Psychonomic Bulletin & Review (2026) 33:54
https://doi.org/10.3758/s13423-025-02847-7

BRIEF REPORT

Preparing to produce (without production) is sufficient to elicit a behavioral and pupillometric production effect

Jonathan M. Fawcett[1] · **Brady R. T. Roberts**[2,3] · **Siyue Hu**[1] · **Cayley Thoms**[1] · **Jedidiah Whitridge**[1] · **Colin M. MacLeod**[4] · **Hannah Willoughby**[5]

Received: 15 September 2025 / Accepted: 12 December 2025

Abstract

The production effect refers to the finding that words read aloud are remembered better than words read silently. Historically, this phenomenon has been explained with reference to distinctive features encoded at study (e.g., auditory and motor elements) being retrieved at test to discriminate between studied and unstudied items, with emphasis placed on features stemming from the act of production itself. Across two experiments, we demonstrate that even anticipation of reading a word aloud is sufficient to improve its memory over silent items. Using a recent variant of the production paradigm involving pupillometry, participants were instructed to withhold their response until a "Go" signal appeared. On "catch" trials this signal never occurred. Despite having not produced the word on a catch trial, participants nonetheless demonstrated both a behavioral (Experiments 1 and 2) and a pupillary (Experiment 2) production effect, although both were of lesser magnitude than on trials requiring actual production. For "Go" trials, the behavioral production effect was evident for both recollection and familiarity; for "catch" trials, the effect was evident only for recollection. These results support recent claims that motivational or attentional factors play a role in the emergence of the production effect, connecting this effect to a broader framework of action-oriented memory enhancement.

Keywords Production effect · Pupillometry · Preparation · Distinctiveness · Attention · Memory

Introduction

The production effect (MacLeod et al., 2010) refers to the well-replicated finding that words read aloud are better remembered than words read silently. This phenomenon has been studied for over six decades (Conway & Gathercole, 1987, 1988); Ekstrand et al., 1966; Hopkins & Edwards, 1972; Murray, 1965), with production known for over a century in popular culture as an effective mnemonic strategy (e.g., Gates, 1917; Herndon & Weik, 1896). Although most research has focused on reading aloud, the effect has been generalized to other modalities including writing (e.g., Forrin et al., 2012), mouthing (e.g., Fawcett et al., 2012; MacLeod et al., 2010), and singing (Quinlan & Taylor, 2013; for a meta-analysis, see Whitridge et al., 2024), and across the lifespan from children (e.g., Icht & Mama, 2015; Pritchard et al., 2020) to older adults (e.g., Lin & MacLeod, 2012).

Since early in the history of the production effect (e.g., Conway & Gathercole, 1987, 1988; Hopkins & Edwards, 1972; Gathercole & Conway, 1988), theoretical accounts have attributed the memory benefit elicited by reading aloud to distinctive encoding (for more on distinctiveness, see Hunt, 2006, 2013). It has been argued that at study a record of production features is stored with each produced item (e.g., MacLeod et al., 2010) and is then used at test to facilitate recall or recognition of aloud items which "stand out" against the backdrop of silent items lacking these features.

Two prominent variants of the distinctiveness account exist, differing largely with respect to the mechanism

✉ Jonathan M. Fawcett
jfawcett@mun.ca

[1] Department of Psychology, Memorial University of Newfoundland, St. John's, NL A1C 5S7, Canada

[2] Department of Psychology, University of Chicago, Chicago, IL 60605, USA

[3] Institute for Mind and Biology, University of Chicago, Chicago, IL 60637, USA

[4] Department of Psychology, University of Waterloo, Waterloo, ON N2L 3G1, Canada

[5] C3 Human Factors Inc., St. John's, NL, Canada

Published online: 21 January 2026

through which the sensorimotor features associated with production (i.e., the *production record*; Fawcett et al., 2012; Fawcett, 2013) are thought to aid memory. The *distinctiveness heuristic account* (e.g., Dodson & Schacter, 2001) argues that participants evaluate whether they believe that they recently said each test word aloud, the logic being that words said aloud recently must have been studied; the *relative distinctiveness account* argues instead that studied words with an associated production record simply "pop out" during recall or recognition due to normal retrieval dynamics (for computational models, see Caplan & Guitard, 2024; Jamieson et al., 2016; Kelly et al., 2022; Saint-Aubin et al., 2021; Wakeham-Lewis et al., 2022). A third form is *statistical distinctiveness* (Icht et al., 2014), which posits that the production effect can also arise from the relative rarity of produced items within a study list. In many experiments, however, including those reported here, aloud and silent items are equally frequent, minimizing this statistical contribution.

Although there is now substantial evidence favoring the role of distinctive encoding, there have been suggestions that motivational or attentional factors also play a role in the production effect (e.g., Bodner et al., 2014; Fawcett, 2013; MacDonald & MacLeod, 1998; Ozubko et al., 2012). Knowing that an overt response is required on a given trial could encourage participants to remain more alert and on-task than they would be for a trial not requiring an overt response. Thus, production may be akin to a desirable difficulty (e.g., Bjork, 1994; see also Bjork & Bjork, 2020), forcing greater engagement and improving later memory (Fawcett et al., 2025; Hourihan & Fawcett, 2024). In debriefing, participants report that they were more attentive during aloud than during silent trials (Fawcett & Ozubko, 2016). Smaller or absent production effects have also been observed when attention is impaired through distraction (e.g., Mama et al., 2018) or a psychiatric condition (e.g., attention-deficit/hyperactivity disorder; Mama & Icht, 2019).

Recently, Fawcett et al. (2025) used pupillometry to demonstrate that participants disengage soon after word onset during silent trials but remain attentive during aloud trials. Pupillometry – measuring changes in pupil size during performance of a cognitive task to gauge attention or processing load – is thought to reflect phasic firing of the locus coeruleus-norepinephrine system (Hess & Polt, 1964; Papesh et al., 2012; Sirois et al., 2014). Larger pupil size reflects increased activation of the norepinephrine system and thus greater engagement with the task. Fawcett et al. (2025) found reading aloud to be associated with larger pupils than reading silently during study. Critically, this *pupillometric production effect* was dissociable from (and partially preceded) the act of production itself. In fact, pupil size diverged between aloud and silent trials upon instruction onset even when the target word was not presented until later in the trial, with actual production inducing a larger, positive deflection superimposed atop this smaller, preparatory deflection. Fawcett et al. interpreted their findings as evidence that preparing an overt response elicits attention focusing that facilitates (or is even a necessary component of) subsequent distinctive encoding of target word features preceding and/or during production.

Whereas Fawcett et al. (2025) manipulated the relative timing of word, instruction, and response onset to facilitate dissociation of the underlying processes, they could not separate these mechanisms fully because each event occurred on every trial (at different times). To separate the unique contributions of the hypothesized attentional "boost" underlying response preparation, apart from the distinctive encoding thought to surround actual production, a condition would be needed wherein participants believed that they would produce a word but never actually did. The present experiments achieved this by modifying the "delayed" production paradigm (e.g., Hassall et al., 2016) used by Fawcett et al. (2025, Experiment 4). In this paradigm, participants see a target word and production indication at trial onset but are instructed to withhold responding until a "Go" signal appears. In the present experiments, we modified this paradigm such that the "Go" signal sometimes did not appear: In these "catch" trials, participants were instructed to do nothing while awaiting the following trial, permitting us to isolate and test the influence of preparation itself on later memory.

We report two identical experiments differing only in that Experiment 1 was purely behavioral whereas Experiment 2 measured pupil size concurrent to the behavioral task. Presuming that response preparation elicits heightened attentional engagement, we expected a pupillometric production effect even during catch trials. Insofar as this attentional component is itself related to the memory benefit observed in a typical production paradigm, we further predicted a behavioral production effect for catch trials despite no overt verbal production.

Method

Participants

Experiments 1 and 2 recruited 32 and 42 participants, respectively. Sample sizes were based on the number of participants that could be tested in a single academic term and two academic terms, respectively, with no upper limit. The sample size for Experiment 2 is comparable to that of Fawcett et al. (2025; Experiment 4). Experiment 1 was completed in Fall 2019; Experiment 2 was partially conducted in Winter 2020 (cut short by the global COVID-19 'students (mostly female, average age ~20–22

years) recruited via the local Psychology Research Experience Pool in exchange for partial course credit. The procedures for these experiments received approval from the Interdisciplinary Committee on Ethics in Human Research at <location redacted for anonymity> (Protocol #20171484-SC).

Materials

The two experiments used identical materials and experiment contexts, with the exception that the eye-tracker, although present, was not activated during Experiment 1. Materials were based on Experiment 4 of Fawcett et al. (2025), with the exception that additional words were needed, and the paradigm was reprogrammed using *OpenSesame* (Version 3.2.5; Mathôt et al., 2011). The experiment was conducted on a MacMini computer running OSX 10.12 and displayed on a 22-in. BenQ monitor at a resolution of 1,024 × 768 pixels. All text was presented in uppercase black, 18-point Courier font against a gray (#808080) background.

As with Fawcett et al. (2025), instructions to read a given word aloud or silently were depicted by a border surrounding each word made up of either +s (read aloud) or Xs (read silently). Two additional borders were created: a post-word waiting screen (a border of squares), and the "Go" signal indicating that a response should be made (a border of equal signs). In all cases, the symbols making up the border were created using four black lines of equal size arranged in different orientations, thereby matching for luminance.

The 225 words used in the experiment were selected from the MRC Psycholinguistic Database (1988). Words were randomly selected and were grouped into nine lists of 25 words. These were counterbalanced across our production (silent, aloud) and delay (early, late, catch) conditions as well as the foil condition (the latter of which contained three lists). Lists were matched for word length (M = 5.47, SD = 1.26) and written word frequency (M = 46.03, SD = 40.16) according to Kucera and Francis (1967).

During Experiment 2, the right pupil of each participant was recorded at either 500 or 1,000 Hz using an Eyelink 1000 Plus eye-tracker (SR Research, Mississauga, Canada) placed beneath the monitor in a desk-mounted configuration, with the participant's head in a stabilizing chin rest to minimize movement. We had intended to record at 500 Hz exclusively, but a technical error resulted in 1,000 Hz for some participants. This was accounted for in our processing scripts, and all data were down-sampled to 50 Hz prior to analysis.

Procedure

Experiments 1 and 2 were identical except that in Experiment 1 participants were able to move their heads freely (i.e., without a chin rest) and no eye-tracking calibration or validation protocols were undertaken. Each phase of Experiment 2 instead began with a standard 13-point calibration and validation protocol, repeated as necessary, and participant heads were stabilized using a chin rest to minimize motion artifacts.

During an initial familiarization phase, participants completed trials in which each production instruction (a border of +s or Xs) was presented simultaneously with the associated encoding task instructions ("read the word aloud"; "read the word silently"). Participants also viewed the "Go" signal (a border of =s), with the understanding that on aloud trials they should withhold their response until this border appeared. Participants were told that not every trial would contain a "Go" signal: When this signal did not appear, they were to do nothing until the next trial began automatically.

Participants completed 18 practice trials split evenly over six different combinations of production (aloud, silent) and delay (early, late, catch). Words presented in this phase were not tested; all timings and trial events were otherwise identical to the study phase. During the actual study phase that followed, participants completed 150 trials divided evenly over the same six intermixed conditions (25 each). Participants were instructed to remember all non-practice words for a later memory test. As depicted in Fig. 1, each trial began with a designated blink period demarcated by a 500-ms blank screen, a 1,000-ms display of "!!!", and another 500-ms blank screen. Participants were instructed to do their best to blink only during this period. After the blink period, a fixation (i.e., a black dot) appeared for 500 ms followed by a word for 3,000 ms surrounded by the appropriate border as determined by condition. Participants then saw an additional fixation screen that varied in duration between 500 ms (early trials), 1,500 ms (late trials), and 3,500 ms (catch trials), followed on some trials by a "Go" signal that remained on the screen for either 3,000 ms (early trials) or 2,000 ms (late trials). Together, the duration of the post-word fixation and the "Go" signal always summed to 3,500 ms. Each trial ended with a borderless fixation (i.e., a dot) for 500 ms. During aloud trials, participants were to say the target word aloud when the "Go" signal appeared; during silent trials, participants were to instead say the target word silently in their head when the "Go" signal appeared. When the "Go" signal did not appear (i.e., a fixation dot remained on screen for the entire 3,500 ms), participants were to do nothing until the following trial.

During the recognition test that followed, trials began with a 500-ms blank screen, a 1,000-ms display of "!!!", and another 500-ms blank screen, again permitting a designated blink period. This was followed by a 500-ms fixation (i.e., dot) screen and a word for 3,000 ms. Next, the word was replaced by a 6-point rating scale with the options: (1) very sure new, (2) mostly sure new, (3) unsure new, (4) unsure

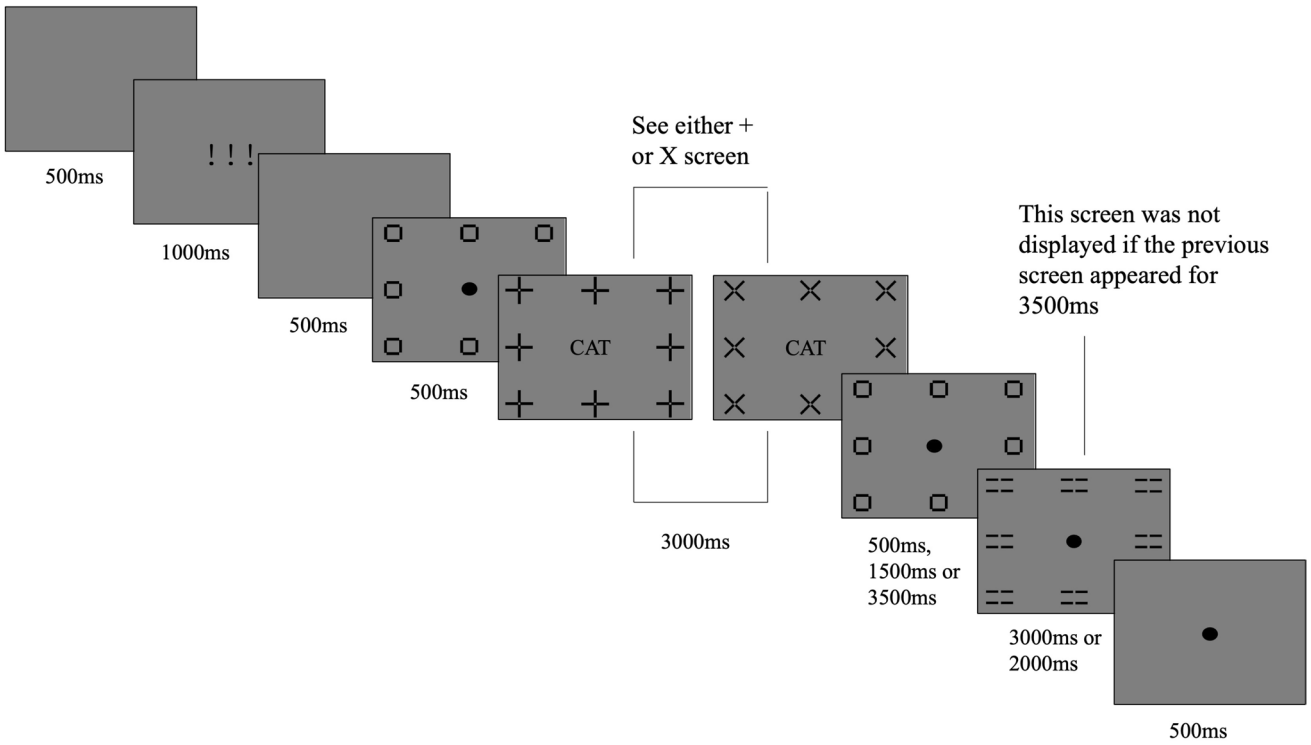

Fig. 1 Schematic representation of study trial events and event timings. Responses were withheld until the "Go" signal (a border made of "=") appeared. See text for further details

old, (5) mostly sure old, and (6) very sure old. Participants pressed the appropriate key and were further instructed that, should they need to blink, they could do so but that they should wait until the confidence scale was presented or do so during the designated period. Responding was self-paced.

Signal processing and statistical approach

Behavioral analyses and correlations. Our statistical approach followed Fawcett et al. (2025), with some exceptions. Recognition data were analyzed as raw confidence ratings using multilevel Bayesian implementations of the unequal variance signal detection (for *d'*) and dual-process signal detection models (for recollection and familiarity; see Fawcett & Ozubko, 2016, or Fawcett et al., 2022, for a discussion), each implemented using the Stan programming language (Carpenter et al., 2017), and was fit using the *cmdstanr* package (v. 0.8.0; Gabry et al. 2025) within *R 4.4.0* (R Core Team, 2024). Given the centrality of these signal detection models, we briefly describe some theoretical background pertaining to both models below. Further details, along with supplementary models, are provided in the Online Supplementary Material (OSM).

Conventional approaches to signal detection analysis of recognition data typically involve the calculation of the criterion (*c*; i.e., the threshold of evidence needed for an "old" response) and discriminability (*d'*; i.e., the distance between the latent evidence distributions for targets and lures) by respectively averaging and taking the difference of the *z*-transformed hit and false alarm rates (Stanislaw & Todorov, 1999). An alternative approach is to estimate the parameters of interest using a generalized linear model (DeCarlo, 1998) that allows for quantification of uncertainty in parameter estimates and integration of group-level effects to account for heterogeneity in clustering variables (e.g., subjects or items; Rouder et al., 2007); we opted to use the latter approach. For a typical recognition paradigm in which participants provide binary old/new responses at test, a binomial model that produces estimates for *d'* and *c* can be fit to the data. When rating data are available, however, it is possible to fit ordinal extensions to the binary signal detection model. In addition to discriminability, these extensions permit estimation of a set of criteria ($K-1$ threshold parameters for a rating task with K categories) as well as additional parameters that cannot be estimated from binary responses due to identifiability constraints. Below, we briefly review several ordinal signal detection models that we fit to our behavioral data.

The *equal variance signal detection model* (EVSD) for rating data is the simplest ordinal extension to the binary

signal detection model (see, e.g., DeCarlo, 1998, 2010). This model estimates discriminability alongside the threshold structure described above. Notably, this model assumes equal variance (conventionally set to unity) across the latent target and lure distributions. Put differently, such a model assumes that the underlying "strength" of memory activation elicited by targets and lures varies to a similar degree and exhibits only a mean difference between these conditions. Contrary to this assumption, it has frequently been shown that memory strength is more variable for targets than for lures; this is supported implicitly through analyses of receiver-operating characteristic (ROC) curves (e.g., Swets, 1986) and directly by analyzing judgments of memory strength for targets and lures (e.g., Mickes et al., 2007). Thus, the *unequal variance signal detection model* (UVSD; Green & Swets, 1966) extends the EVSD by allowing the standard deviation of the target distribution to vary freely (while setting the standard deviation of the lure distribution to 1). We opted to report the UVSD here because it typically provides a superior fit to the data. Moreover, when targets and lures do share the same variability, the UVSD is equivalent to the EVSD. For completeness, the EVSD (and other ordinal models) are reported in the OSM.

Another extension to the EVSD derives from dual process theory. Broadly, this theory contends that recognition is driven by two processes: (a) familiarity, a nonspecific feeling of knowing a stimulus, and (b) recollection, a specific and vivid re-experiencing of the stimulus (for reviews, see Wixted, 2007; Yonelinas, 2002). Competing schools of thought exist within the realm of dual process theory: Some authors view recollection as a continuous signal (e.g., Wixted & Mickes, 2010), whereas others view recollection as an all-or-none process, wherein the stimulus is either recollected or not (e.g., Yonelinas, 1996). The *dual process signal detection model* (DPSD; Yonelinas, 1994) employed here conforms to the assumptions of the latter perspective. This variant of the DPSD assumes that responses derive from one of two discrete states (implemented statistically as a mixture model): For target items, recollection occurs with probability R. When recollection occurs, participants will always choose the maximum confidence response. If recollection does not occur (with probability $1 - R$), the rating given is based on familiarity and follows a standard cumulative probit likelihood (like the EVSD described earlier). This model is reported in text to further evaluate the degree to which the production effect is driven by recollection or familiarity, as has been investigated in earlier work (e.g., Fawcett & Ozubko, 2016; Whitridge et al., 2024).

In all cases, mildly informative Bayesian priors comparable to those from that earlier work were used, and our models included random slopes and intercepts for participants as appropriate (we excluded random intercepts and slopes for item owing to the computational demands of our new approach, and to maintain consistency with our pupillary analyses; see below). Bayesian correlations were fit using *brms* after evaluating for outliers using both the minimum covariance determinant estimator (MCD; Fauconnier & Haesbroeck, 2009; Hubert & Debruyne, 2009; Rousseeuw, 1984; Rousseeuw & Van Driessen, 1999) implemented within the *Routliers* package (Delacre & Klein, 2019) and the more modern isolation forest method using the *isotree* package (Liu et al., 2008; Cortes, 2025). Only participants categorized as outliers by both techniques were dropped.

Because the present work focused on pupil size during the study phase response period, pupil-behavior correlations were calculated using a pupillary production effect measured as the average of a window starting at 3,500 ms (post word onset) and extending to the end of the trial. However, similar conclusions were reached using the entire trial. Correlation models were fit twice, once using regularizing priors (which assume no effect and pull the correlation toward 0), and once using priors based on the analyses reported by Fawcett et al. (2025; combining across experiments), the latter providing a basic meta-analytic estimate combining present with existing data, although further experiments are needed (Gelman et al., 2021).

Statistical rationale

Bayesian multilevel models offer important advantages over traditional ANOVAs. One key benefit is that Bayesian estimation permits informative-yet-unoffensive priors that help stabilize parameter estimation and improve convergence, especially in complex designs. These priors act as a form of weak regularization, constraining estimates to plausible ranges without unduly influencing the results. For example, our analyses of *d'* use priors that suggest mean *d'* for a typical participant in each condition ought to fall between −1 and 3 (this corresponds to a normal distribution centered on 1 with a standard deviation of 1). Few cognitive experiments would expect mean performance (for a "typical" participant) to fall outside that range; however, frequentist models are equivalent to adopting uniform priors, viewing an overall mean *d'* of −3 (reflecting a strong bias *against* recognizing studied over unstudied items) to be as likely as an overall mean *d'* of 1 or 2. Importantly, none of our conclusions (beyond the one explicitly using past priors) hinge specifically on our choice of priors, and a wide range of reasonable priors (e.g., suggesting mean *d'* for a typical participant in each condition ought to fall between -2 and 4) produce comparable conclusions. They also remain entirely agnostic as to the magnitude of any key differences (e.g., the production effect), tacitly assuming no difference (and pulling each production effect toward 0). Our code and data are available so that those interested in exploring how different priors might influence our conclusions can do so.

The stabilizing properties of mildly informative priors are especially important in the context of complex models, such as the multilevel models below. In fact, Bayesian methods make it possible to fit multilevel models that are difficult or impossible to estimate reliably using frequentist approaches, particularly when (a) parameters are correlated, (b) data are missing or unevenly distributed, or (c) variance components are very small or large. In the case of our ROC curve models, each participant has few trials per condition relative to what is typical for this type of model in other literatures (e.g., perception). This can result in malformed individual ROC curves (e.g., if by chance a participant did not distribute their numeric responses evenly across conditions or if they tended toward very few false alarms or near perfect hits). A malformed ROC curve (e.g., one partially flattened or inverted) will result in poor estimation of the underlying parameters. For example, within dual-process signal-detection models, recollection is particularly prone to error when estimated via common optimization approaches (as used by frequentist models). This sometimes results in recollection being (incorrectly) set to 0 in cases of poor model fit. These issues are worsened because non-multilevel approaches involve a "two-step" process whereby the ROC curve is fit to each participant separately, with the resulting participant-level parameters fed into later analyses (e.g., ANOVAs). This approach discards uncertainty in the participant-level estimates, tacitly treating those parameters (e.g., the *d'* or recollection value for a given participant) as "known" with absolute certainty, despite having been estimated with varying precision. Multilevel models (especially Bayesian ones) instead propagate that uncertainty and mitigate the influence of malformed individual ROC curves by informing individual participant-level parameters based on group-level parameters.

Finally, our use of Bayesian multilevel models (rather than strictly ANOVAs and *t*-tests) permits us to model our ordinal confidence ratings without assuming homoscedastic or Gaussian errors. Although common, it is well known that the application of ANOVAs to categorical, ordinal, or bounded outcomes can be misleading. For example, Liddel and Kruschke 2018), Dixon 2008), and Jaeger 2008) each have provided simulations showing how the analysis of binomial (e.g., yes-no recognition) or ordinal (e.g., our confidence ratings) data using ANOVAs distorts error terms, biases estimation, and provides misleading inferences. These methodologists recommend adopting generalized multilevel modelling in situations such as ours. There are few disadvantages beyond complexity, and such models frequently are shown to produce better estimates and inference. In short, whereas we certainly do not wish to engage in *methodological imperialism* (insisting that everyone adopt our methods), we believe that the present models reflect best practice (for a related discussion pertaining to Bayesian modelling applied to similar data, see also Fawcett & Ozubko, 2016; Fawcett et al., 2016).

Pre-processing and analysis of pupillometry data. Pupillary data from Experiment 2 were pre-processed using the same pipeline as reported by Fawcett et al. (2025). This process involved: (1) de-blinking via Eyelink's built-in algorithm with 100 ms of padding on either side of each blink; (2) artifact detection (e.g., a near stationary pupillary signal for at least 500 ms or a signal that moved unnaturally fast); (3) identification of gaze too far from center; (4) interpolation (via a piecewise cubic Hermite approach); (5) a 4-Hz low-pass Butterworth filter (to reduce high-frequency noise); (6) epoching (using -200 to 7,000 ms relative to word and instruction onset); (7) subtractive baselining; (8) down-sampling to 50 Hz; and (9) *z*-score normalizing within participant to control for individual differences in response scale and variation in the lighting conditions during testing. Following pre-processing, trials were flagged as "bad" if 30% of the samples were marked as missing for any reason and participants were dropped if they were missing more than 50% of trials on this basis.

The pupil data were analyzed in two ways. First, we used frequentist generalized additive mixed models (GAMM) with thin-plate regression splines in the *mgcv* package (Wood, 2017). Random curves were included for participant by condition and autocorrelation was also modelled (*rho* = 0.9). Differences between aloud and silent waveforms were estimated using the *itsadug* package (van Rij et al., 2022). Next, we conducted an exploratory mass univariate analysis comparing the aloud and silent waveforms for each delay at each time point. Finally, we determined whether a 100-ms moving window surrounding that time point predicted subsequent memory. For the latter, multilevel (logistic) models were used with a random intercept and slope for each participant. We corrected these models for multiple comparisons using cluster-based *p*-values derived by simulating time-series from a hypothetical *Null* distribution for the relevant test statistics. See Fawcett et al. (2025) for further details.

Results

Prior to data processing, three participants were excluded from Experiment 2 due to excessive drowsiness, excessive blinking, and technical failure, respectively. This resulted in final samples of 32 participants in Experiment 1 and 39 participants in Experiment 2.

Recognition memory

Empirical "hit" and "false alarm" rates are presented in Table 1. As depicted in Fig. 2, both experiments

demonstrated credible behavioral production effects (aloud > silent) for *d'* at each delay, with the magnitude of the effect roughly halved for "catch" trials compared to early or late "Go" trials. Recollection and familiarity, analyzed next, suggested that the production effect was driven by both recollection and familiarity in both the early and the late "Go" conditions (e.g., see Fawcett & Ozubko, 2016; Ozubko et al., 2012), but only by recollection during "catch" trials. Corresponding frequentist models (i.e., ANOVAs) are also reported in the OSM (with data and code provided on the Open Science Framework (OSF) and GitHub). In all cases, frequentist ANOVAs supported the conclusions drawn from the Bayesian models, except that the production effect as measured by recollection was only marginally significant for Experiment 2 catch trials (although this comparison was significant both for Experiment 1 and for catch trials combined across experiments).

Recognizing that differences between Experiments 1 and 2 (e.g., use of a chin rest, avoidance of blinking) may have influenced the magnitude of the production effect, contrasts were calculated comparing across experiments for each dependent variable and delay condition in Fig. 2. These are detailed in the OSM and summarized in Table S6 (OSM). Although frequentist models found no significant interaction

Table 1 Experiments 1 and 2 and combined: Mean percent hits for each combination of encoding condition (aloud, silent) and delay (early, late, catch), as well as mean percent false alarms

	Early		Late		Catch		False alarms
	Aloud	Silent	Aloud	Silent	Aloud	Silent	
Experiment 1 ($n = 32$)	78.1 (2.5)	57.1 (2.8)	80.9 (1.9)	60.4 (3.1)	69.8 (2.9)	61.6 (2.5)	21.8 (2.1)
Experiment 2 ($n = 39$)	77.6 (2.0)	62.1 (2.4)	77.3 (2.1)	60.0 (2.4)	67.5 (2.6)	58.2 (2.4)	33.8 (2.4)
Combined ($n = 71$)	77.9 (1.6)	59.8 (1.8)	78.9 (1.4)	60.1 (1.9)	68.5 (1.9)	59.7 (1.7)	28.4 (1.7)

Each experiment used a single false alarm rate for all conditions. Values in parentheses are standard errors

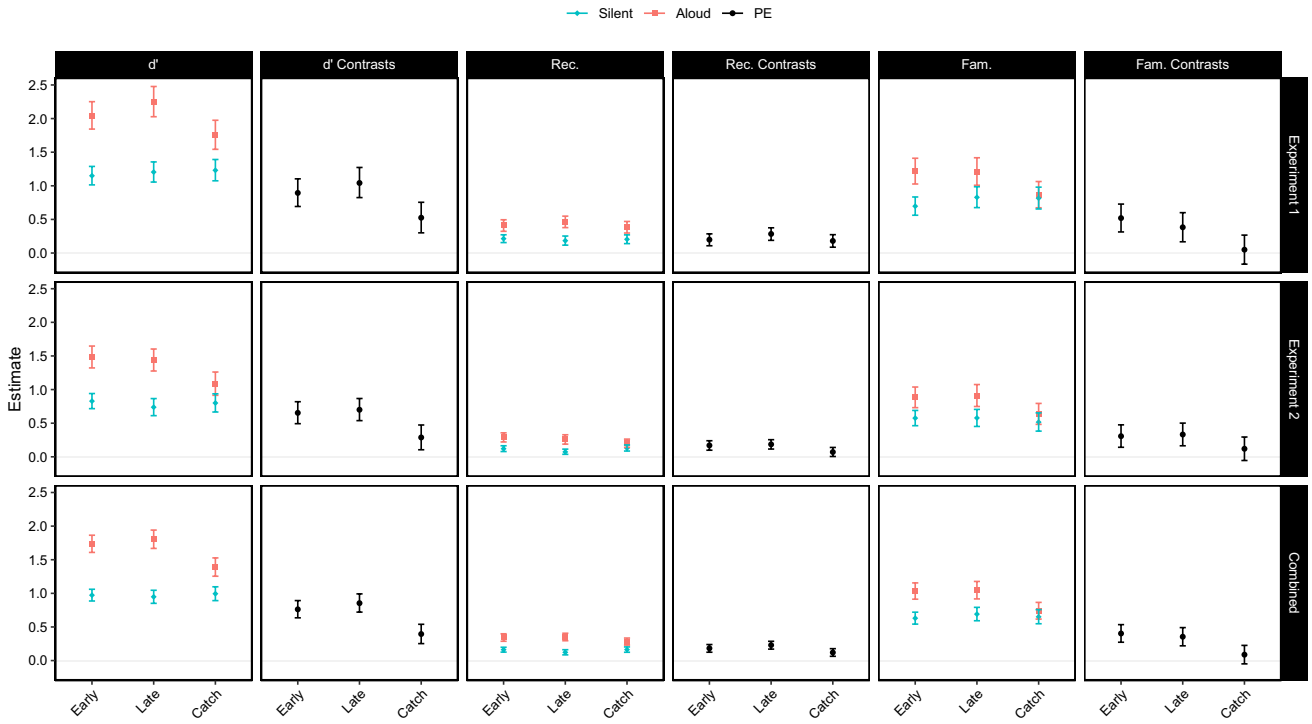

Fig. 2 Memory sensitivity (*d'*), familiarity, recollection, and associated contrasts as a function of production (aloud, silent) and delay (early, late, catch). Error bars represent 95% confidence intervals. Comparisons are considered credible when the confidence interval for a contrast does not cross 0. Rec. = recollection, Fam. = familiarity, PE = Production Effect (aloud – silent)

between Experiment and any other variable, within our primary Bayesian models the magnitude of the production effect (averaged across delay condition) was credibly smaller in Experiment 2 than in Experiment 1 for both the unequal variance model (*d'*) and for recollection within our dual-process model. Insofar as certain features of Experiment 2 might be considered liable to be distracting, this pattern is not surprising. Importantly, data are provided separately for each experiment (as well as combined) in Fig. 2, and the pattern of results (and conclusions) is similar for each.

Pupil size by condition

As depicted in Fig. 3, both mass univariate and GAMM models demonstrated sizable pupillary production effects (aloud > silent) for all conditions, with a credible difference emerging for the early and "catch" trials roughly 1,000 ms following word and instruction onset and persisting throughout the remainder of the trial. For late trials, this difference was delayed but still preceded the response period. During the response period, a large peak was observed for both early and late aloud trials, replicating Experiment 4 of Fawcett et al. (2025). Notably, the late condition peaked ~1,000 ms later than the early condition, reflecting the difference in "Go" signal onset between these conditions. Of greater interest, a smaller pupillary production effect was also observed for the "catch" trials around the time that production would normally commence, despite no production occurring.

Predicting memory outcomes using study phase pupil dilation

Using aggregate pupil size (i.e., averaged within each trial) to predict recognition performance on a per-item basis for aloud and silent trials at each delay failed to produce credible effects. Likewise, as depicted in Table 2, none of the correlations between the aggregate behavioral (in *d'*) and pupillary production effects were credible when using a regularizing prior, but the correlations were credible for both the late and "catch" trials when using priors informed by Fawcett et al. (2025). Incorporating priors derived from past studies constitutes a form of meta-analysis and reflects a cumulative approach to scientific inference (e.g., Gelman et al., 2021). Importantly, given that the reported correlations between the behavioral and pupillometric production effects were medium in size ($r \approx .3$; Fawcett et al., 2025), a much larger sample would be needed to definitively resolve this association in a single study. This is especially true for the "catch" trials, for which the correlation might be expected to be smaller (e.g., $r \approx .2$, requiring $n \approx 194$ for 80% power). Given the labor-intensive nature of pupillometry, larger-scale data collection across multiple studies and future meta-analytic aggregation will likely be required to definitively characterize the behavioral-pupillometric relation.

Our exploratory window analysis revealed that pupil size for aloud trials predicted subsequent memory (larger pupils associated with a greater probability of recognizing the word) within the range 5,380–6,950 ms for early trials and 6,430–6,900 ms for late trials. No windows predicted later memory for silent trials or for "catch" trials.

Discussion

Across two experiments we isolated the processes involved in response preparation from those associated with producing the target word in a production paradigm using delayed production and "catch" trials where participants expected to produce a word but did not do so. Despite no production during "catch" trials, both a pupillary and a behavioral production effect were observed, albeit of lesser magnitude relative to early or late "Go" trials. Notably, for early and late "Go" trials, the behavioral production effect was driven by both recollection and familiarity (e.g., Fawcett & Ozubko, 2016; Ozubko et al., 2012), but for "catch" trials it was driven only by recollection. Beyond these novel findings, the early and late trials replicate Fawcett et al.'s (2025) final experiment, again exhibiting a pupillary production effect with onset prior to actual production. Together, these findings challenge certain distinctiveness-based accounts and provide further support for the dual-process perspective (e.g., Fawcett & Ozubko, 2016; Ozubko et al., 2012) that the production effect involves multiple mechanisms (as has been argued for related paradigms such as enactment and generation; see Fawcett et al., 2022).

Both the silent and aloud conditions were associated with an initial increase in pupil size that was larger and more sustained during aloud than silent trials. For silent trials, pupil size diminished throughout the trial. For both early and "catch" conditions, pupil size was greater during aloud than during silent trials from soon after word and instruction onset until the trial ended; for late trials, this difference was delayed. Given that early and late trials were identical until the response window, we view this delayed pattern as arising from statistical noise (see also Fawcett et al., 2025).

For aloud trials, two additional peaks were observed. One was associated with the response itself (present during early and late trials following "Go" signal onset but absent during "catch" trials where there was no "Go" signal or response). A second, smaller peak was associated with onset of the prospective response period (observed fully during "catch" trials, partially occluded during late trials, and fully overlaid by the response peak during early trials). Our interpretation is that, when participants are alerted that they are about to speak during instruction onset, there is an

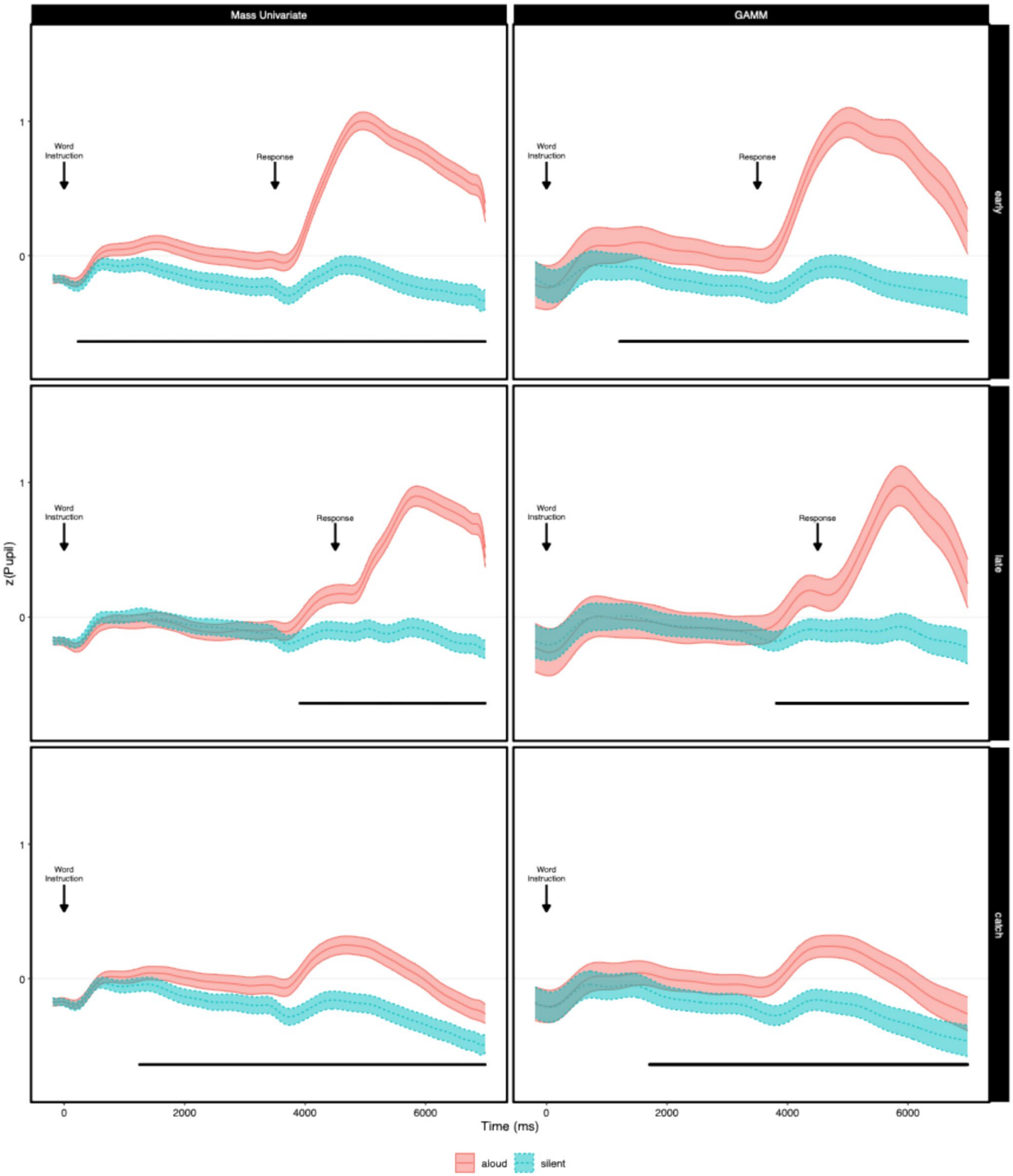

Fig. 3 Normalized pupil size as a function of time (ms), modelling approach (mass univariate, generalized additive mixed model), condition (aloud, silent, control), and experiment. The mass univariate model reflects the empirical mean and its 95% confidence interval, with rug plots (black horizontal bars) indicating timepoints where paired cluster corrected *t*-tests identified significant differences between the indicated conditions. GAMM reflects predictions from a multilevel generalized additive mixed model and its 95% confidence interval, with rug plots indicating timepoints where this model predicted credible differences between the indicated conditions. Onsets of trial events are noted by arrows. The response period was absent for "catch" trials

Table 2 Experiment 2: Correlations (*r*) between the behavioral and pupillary production effects (after outlier removal) as a function of delay, using either regularizing priors or informative priors based on Fawcett et al. (2025). Frequentist Pearson correlations are also provided for comparison

	Outliers	*r* *(regularizing priors)*	*r* *(informative priors)*	*r* *(frequentist)*
Early	3	-.15 (-.46,.17)	.09 (-.11,.30)	-.16 (-.47,.17)
Late	1	.26 (-.05,.56)	.27 (.07,.47)	.29 (-.04,.55)
Catch	1	.14 (-.17,.44)	.21 (.01,.42)	.15 (-.17,.45)

Values in parentheses are 95% confidence intervals. Outliers were defined as participants detected as anomalous by both the minimum covariance determinant estimator (MCD) and isolation forest methods

initial attention-focusing effect (reflecting activation of the norepinephrine system) that is magnified as the anticipated response deadline approaches. Following this preparation time, overt verbal production is then associated with a period of distinctive encoding (Fawcett et al., 2025). The present findings are the first to provide separable psychophysiological markers of each process, demonstrating that preparing to produce a word elicits focused attention separable from production itself.

With respect to the behavioral production effect, performance during "catch" trials provides key evidence that intending to produce a word aloud explains roughly half the magnitude of the production effect in memory (as evidenced by a behavioral production benefit of roughly half the magnitude during those trials compared to the "early" and "late" delay conditions). This challenges the viability of the *distinctiveness heuristic account* (Dodson & Schacter, 2001), as in the case of "catch" trials participants are unable to use having said a word at study to guide their test responses. Also novel is the finding that the production effect for "catch" trials appears to be driven by recollection, not familiarity. The dissociation between familiarity and recollection supports a dual-process perspective (Fawcett & Ozubko, 2016). Previous research established that the production effect is driven by both familiarity and recollection when manipulated within-subject (Ozubko et al., 2012; Fawcett & Ozubko, 2016) but by familiarity alone when manipulated between-subjects. Fawcett and Ozubko (2016) attributed the reduced magnitude of the production effect for between-subjects (vs. within-subject) designs to the absence of the recollective component (for meta-analyses, see Fawcett, 2013; Fawcett et al., 2023).

Whereas both Ozubko et al. (2012) and Fawcett and Ozubko (2016) speculatively aligned distinctive encoding with recollection and heightened attention with familiarity, the fact that the use of pure-list designs minimizes the contribution of recollection (but not familiarity) and that withholding production on "catch" trials minimizes the contribution of familiarity (but not recollection) may suggest a more complex interpretation. Specifically, that the preparatory (at instruction onset) and anticipatory (at response window onset) pupillary components occur during "catch" trials suggests that the recollection-based component is driven instead by attentional or motivational factors whereas the familiarity-based component is driven by distinctive encoding. Of course, neither measure is likely process-pure (e.g., Surprenant & Neath, 2009). For example, the production-driven recollective benefit for "catch" trials may reflect distinctive encoding associated with response preparation (e.g., activation of phonological representations and motor planning), which itself may partially explain the need for focused attention. Consequently, the *relative distinctiveness account*, which relies on accessibility to – but not strategic use of – the production record is not necessarily incompatible with the present findings.

An alternative interpretation is that catch trials elicit processes associated with a violation of expectations, rather than with a subset of processes akin to those engaged during produced trials. Although we cannot entirely rule this out, we view it as less likely in the present paradigm. Participants were explicitly informed that approximately one-third of trials would not contain a "Go" signal and were instructed that, when this occurred, they should simply wait for the next trial to begin. Thus, the absence of a "Go" signal was an expected and fully instructed event, rather than an unanticipated deviation from task structure. This differs from paradigms in which unexpected events trigger adaptive shifts in strategy or processing mode (e.g., Chen et al., 2019). Further, inspection of Fig. 3 reveals a similar pupillometric deflection occurring for the late trials – preceding and partially summating with the deflection associated with speaking aloud – to the one observed for catch trials. This is even though, during late trials, expectations are minimally violated as participants are accustomed to the "Go" signal often being delayed. Nevertheless, future work could test this *expectation violation account* more directly by incorporating truly

unexpected omissions or task changes to determine whether such expectation violations produce memory outcomes (and pupillary deflections) comparable to those observed here.

Surprisingly, we did not observe a larger production effect for late relative to early delay conditions. Mama and Icht (2018) found that asking participants to delay production resulted in a larger effect on free recall. The absence of a *delayed production effect* in our experiments may be attributable to three factors. First, we did not include an immediate condition: The early condition incorporated a 500-ms delay before a response was made. Second, because Mama and Icht provided the word prior to the production instruction, participants were unaware whether a given word would be produced or read silently until their delayed instruction appeared. And third, we used a recognition test whereas they used a free-recall test. Interestingly, the production effect in our study as measured by recollection (the measure most like free recall) tended to be larger for the late as opposed to the early or "catch" conditions, although this was only a trend in the former case. Further research – possibly using free recall – may determine whether delaying production and/or the instruction is necessary for the delayed production effect to emerge.

Owing to our emphasis on preparatory processes, our work is also thematically related to the performance anticipation hypothesis described by Forrin et al. (2019). Using a production paradigm with a predictable trial order, they observed impaired memory for silent items when participants anticipated production on the following trial. They attributed this to heightened anxiety or reduced attention to the silent items. This finding was later challenged by Gionet et al. 2025), who showed that manipulations intended to heighten performance anxiety (e.g., whether the researcher was in the room) did not influence the production effect, and that the seemingly anticipatory impairment could instead be explained via rehearsal disruption (for further discussion of performance anxiety in this paradigm, see Wakeham-Lewis et al., 2022; Whitridge et al., 2024). Importantly, those studies addressed how anticipating an aloud trial might influence performance on the *preceding* trial whereas the present work addresses preparation to read aloud (and its influence on performance) within the *current* trial. Both the preceding and the present work hypothesize a pre-occupation with items soon to be read aloud but differ in that our proposed preparatory mechanisms focus attention on the *current* word and contribute to efficient encoding, rather than impairing performance on other trials. That we used a randomized (and therefore unpredictable) trial order means that participants could not anticipate whether they would read aloud or silently on future trials, minimizing the possibility that a preoccupation with future production might negatively influence encoding of silent items.

The present findings also shed light on the *offline production effect* described by Jamieson and Spear (2014). In their study, asking participants to *imagine* producing a word was sufficient to elicit a production effect, albeit of lesser magnitude than that elicited by standard "online" production.[1] Their study would appear aligned with ours, with one key exception: Their participants were asked to actively imagine producing the item whereas our participants were told that when a "Go" signal was not presented they were to do nothing until the next trial. We believe it unlikely that participants would actively imagine producing words during "catch" trials without being instructed to do so. It is plausible, however, that Jamieson and Spear's (2014) findings were driven by the mechanisms observed in our experiments. If so, this would suggest that it was not offline production but attention and heightened encoding elicited in preparation for and anticipation of production that drove their effect. Future research using psychophysiological measures (with indices of recollection and familiarity) should be better able to compare our paradigm to theirs. Indeed, "catch" trials could be incorporated into the paradigm reported by Jamieson and Spear (2014) to investigate whether interfering with the imagined response would influence their findings.

The idea that planning to act may itself improve memory independent of the execution of any intended action also connects the current production study to a broader principle within human memory. A similar pattern has been observed for both the drawing effect (Wammes et al., 2018) and the enactment effect (Helstrup, 1996; for a meta-analysis, see Roberts et al., 2022) in paradigms where participants were instructed to prepare but not to execute the associated response at encoding. However, whereas those earlier studies involved planning of complex, multistep behaviors (e.g., creating an image of a provided word, or planning the steps to achieve a specific goal such as breaking a matchstick), our work is the first to address a situation as simple as reading aloud. Further, our work both links the observed memory benefits to psychophysiological markers of focused attention and – for the first time – proposes that planning to act may operate largely by improving recollection.

Further work is needed to evaluate this preparatory stage across encoding techniques (e.g., drawing, enactment, production) to judge whether they share a common process. Uniting these methods would constitute a theoretical step forward given that the drawing, enactment, and production effects often are viewed as related but independent phenomena. Future work could use a paradigm like the one reported here but including separate drawing,

[1] Jamieson and Spear used actual typing versus imagined typing; MacLeod, Ozubko, and Major (n.d.) have data showing a similar difference for aloud versus imagined aloud conditions.

enactment, and production groups to permit direct comparison of their respective pre-task planning processes, their psychophysiological markers, and their unique influence on familiarity and recollection.

In conclusion, the present experiments provide compelling evidence that response preparation alone – without any overt action – is sufficient to elicit both behavioral and pupillometric signatures of the production effect, albeit reduced relative to actual production. Memory performance during catch trials suggests that as much as half of the memory benefit usually attributed to production itself may arise from preparatory processes that enhance attention and later recollection, independent of the act of speaking. By isolating the attentional and distinctive encoding phases of production, this work not only challenges the sufficiency of existing distinctiveness-based accounts, but also reframes the production effect as a multifaceted phenomenon. Our results further bridge the production, drawing, and enactment literatures, suggesting a common cognitive mechanism whereby preparing to act benefits memory, possibly through a shared combination of enhanced attention and distinctive (preparatory) encoding.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13423-025-02847-7.

Acknowledgements This research was supported by Natural Sciences and Engineering Research Council (NSERC) of Canada Discovery Grant RGPIN-2017-05250 to JMF and an NSERC postdoctoral scholarship to BRTR. This work comprised the honors thesis completed by Cayley Thoms. Parts of this work were presented at the annual meeting of the Canadian Society for Brain, Behaviour, and Cognitive Science (in 2025) and at multiple in-house conferences at Memorial University of Newfoundland.

Authors' contributions *JMF:* Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing – original draft, writing – review and editing, visualization, supervision, project administration, funding acquisition. *BRTR:* Writing – original draft, writing – review and editing. *SH:* Validation, investigation, data curation, writing – original draft, writing – review and editing, visualization. *CT:* Conceptualization, methodology, software, investigation, data curation, writing – original draft, writing – review and editing, visualization. *JW:* Software, formal analysis, data curation, writing – original draft, writing – review and editing, visualization, supervision. *CMM:* Writing – original draft, writing – review and editing. *HW:* Conceptualization, methodology, software, data curation, writing – original draft, writing – review and editing, supervision.

Funding This research was supported by Natural Sciences and Engineering Research Council (NSERC) of Canada Discovery Grant RGPIN-2017-05250 to JMF and an NSERC postdoctoral scholarship to BRTR.

Availability of data and materials The data for all experiments are available on the project OSF (https://osf.io/dj3nu/) or GitHub (https://github.com/jmfawcet/proddelay_public) pages. No studies were pre-registered.

Code availability The code for all analyses are available on the OSF (https://osf.io/dj3nu/) or GitHub (https://github.com/jmfawcet/proddelay_public) pages

Declarations

Ethics approval The procedures for these experiments received approval from the Interdisciplinary Committee on Ethics in Human Research at <location redacted for anonymity> (Protocol #20171484-SC).

Consent to participate All participants included in our analyses consented to participate in this study following local policies and guidelines.

Consent for publication All participants included in our analyses consented to have work derived from their data published (and their anonymized data made publicly available) and all authors have read and consented to publication.

Conflicts of interest/Competing interests The authors report no conflicts of interest or competing interests.

Open Practices Data are available via the Open Science Framework (OSF) at: https://osf.io/dj3nu/ or GitHub at: https://github.com/jmfawcet/proddelay_public pages for this project. No studies were pre-registered.

Author Note This research was supported by Natural Sciences and Engineering Research Council (NSERC) of Canada Discovery Grant RGPIN-2017-05250 to JMF and an NSERC postdoctoral scholarship to BRTR. This work comprised the honors thesis completed by Cayley Thoms. Parts of this work were presented at the annual meeting of the Canadian Society for Brain, Behaviour, and Cognitive Science (in 2025) and at multiple in-house conferences at Memorial University of Newfoundland.

References

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

Bjork, R. A., & Bjork, E. L. (2020). Desirable difficulties in theory and practice. *Journal of Applied Research in Memory and Cognition, 9*(4), 475–479. https://doi.org/10.1016/j.jarmac.2020.09.003

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

Caplan, J. B., & Guitard, D. (2024). A feature-space theory of the production effect in recognition. *Experimental Psychology, 71*(1), 64–82. https://doi.org/10.1027/1618-3169/a000611

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., et al. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software, 76*(1), 1–32.

Chen, H., Yan, N., Zhu, P., Wyble, B., Eitam, B., & Shen, M. (2019). Expecting the unexpected: Violation of expectation shifts strategies toward information exploration. *Journal of Experimental Psychology. Human Perception and Performance, 45*(4), 513–522. https://doi.org/10.1037/xhp0000622

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

Cortes, D. (2025). *isotree: Isolation-based outlier detection*. R package version 0.6.1-4, https://CRAN.R-project.org/package=isotree

Delacre, M., & Klein, O. (2019). *Routliers: robust outliers detection*. R package version 0.0.0.3. https://CRAN.R-project.org/package=Routliers

DeCarlo, L. T. (2010). On the statistical and theoretical basis of signal detection theory and extensions: Unequal variance, random coefficient, and mixture models. *Journal of Mathematical Psychology, 54*(3), 304–313.

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods, 3*(2), 186–205.

Dixon, P. (2008). Models of accuracy in repeated-measures designs. *Journal of Memory and Language, 59*(4), 447–456.

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

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

Fauconnier, C., & Haesbroeck, G. (2009). Outliers detection with the minimum covariance determinant estimator in practice. *Statistical Methodology, 6*(4), 363–379. https://doi.org/10.1016/j.stamet.2008.12.005

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

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

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

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

Fawcett, J. M., Baldwin, M. M., Whitridge, J. W., Swab, M., Malayang, K., Hiscock, B., ... & Willoughby, H. V. (2023). Production improves recognition and reduces intrusions in between-subject designs: An updated meta-analysis. *Canadian Journal of Experimental Psychology, 77*(1), 35–44. https://doi.org/10.1037/cep0000302

Fawcett, J. M., Roberts, B. R. T., Willoughby, H., Tiller, J., Hourihan, K. L., & MacLeod, C. M. (2025). The pupillometric production effect: Evidence for enhanced processing preceding, during, and following production. *Cognition, 266*, Article 106326. https://doi.org/10.2139/ssrn.5210001

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/s1321-012-0210-8

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*, 104050. https://doi.org/10.1016/j.jml.2019.104050

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology, 40*, 1–104.

Gabry, J., Češnovar, R., Johnson, A., & Bronder, S. (2025). cmdstanr: R interface to CmdStan (Version 0.9.0) [Computer software]. https://mc-stan.org/cmdstanr/

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

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2021). *Bayesian data analysis* (3rd ed). Chapman & Hall/CRC.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Wiley.

Gionet, S., Guitard, D., Poirier, M., Yearsley, J. M., & Saint-Aubin, J. (2025). Distinctiveness and interference in free recall: A test with the production effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition. Advance online publication.* https://doi.org/10.1037/xlm0001504

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

Helstrup, T. (1996). Not only encoding and retrieval: The enactment effect as a function of task preparation. *Scandinavian Journal of Psychology, 37*(4), 407–415. https://doi.org/10.1111/j.1467-9450.1996.tb00672.x

Herndon, W. H., & Weik, J. W. (1896). *Abraham Lincoln: The true story of a great life* (Vol. 2). accessed via http://www.gutenberg.org/files/38484/38484-h/38484-h.htm

Hess, E. H., & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem solving. *Science, 143*(3611), 1190–1192. https://doi.org/10.1126/science.143.3611.1190

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

Hourihan, K. L., & Fawcett, J. M. (2024). It's all about that case: Production and reading fluency. *Experimental Psychology, 71*(2), 83–96. https://doi.org/10.1027/1618-3169/a000615

Hubert, M., & Debruyne, M. (2009). Minimum covariance determinant. *Wiley Interdisciplinary Reviews: Computational Statistics, 2*(1), 36–43. https://doi.org/10.1002/wics.61

Hunt, R. R. (2006). The concept of distinctiveness in memory research. In R. R. Hunt & J. Worthen (Eds.), *Distinctiveness and memory* (pp. 3–25). Oxford University Press.

Hunt, R. R. (2013). Precision in memory through distinctive processing. *Current Directions in Psychological Science, 22*(1), 10–15. https://doi.org/10.1177/0963721412463228

Icht, M., Mama, Y., & Algom, D. (2014). The production effect in memory: Multiple species of distinctiveness. *Frontiers in Psychology, 5*. https://doi.org/10.3389/fpsyg.2014.00886

Jamieson, R. K., Mewhort, D. J. K., & Hockley, (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

Jamieson, R. K., & Spear, J. (2014). The offline production effect. *Canadian Journal of Experimental Psychology, 68*(1), 20–28. https://doi.org/10.1037/cep0000009

Jaeger, T. F. (2008). Categorical Data Analysis: Away from ANOVAS (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*(4), 434–446.

Kelly, M. O., Ensor, T. M., Lu, X., MacLeod, C. M., & Risko, E. F. (2022). Reducing retrieval time modulates the production effect: Empirical evidence and computational accounts. *Journal of Memory and Language, 123*(1), Article 104299. https://doi.org/10.1016/j.jml.2021.104299

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

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology, 79*, 328–348.

Lin, O.Y.-H., & MacLeod, C. M. (2012). Aging and the production effect: A test of the distinctiveness account. *Canadian Journal of Experimental Psychology, 66*(3), 212–216. https://doi.org/10.1037/a0028309

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. *Proceedings of the 8th IEEE International Conference on Data Mining* (ICDM), 413–422.

MacDonald, P. A., & MacLeod, C. M. (1998). The influence of attention at encoding on direct and indirect remembering. *Acta Psychologica, 98*(2–3), 291–310. https://doi.org/10.1016/S001-6918(97)00047-4

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

MacLeod, C. M., Ozubko, J. D., & Major, J. C. (n.d.). *The production effect: Intentional forgetting and imagining* [Unpublished manuscript]. Department of Psychology, University of Waterloo.

Mama, Y., Fostick, L., & Icht, M. (2018). The impact of different background noises on the production effect. *Acta Psychologica, 185*(3), 235–242. https://doi.org/10.1016/j.actpsy.2018.03.002

Mama, Y., & Icht, M. (2018). Production on hold: Delaying vocal production enhances the production effect in free recall. *Memory, 26*(5), 589–602. https://doi.org/10.1080/09658211.2017.1384496

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

Mathôt, S., Schreij, D., & Theeuwes, J. (2011). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods, 44*(2), 314–324. https://doi.org/10.3758/s13428-011-0168-7

Mickes, L., Wixted, J. T., & Wais, P. E. (2007). A direct test of the unequal-variance signal detection model of recognition memory. *Psychonomic Bulletin & Review, 14*(5), 858–865.

Murray, D. J. (1965). Vocalization-at-presentation and immediate recall, with varying presentation rates. *The Quarterly Journal of Experimental Psychology, 17*(1), 47–56. https://doi.org/10.1080/17470216508416407

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

Papesh, M. H., Goldinger, S. D., & Hout, M. C. (2012). Memory strength and specificity revealed by pupillometry. *International Journal of Psychophysiology, 83*(1), 56–64. https://doi.org/10.1016/j.ijpsycho.2011.10.002

Pritchard, V. E., Heron-Delaney, M., Malone, S. A., & MacLeod, C. M. (2020). The production effect improves memory in 7 to 10-year-old children. *Developmental Psychology, 91*(3), 901–913. https://doi.org/10.1111/cdev.13247

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

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Roberts, B. R., MacLeod, C. M., & Fernandes, M. A. (2022). The enactment effect: A systematic review and meta-analysis of behavioral, neuroimaging, and patient studies. *Psychological Bulletin, 148*, 397–434. https://doi.org/10.1037/bul0000360

Rouder, J. N., Lu, J., Sun, D., Speckman, P., Morey, R., & Naveh-Benjamin, M. (2007). Signal Detection Models with Random Participant and Item Effects. *Psychometrika, 72*(4), 621–642.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association, 79*(388), 871–880. https://doi.org/10.2307/2288718

Rousseeuw, P. J., & Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics, 41*(3), 212–223. https://doi.org/10.1080/00401706.1999.10485670

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

Sirois, S., & Brisson, J. (2014). Pupillometry. *Wiley Interdisciplinary Reviews: Cognitive Science, 5*(6), 679–692. https://doi.org/10.1002/wcs.1323

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

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31*(1), 137–149.

Swets, J. A. (1986). Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychological Bulletin, 99* (2), 181–198. https://doi.org/10.1037/0033-2909.99.2.181

van Rij, J., Wieling, M., Baayen, R., & van Rijn, H. (2022). itsadug: Interpreting time series and autocorrelated data using GAMMs. *R package version, 2*(4), 1.

Wakeham-Lewis, R. M., Ozubko, J. D., & 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., Roberts, B. R. T., & Fernandes, M. A. (2018). Task preparation as a mnemonic: The benefits of drawing (and not drawing). *Psychonomic Bulletin & Review, 25*(6), 1–8. https://doi.org/10.3758/s13423-018-1477-y

Wixted, J. T., & Mickes, L. (2010). A continuous dual-process model of remember/know judgments. *Psychological Review, 117*(4), 1025–1054.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114*(1), 152–176.

Whitridge, J. W., Huff, M. J., Ozubko, J. D., Bürkner, P. C., Lahey, C. D., & Fawcett, J. M. (2024). Singing does not necessarily improve memory more than reading aloud. *Experimental Psychology, 71*(1), 33–50. https://doi.org/10.1027/1618-3169/a000614

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed). Chapman and Hall/CRC.

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

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

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