Canadian Journal of Experimental Psychology / Revue canadienne de psychologie expérimentale

Production Improves Recognition and Reduces Intrusions in Between-Subject Designs: An Updated Meta-Analysis

Jonathan M. Fawcett, Maddison M. Baldwin, Jedidiah W. Whitridge, Michelle Swab, Kyla Malayang, Brooke Hiscock, Dalainey H. Drakes, and Hannah V. Willoughby

Online First Publication, December 15, 2022. https://dx.doi.org/10.1037/cep0000302

CANADIAN PSYCHOLOGICAL ASSOCIATION
SOCIÉTÉ CANADIENNE DE PSYCHOLOGIE

Canadian Journal of Experimental Psychology / Revue canadienne de psychologie expérimentale

ISSN: 1196-1961

https://doi.org/10.1037/cep0000302

Production Improves Recognition and Reduces Intrusions in Between-Subject Designs: An Updated Meta-Analysis

Jonathan M. Fawcett, Maddison M. Baldwin, Jedidiah W. Whitridge, Michelle Swab, Kyla Malayang, Brooke Hiscock, Dalainey H. Drakes, and Hannah V. Willoughby
Department of Psychology, Memorial University of Newfoundland

The production effect refers to the finding that words read aloud are better remembered than those read silently. This pattern has most often been explained as arising from the incorporation of sensorimotor elements into the item representation at study, which could then be used to guide performance at later test. This theoretical framework views aloud items as being distinctive in relation to silent items, and thus the effect was thought to emerge only when production was manipulated within-subjects. This claim was later challenged, and a reliable (albeit smaller) between-subject production effect has since been shown in recognition memory. Across a series of meta-analyses, we extend this earlier work, replicating the between-subject production effect for recognition, and demonstrating no such effect for overall target recall. However, supporting recent theoretical claims, we further observed an interaction between the production effect and serial position within recall, such that a production effect was observed for late time points but not early time points (a similar, albeit smaller and noncredible trend was observed for recognition). Finally, we provide evidence that production reduces off-list intrusions. In summary, production has a reliable impact on recognition memory when manipulated between-subjects, but a more complex relationship with recall performance.

Public Significance Statement
When studying, reading some passages aloud and others silently has been shown to help you remember the ones you read aloud. The present work shows that reading *everything* aloud provides a similar benefit, but only if you are tested using recognition memory (e.g., multiple choice). If you are tested using recall (e.g., short answer), reading everything aloud will make it easier to remember the last few passages but harder to remember the first few passages.

Keywords: production effect, distinctiveness, between-subjects, recall, meta-analysis

Our ability to selectively retain information is central to most aspects of our lives, granting us both the potential to cultivate new skills and maintain a sense of personal stability (Fawcett & Hulbert, 2020). Indeed, there are few areas of human behaviour in which memory does not play at least a supporting role. But not all information is worth retaining, at least in the long run. This has driven scientific efforts to identify ways in which important information might be highlighted to ensure that it remains accessible over time. One approach identified has been that of production. Since at least the 1970s, it has been known that producing something—for example, by reading aloud (e.g., Hopkins & Edwards, 1972), mouthing (e.g., Gathercole & Conway, 1988), or singing (e.g., Quinlan & Taylor, 2013)—improves retention of that information relative to nonproduced information. MacLeod et al. (2010) revived interest in this topic, rebranding it as the *production effect*. Building on earlier work by Conway and Gathercole (1987), MacLeod et al. (2010) attributed this effect to distinctiveness, whilst favouring the notion that participants used memory of having produced the words at study as a means of rejecting unstudied words at test (Dodson & Schacter, 2001).

However, a strong claim made by proponents of the distinctiveness account was that the production effect should be limited to

Jonathan M. Fawcett https://orcid.org/0000-0002-4248-5371
Hannah V. Willoughby https://orcid.org/0000-0002-3597-5532

The authors would like to thank Colin MacLeod, Kathleen Hourihan, Glen Bodner, and Noah Forrin for discussions relating to an earlier draft of this article. The authors would also like to thank Noah Pevie for helping with the search. Jonathan M. Fawcett was funded by a discovery Grant from the Natural Sciences and Engineering Research Council of Canada (ID0EFABG877).

Jonathan M. Fawcett and Hannah V. Willoughby conceptualized the project. Michelle Swab coordinated and managed the search. All authors contributed to the search and coding of included studies. Jonathan M. Fawcett conducted all analyses. All authors contributed to the article.

The present meta-analysis was not preregistered, having begun prior to the adoption of such practises in our laboratory (which is in transition even at this time). All coded data and analysis scripts are available on the Open Science Framework (https://osf.io/rsu6w/) with further details or clarification available from the Jonathan M. Fawcett (jfawcett@mun.ca).

Correspondence concerning this article should be addressed to Jonathan M. Fawcett, Department of Psychology, Memorial University of Newfoundland, St. John's, NL A1B 3X9, Canada. Email: jfawcett@mun.ca

within-subject designs wherein participants studied words both aloud and silently (e.g., MacLeod et al., 2010). According to this perspective, aloud items should be distinctive only against a backdrop of silent items. At first, this prediction was supported, with between-subject designs (wherein participants studied only items read aloud or silently) showing no effect (e.g., Hopkins & Edwards, 1972; MacLeod et al., 2010). However, Fawcett (2013) later demonstrated that when meta-analyzed, the literature produced a small but surprisingly consistent production effect for recognition memory. Although this finding might have been viewed as a challenge to the distinctiveness account, this account has since been revised to view the smaller effect between-subjects as supportive of the distinctiveness account (e.g., Bodner et al., 2020).

Notably, Fawcett (2013) was unable to evaluate the existence of a production effect in recall memory because too few studies used that measure. A decade later, several studies have reported the absence of a between-subject production effect using recall (e.g., Forrin & MacLeod, 2016; Jones & Pyc, 2014; Lambert et al., 2016). However, before accepting this conclusion, we must interrogate the evidence: As noted above, the analogous effect for recognition was long thought absent until meta-analyzed. Further, recent theorists have qualified claims pertaining to the absence of a between-subject production effect in recall, pointing to a possible interaction with serial position (e.g., Gionet et al., 2022; Saint-Aubin et al., 2021). For that reason, the primary goal of the present article was to synthesize those studies testing the between-subject production effect in recall. We also opted to update the analyses conducted by Fawcett (2013), as nonsignificant effects continue to be reported as evidence in support of a distinctiveness account (e.g., Quinlan & Taylor, 2019).

Search and Coding Procedures

A search of the online resources *PsycINFO*, *PsychARTICLES*, *Web of Science*, and *Scopus* was conducted using the Boolean search phrase: "production effect." We further used *Web of Science* to review all articles citing MacLeod et al. (2010) and evaluated any article cited by Fawcett (2013) original meta-analysis to conduct both forward and backward snowball searches of all eligible studies. Advertisements were also forwarded to relevant societies (e.g., the Canadian Society for Brain, Behaviour, and Cognitive Science) seeking additional data. Finally, all senior authors with an eligible study were contacted to ask whether they knew of potentially unpublished work—and whether they were willing to provide raw data. Any article containing a between-subject production effect measured using recognition or recall within a healthy, young adult sample of greater than 10 participants per group was included. Only articles published after 1970 were considered for inclusion; this was because studies published around and prior to that date rarely provided sufficient data for our analyses, and excluding them allowed us to keep our snowball search manageable.

During our search, we became aware of studies using pure-lists in within-subject designs. Because our focus was on the between-subject production effect, we included such studies only in cases where (a) the task used study-test blocks rather than the provision of all study items followed by a singular test phase; and, (b) we had sufficient information to code the initial block, effectively treating the study as between-subjects. It is not our intention to adopt a stance as to whether production has a differential effect on pure-lists when manipulated within- or between-subjects (we suspect not), but rather to ensure the simplicity of our analyses and authenticity to our intended goals. Finally, although included in some analyses by Fawcett (2013), we excluded list discrimination (Ozubko & Macleod, 2010) and frequency judgement (Hopkins & Edwards, 1972) tasks from our present analyses on the basis that (a) there were few studies using either measure, and (b) that they may tap different underlying processes. We also excluded one study using drawing (Wammes et al., 2016) on the basis that some theorists have argued for a distinction between drawing and verbal production. When recall data were scored in multiple ways, we preferred the method closest to free recall.

Because our focus was on the production effect itself, rather than on modulations thereof, in studies containing a secondary, within-subject encoding manipulation (e.g., generation, imagery; Bodner et al., 2020) we included only the standard (i.e., nonelaborative) trials, where able. Samples for which no such nonelaborative condition was available (e.g., Taikh & Bodner, 2016, Experiment 2) were still included, although sensitivity analyses were undertaken verifying that their exclusion did not affect our conclusions. One experiment (Bodner et al., 2016, Experiment 2) was excluded prior to analysis because the production manipulation was intentionally confounded with study time to undermine the effect. One unpublished study (Maddox, 2019) was excluded due to there being too few usable first-block participants in one experiment (8 in one group), having used a read aloud manipulation in a group setting, and including practise trials for each pure-list production prior to the initial list in their other experiment (inclusion of this study did not alter our conclusions). Following initial (but prior to final) analysis, we also excluded studies using two-alternative forced choice (MacLeod et al., 2010, Experiment 3, and Hopkins & Edwards, 1972, Experiment 1) because they were uncommon and their inclusion complicated interpretation pertaining to the division between hits and false alarms. In this case, either sample demonstrated a numeric production effect of ~3%, meaning that inclusion yielded stronger support for our conclusions.

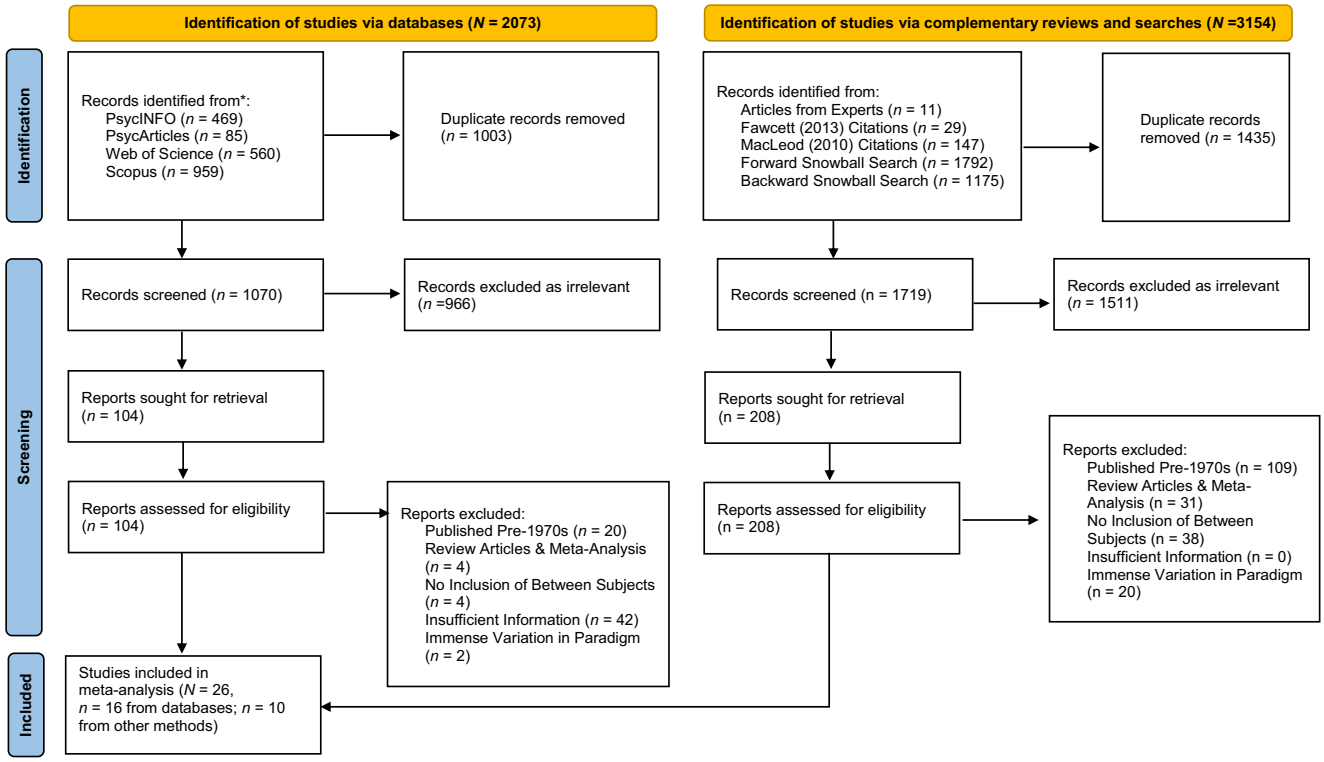
After exclusions and removal of duplicates, a total of 39 recognition and 22 recall effects from 26 articles were analyzed (as opposed to the 12 recognition effects from eight articles analyzed by Fawcett, 2013); a flowchart summarizing inclusion decisions is provided in Figure 1. Articles contributing one or more effect sizes are indicated in the reference section by an asterisk (*). Data were coded for measures of yes–no recognition and recall as the percent correct responses for the target items. Sensitivity (d') was calculated where appropriate and intrusions were calculated as the mean number of unstudied words recalled in each condition. Production method (e.g., reading aloud, mouthing), number of study items, presentation time at study and study-test delay were coded as exploratory moderator variables for hits, false alarms and d'; test duration (coded as short = 2 min or less and long = greater than 2 min), average intrusions (averaged across the study without regard to production condition) and per-subject recall performance were coded as exploratory moderator variables for intrusions. Continuous moderators were scaled. Where possible, effect sizes were calculated based on serial position with separate effects for early (i.e., the first three) and late (i.e., the last three) items, permitting us to evaluate claims of a production effect for the latter but not the former in studies using recall (e.g., Gionet et al., 2022; Saint-Aubin et al., 2021).

Effect Size Calculations and Statistical Approach

Effect sizes were calculated as raw mean differences via the *escalc* function of the *metafor* package (Viechtbauer, 2010) in the *R*

Figure 1

Meta-Analysis Inclusion Flowchart



Note. See the online article for the color version of this figure.

statistical programming language (R Core Team, 2018). In cases where the variability within a given group was unavailable, they were imputed from other studies using the same measure. All models were fit using the *brms* package (Bürkner, 2017, 2018) under assumptions analogous to a random-effects model; random effects likewise permitted inclusion of multiple dependent effects (e.g., due to the use of a common comparison group) from the same study, by modelling a single latent estimate for those effects (this decision impacted, e.g., Forrin & MacLeod, 2018; Gathercole & Conway, 1988; Quinlan & Taylor, 2019). In particular, our models corresponded to a three-level meta-analysis with random effects for sample and effect. Bayesian models were preferred because they (a) permit direct interpretation of our effect sizes (whereas Frequentist models permit confidence intervals to be used only inferentially; e.g., Hoekstra et al., 2014); (b) naturally propagate uncertainty from all parameters (including τ) across all other parameters; and, (c) permit the incorporation of regularizing, prior knowledge (e.g., that the aggregate production effect is not greater than 50% in magnitude as measured by hits). Readers interested in Bayesian modelling and its benefits over Frequentist statistics are referred to textbooks by Kruschke (2015) or McElreath (2020).

As we were able to access raw data for each study reporting intrusions, we conducted models directly on those raw data. Visual inspection of the intrusion data revealed them to be skewed toward 0, as is expected with rare events. This, combined with the fact that they were inherently categorical in nature, led us to adopt a Poisson regression (with a log-link function), as is considered best practise in such cases (Cameron & Trivedi, 1998). These models included random slopes and intercepts for each sample alongside the preexisting random-effects structure.

Each model used uninformative, mildly regularizing priors for all parameters. For our meta-analytic analyses of hits, false alarms and recall, our priors reflected the expectation that the production effect in a typical study would likely range between −20% and 20%, with individual studies ranging between −30% and 30%; for our analysis of d', priors reflected our belief that the production effect in a typical study would likely range between −1 and 1, with individual studies ranging between −1.5 and 1.5. For intrusions, priors were calibrated to the belief that the average number of intrusions could range from < 0.05 to 20, with individual participant performance ranging from 0 to 1,000 (the latter being impossibly high). These priors were broad because we were unaware of any basis to inform expectation at the time the model had been run. Importantly, their broad nature meant only that the posterior distribution would be more sensitive to the data. Parallel zero-inflated Poisson models were fit alongside each of the Poisson models reported in-text. In each case, the zero-inflated models produced similar conclusions, although the simpler models were generally supported via model comparison metrics, such as cross-validation. For those reasons, and for the sake of brevity, we reported only the former. Priors for slopes within each model were designed to be uninformative.

Results and Discussion

Apart from those discussed in text, no moderators were credible. For the sake of brevity, those models are not discussed further.

Recognition

Hits and False Alarms

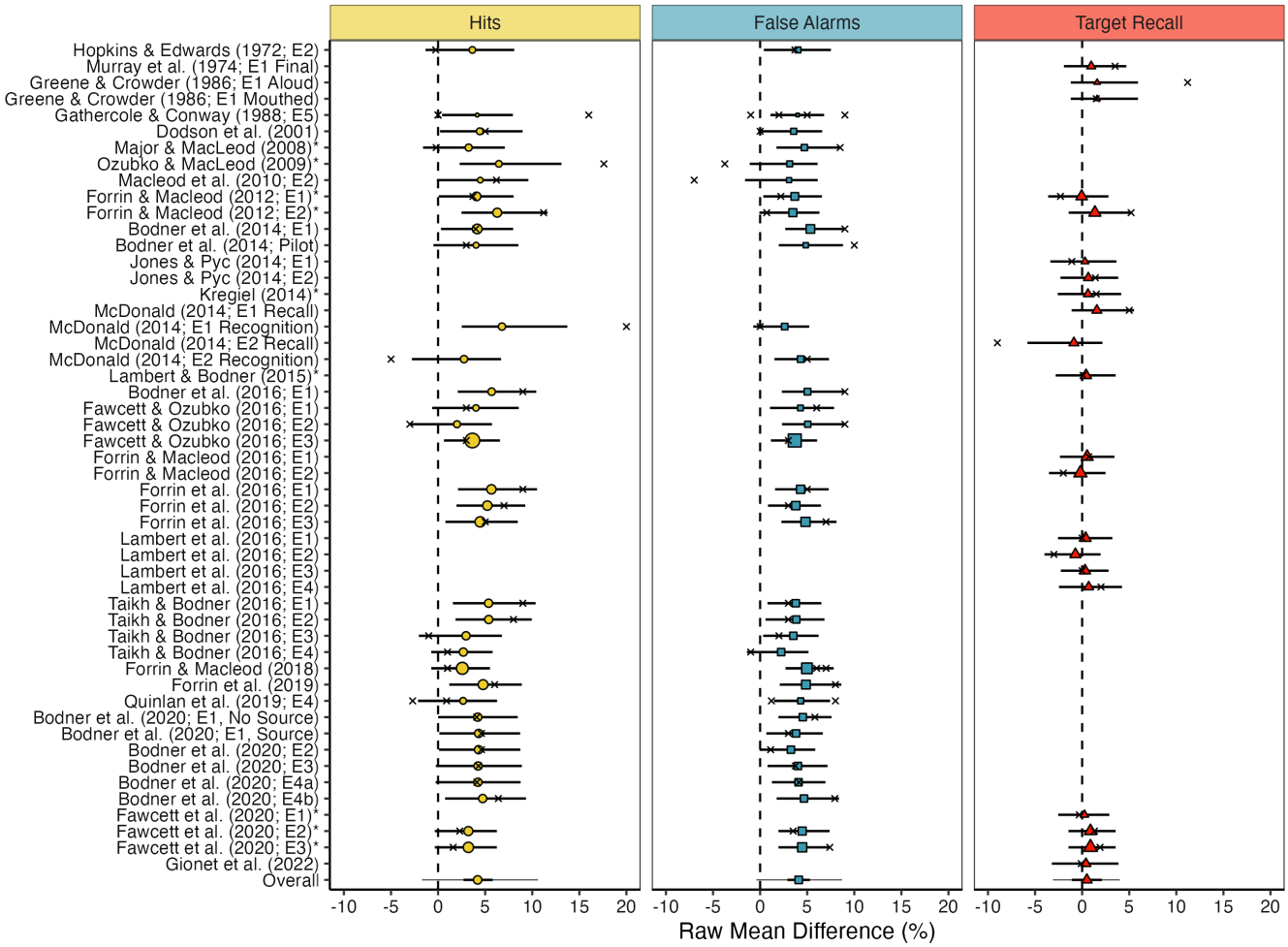
As depicted in Figure 2, an aggregate production effect of similar magnitude was observed for both hits, 4.3%, 95% CI [2.7%, 5.9%], and false alarms, 4.1%, 95% CI [2.9%, 5.3%]. To evaluate claims that the production effect is more reliable in the former than the latter (e.g., Forrin & MacLeod, 2018), an exploratory model was also fit comparing the production effect for each measure. This analysis revealed a difference of only 0.2%, 95% CI [−1.7%, 2.2%], suggesting that this claim is currently without empirical basis. Both models demonstrated some evidence of between-study variability, with prediction intervals ranging from −1.7% to 10.5% for hits and from −0.6% to 8.7% for false alarms. These intervals reflect the range of plausible "true" effects one might expect in a new study and estimate with similar methods (e.g., after removing sampling error), meaning that current evidence allows that some studies might demonstrate roughly no effect in one measure or the other. However, foreshadowing our next analysis, we believe it probable that this reflects a trade-off such that the effect is spread between these measures and is captured only imperfectly in either; in fact, this is the typical argument favouring the use of signal detection metrics such as d'. Even so, the occurrence of a "true" effect equal to or less than 0% is expected in only ~5% of studies, suggesting that most *Null* findings are likely Type II errors.

Sensitivity

As depicted in Figure 3, our analysis of d' produced a similar pattern, with an aggregate effect of 0.34, 95% CI [0.27, 0.40].

Figure 2

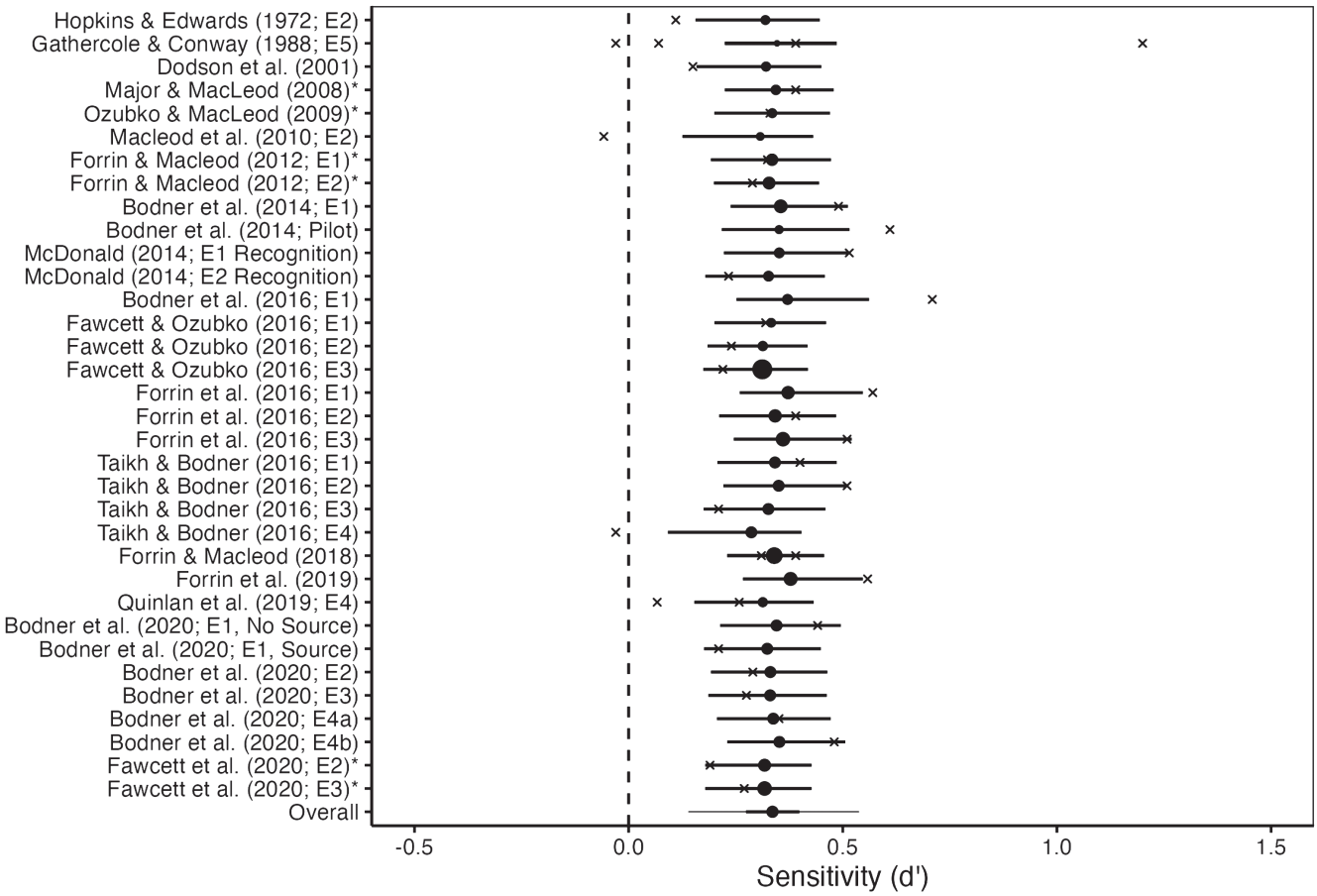
Raw Mean Differences (%) and Confidence Intervals for Hits (Yellow Circles), False Alarms (Blue Squares) and Target Recall (Red Triangles) Estimated From the Models Reported in Text



Note. The empirical values reported by each study are represented by an "X" and relative sample size is depicted by the size of the marker. The final entry in each column represents the estimated effect in a typical study and the thin line radiating from that point represents the prediction interval. Unpublished work is marked with an asterisk (*). See the online article for the color version of this figure.

Figure 3

Mean Differences and Confidence Intervals for d′ Estimated From the Model Reported in Text



Note. The empirical values reported by each study are represented by an "X" and relative sample size is depicted by the size of the marker. The final entry represents the estimated effect in a typical study and the thin line radiating from that point represents the prediction interval. Unpublished work is marked with an asterisk (*).

Supporting our earlier speculation, the production effect was less variable across studies using d', producing a prediction interval ranging from 0.14 to 0.54. Notably, this interval excludes zero, meaning that all well-powered studies should produce a between-study production effect measured using d'. This supports earlier arguments favouring the use of signal detection measures in this literature (e.g., Fawcett et al., 2012).

Power Analyses

Having replicated Fawcett's (2013) earlier analyses, our next goal was to provide guidance as to the sample size required to uncover such an effect reliably. To do so, we calculated a series of power analyses using the *pwr* package (Champely, 2020) based on effect sizes of $d = 0.30$ and 0.50 ($d \sim 0.30$ and 0.50 for hits/false alarms, and d', respectively). These calculations suggest sample sizes of 176 and 64, respectively, *per group* to ensure 80% power. More typical sample sizes of 18, 24, or 36 per group have statistical power of ($d =$ 0.30/$d = 0.50$) 14%/31%, 17%/40% and 24%/55%, respectively. For comparison, the median sample size in this analysis (∼32 per group) produced statistical power of ∼22%/∼50%.

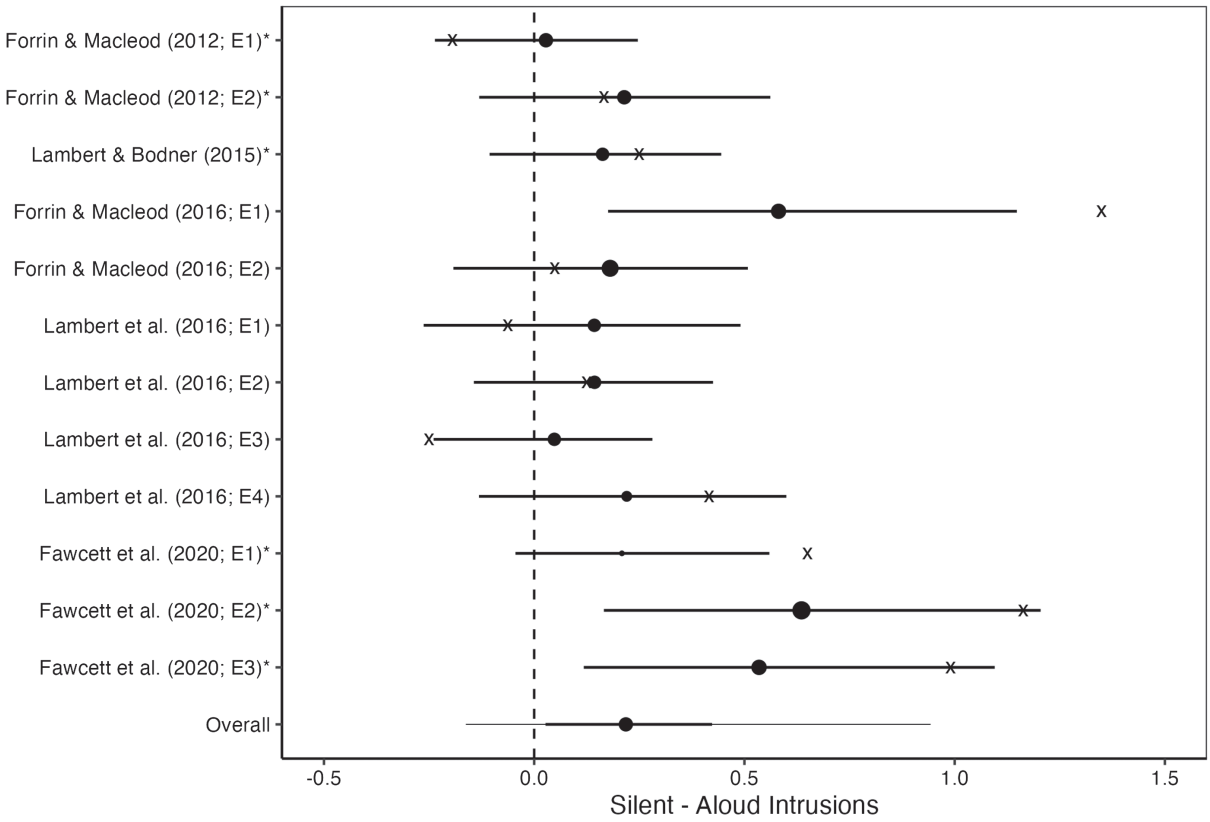
Given the resources required to produce a reliable between-subject effect, it is no surprise that such studies are both underpowered and prone to nonsignificance, even in the modern literature (the median sample size remains 32 in recent years and some studies continue using 20 or less). Our general advice to the field is to cease drawing strong conclusions as to the absence of a between-subject effect based on a nonsignificant statistical test unless the study has been weighed against the now abundant meta-analytic support.

Publication Bias

To evaluate the possibility of publication bias, multilevel regression models were undertaken using the (scaled) standard error and average sample size of each study as moderators in separate models. This approach is comparable to the *regtest* function of the *metafor* package (Viechtbauer, 2010). These analyses were conducted separately for hits, false alarms, and d'. No association was observed between either standard error or sample size and any of our dependent measures.

Figure 4

Back-Transformed Mean Differences and Confidence Intervals for Recall Intrusions Estimated From the Model Reported in Text



Note. The empirical values reported by each study are represented by an "X" and relative sample size is depicted by the size of the marker. The final entry represents the estimated effect in a typical study and the thin line radiating from that point represents the prediction interval. Unpublished work is marked with an asterisk (*).

Recall

Target Recall

Unlike earlier models, the comparison within target recall revealed a difference of only 0.5%, 95% CI [−1.1%, 2.1%], with confidence intervals spread around 0%. Prediction intervals including both positive and negative values, ranging from −4.0% to 5.1%, although much of this range derived from uncertainty (owing to few studies and our inclusion of two random effects) rather than actual variability. This outcome provides empirical support for the notion that the production effect does not benefit target recall, at least in aggregate.[1] As a final model, to verify empirically that the production effect is in fact *smaller* for recall than recognition, we combined the recall data with recognition hits and compared them directly. This analysis revealed a difference of 3.7%, 95% CI [1.6%, 5.7%]. An analysis using a standardized effect size comparing recall to d' demonstrates the same.

Intrusions

As depicted in Figure 4, most studies reporting intrusions appear congruent with the presence of a between-subject effect. A typical study would be expected to show a difference of 0.22, 95% CI [0.03, 0.43]. However, prediction intervals ranged from −0.16 to 0.93, indicating that the "true" magnitude of the effect varies across studies and may be absent in some cases.

Exploratory moderator analyses revealed the effect of production on intrusion ratings to be greatest in studies with a large number of average intrusions, interaction slope = −0.23, 95% CI [−0.43, −0.04], and a trend favouring the effect amongst participants with relatively worse target recall, interaction slope = 0.15, 95% CI [−0.04, 0.29]. However, given the small number of studies investigating this phenomenon, more research is warranted.

Publication Bias

Similar analyses of publication bias were undertaken for the recall data, and once again, no association was observed between either standard error or sample size and any of our dependent measures.

[1] Although our strong preference is in favour of the parameter estimation approach to Bayesian statistics (e.g., Kruschke, 2015), particularly considering the goals of meta-analysis, given the theoretical relevance of our claims pertaining to equivalence, we also calculated a Bayes Factor (BF) by refitting the model setting the aggregate effect to 0 and comparing the two using the *bayes.factor* function. This produced moderate to strong evidence favouring the *Null* model (BF_{01} = 10.6). However, as our analysis of serial position demonstrates below, this is likely in part due to trade-off effects that occur as a function of serial position.

Analyses of Serial Position

Having both established the presence of a production effect for recognition and observed evidence *against* a production effect for recall, our final analyses were focused on evaluating recent claims that the production effect in pure lists (including between-subject designs) interacts with serial position, such that a typical effect emerges for later positions with a reverse effect for earlier positions (e.g., Gionet et al., 2022; Saint-Aubin et al., 2021). This analysis required studies matching our inclusion criteria for which performance was reported as a function of serial position. Including raw data, this resulted in a total of nine recognition and 11 recall effects from seven articles or data sets. As noted earlier, we calculated mean performance for two time points: The initial three items (early) and the final three items (late). The number of time points included in each condition (three) was selected arbitrarily based on where effects are typically observed. These effects were then analyzed using a multilevel model with time (early, late) as a predictor (and as a random slope). This analysis was conducted separately for recognition and recall, after which a new model was fit combining these dependent measures to explore the possibility of an interaction between time and measure.

For the recall data, a reverse production effect was observed for the early time points, $M = -9.0\%$, 95% CI [−17.8, −0.5], with a complementary positive production effect observed for the late time points, 10.3%, 95% CI [1.1, 19.2], with a difference of $M = 19.2\%$, 95% CI [5.4, 33.0]. This pattern largely supports the assertion that production might—at times—interfere with encoding (Gionet et al., 2022; Saint-Aubin et al., 2021; see also, Wakeham-Lewis et al., 2022). We next evaluated whether a similar trend would be observed for d'. Unlike for recall, there was a sizable production effect observed for the early time points, $M = 0.34$, 95% CI [0.09, 0.60], although the effect remained numerically larger for the later time points, $M = 0.51$, 95% CI [0.20, 0.80], with a difference of $M = 0.17$, 95% CI [−0.17, 0.50], which failed to exclude zero as a credible value. Our final model compared recall to recognition hits. The pattern observed for recall was unsurprisingly identical to the above, and although for recognition hits there was a trend favouring a larger production effect for later time points, $M = 7.8\%$, 95% CI [1.1, 14.8], as opposed to early time points, $M = 3.1\%$, 95% CI [−2.4, 8.7], with a difference of $M = 4.6$, 95% CI [−4.3, 14.0], time and measure interacted such that this difference was smaller for recognition hits than for recall, slope of the interaction = −14.8%, 95% CI [−25.4, −4.3]. The same pattern is observed if we instead used standardized effect sizes and compared recall to d'.[2]

Conclusions

With respect to recognition, our findings provide a cautionary note concerning underpowered studies: Because the effect is small, future studies must either ensure a sufficiently large sample, or resist interpreting the absence of a significant difference as evidence favouring the absence of an underlying effect. To maximize statistical power, we urge researchers to focus on d' rather than hits or false alarms: Whereas researchers have at times largely ascribed the production effect to a reduction in false alarms (e.g., Forrin & MacLeod, 2018), present analyses reveal the effect to be equally robust in hits. Our present view is that the effect is spread between these metrics, meaning that a focus on either may distort the influence of production. This perspective is supported by prediction intervals demonstrating the effect to be variable in either measure, but stable using d'. Although this is not to say the effect is necessarily mechanistically equivalent between hits and false alarms, one theoretical implication is that the distinctiveness heuristic as commonly applied must be expanded to ease its focus on the reduction of false alarms, as there is no empirical basis to claim that they are privileged over hits. Indeed, if anything, this finding favours a view of distinctiveness more aligned with Jamieson et al. (2016), which neither requires strategic intent nor elevates either measure.

Present evidence also supports the conclusion that production does not improve target recall in aggregate but may at times reduce intrusions and benefit the final items. The former is surprisingly consistent in Figure 2: To date, there have been very limited reports of a between-subject production effect using recall, and if such an effect existed, we would expect it to be very small based on present evidence. Importantly, whereas some researchers might dismiss this conclusion as general knowledge, it is a conclusion that could not have been drawn with any certainty prior to our meta-analysis. As had been the case with recognition a decade earlier, this conclusion had previously been grounded on a handful of underpowered studies, and only through their meta-analytic synthesis could this theoretically important finding be supported to any great extent.

Perhaps more promising is the potential for production to reduce recall intrusions or benefit memory for the last few items in a list. Here, we resolve a discrepancy within the literature wherein only a single study had previously observed such a reduction for intrusions (Forrin & MacLeod, 2016), with others showing no difference (Lambert et al., 2016). Even so, this finding would appear to be inconsistent across studies. Preliminary evidence suggests the effect tends to emerge under circumstances where intrusions are common and target recall poor. Within the tradition of generate–recognize models (Anderson & Bower, 1972; Bahrick, 1970), recall begins with the generation of candidate responses, which are then evaluated via a recognition process prior to output. Items lacking in familiarity or failing to match the list-context are subsequently excluded as part of the latter process; it is possible that the effect of production on familiarity (e.g., Fawcett & Ozubko, 2016) might have been used to filter intrusions at this stage. This proposal has yet to be evaluated and represents an excellent target for future investigation.

Likewise, we provide initial meta-analytic evidence that the production effect interacts with serial position, giving rise to a reverse production effect for early items and a typical production effect for late items in recall, as predicted by Saint-Aubin et al. (2021) and Gionet et al. (2022; who also provide a nonempirical review). The aforementioned authors have adopted the revised feature model (RFM; Saint-Aubin et al., 2021), based on Nairne's (1990) feature model of immediate memory, to explain this interaction. Broadly, this framework contends that the probability of retrieving studied items depends on two types of features associated with the item: Modality-independent features, which relate to intrinsic phonological and semantic categorization processes, and modality-dependent

[2] An exploratory analysis of the "middle" time points revealed a production effect for d', $M = 0.36$, 95% CI [0.22, 0.50], but not recall, $M = 0.2\%$, 95% CI [−3.0, 3.2]. A corollary of the proposition that the production effect emerges only for—or is larger at—later timepoints is that the magnitude of the production effect ought to be larger for shorter lists, where these last few items hold greater sway. Although the number of study items failed to credibly predict the magnitude of the production effect for either d' or recall, the data trended in this direction in either case, with directional Bayesian p-values of .95 and .90, respectively.

features, which relate to characteristics of the specific presentation modality. Features can be overwritten only by interference from similar features of the same type, and overwritten features can be restored through rehearsal (for this addition, see Cyr et al., 2021; Saint-Aubin et al., 2021). According to Saint-Aubin et al. (2021), production disrupts rehearsal, interfering with the maintenance of features for early items and hindering retrieval. Conversely, late items are primarily subject only to modality-independent interference, leaving a larger number of modality-dependent features intact relative to early items. Because production is thought to encode more modality-dependent features relative to silent reading, produced items occurring late in a list should be better retrieved than unproduced items despite disruption of rehearsal. The results of our analyses, then, appear to validate predictions of the RFM. Although other authors have failed to observe a between-subject production effect in recall (e.g., Jones & Pyc, 2014), our findings lend support to the hypothesis that the aggregate effect is mediated by opposing effects at early and late list positions, rather than wholly absent (Gionet et al., 2022; Saint-Aubin et al., 2021).

We observed a different pattern of results for recognition memory: Whilst the production effect was numerically larger for later items relative to early items, we observed a typical benefit for the early items as well. Thus far, the RFM has been applied only to the production effect in recall; predictions regarding interactions between the effect and serial position in recognition have yet to be made. Because recognition paradigms qualitatively differ from recall paradigms, it is unclear if our findings can be interpreted as congruent with the RFM as it presently exists. Whilst it has been established that the predicted interaction between production and serial position persists in long-term memory tasks (Cyr et al., 2021; Gionet et al., 2022), it is yet to be determined how other recognition-specific task elements (e.g., random presentation order of items at test) might impact this interaction. Presently, we can conclude only that the between-subject production effect in recognition appears to be marginally more pronounced for late items relative to early items.

With respect to other theoretical perspectives, our present findings might also be interpreted through the lens of either the dual-process (Fawcett & Ozubko, 2016) or item-order account (Jonker et al., 2014). With respect to the former, the production effect has been shown to derive from both familiarity and recollection in within-subject designs, but only from familiarity in between-subject designs. Given neural and behavioural evidence (e.g., Okada et al., 2012; Quamme et al., 2004; Yonelinas, 2002) suggesting that recollection and recall are reliant on a common underlying mechanism dissociable from familiarity, this account would predict the lack of a between-subject effect for recall. It might also explain the reduction in intrusions as described above, with familiarity serving as a filter. However, it would not specifically predict the interaction with serial position. With respect to the item-order account, although aloud items are thought to benefit from greater item-specific encoding, silent items have been shown to benefit from superior relational encoding, resulting in a typical production effect for recognition but equivalent performance for recall (Forrin & MacLeod, 2016; Jones & Pyc, 2014; Jonker et al., 2014). It is less clear how the latter account would explain the reduction in intrusions, although enhanced item-specific encoding for aloud items could perhaps provide a similar filter. Of the two, we presently prefer the dual-process interpretation. Although the item-order account provides an elegant mechanistic explanation for the absence of a between-subject effect in recall, its core assumptions have recently been challenged on empirical grounds (Cyr et al., 2021; Saint-Aubin et al., 2021) and it cannot explain the absence of the effect in recollection: Recollection judgments offered in the context of a recognition task should not depend strongly on item-order information, since the tested items are provided, making generation via relational associations unnecessary. The explanation provided by the dual-process account naturally addresses either finding, although it remains to be determined how such an account might contend with the serial order effects.

In conclusion, our findings provide a novel empirical basis for theoretical claims made in the broader literature, whilst also establishing the role of production in reducing off-list intrusions during recall. The effect of production on recognition memory has also been updated (and shown to be robust in both hits and false alarms), and guidance has been provided concerning appropriate sample sizes for future studies.

Résumé

L'effet de production fait référence au fait que les mots lus à haute voix sont mieux mémorisés que ceux lus en silence. Cette tendance a le plus souvent été expliquée comme étant le résultat de l'incorporation d'éléments sensorimoteurs dans la représentation de l'item lors de l'étude, qui pourrait ensuite être utilisée pour guider la performance lors d'un test ultérieur. Ce cadre théorique considère que les éléments lus à haute voix sont distincts des éléments silencieux, et l'on pensait donc que l'effet n'apparaissait que lorsque la production était manipulée au sein des sujets. Cette affirmation a été contestée par la suite, et un effet de production entre sujets fiable (bien que plus petit) a depuis été démontré dans la mémoire de reconnaissance. À travers une série de méta-analyses, nous élargissons ces travaux antérieurs, en reproduisant l'effet de production entre sujets pour la reconnaissance, et en démontrant l'absence d'un tel effet pour le rappel global de la cible. Or, à l'appui de récentes affirmations théoriques, nous avons également observé une interaction entre l'effet de production et la position sérielle dans le rappel, de sorte qu'un effet de production a été observé pour les points de temps tardifs, mais pas pour les points de temps précoces (une tendance similaire, bien que plus petite et non crédible, a été observée pour la reconnaissance). Enfin, nous fournissons des éléments probants que la production réduit les intrusions hors liste. En résumé, la production a un impact fiable sur la mémoire de reconnaissance lorsqu'elle est manipulée entre sujets, mais une relation plus complexe avec la performance de rappel.

Mots-clés : effet de production, caractère distinctif, entre sujets, rappel, méta-analyse

References

References marked with an asterisk indicate studies included in the meta-analysis.

Anderson, J. R., & Bower, G. H. (1972). Recognition and retrieval processes in free recall. *Psychological Review*, *79*(2), 97–123. https://doi.org/10.1037/h0033773

Bahrick, H. P. (1970). Two-phase model for prompted recall. *Psychological Review*, *77*(3), 215–222. https://doi.org/10.1037/h0029099

*Bodner, G. E., Huff, M. J., & Taikh, A. (2020). Pure-list production improves item recognition and sometimes also improves source memory. *Memory & Cognition*, *48*(7), 1281–1294. https://doi.org/10.3758/s13421-020-01044-2

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

*Bodner, G. E., Taikh, A., & Fawcett, J. M. (2014). Assessing the costs and benefits of production in recognition. *Psychonomic Bulletin & Review*, *21*(1), 149–154. https://doi.org/10.3758/s13423-013-0485-1

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411. https://doi.org/10.32614/RJ-2018-017

Cameron, A. C., & Trivedi, P. K. (1998). *Regression analysis of count data*. Cambridge Press. https://doi.org/10.1017/CBO9780511814365

Champely, S. (2020). *pwr: Basic functions for power analysis*. R package Version 1.3-0. https://CRAN.R-project.org/package=pwr

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

Cyr, V., Poirier, M., Yearsley, J. M., Guitard, D., Harrigan, I., & Saint-Aubin, J. (2021). The production effect over the long term: Modeling distinctiveness using serial positions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. https://doi.org/10.1037/xlm0001093

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

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

Fawcett, J. M., & Hulbert, J. C. (2020). The many faces of forgetting: Toward a constructive view of forgetting in everyday life. *Journal of Applied Research in Memory and Cognition*, *9*(1), 1–18. https://doi.org/10.1016/j.jarmac.2019.11.002

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

Fawcett, J., Baldwin, M., Whitridge, J., Swab, M., Malayang, K., Hiscock, B., Drakes, D., & Willoughby, H. (2022, October 18). *Production improves recognition and reduces intrusions in between-subject designs: An updated meta-analysis*. http://osf.io/rsu6w

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. https://doi.org/10.1080/09658211.2012.693510

*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. https://doi.org/10.1037/xlm0000214

*Forrin, N. D., & MacLeod, C. M. (2012). *The costs and benefits of production* [Unpublished raw data].

*Forrin, N. D., & MacLeod, C. M. (2016). Order information is used to guide recall of long lists: Further evidence for the item-order account. *Canadian Journal of Experimental Psychology*, *70*(2), 125–138. https://doi.org/10.1037/cep0000088

*Forrin, N. D., & MacLeod, C. M. (2018). Cross-modality translations improve recognition by reducing false alarms. *Memory*, *26*(1), 53–58. https://doi.org/10.1080/09658211.2017.1321129

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

*Forrin, N. D., Ralph, B. C. W., Dhaliwal, N. K., Smilek, D., & MacLeod, C. M. (2019). Wait for it … performance anticipation reduces recognition memory. *Journal of Memory and Language*, *109*, Article 104050. https://doi.org/10.1016/j.jml.2019.104050

*Gathercole, S. E., & Conway, M. A. (1988). Exploring long-term modality effects: Vocalization leads to best retention. *Memory & Cognition*, *16*(2), 110–119. https://doi.org/10.3758/BF03213478

*Gionet, S., Guitard, D., & Saint-Aubin, J. (2022). The production effect interacts with serial positions: Further evidence from a between-subjects manipulation. *Experimental Psychology*, *69*(1), 12–22. https://doi.org/10.1027/1618-3169/a000540

*Greene, R. L., & Crowder, R. G. (1986). Recency effects in delayed recall of mouthed stimuli. *Memory & Cognition*, *14*(4), 355–360. https://doi.org/10.3758/BF03202514

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*(5), 1157–1164. https://doi.org/10.3758/s13423-013-0572-3

*Hopkins, R. H., & Edwards, R. E. (1972). Pronunciation effects in recognition memory. *Journal of Verbal Learning and Verbal Behavior*, *11*(4), 534–537. https://doi.org/10.1016/S0022-5371(72)80036-7

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

*Jones, A. C., & Pyc, M. A. (2014). The production effect: Costs and benefits in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(1), 300–305. https://doi.org/10.1037/a0033337

Jonker, T. R., Levene, M., & Macleod, C. M. (2014). Testing the item-order account of design effects using the production effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(2), 441–448. https://doi.org/10.1037/a0034977

*Kregiel, B. (2014). *The production effect and item-order encoding* [Unpublished honours thesis, John Carroll University]. http://collected.jcu.edu/honorspapers/38

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, Jags, and Stan*. Academic Press.

*Lambert, A., & Bodner, G. (2015). *Between-subjects production effect in categorized lists* [Unpublished raw data].

*Lambert, A. M., Bodner, G. E., & Taikh, A. (2016). The production effect in long-list recall: In no particular order? *Canadian Journal of Experimental Psychology*, *70*(2), 165–176. https://doi.org/10.1037/cep0000086

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

Maddox, G. (2019). *Do drawings speak louder than words? Comparing the effects of production and drawing on memory for concrete and abstract words*. Senior Honours Project.

*Major, J. C., & MacLeod, C. M. (2008). *The production effect: Strength and distinctiveness* [Unpublished raw data].

*McDonald, D. L. L. (2014). *Proportional manipulation of produced words tests the distinctiveness and strength accounts of the production effect* [Unpublished honours thesis, Kwantlen Polytechnic University]. https://kora.kpu.ca/islandora/object/kora:23/datastream/PDF/view

McElreath, R. (2020). *Statistical rethinking*. CRC Press. https://doi.org/10.1201/9780429029608

*Murray, D. J., Leung, C., & McVie, D. F. (1974). Vocalization, primary memory and secondary memory. *British Journal of Psychology*, *65*(3), 403–413. https://doi.org/10.1111/j.2044-8295.1974.tb01414.x

Nairne, J. S. (1990). A feature model of immediate memory. *Memory & Cognition*, *18*(3), 251–269. https://doi.org/10.3758/BF03213879

Okada, K., Vilberg, K. L., & Rugg, M. D. (2012). Comparison of the neural correlates of retrieval success in tests of cued recall and recognition memory. *Human Brain Mapping*, *33*(3), 523–533. https://doi.org/10.1002/hbm.21229

Ozubko, J. D., & Macleod, C. M. (2010). The production effect in memory: Evidence that distinctiveness underlies the benefit. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(6), 1543–1547. https://doi.org/10.1037/a0020604

*Ozubko, J. D., & MacLeod, C. M. (2009). *Recognition test performance in production* [Unpublished raw data].

Quamme, J. R., Yonelinas, A. P., Widaman, K. F., Kroll, N. E. A., & Sauvé, M. J. (2004). Recall and recognition in mild hypoxia: Using covariance structural modeling to test competing theories of explicit memory. *Neuropsychologia*, *42*(5), 672–691. https://doi.org/10.1016/j.neuropsychologia.2003.09.008

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

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

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

Saint-Aubin, J., Yearsley, J. M., Poirier, M., Cyr, V., & Guitard, D. (2021). A model of the production effect over the short-term: The cost of relative distinctiveness. *Journal of Memory and Language*, *118*, Article 104219. https://doi.org/10.1016/j.jml.2021.104219

*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. https://doi.org/10.1037/cep0000083

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Wakeham-Lewis, R. M., Ozubko, J., & Fawcett, J. M. (2022). Characterizing production: The production effect is eliminated for unusual voices unless they are frequent at study. *Memory*, *30*(10), 1319–1333. https://doi.org/10.1080/09658211.2022.2115075

Wammes, J. D., Meade, M. E., & Fernandes, M. A. (2016). The drawing effect: Evidence for reliable and robust memory benefits in free recall. *Quarterly Journal of Experimental Psychology*, *69*(9), 1752–1776. https://doi.org/10.1080/17470218.2015.1094494

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

Received May 16, 2021
Revision received October 19, 2022
Accepted October 20, 2022 ■