormpa et al., 2019) does not mitigate the production effect. Nonetheless, in these studies, compliance with the deep task on the silent trials could not be checked (to avoid production of a response), and participants knew they would also be performing the production task. Therefore, participants may have differentially processed items in the deep task based on whether they needed to produce them. Specifically, participants may have engaged in less semantic processing for aloud items than silent items in the context of a semantic encoding task, thus masking a reduction in the production effect. This possibility warrants further investigation.

Importantly, our claim that production can enhance semantic encoding is not incompatible with the role of distinctiveness of the production trace—it simply challenges models based solely on this mechanism. In the context of a formal computational model, such as the one proposed by Jamieson et al. (2016), the present findings can be accommodated by recognizing that the production trace (i.e., those additional features appended to the representation of aloud items) may include semantic components alongside sensorimotor components. In this manner, additional semantic encoding would result in a form of *semantic* distinctiveness within the model. However, if the production trace contains unique sensorimotor components that are used to guide recognition decisions, then it remains surprising that the production effect was undiminished in the homophone-lure group in Experiment 1.[3]

An alternate interpretation is that the production effect does not arise from a singular mechanism. Researchers studying related phenomenon (e.g., enactment) have often commented on the fact that many difficulties in developing a coherent theoretical narrative are attributable to the field's desire that effects be explained by a single process. However, tasks in human memory are rarely process pure (e.g., Surprenant & Neath, 2009). In discussing the enactment effect, Russ et al. (2003) point out that processing an action phrase for enactment "involves a strong self-involvement, the formation of an intention to act, an obligatory activation of the action schema, and object knowledge" that results in "cognitive processing at a much higher complexity level than provided by the primary motor cortex functions" (p. 498). Rosner et al. (2013) echo a similar sentiment, arguing that generation might "promote increases in attention, cognitive effort, item-distinctiveness, semantic processing and conceptual processing" (p. 6). These arguments fit well within a dual-process account of the production effect that posits roles for both recollection and familiarity (e.g., Fawcett & Ozubko, 2016; Ozubko et al., 2012), as well as attentional involvement in the effect (e.g., Mama et al., 2018; Mama & Icht, 2018). Complex encoding manipulations—including production—stand to invoke multiple processes, and which of these contribute to memory performance will depend on task and contextual factors (e.g., Morris et al., 1977). Given that production shares similarities with both enactment (production involves enacting a response) and generation (production involves generating a response), this argument seems reasonable to us. Determining these multiple processes remains an important area for future work on the production effect.

In a similar vein, it is worth pointing out that although the present investigation used only a single production modality (i.e., reading aloud), we would predict similar results had participants sang or written the produced words (rather than read them aloud). This is because such production modalities presently share a common theoretical framework and should evoke similar semantic encoding based on the logic laid out in the introduction. Even so, whether a similar pattern would be observed using a different form of production—or even a different memory task (e.g., matched foils in a yes-no or remember-know recognition paradigm) or between-subject design—remain an important future direction that we are presently investigating.

Conclusion

The present study provides a key theoretical challenge to the notion that the production trace is the sole driver of the production effect on recognition memory. Production can benefit memory even when memory of saying items aloud is not diagnostic. Further, manipulations that modulate the utility of the semantic features of items differentially impact aloud and silent items, mitigating the production effect's magnitude. We suggest that items read aloud are distinctive not only due to the inclusion of sensorimotor elements, but also because the act of production encourages broader conceptual encoding.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13423-022-02140-x.

[3] According to the "triangle" model of word recognition (e.g., Seidenberg & McClelland, 1989) visual word presentation activates three types of representations: phonological, semantic, and orthographic. Whereas we have already addressed phonological and semantic representations, one might wonder whether the production effect is driven by the augmentation of orthographic representations. Although not tested directly, our findings suggest otherwise: If participants were using orthography to facilitate recognition, we would not expect an impact of lure type in either experiment as our lures always differed orthographically from their matched targets (particularly in Experiment 2).

References

Bailey, L. M., Bodner, G. E., Matheson, H. E., Stewart, B. M., Roddick, K., O'Neil, K., ... Fawcett, J. M. (2021). Neural correlates of the production effect: An fMRI study. *Brain and Cognition, 152*, Article 105757.

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

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1–28.

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal, 10*(1), 395–411.

Conway, M. A., & Gathercole, S. E. (1987). Modality and long-term memory. *Journal of Memory and Language, 26*, 341–361.

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

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

Fawcett, J. M., & Ozubko, J. (2016). Familiarity, but not recollection, supports the between-subject production effect. *Canadian Journal of Experimental Psychology, 70*, 99–115.

Fawcett, J. M., Quinlan, C. K., & Taylor, T. L. (2012). Interplay of the production and picture superiority effects: A signal detection analysis. *Memory, 20*(7), 655–666.

Fawcett, J. M., Lawrence, M. A., & Taylor, T. L. (2016). The representational consequences of intentional forgetting: Impairments to both the probability and fidelity of long-term memory. *Journal of Experimental Psychology: General, 145*, 56–81.

Fawcett, J. M., Baldwin, M. M., Drakes, D. H., & Willoughby, H. V. (2022). *Production improves recognition and reduces intrusions in between-subject designs: An empirical and meta-analytic investigation.* Manuscript submitted for publication.

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

Forrin, N. D., Jonker, T. R., & MacLeod, C. M. (2014). Production improves memory equivalently following elaborative vs non-elaborative processing. *Memory, 22*(5), 470–480.

Forrin, N. D., Groot, B., & MacLeod, C. M. (2016). The d-Prime directive: Assessing costs and benefits in recognition by dissociating mixed-list false alarm rates. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(7), 1090–1111.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multi-level/hierarchical models*. Cambridge University Press.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Chapman & Hall.

Herndon, W. H., & Weik, J. W. (1896). *Abraham Lincoln: The true story of a great life* (Vol. 2). Project Gutenberg.

Herzfeld, D., Vaswani, P., Marko, M., & Shadmehr, R. (2014). A memory of errors in sensorimotor learning. *Science, 345*(6202), 1349–1353.

Hopkins, R. H., & Edwards, R. E. (1972). Pronunciation effects in recognition memory. *Journal of Verbal Learning and Verbal Behavior, 11*(4), 534–537.

Icht, M., Mama, Y., & Algom, D. (2014). The production effect in memory: Multiple species of distinctiveness. *Frontiers in Psychology, 5*, 886.

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

Macleod, C., & Bodner, G. (2017). The production effect in memory. *Current Directions in Psychological Science, 26*(4), 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*(3), 671–685.

Mama, Y., & Icht, M. (2018). Production effect in adults with ADHD with and without methylphenidate (MPH): Vocalization improves verbal learning. *Journal of the International Neuropsychological Society, 25*(2), 230–235.

Mama, Y., Fostick, L., & Icht, M. (2018). The impact of different background noises on the production effect. *Acta Psychologica, 185*, 235–242.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16*(5), 519–533.

Ozubko, J. D., & Macleod, C. M. (2010). The production effect in memory: Evidence that distinctiveness underlies the benefit. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 1543–1547.

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

Paulewicz, B., & Blaut, A. (2020). The bhsdtr package: A general-purpose method of Bayesian inference for signal detection theory models. *PsychArchives*. https://doi.org/10.23668/PSYCHARCHIVES.2725

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., ... Lindeløv, J. (2019). PsychoPy2: experiments in behavior made easy. *Behavior Research Methods, 51*, 195–203.

Pillemer, D. (2003). Directive functions of autobiographical memory: The guiding power of the specific episode. *Memory, 11*(2), 193–202.

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

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing https://www.R-project.org/

Rosner, Z. A., Elman, J. A., & Shimamura, A. P. (2013). The generation effect: Activating broad neural circuits during memory encoding. *Cortex, 49*(7), 1901–1909.

Russ, M. O., Mack, W., Grama, C.-R., Lanfermann, H., & Knopf, M. (2003). Enactment effect in memory: Evidence concerning the function of the supramarginal gyrus. *Experimental Brain Research, 149*(4), 497–504.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96*(4), 523.

Surprenant, A. M., & Neath, I. (2009). *Principles of memory*. Psychology Press.

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.

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.

Wilson, A., & Ross, M. (2003). The identity function of autobiographical memory: Time is on our side. *Memory, 11*(2), 137–149.

Zormpa, E., Brehm, L. E., Hoedemaker, R. S., & Meyer, A. S. (2019). The production effect and the generation effect improve memory in picture naming. *Memory, 27*(3), 340–352.

Open practices statements None of the data or materials for the experiments reported here are available elsewhere (except by request), and none of the experiments were preregistered.