Cognition 266 (2026) 106326

Contents lists available at ScienceDirect

Cognition

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

The pupillometric production effect: Evidence for enhanced processing preceding, during, and following production

Jonathan M. Fawcett [a,*], Brady R.T. Roberts [b,c,*], Hannah V. Willoughby [d], Jenny C. Tiller [a], Kathleen L. Hourihan [a], Colin M. MacLeod [e]

[a] Department of Psychology, Memorial University of Newfoundland, St. John's, NL A1C 5S7, Canada
[b] Department of Psychology, University of Chicago, Chicago, IL 60637, USA
[c] Institute for Mind and Biology, University of Chicago, Chicago, IL 60637, USA
[d] C3 Human Factors Inc., St. John's, NL, Canada
[e] Department of Psychology, University of Waterloo, Waterloo, ON N2L 3G1, Canada

ARTICLE INFO

Keywords:
Production effect
Pupillometry
Distinctiveness
Attention
Memory

ABSTRACT

The production effect refers to superior memory performance for words read aloud than for those read silently. This finding has usually been attributed to the incorporation of distinctive sensorimotor information into the memory record of items read aloud, facilitating their successful retrieval during the memory test. Less research has explored other cognitive or motivational differences between the aloud and silent conditions. Here we used pupillometry to explore the time course of attention allocated during aloud, silent, and control (say "check") study trials. Across four experiments, instructions were presented either concurrently with or preceding the word. To permit evaluation of preparatory processing independent of a verbal response, we explored the case where responses had to be withheld until a "Go" signal appeared. In addition to the typical behavioral production effect in memory, each experiment also revealed a pupillometric production effect (greater pupil dilation for aloud than for silent words) that—while separable from the act of speaking itself—was correlated with the size of the memory benefit. Critically, this pupillometric-behavioral correlation did not occur for control (say "check") trials. We interpret these findings as support for an initial attention-focusing effect that comes from preparing for and executing vocalization during both aloud and control trials, followed by a phase of distinctive processing of target word features that is unique to aloud trials.

The finding that reading words aloud results in better memory than reading words silently is often credited to Hopkins and Edwards (1972)) (for earlier efforts, see Ekstrand et al., 1966; Murray, 1965; and for evidence that the strategy was understood and popularly used even earlier, see Gates, 1917; Herndon & Weik, 1896). Over the decades to follow, relatively few studies investigated this phenomenon, and it was referred to by several names, including the 'pronunciation effect' (e.g., Hopkins & Edwards, 1972) and the 'modality effect' (e.g., Conway & Gathercole, 1987; Gathercole & Conway, 1988), complicating connections across the literature.

In the past 15 years, however, research interest in this topic has increased substantially following publication of an influential paper by MacLeod et al. (2010). Across eight experiments, they delineated the phenomenon, providing an updated theoretical framework and identifying the breadth and boundaries of what they labeled the *'production effect.'* The phenomenon has since been generalized to a range of productive acts, including writing or typing (e.g., Forrin et al., 2012), singing (e.g., Quinlan & Taylor, 2013; but see Whitridge et al., 2024), and even imagining producing words in different ways (e.g., Jamieson & Spear, 2014). MacLeod and Bodner (2017) provide a brief review of the literature.

1. The distinctiveness account

The currently dominant theoretical account argues that producing a word encourages distinctive encoding (for a review of distinctiveness as a concept, see Hunt, 2006), making words that were produced at study more memorable than words that were not produced at study on a later

* Corresponding authors at: Department of Psychology, Memorial University of Newfoundland, St. John's, NL A1C 5S7, Canada.
E-mail addresses: jfawcett@mun.ca (J.M. Fawcett), bradyr@uchicago.edu (B.R.T. Roberts), hannah.willoughby@c3hf.com (H.V. Willoughby), khourihan@mun.ca (K.L. Hourihan), cmacleod@uwaterloo.ca (C.M. MacLeod).

https://doi.org/10.1016/j.cognition.2025.106326
Received 1 April 2025; Received in revised form 4 September 2025; Accepted 12 September 2025
Available online 16 September 2025

test. The distinctiveness idea has its roots in early work such as Hopkins and Edwards (1972) and especially in the work of Conway and Gathercole (1987) (Gathercole & Conway, 1988), with an update provided by MacLeod et al. (2010). Briefly, the idea is that reading aloud results in additional sensorimotor (e.g., auditory, articulatory) features being incorporated into the encoding episode at study. These features, sometimes referred to as the production record (e.g., Fawcett, Quinlan, & Taylor, 2012; MacLeod et al., 2010), permit words previously read aloud to "stand out" at test against a backdrop of previously read silent words that lack these features.

There are two variants of this distinctiveness perspective differing in how these features benefit memory at test. The *distinctiveness heuristic account* (e.g., Dodson & Schacter, 2001) argues that, for each word, participants consider whether they believe that they said the word recently: If so, this is evidence that it must have been studied. Such evidence is not available for words that were read silently. This approach uses access to the production record to discriminate between old and new words. The *relative distinctiveness account* is similar but argues that spoken words have distinctive features in memory, features not present for words read silently, and so seem to "pop out" during recall or recognition tests due to implicit retrieval dynamics, as captured in a variety of computational models (e.g., Caplan & Guitard, 2024; Jamieson et al., 2016; Kelly et al., 2022; Saint-Aubin et al., 2021; Wakeham-Lewis et al., 2022). This account does not necessitate strategic—or even conscious—use of the production record: The benefit emerges as a routine consequence of how encoding and retrieval operate.

The primacy of distinctiveness as an explanation for the production effect is backed by studies using source or list discrimination tasks, the outcomes of which suggest that participants do access distinctive information attached to the encoding record—such as motor, auditory, or semantic information—during the memory test and that they use this information to discriminate between test items (Conway & Gathercole, 1987; Ozubko et al., 2012; Ozubko et al., 2013). Participants also report using such a strategy when asked directly (e.g., Fawcett & Ozubko, 2016; Online Supplement). Very early on, Hopkins and Edwards (1972) observed the production memory benefit in a mixed-list design, where each participant studied *both* aloud and silent words, but not in a pure-list design, where each participant studied *only* aloud or *only* silent words. That the benefit of production emerged only for mixed-list designs was also used as evidence favoring distinctiveness: In the absence of silent words, the aloud words would lack a comparator from which to "pop out" (e.g., MacLeod et al., 2010). In a series of meta-analyses (e.g., Bodner et al., 2014; Fawcett, 2013; Fawcett et al., 2023), however, and in subsequent empirical work (Bodner et al., 2014; Bodner et al., 2016; Forrin et al., 2019; Forrin & MacLeod, 2016; Taikh & Bodner, 2016), consistent evidence has now shown that the production effect *does* occur in between-subjects designs, albeit with reduced magnitude compared to within-subject designs.

The magnitude of the production effect has also been found to 'scale' with increasingly elaborative forms of production (e.g., Conway & Gathercole, 1987; for a more recent perspective, see Forrin et al., 2012). For example, compared to silently mouthing a word, reading a word aloud leads to a larger production effect (e.g., Conway & Gathercole, 1987), possibly due to the exclusion or mitigation of certain features such as auditory feedback. In general, most modern distinctiveness-based accounts subscribe to the *sensorimotor scaling hypothesis* (Whitridge et al., 2024) that the production effect ought to scale proportional to the number of distinctive processes which occur during the study phase (e.g., Forrin et al., 2012).

2. Other potential mechanisms

Alternative accounts of the production effect tend to focus on enhanced encoding or "strengthening" of the episodic memory for words read aloud (Bodner & Taikh, 2012; see also Craik & Lockhart, 1972, for a related concept). One mechanism through which this might be achieved is increased attention to those words (Bodner et al., 2014; Fawcett, 2013; MacDonald & MacLeod, 1998; Ozubko et al., 2012). By making processing of the word more challenging and forcing engagement, production may also function as a form of desirable difficulty (e.g., Bjork, 1994; see also Bjork & Bjork, 2020), akin to how testing knowledge rather than re-studying it produces a greater benefit (e.g., Roediger & Karpicke, 2006; although production does not necessarily interact with other forms of desirable difficulties, see Hourihan & Fawcett, 2024).

Supporting the idea that they are more attentive during production than during silent reading, participants self-report greater engagement with aloud words (Fawcett & Ozubko, 2016) and are less likely to mind-wander while reading aloud than while reading silently (Varao Sousa et al., 2013). Further, distracting attention using auditory nonsense syllables (but not white noise) eliminates the effect (Mama et al., 2018), and for people with attention-deficit/hyperactivity disorder the effect has been observed only in medicated but not in unmedicated individuals (Mama & Icht, 2018a). The influence of either auditory distraction or of an attentional deficit is observed predominantly within the aloud condition, suggesting that unimpaired and unimpeded attention is a necessary component for the memory advantage observed for aloud words.

On a recognition test, judging a studied word to be "familiar" reflects a quick, almost automatic sense of having previously encountered that word, but without specific details; in contrast, judging it to be "recollected" requires retrieval of specific contextual details more aligned with vivid remembering (Yonelinas, 2002). This distinction is informative in the case of the production effect. In a mixed-list design, producing words aloud during the study phase benefits both recollection and familiarity whereas in a pure-list design only familiarity shows a benefit. This pattern led Fawcett and Ozubko (2016) to argue that the production effect may arise due to multiple processes (e.g., see Fawcett et al., 2022; MacLeod & Bodner, 2017 for further discussion of this idea).

3. Beyond behavioral measures

Not surprisingly, over the past decade, interest has also grown in using psychophysiological and neuroimaging approaches to explore the mechanisms underlying the production effect. Two studies have used functional magnetic resonance imaging (fMRI; Bailey et al., 2021; Nakamura et al., 2023) and two have used electroencephalography (EEG; Hassall et al., 2016; Zhang et al., 2023) to study neural markers of this phenomenon.

Bailey et al. (2021) first used fMRI to identify brain regions and networks critical to the production effect. They reasoned that if production resulted in the encoding of additional distinct elements (i.e., motoric, semantic, etc.) then the areas responsible should be more active during study and should also be re-activated at test. They did observe greater activation in brain regions associated with sensorimotor and semantic processing during aloud than silent study trials. Moreover, this neural activity predicted memory performance for words studied aloud. Bailey et al. also hinted at additional processing that could be involved, such as enhanced engagement or conceptual encoding during aloud trials (see also Fawcett et al., 2022; Lu et al., 2025). Later, Nakamura et al. (2023) used fMRI to record neural activity during the test phase one week after study. They largely replicated the findings of Bailey et al. (2021), notably showing significant aloud-silent differences in the right lingual gyrus (involved in visual processing and potentially the reactivation of visually encoded word representations) and the left and right precentral gyri (involved in motor control, including articulation-related processes).

fMRI is limited in temporal resolution and so cannot completely characterize how an effect unfolds over time. Hassall et al. (2016) used EEG instead, measuring the P300—a positive deflection over the frontal and parietal lobes. Because it had previously been associated with

distinctive encoding in other paradigms (Fabiani & Donchin, 1995; Kamp et al., 2012), Hassall et al. inferred that the P300 amplitude should be greater for aloud than for silent words, which is what they observed. They interpreted this as neural evidence of distinctive encoding for aloud items. A more recent EEG study by Zhang et al. (2023) largely replicated the results of Hassall et al. (2016)—a larger P3b response during preparatory phases before reading aloud relative to reading silently—but they argued that the larger P3b response for aloud items could be due to attention rather than distinctiveness. Therefore, it remains an open question to what extent aloud items benefit from further attentional processing or increased distinctiveness of the target word in memory.

Although these recent EEG studies benefit from good temporal resolution to quantify production and related processing as it unfolds, they share a common limitation: To avoid contamination of the EEG signal by movement during vocalization, target words and trial-type instructions were presented while neural measures were recorded, all before the actual response window. That is, ERPs were recorded during the pre-production intention phases rather than during the actual speaking time itself. ERP differences observed prior to the act of production suggest that aloud and silent words diverge in their processing even before participants say a word aloud.

To ensure measurement of psychophysiological responses with good temporal resolution throughout the productive act—and to provide another novel approach—we opted to use pupillometry for the first time. Although pupil size can be modulated by task-irrelevant motor activity, it is primarily considered a proxy for activity within a distributed set of brain regions supporting focused attention (Strauch et al., 2022). This capturing of attentional processes throughout the act of production was crucial for testing our predictions.

4. The present research

The purpose of the present study is, then, to use pupillometry to unobtrusively quantify changes in processing, mental load, and engagement throughout the entire study phase of a typical production effect paradigm. Specifically, pupils dilate in response to phasic firing in the locus coeruleus (LC; Sirois & Brisson, 2014), reflecting moment-to-moment changes in attention or processing load. Pupils will be largest during difficult tasks (e.g., a hard math problem) and smallest during easy tasks (e.g., an easy math problem; Hess & Polt, 1964). Pupillometry has been used to unobtrusively study cognitive activity, including fundamental processes involved in attention (Unsworth & Robison, 2018), and working memory (Keene et al., 2022; Unsworth & Robison, 2015). Most relevant to our work, it has also been used as a tool to investigate recognition memory processes (Geller et al., 2016). For example, pupil size at encoding has been shown to reflect later retrieval success for words (Võ et al., 2008), images of manmade and natural objects (Kafkas & Montaldi, 2011), pictures of natural scenes (Naber et al., 2013), emotional memories (Sterpenich et al., 2006), and voices (Papesh et al., 2012).

Across four experiments, we applied pupillometry in standard mixed-list production paradigms where participants read some words aloud and some words silently. We expected that relative to silent reading, production ought to be associated with greater pupil dilation. This finding would provide a psychophysiological signature of the production effect reflective of additional attention dedicated to processing aloud words. In the taxonomy of attentional processes indexed by pupillometry as laid out by Strauch et al. (2022), this attention would likely be primarily at the executive control level due to coordination of motor output when speaking, monitoring of pronunciation, auditory feedback from one's own voice, and so on. Put simply, reading aloud should influence state-level fluctuations of focused attention on the word being pronounced.

Experiment 1 provided a simple demonstration of this physiological signature using just the aloud and silent conditions. Subsequent experiments built on this finding. Experiment 2 added a control condition in which participants repeated the same irrelevant word on each trial. In Experiment 3, pre-cuing with the task instruction prior to target word onset permitted observation of how participants react when they know whether the upcoming word will be produced aloud. Finally, Experiment 4 required participants to withhold responding until a "Go" signal was presented, allowing separation of processes involved prior to versus during the productive response itself.[1]

5. Experiment 1: reading aloud versus silently

In Experiment 1, we measured variation in pupil size during a production task to determine whether aloud trials involved greater attention than did silent trials. If production enhanced attention, then pupil size should be greater on aloud trials than on silent trials. Further, if participants often are inattentive during silent trials, then there should be a negative drift in pupil size following word onset, possibly indicating mind-wandering or other task-irrelevant thoughts (Grandchamp et al., 2014; Smallwood et al., 2011; Smilek et al., 2010; although see Franklin et al., 2013). Should the predicted pupillometric production effect represent processes necessary for the behavioral memory benefit to occur, we predicted a correlation between the aloud – silent difference in recognition performance and the aloud – silent difference in pupil size. We planned to explore this association both as an individual difference measure across participants and at the level of individual study trials.

5.1. Method

Participants. A minimum target sample size was set at 48 participants but data collection was set to continue until the end of the academic semester. Participants self-selected to participate in the study. All reported normal vision without requiring glasses or contact lenses. In the end, 58 University of Waterloo undergraduate students each took part in a single session in exchange for course credit.[2] These data were collected in the Winter of 2019.[3] The procedures received approval from the Office of Research Ethics (Protocol #32011).

Materials and Apparatus. Word stimuli were obtained from the English Lexicon Project (Balota et al., 2007). A master list was formed containing 160 words, each five letters long and most being one syllable ($M = 1.35$, $SD = 0.52$). The selected words were of average to high frequency ($M = 177$, $SD = 200$; Kučera & Francis, 1967), and were fairly concrete ($M = 3.68$, $SD = 0.94$). All words were presented in upper case using Arial 20 pt. font and were centered on a grey background. Fixation crosses were presented in Times New Roman 55 pt. font.

Colored fonts were used to indicate whether a participant should read a word aloud or silently during the study phase: either purple (hex color #950064) or green (hex color #005A00). Fixation crosses and words on the recognition test were presented in blue (hex color #0D00FF). These colors were chosen specifically to maintain a 2.16 luminescence contrast ratio relative to the grey (hex color #808080) background used throughout the experiment. The colors used to indicate aloud versus silent study conditions, as well as the assignment of a given

[1] Experiment 1 was designed by the second and last author; Experiments 2, 3, and 4 were designed by the remaining authors. This explains the method changes between Experiments 1 and 2. That Experiments 1 and 2 were run independently using different methods and populations, yet produced substantially the same results, is a strength of the present work.

[2] Demographics were gathered but are no longer available, having been lost for all four experiments because of team members changing positions and research groups moving laboratory spaces in the period surrounding the global pandemic.

[3] Although data collection was completed after Experiment 2, this is presented as the first experiment in the series because its design differed from those of Experiments 2–4, all of which were carried out at a different institution.

word to serve as a target or a lure, were counterbalanced across participants.

Eye-tracking and pupillometry recording was conducted using an SR Research EyeLink 1000 Plus (v.508) desktop model with a monocular lens, sampling at 2000 Hz. For the duration of the experiment, participants remained in a chin rest and forehead brace about 45 cm from the screen. Eye-tracking monitored the dominant eye, ipsilateral to the participant's self-reported dominant hand. An invisible but centrally placed 170 × 80-pixel rectangular interest region was used for trial progression and to track whether the participant's gaze was focused on trial words. The experiment was presented on a 1920 x 1080p screen refreshing at 60 Hz. Lighting in the testing room was kept to a dim level to allow for proper expansion and contraction of pupils. Responses were made using a standard QWERTY keyboard.

Procedure. Following informed consent, participants were told that they would be studying a list of words for a later memory test. They were instructed to read words presented in one color aloud and words presented in the other color silently without moving their lips. Participants were encouraged to remember all words regardless of their color. After basic task instructions were conveyed, calibration was performed (HV9 type with 1-s pacing interval). Following calibration, a final reminder about color-task assignments was provided.

During the study phase, 80 words were presented one at a time in a fully randomized order, with trial types randomly intermixed; all reported experiments were similarly randomized. At the beginning of each study trial, a blue fixation cross appeared at the center of the screen. Once a participant's gaze was in the interest area surrounding the fixation cross (but after at least 250 ms had elapsed), the trial proceeded and the fixation cross was immediately replaced with a target word for 2 s. Depending on the color of the target word, the participant read the word aloud or silently. Each trial then ended with a 500-ms blank screen. A schematic representation of study phase events and timings is provided in Fig. 1.

Once all study words had been presented, participants completed another eye-tracking calibration and then began the recognition test phase. Here, participants were presented with all 160 words from the master stimulus list one at a time in a fully randomized order (80 targets plus the remaining 80 words that served as lures) and were tasked with responding with the 'n' key if the word was 'new' (i.e., not studied previously), or the 'm' key if the word was 'old' (i.e., remembered from the study phase). Each test trial began with a blue fixation cross at the center of the screen. Once a participant's gaze was in the interest area surrounding the fixation cross (but after at least 250 ms had elapsed), the test trial proceeded. The fixation cross was immediately replaced with a target word that remained on the screen until a keyboard response was made. Each recognition trial ended with a 250-ms blank screen.

Pupil Signal Processing. Specifics pertaining to pre-processing steps associated with pupillary data are discussed in detail in the Online Supplement. In short, our workflow followed the same general progression for each experiment: blink removal, artifact rejection, rejection of samples (i.e., individual timepoints within the continuous data set) where participants were looking too far from center, interpolation of missing samples, low-pass filtration, epoching, subtractive baselining,

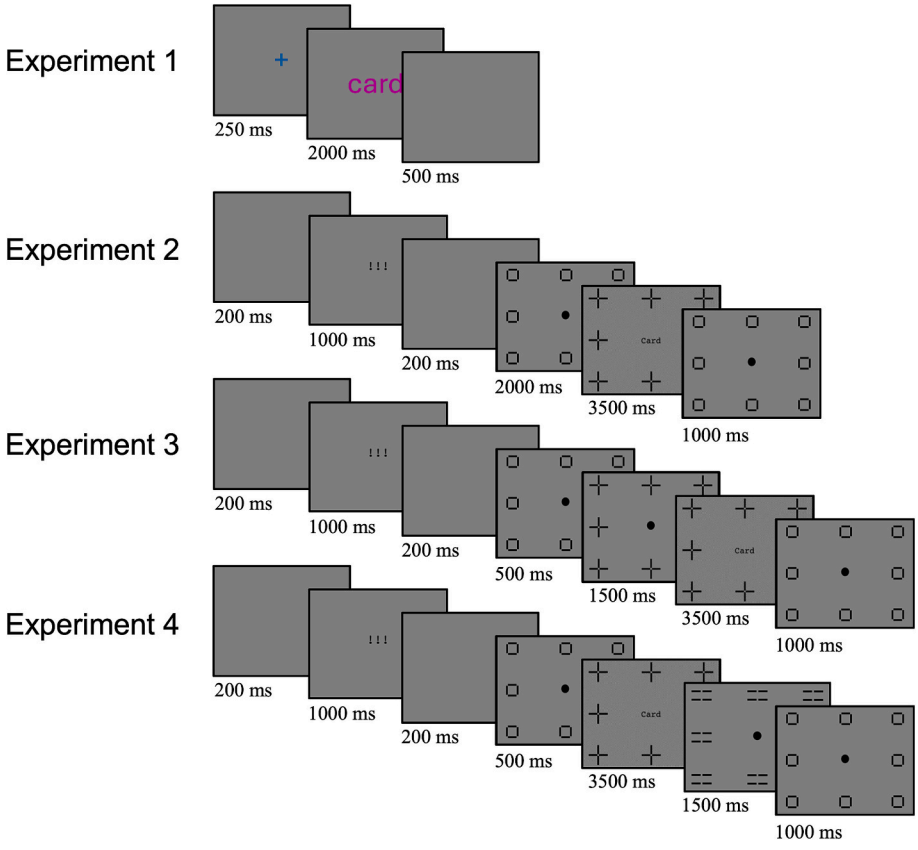

Fig. 1. Schematic representation of study trial events and event timings for the aloud condition in each of the four experiments.
Note: In Experiment 1, equiluminant word color (purple/green) indicated whether to read the word aloud or silently; further, the duration of the fixation period at the start of each trial varied, lasting until fixation was maintained for at least 250 ms. In Experiments 2–4, "!!!" denoted a blink period, during which participants were encouraged to blink; a border made of boxes was used to maintain equiluminance before and after critical trial screens; "+" signalled participants to read the word aloud, whereas similar borders made up of "X" signalled participants to read silently, and "✓" signalled participants to say "Check" aloud. In Experiment 4, participants withheld any response until the "Go" signal represented by a border made of "=" appeared. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

down-sampling to 50 Hz, within-subject *z*-score normalization, "bad" trial rejection (owing to too much missing data), and participant removal if too many trials were flagged as "bad." All statistical models across all experiments were conducted using these *z*-score normalized data.

Statistical Approach. Details of our statistical approach are presented in the Online Supplement. To summarize, recognition data were modelled using multilevel Bayesian probit regression via the *brms* package (v. 2.22.0; Bürkner, 2017, 2018, 2021) within *R* (v. 4.4.0; R Core Team, 2024). This permitted us to report our findings as memory sensitivity (*d'*) while still taking participant and item variation into account. Experiments 2–4 collected confidence ratings at test, permitting us to estimate (and therefore analyze) both familiarity and recollection using Receiver Operating Characteristic (ROC) curves, following the methods popularized by Yonelinas (1994, 1997, 2001) (for an application within the production effect literature, see Fawcett & Ozubko, 2016). In this approach, asymmetry in ROC curves within recognition memory is attributed to a threshold recollection process that contributes disproportionately to high-confidence recognition decisions, whereas the otherwise graded shape of the ROC reflects a continuous familiarity signal. Because the familiarity and recollection estimates from these ROC models almost always agreed with our analysis of *d',* they are reported in our figures but otherwise mentioned only when they deviated from our main analyses.

Bayesian correlations were also fit comparing the magnitude of the pupillary production effect (aloud > silent based on mean pupil size) to the behavioral production effect (aloud > silent in *d'*) using *brms* following evaluation for outliers using two modern approaches (with data excluded only when a given point was identified as anomalous by both). Based on reviewer feedback, additional correlations compared average pupil size in each condition to average memory performance in that condition (as measured via *d'*). Note that in our Bayesian models a difference is considered credible if 0 is excluded from the confidence intervals of the relevant difference, akin to using confidence intervals to gauge significance in frequentist models.

Study phase pupil data were analyzed using frequentist generalized additive mixed models (GAMMs) with thin-plate regression splines via the *mgcv* package (v. 1.9–1; Wood, 2017) with random curves for both participant and item. For each model, we used a *k* (basis dimension) of 12 and verified model adequacy via visual inspection of smooths and residual plots. We confirmed that effective degrees of freedom (*EDF*) were below their specified *k* values, indicating no evidence of overfitting. Underfitting was similarly ruled out using the *gam.check* function. Overall, this is a more conservative approach than often is used in this literature but again permits confidence in generalizing to new participants and items (assuming properties like those used in this study). For each experiment, an additional model was fit exploring the waveforms separated not only by condition but also by test phase accuracy to determine whether, within a given condition, the waveforms differed as a function of accuracy. However, these models failed to observe significant differences between recognized and non-recognized trials (within a given production condition) for any experiment and therefore are not reported below.

Finally, an exploratory mass univariate approach was used comparing each condition at each time point and determining whether a 100-ms moving average window surrounding that time point predicted subsequent memory outcomes. For the latter, multilevel (logistic) models were again used, incorporating random intercepts and slopes for participant and word. Both were corrected for multiple comparisons using cluster-based *p*-values derived by simulating time-series from a hypothetical *Null* distribution for the relevant test statistics (correcting for false discovery rate produced similar, albeit far more conservative, results).

5.2. Results

Prior to analysis, 13 participants were excluded for having too many 'bad' trials. Although this exclusion rate seems high, it is directly attributable to the short duration of the trial and the lack of a suitably long blink period. Our subsequent experiments addressed this issue by including an explicit blink period. The remaining 45 participants were included in our analyses. For those participants that remained, an average of 14.1 % (SD = 14.7 %) of trials were excluded as 'bad' and on average 13.0 % (SD = 5.7 %) of the samples within each retained trial were imputed owing to missing data (e.g., blinks).

For each experiment, all raw behavioral and study-phase pupillary data, as well as our pre-processing and analysis scripts, are available on the Open Science Framework (OSF; https://osf.io/9mqhd/).

5.2.1. Recognition memory

Hits and false alarms are provided in Table 1 for all experiments. As depicted in Fig. 2, a consistently credible behavioral production effect was observed: Participants exhibited greater memory sensitivity (*d'*) for aloud words than for silent words in all four experiments.

5.2.2. Pupil size by condition

We next conducted our generalized additive mixed model and mass univariate analysis of the study phase pupil data. As depicted in the rug plots of Fig. 3, both analyses showed an initial increase in pupil size as the word and instruction were processed, followed by a gradual decline in pupil size for silent trials and a sharp, positive deflection for aloud trials. This resulted in a large pupillary production effect (aloud > silent) that was significant beginning around 500 ms following word and instruction onset and that lasted until the end of the trial. That the effect emerged ~500 ms following stimulus onset aligns with the typical time course expected of changes in pupil size related to cognitive mechanisms, in particular attention.

5.2.3. Predicting memory outcomes using study phase pupil dilations

We next calculated mean pupil size for each trial from 500 ms after instruction onset until trial end and used this to predict later memory accuracy (i.e., "hits") on a trial-by-trial basis (i.e., relating pupil size on a given study trial to the probability of later recognizing the same word during the memory test). Using the mean of the entire trial in this way failed to produce a significant effect for either aloud, $B = -0.02$, 95 % CI [−0.10, 0.06], or silent, $B = 0.01$, 95 % CI [−0.05, 0.08] trials. Our exploratory model, using a moving 100-ms window, did produce a window from 380 ms to 670 ms for which larger pupils were predictive of subsequently recognizing the word for silent trials. No time window was predictive of memory for aloud words.

5.2.4. Correlating the behavioral and pupillometric production effects

Our final analysis evaluated whether there was a correlation between the behavioral production effect (defined as the difference in *d'* between words studied aloud versus words studied silently) and the pupillometric production effect (defined as the difference in mean pupil size between aloud and silent study phase trials) at the participant level. Whereas the models reported in the preceding paragraph were conducted using trial-level pupil dilation to predict later "hits," here the models evaluated the association between mean differences in pupil dilation and mean

Table 1
Experiments 1, 2, 3, and 4: Mean hits (%) for each condition (aloud, silent, control) as well as overall false alarms (%) (with standard errors in parentheses).

Experiment	Aloud	Silent	Control	False Alarms
Experiment 1	74.3 (1.9)	56.3 (2.2)	–	23.7 (1.5)
Experiment 2	82.0 (1.6)	62.0 (2.1)	57.3 (2.1)	34.6 (2.2)
Experiment 3	83.6 (1.4)	66.1 (1.7)	60.1 (1.8)	29.1 (1.3)
Experiment 4	80.3 (2.1)	61.4 (2.5)	58.8 (2.0)	34.9 (1.9)

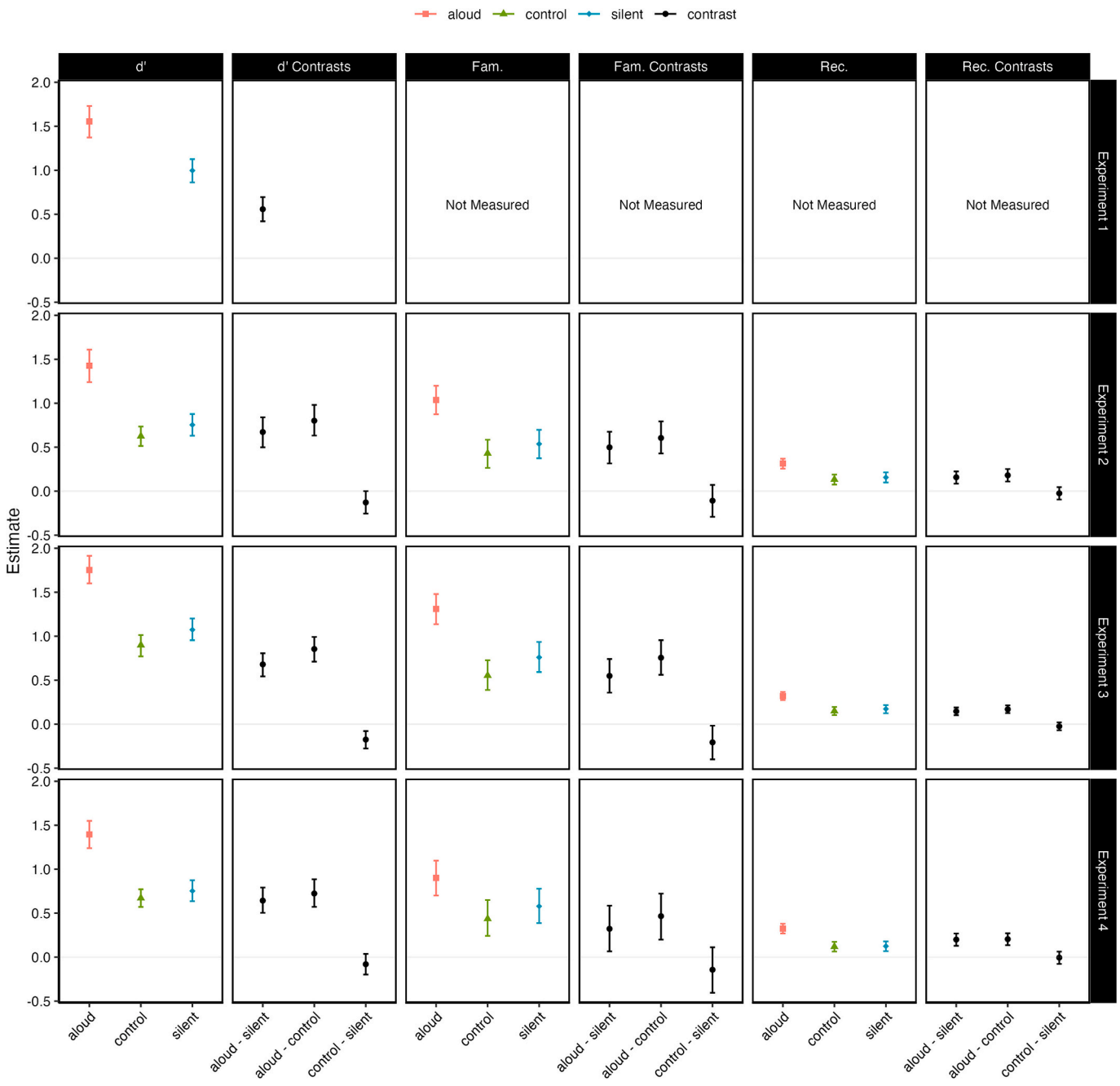

Fig. 2. Experiments 1, 2, 3, and 4: Memory sensitivity (*d'*) and associated contrasts estimated from a multilevel probit regression model as a function of condition (aloud, silent, control) and experiment.

Note. Error bars represent 95 % confidence intervals. Comparisons are considered credible if the confidence intervals for the respective contrast do not cross 0. Rec. = recollection, Fam. = familiarity.

differences in memory performance (*d'*) averaged across trials within a given participant. As depicted in Fig. 4, participants with larger pupillometric production effects also exhibited larger behavioral production effects, r = .37, 95 % CI [.07, .66]. To decompose this association further, we next correlated mean memory performance (*d'*) for each condition with the corresponding mean pupil dilation during their matching encoding trials. As depicted in Table 2, the correlation between behavioral and pupillary production effects appears to be driven by a positive association between memory performance and pupil size during aloud trials, r = .32, 95 % CI [.03, .60]; there was no evidence of association between memory performance and pupil size during silent trials, r = −.02, 95 % CI [−.31, .27].

5.3. Discussion

Experiment 1 provided, to our knowledge, the first demonstration of a pupillometric production effect. During aloud trials, a large positive deflection was observed starting at roughly 500 ms following word onset. This timeframe is typical of changes in pupil size related to cognitive processes and likely reflects an increase in engagement or processing load associated with encoding or producing a word in the aloud condition. Conversely, for silent words, there was an initial increase in pupil size followed by a precipitous decline for the remainder of the trial. We interpret the increase in pupil size during aloud study trials as reflecting greater attention paid to aloud words, whereas the drop in pupil size for silent words (well below baseline levels) may

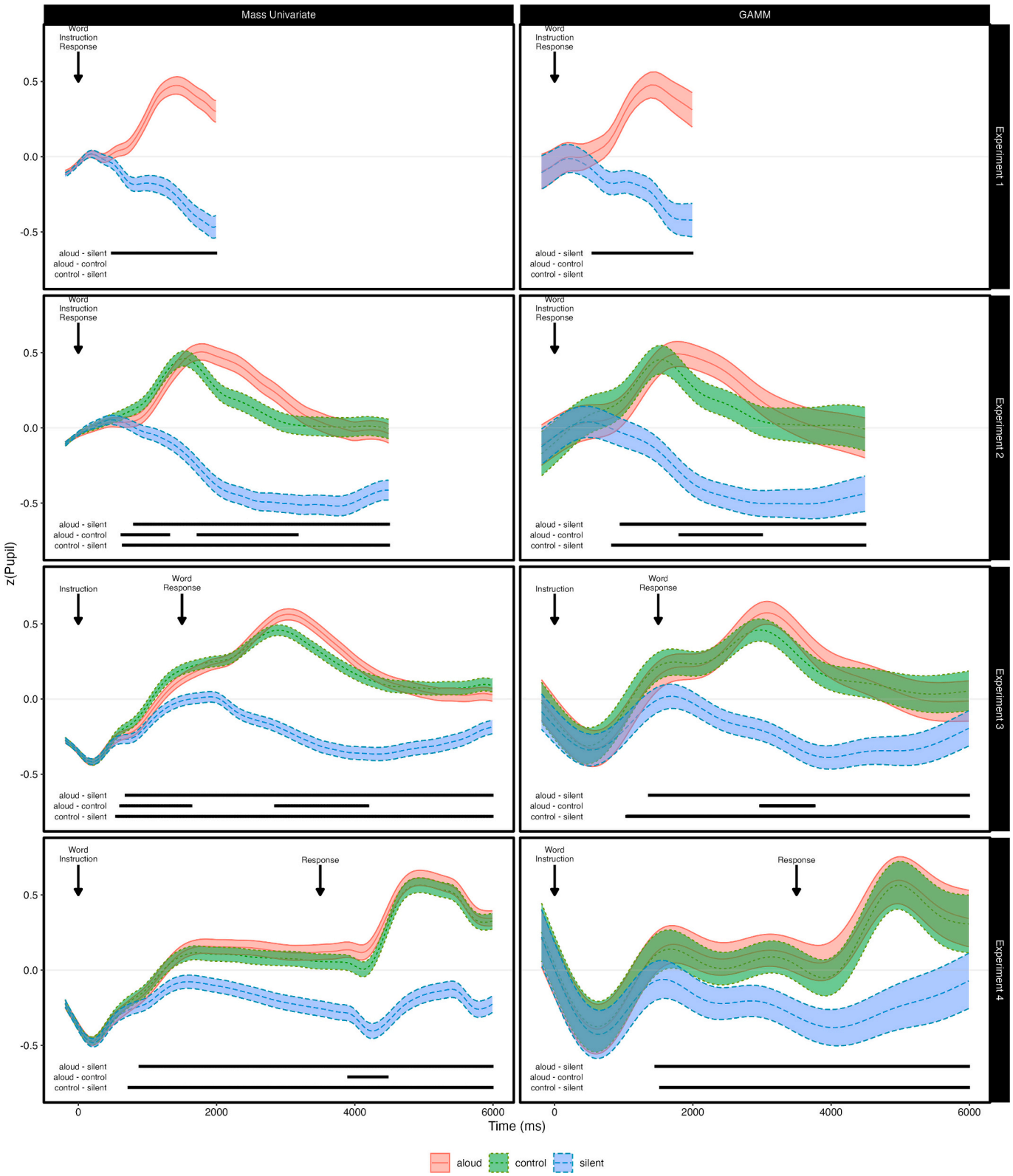

Fig. 3. Experiments 1, 2, 3, and 4: Normalized pupil size as a function of time (ms), modelling approach (mass univariate, generalized additive mixed model), condition (aloud, silent, control), and experiment.

Note. Columns reflect two different analysis approaches; rows represent individual experiments. 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. Generalized additive mixed models (GAMM) reflect predictions from a multilevel 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.

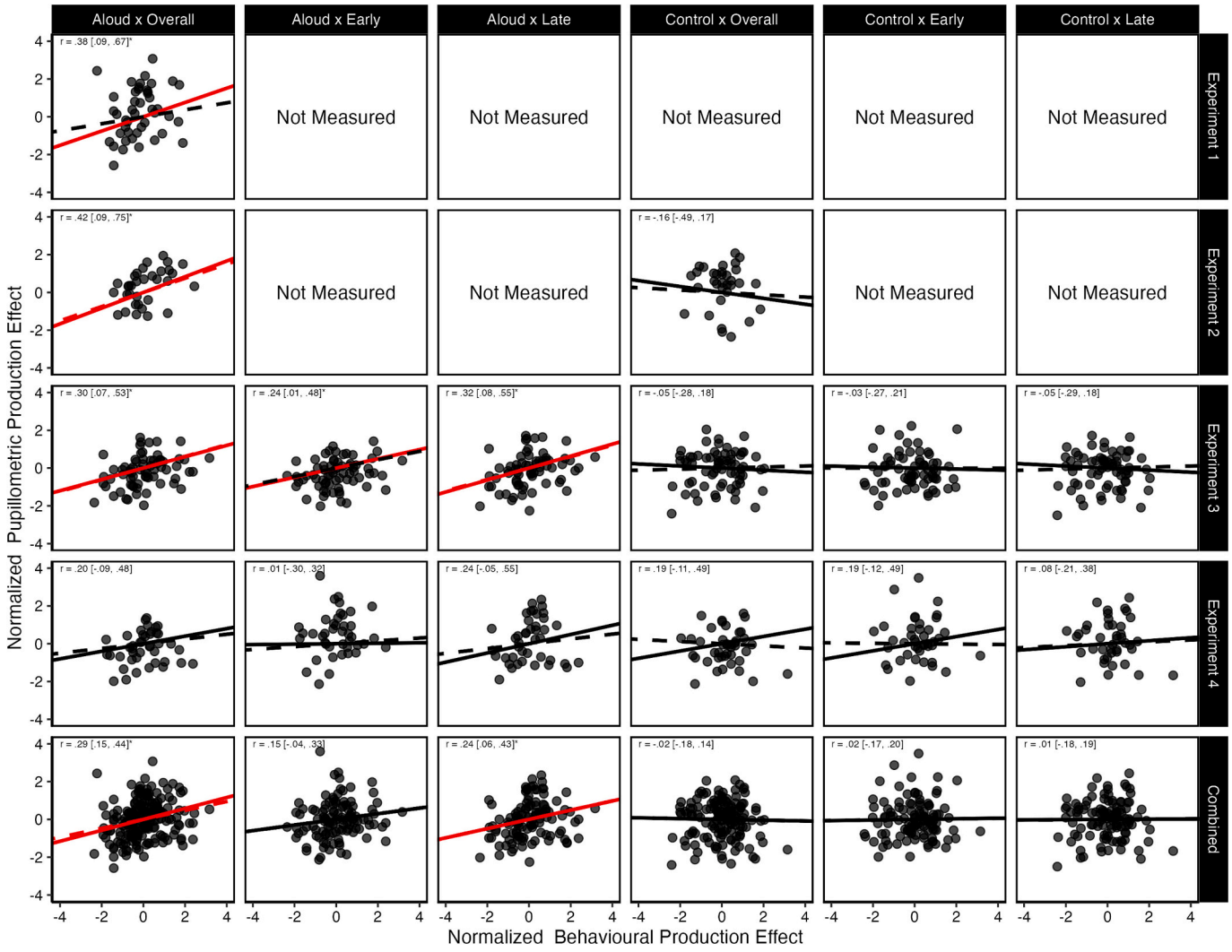

Fig. 4. Experiments 1, 2, 3, 4, and combined: Normalized behavioral (measured as *d'*) and pupillometric (measured as normalized pupil size) production effects and corresponding control effects as a function of measurement window (overall, early, late) and experiment.
Note. Columns are titled based on whether the production effect (aloud – silent) or control effect (control – silent) is reported, and whether the pupillometric effect is based on the whole trial (overall), early time window, or late time window. For Experiment 3, early refers to 500–1500 ms (instruction period) and late refers to 1500–5000 ms (word and response period). For Experiment 4, early refers to 500–3500 ms (word and instruction period) and late refers to 3500–6000 ms (response period). Correlations following outlier correction are stated in the top left corner (* indicates a credible relation) and plotted as a solid line; correlations preceding outlier correction are depicted via the dotted lines. All credible correlations are plotted in red.

indicate that participants became relatively disengaged during the silent trials. Given that this is the first demonstration of the pupillometric production effect, we reserve further consideration until the General Discussion.

6. Experiment 2: reading aloud, reading silently, or making an unrelated response

Experiment 1 provided a basic test of our hypothesis that pupil size would increase during the encoding of aloud trials, while simultaneously revealing a precipitous decline in pupil size for silent trials. This pupillometric production effect suggests that participants pay more attention when reading aloud than when reading silently. Confirming this, the pupillometric pattern was significantly correlated with the magnitude of the behavioral production effect.

Experiment 2 replicated this finding in a paradigm incorporating two enhancements. First, in addition to aloud and silent trials, control trials were added in which participants studied individual target words but spoke the same irrelevant word ("check") on each trial. This control condition was included to account for task-irrelevant sensorimotor stimulation (cf. Bailey et al., 2021; MacDonald & MacLeod, 1998). We were interested in how a repeated spoken response, in lieu of the actual word presented on that trial, would influence the pupillometric pattern, and whether this pattern would predict later memory performance. We predicted that the control condition would exhibit behavioral performance comparable to the silent condition (cf. MacLeod et al., 2010) and a pupillometric pattern smaller in magnitude than that in the aloud condition. The second modification involved extending trial duration and incorporating a blink period. Lengthening the trials permitted more time for the pupillary response to form given that in Experiment 1 the production-related deflection appeared truncated; adding the blink period also has the advantage of reducing exclusions based on 'bad' trials by allowing the participant to blink prior to trial initiation.

6.1. Method

Participants. Sample size was determined by the number of participants that could be recruited during an academic term, with no

Table 2
Experiments 1, 2, 3, and 4: Correlations (with 95 % confidence intervals in parentheses) between *d*' and mean pupil dilation for each condition (aloud, silent, control). *N* reflects total sample size and *n* reflects the number of participants included in each model following exclusion of outliers.

	Aloud			Control			Silent		
Experiment	Overall	Early	Late	Overall	Early	Late	Overall	Early	Late
Experiment 1 (*N* = 45)	.32 [.03, .60] *n* = 45	–	–	–	–	–	-.02 [-.31, .27] *n* = 42	–	–
Experiment 2 (*N* = 33)	.45 [.12, .77] *n* = 31	–	–	.31 [-.03, .64] *n* = 32	–	–	-.27 [-.60, .05] *n* = 32	–	–
Experiment 3 (*N* = 68)	.20 [-.03, .43] *n* = 67	.36 [.12, .60] *n* = 64	.11 [-.13, .35] *n* = 65	-.19 [-.43, .05] *n* = 64	.14 [-.10, .37] *n* = 65	-.25 [-.49, -.01] *n* = 65	.11 [-.12, .35] *n* = 66	.16 [-.08, .40] *n* = 64	.02 [-.21, .25] *n* = 66
Experiment 4 (*N* = 42)	.17 [-.12, .46] *n* = 41	-.09 [-.39, .20] *n* = 41	.29 [.01, .59] *n* = 41	.40 [.10, .71] *n* = 38	-.02 [-.02, .26] *n* = 41	.09 [-.21, .38] *n* = 41	-.12 [-.42, .18] *n* = 41	-.05 [-.34, .25] *n* = 41	-.12 [-.41, .17] *n* = 41
Combined (*N* = 188)	.27 [.12, .41] *n* = 182	.12 [-.06, .32] *n* = 188	.11 [-.07, .30] *n* = 188	.04 [-.13, .20] *n* = 188	.20 [.01, .39] *n* = 188	-.19 [-.37, -.01] *n* = 188	-.07 [-.21, .08] *n* = 182	.17 [-.01, .37] *n* = 188	-.06 [-.25, .13] *n* = 188

Note: Sample sizes used for each model differ based on which data were identified as outliers based on the automated approach described in-text. Credible correlations excluding 0 are highlighted.

minimum. In the end, 36 undergraduate students at Memorial University of Newfoundland (MUN) participated for an hour in exchange for either 1 credit point toward a psychology class or $10 CAD. Participants were asked not to wear facial or eye makeup which interfere with the accuracy of the pupillometric equipment. Unlike in Experiment 1, participants were allowed to wear glasses. Data from three participants were excluded prior to processing, two because their glasses interfered with the eye-tracker (both also had elevated 'bad' trial counts) and one because the tracker conflated their thick eyelashes with the pupil. These data were collected between Fall 2017 and Winter 2018. The procedures for this and all following experiments received approval from the Interdisciplinary Committee on Ethics in Human Research (ICEHR; Protocol #20171484-SC).

Stimuli and Apparatus. This experiment was created using the Experiment Builder (SR Research, 2020) software package developed for the EyeLink 1000 Plus (SR Research, 2010). The experiment was displayed using a MacMini computer running OSX 10.12 with a 22″ 1020 × 768-resolution BenQ monitor. All stimuli were presented in black (RGB: 0, 0, 0) size 24 Courier font on a grey (RGB: 128, 128, 128) background.

Pupillary data from the participant's right eye were recorded at a rate of 500 Hz using a desk-mounted EyeLink 1000 Plus comparable to the one used in Experiment 1. The eye-tracking hardware was placed below the monitor and a chin rest was installed on the desk to minimize head movement during the study. The chin rest was positioned so that the participant's eyes were 105 cm away from the display screen and their forehead was against the forehead rest. The distance from the eye-tracker camera to the chin rest was 50 cm. Participants responded using a standard QWERTY keyboard.

Stimuli were 200 words taken from the MRC Psycholinguistic Database (Coltheart, 1981). These words were subdivided into five randomized lists of 40 words. Lists were matched for word length (M = 5.42, SD = 1.28) and Kucera-Francis written word frequency (M = 63.76, SD = 81.02; Kučera & Francis, 1967). To control for word-level variability, the five lists were counterbalanced across all conditions (aloud, silent, control, and two foil lists). Counterbalancing and randomization were programmed using a custom script run using *PsychoPy3* (Peirce et al., 2019).

Production instructions (aloud, silent, check) were delivered via a box (720 pixels × 540 pixels) at the center of the screen. The perimeter of the box was made up of a series of 'x' for silent trials, '+' for aloud trials, or '✓'for "check" trials. Each symbol was comprised of four identical straight lines, reconfigured to create the desired shape, ensuring that they were matched for luminance (see Fig. 1). A fourth box was made from small squares, also matched for luminance, and acted as a neutral interstimulus placeholder between trials to maintain equal luminance between screen changes.

Procedure. The experiment consisted of four phases: a familiarization phase, a practice phase, a study phase, and a test phase. Calibration of the eye-tracking device was completed prior to each phase to ensure precise measurement of the pupil. Calibration and validation used the same protocol as in Experiment 1.

In the familiarization phase, participants were acquainted with the study procedure. Each instruction cue (i.e., the box made up of symbols) was presented with an instruction in the center noting its meaning: 'x' with the instruction "Read the word silently," '+' with the instruction "Read the word out loud," and '✓' with the instruction "Say 'Check'." For all conditions, participants were asked to make the respective response once. Each of the instruction cues was presented three times for 6 s. In the "check" condition, participants were also orally instructed to read the word on the screen silently to themselves in addition to making the overt response.

Next came the practice phase. Participants were asked to restate the instructions from the familiarization phase to ensure understanding. If they could not remember the instructions, the researcher repeated them. The practice phase followed a format identical to that of the study phase described below, with calibration and validation completed prior to the presentation of the practice list. The practice phase was shorter than the study phase, containing only 9 three-letter words (e.g., 'CAT, 'CAR', and 'BAR') each repeated three times. This shorter list was presented to ensure that the participant understood the instructions and could perform the tasks.

Immediately preceding the study phase, the researcher recalibrated the eye tracker. As depicted in Fig. 1, each study phase trial began with two blank screens for 200 ms interposed by a 1-s "blink" screen—three exclamation marks ('!!!') presented at the center of the screen. Participants were instructed to blink if needed during this period (and to avoid blinking outside this period) so that they could keep their eyes open for the remainder of the trial. The target word was presented inside an instruction cue box (i.e., '+', 'x', '✓') for 3500 ms. Participants were to implement the appropriate task during this time. To control for changes in luminance that would affect pupil dilation, a placeholder screen with a fixation dot that matched the pixilation of the cue screen was displayed for 2 s before the pre-cue and 1 s after stimulus presentation. The fixation screen served as a baseline period for analysis purposes. To foreshadow, although trial events differed across our experiments, the same fixation period was used in each case to ensure cross-condition and cross-experiment comparability because participants are thought to be engaged in similar processing during these standard baseline periods.

The study phase was followed by a brief break during which the experimenter provided instructions for the test phase, and recalibration

and validation were again completed with the eye tracker. Each test trial began with two blank 200-ms screens surrounding a 1-s blink screen during which participants were again instructed to blink if needed. Next, a fixation dot was displayed in the center of the screen for 2 s and then replaced by the test word for 3 s. After the test word disappeared, participants made their responses using the numerical keys (1–6) at the top of the keyboard. Participants were asked to indicate whether they had seen the word before using the following scale: 1 – very sure new, 2 – mostly sure new, 3 – unsure new, 4 – unsure old, 5 – mostly sure old, and 6 – very sure old. The scale was displayed on each trial following the removal of the target word and remained until the participant made a self-paced response.

6.2. Results

All participants that would have been excluded owing to too many 'bad' trials were already excluded for other reasons. Of the 33 participants that remained, an average of 1.8 % (SD = 3.7 %) of trials were excluded as 'bad'; on average, 5.9 % (SD = 4.6 %) of the samples within each retained trial was imputed owing to missing data (e.g., blinks).

6.2.1. Recognition memory

As in Experiment 1, we first analyzed recognition phase performance. There were two major differences from Experiment 1. First, the response was now a confidence-based recognition judgment. As a result, responses were dichotomized such that 1, 2, or 3 became "No "and 4, 5, or 6 became "Yes." Confidence ratings were further used to estimate recollection and familiarity. As depicted in Fig. 2, there was a credible behavioral production effect: Participants exhibited greater sensitivity (d') for aloud words than for words from silent or control trials. Unexpectedly, however, participants exhibited credibly *worse* performance for words from control trials than for silent words (this was only a trend for our analyses of recollection and familiarity).

6.2.2. Pupil size by condition

As depicted in Fig. 3, analyses of the aloud and silent waveforms replicated the results observed in Experiment 1, with a slight increase in pupil size coinciding with word onset in both conditions followed by a rapid decline for silent trials and a rapid increase for aloud trials. The control condition presented similarly to the aloud condition, differing in peaking earlier and more shallowly and dropping off more quickly. Whereas the silent condition differed significantly from both the aloud and control conditions for most of the trial, the aloud and control conditions differed only from ~1800 ms to ~3000 ms.

6.2.3. Predicting memory outcomes using study phase pupil dilations

Using average pupil size across an entire trial to predict memory performance (i.e., "hits") for individual words during aloud, silent, and control trials once again failed to produce a significant effect for any condition, $B_{aloud} = 0.05$, 95 % CI [−0.04, 0.14], $B_{silent} = 0.02$, 95 % CI [−0.06, 0.09], $B_{control} = -0.01$, 95 % CI [−0.09, 0.07].

As in Experiment 1, we next conducted a similar model using a 100-ms moving window. Here, pupil size at encoding predicted later memory for silent trials between 210 and 270 ms, and for aloud trials between 2340 and 2690 ms, each such that larger pupils were associated with a higher probability of recognizing the word on that trial. In this case the window for silent trials failed to survive cluster-level correction.

6.2.4. Correlating the behavioral and pupillometric production effects

Finally, we correlated the size of the behavioral production effect (aloud - silent) as measured via d' for each participant with the size of that participant's pupillometric production effect (averaging across trials). For comparison, we did the same for the control effect (control - silent). As depicted in Fig. 4, whereas the correlation between the control < silent effect and its corresponding pupil deflection was non-credible and, if anything, negatively associated, $r = -.15$, 95 % CI [−.49, .19], the correlation between the behavioral and pupillary production effects was again credible: Participants exhibiting a larger aloud > silent deflection in pupil size also showed a larger production effect in memory as measured using d', $r = .42$, 95 % CI [.08, .74]. This correlation of behavioral and pupillary effect sizes was again driven primarily by a positive association between memory performance and pupil size during aloud trials, $r = .45$, 95 % CI [.12, .77], although in this case there was also a non-credible trend toward a negative association between memory performance and pupil size during silent trials, $r = -.27$, 95 % CI [−.60, .05].

Interestingly, a further exploratory correlation showed that the pupillary production effect (aloud > silent) and the corresponding control > silent pupillary effect were strongly correlated, $r = .77$, 95 % CI [.55, .87]. This suggests that participants who demonstrate especially large pupillary deflections during aloud (as compared to silent) study trials also demonstrate especially large pupillary deflections during control (as compared to silent) trials, despite the latter providing no memory advantage for the words on those trials. Because we interpret the former as reflecting attention to the target word being pronounced aloud, and the latter as reflecting attention directed toward a competing verbal response (i.e., say 'check') at the expense of the target word, we see this pattern as compatible with our hypotheses. That said, there are convergent but not mutually exclusive reasons why this correlation exists. For instance, it could arise from shared sub-processes of vocalization (e.g., speech motor effort, response anticipation, or action monitoring), or from measurement artifacts owing to jaw and cheek movements subtly altering pupil size estimates.

6.3. Discussion

As in Experiment 1, our predictions were supported. There was a pupillometric production effect—larger pupils on aloud trials than on silent trials. Again, production was associated with a large, positive deflection whereas silent reading was associated with a precipitous decline to well below baseline. The former is consistent with enhanced attention during the aloud trials; the latter is consistent with attentional disengagement or mind-wandering (e.g., Unsworth & Robison, 2018). These findings also gain support from the moving window analysis. For silent trials, greater pupil size following word onset (i.e., remaining attentive after the word was initially encoded) was predictive of later memory for the word; for aloud trials, greater pupil size near the end of the trial (i.e., remaining attentive after the word was read aloud, possibly rehearsing) was predictive of later memory for the word. It is also worth noting that the aloud and silent conditions diverged later in Experiment 2 than in Experiment 1, potentially owing to the use of symbolic cues rather than color cues as the trial type instruction. Symbolic cues may take longer to decode although it is also the case that there were now three possible trial types.

Our predictions concerning control trials were mostly supported. Compared to aloud trials, the pupillary response for control trials peaked earlier and was smaller and of shorter duration. This could be because on control trials participants were able to activate the associated response more readily given that it was being made repeatedly; in contrast, for aloud trials, greater effort was required to process the unique word for that specific trial. More speculatively, it is possible that participants remain engaged with the response longer on aloud trials because control trials reflect a dual task, with time spent lingering on or processing the control response detracting from time spent studying the actual target word. The behavioral data are consistent with this interpretation as it appears that producing the control word impaired memory for the target word on that trial. This behavioral finding was unexpected given that previous research (e.g., Bailey et al., 2021; MacLeod et al., 2010) tended to observe no difference between control and silent trials.

Although the pupillary production effect was associated with a larger behavioral production effect, the pupillary control effect was not credibly predictive of the behavioral control-silent difference; if anything, it

was negatively associated, suggestive again that the dual task interfered with encoding the word. The pupil-memory correlation observed for the pupillary production effect—also observed in Experiment 1—might suggest that participants who are most engaged during aloud trials and least engaged during silent trials (the pupillary production effect), also tend to show the largest production effect in memory.

The control finding is also novel and might potentially speak against a broad attentional account. Specifically, if production simply served to 'alert' participants during trials in which a response was made, similar benefits might be expected for the aloud and control conditions. The fact that repeating the control word produced a similar positive deflection suggests that rather than a general increase in attention or alertness, it matters what attention is being allocated *toward* for memory improvements to occur. In the case of the aloud condition, participants are forced to engage with the target word on a deeper level and may be encouraged to linger on that word, potentially enhancing other forms of processing. In contrast, for control trials, participants are forced to engage with the control word ("check") which seems to harm encoding of the target word by diverting processing away from that word at a critical moment.

Having now observed the pupillary production effect twice, we were next interested in better understanding its constituent components. To do so, we developed two additional experiments designed to separate the processing of the cue, word, and response so that we could consider them separately.

7. Experiment 3: separating instruction from word onset and response

In Experiments 1 and 2, we demonstrated changes in pupil size related to the production condition which suggested that reading aloud is more demanding than reading silently. Additionally, we demonstrated a relation between this pupillary production effect and the behavioral production effect (but no association between the analogous control effects). These first two experiments provide a temporally precise evaluation of production by capturing pupillometric signals of underlying cognitive processes. Nonetheless, these experiments are limited in their ability to support firm conclusions pertaining to the observed pupillometric effects because any putative differences in attentional allocation necessarily overlapped temporally with effects driven by production itself. To further isolate the mechanisms involved, the next two experiments sought to temporally separate the act of speaking from other cognitive processes.

Experiment 3 did this by implementing a pre-cueing procedure based on Bailey et al. (2021) which highlights preparatory processes that might occur when participants know that they will be producing something. By separating motoric production from the initial encoding of the word, we can address whether processing differs in response to distinctive words and whether there also are preparatory differences in initial encoding when a participant knows that they will be producing the word later. Additionally, by separating vocal responses from the word presentation, we can elucidate attentional differences that might occur in response to words that are produced versus those that are not produced. Participants might engage differently with the trials depending on the action that they will be taking. For example, they might show more attentional engagement on trials where they know that they will be speaking, even before they see the word that they are to say (and therefore, prior to the possibility of any form of distinctive processing for the target word).

7.1. Method

Participants. Sample size was determined by the number of participants that could be recruited during two academic terms with no minimum. The recruitment duration was doubled in this case for pragmatic purposes: For Experiment 2, we spent a term programming the task leaving only one term for recruitment, whereas for Experiment 3 the task was already implemented. In the end, 71 students at MUN participated for an hour in exchange for 1 credit point toward a psychology class or $10 CAD. Data from three participants were excluded prior to processing: One became ill during the session, one confused the instructions, and one had an eye condition that led them to close their eyes for most of the study phase (this participant would also have been excluded based on missing trials). Thus, the data of 68 participants were included. These data were collected between Fall 2018 and Winter 2019.

Materials and Procedure. The experimental setup and words were identical to those used in Experiment 2, except that a pre-cueing procedure was implemented. During this pre-cueing phase, a given symbol border was presented around a fixation dot to indicate the trial type to be performed with the target word; 1500 ms later, the word appeared on the subsequent screen. See Fig. 1 for a visual depiction and for the alterations to trial timings. Participants were also explicitly instructed to hold their response on control trials until the word appeared on the screen.

7.2. Results

All pupil pre-processing steps were the same as for Experiment 2 except that epochs were time-locked to instruction onset and ran from −200 ms to 6 s. Once again, participants who would have been excluded owing to too many 'bad' trials were already excluded for other reasons. For the remaining participants, an average of 1.5 % ($SD = 5.7$ %) of trials were excluded as 'bad' and on average 5.0 % ($SD = 4.4$ %) of the samples within each retained trial were imputed owing to missing data (e.g., blinks).

7.2.1. Recognition memory

As depicted in Fig. 2, a credible production effect (aloud > silent) was observed along with a credible control effect (control < silent) as in past experiments. In this case, the control effect was credible for familiarity but not for recollection.

7.2.2. Pupil size by condition

With the instruction now preceding word onset and response, there are two apparent peaks in the waveforms depicted in Fig. 3. Once instructions appeared, there was an initial positive deflection for all conditions. Whereas the mass univariate approach almost immediately identified differences among the conditions (within ~500 ms), the more conservative GAMM models observed a control > silent difference on a similar timeframe but an aloud > silent difference only starting at ~1400 ms (just preceding word onset). Because pupil change resulting from cognitive processes is slow, the differences observed in the period surrounding word onset likely reflect a cognitive response to the preparatory instruction phase, especially given that even perception of the word (let alone higher cognitive processes) would require more time to produce any such change. Following word onset, pupil size in the silent condition dropped precipitously whereas both the aloud and control conditions exhibited sustained peaks associated with the response, again earlier and smaller in the control condition than in the aloud condition.

7.2.3. Predicting memory outcomes using study phase pupil dilations

Using the mean pupil size across each individual trial at encoding to predict later memory performance (i.e., "hits") once again failed to produce a credible effect for any condition, $B_{aloud} = 0.03$, 95 % CI [−0.04, 0.09], $B_{silent} = 0.02$, 95 % CI [−0.03, 0.07], $B_{control} = -0.01$, 95 % CI [−0.06, 0.05]. Our exploratory window analysis did reveal that pupil size for silent trials within ranges from 50–80 ms, 650–880 ms, and 1080–1990 ms (largely the period preceding and surrounding word onset) was predictive of later recognizing the word; however, only the final period survived cluster correction. The same finding emerged in a model separately including the early and late portions of the trial as predictors, showing that—in aggregate—pupil size in the period preceding and surrounding the word credibly predicted memory for silent

words. Neither window predicted later memory for aloud or control trials.

7.2.4. Correlating the behavioral and pupillometric production effects

As depicted in Fig. 4, the pupillary production effect was again correlated with the behavioral production effect, $r = .29$, 95 % CI [.05, .53] whereas the negative control behavioral effect was not correlated with the control > silent pupil effect, $r = -.05$, 95 % CI [−.29, .18]. In this case, the correlation between memory performance and average pupil size trended positive overall for the aloud condition, $r = .20$, 95 % CI [−.03, .43] but was only credible for a time window around instruction onset, $r = .36$, 95 % CI [.12, .60]. No credible association was observed for any of the silent comparisons (see Table 2). A trend favored a negative association between memory performance and average pupil size for control trials, $r = -.19$, 95 % CI [−.43, .05] but this was only credible for a time window surrounding the response, $r = -.25$, 95 % CI [−.49, −.01]. The aloud > silent and control > silent pupil effects were again highly correlated, $r = .69$, 95 % CI [.55, .80].

7.3. Discussion

Experiment 3 replicated the behavioral and pupillary production effects observed in Experiments 1 and 2. Further, the unexpected pattern in Experiment 2 of worse memory performance for control words than for silent words also replicated. This differs from the prior literature. For example, MacLeod et al. (2010) had previously examined production with a nonunique vocal response ("yes") and had observed no such difference in recognition scores (cf. Bailey et al., 2021). We return to this point in the General Discussion.

In the pupillary data, the initial peak associated with instruction onset in each condition suggests that participants engaged in some form of processing preceding word onset. Given that the word was yet to appear, this presumably reflected heightened alertness or preparation. This peak was larger and more sustained for the aloud and control conditions than for the silent condition. Here, differences emerged almost immediately for the control – silent contrast, but more slowly for the aloud – silent contrast (just prior to word onset). It is well known that cognitive effects appear after a delay of ~500–1000 ms in the pupillary signal (Hoeks & Levelt, 1993; Verney et al., 2004), meaning that these differences are attributable to the instruction itself rather than to the word or the response (pupillary effects of which would be observed ~500 ms later). This initial peak was also larger and earlier for control trials than for aloud trials. This is likely because, although participants withheld responding, they already knew the response on control trials and could prepare that response during this period. For aloud trials, they had to wait for the unique word onset to do so.

The fact that differences emerged in response to the pre-cue alone is remarkable from a theoretical perspective because at this stage there is no distinctive target word information to encode. Yet pupillary differences during this period were also, on their own, predictive of the behavioral production effect (see Fig. 4). This supports the idea that these early pupillary responses were preparatory and most likely attentional in nature. Knowing that a response will be required may itself induce a state of heightened engagement, facilitating encoding and a subsequent production effect in memory. Evidence is mixed, though, as to whether this attentional state offers a general memory benefit: In the control condition, pupillary effects have thus far (and in the final experiment) been non-credibly negatively correlated with behavioral performance.

Upon word onset, we again observed a characteristic positive deflection for both aloud and control trials (with aloud trials exhibiting a greater, later peak than control trials) in contrast to a rapid decline in pupil size for silent trials. This again aligns with the general account that participants disengage more quickly and fully from words that they read silently than they do from words that they read aloud. These findings also broadly align with the possibility that during silent trials participants may be engaged in off-task thoughts or mind-wandering, which are associated with a decrease in pupil size (Unsworth & Robison, 2018). On silent trials, participants maintain a minimum level of engagement until word onset and then lapse soon after reading the word.

8. Experiment 4: separating instruction and word onset from response

In Experiment 3, we separated processing of the instruction from processing of the word and response to evaluate whether preparatory mechanisms contribute to the production effect when participants are aware that production is upcoming. Knowing that you will soon produce a word was associated with heightened engagement (for a related idea, see Forrin et al., 2019). However, word onset remained confounded with response, preventing separate evaluation of the pupillometric signal associated with the productive act. In Experiment 4, we modified our task to use delayed production (Hassall et al., 2016; Mama & Icht, 2018b) to separate the processing that occurs during production from the productive act itself. Here, participants received the instruction and word concurrently but had to withhold responding until a "Go" signal was presented.

We sought to address two issues. First, we wanted to separately quantify differences observed *preceding* the productive act from differences observed *during* the productive act to determine which best predicted the behavioral memory benefit. Second, we wondered whether engagement would "drop off" for silent words during the holding period. For silent words, participants were instructed to say the word quietly in their head at the "Go" signal so it was also plausible that a peak would be observed during silent trials reflecting this covert action. Note that, like the design used by Mama and Icht (2018b), the word disappeared from the screen during the Go cue so that participants had to hold the word in working memory prior to producing it. Hassall et al. (2016) previously demonstrated a greater amplitude of the P300b signal in response to the cue and stimuli in a similarly designed delayed production procedure. Given that, and our finding in earlier experiments that pupil size differed across the three conditions (aloud, control, silent) prior to any motoric production, we expected a similar divergence in the pupillary response during the holding period prior to vocalization.

8.1. Method

Participants. Sample size again was determined by the number of participants that could be recruited during an academic term with no minimum. We had intended to recruit again for two terms, but the global pandemic prevented this, and following the pandemic the researchers involved had taken on new positions elsewhere. Thus, 45 students enrolled at MUN either were recruited through the university's psychology pool and received one course credit or were recruited by poster advertisement and were paid $10 CAD. Data from three participants were excluded prior to processing: One accidently started the task prior to calibrating with the eye-tracker, one had thick glasses making it difficult to calibrate, and one was simply missing the file containing their pupil data. The data of the remaining 42 participants were included in the analyses. These data were collected in Fall 2019 and Winter 2020 (ending with the onset of the pandemic).

Materials and Procedure. All materials and procedures were identical to those of Experiment 2 except that an additional "Go" signal was presented following word and instruction onset. Participants were instructed to withhold their verbal response until this signal appeared, or on silent trials to say the word in their head again upon its presentation. See Fig. 1 for a depiction and for timings.

8.2. Results

No participants were flagged as having too many 'bad' trials. Overall,

an average of 1.1 % (SD = 2.9 %) of trials were excluded as 'bad'; on average 4.7 % (SD = 3.7 %) of the samples within each retained trial were imputed owing to missing data (e.g., blinks).

8.2.1. Recognition memory

Replicating our prior experiments, both a credible behavioral production effect (aloud > silent) and a credible control effect (control < silent) were observed. The control effect was not credible for familiarity or recollection separately.

8.2.2. Pupil size by condition

With the overt response now separated from word and instruction onset, two peaks were again observed in the waveforms depicted in Fig. 3. Once the instruction and word appeared, there was an initial positive deflection for all conditions that was greater for aloud and control trials than for silent trials. This was followed by a sustained positive deflection for the aloud and control conditions (likely reflecting retention of the word in working memory until it could be produced), whereas for silent trials pupil size dropped off rapidly. Because within the control condition the memory component (i.e., retaining a word that was not going to be produced) is comparable to the silent condition, this might best be attributed to preparation for the response. This was then followed by a positive deflection for both aloud and control trials upon receipt of the "Go" signal, reflecting the response itself. For silent trials, where participants were to "say the word in their head" upon receipt of the "Go" signal, there was also a modest positive deflection of lesser magnitude, superimposed on the negative deflection generally observed for those trials. For this experiment, no significant differences in pupil size were observed between aloud and control trials for our more conservative GAMM model, although our mass univariate analysis revealed a period of significance from ~3900 ms to ~4500 ms. Pupil size on aloud and control trials was significantly greater than on silent trials throughout.

8.2.3. Predicting memory outcomes using study phase pupil dilations

Using aggregate per-trial pupil size to predict later memory performance for the aloud, silent, and control trials again failed to produce a significant effect for any condition, B_{aloud} = −0.03, 95 % CI [−0.10, 0.05], B_{silent} = 0.04, 95 % CI [−0.02, 0.11], $B_{control}$ = −0.03, 95 % CI [−0.10, 0.03]. Our exploratory window analysis did reveal that pupil size for silent trials within the range 770–1160 ms and for control trials within the range 880–1680 ms was predictive of memory for words in those respective trials. Interestingly, whereas for silent words greater pupil size predicted better memory (as in our prior studies), for control trials greater pupil size predicted worse memory. No windows were predictive of later memory for aloud trials.

8.2.4. Correlating the behavioral and pupillometric production effects

As depicted in Fig. 4, the correlation between the pupillary production effect and the behavioral production effect trended in the predicted direction, but this time was not credible, r = .20, 95 % CI [−.10, .50]. In this case, the per condition correlations again demonstrated a trend toward a positive association for the aloud condition, r = .17, 95 % CI [−.12, .46], with a credible association observed only for a time window surrounding the response, r = .29, 95 % CI [.01, .59], and a negative trend for the silent condition overall, r = −.12, 95 % CI [−.42, .18]. As in past studies, the control effect was not credibly correlated with the corresponding control > silent pupil effect, r = .20, 95 % CI [−.11, .50], although a positive association was observed between memory performance and pupil size in the control condition overall, r = .40, 95 % CI [.10, .71]. The aloud > silent and control > silent pupil effects were again highly correlated themselves, r = .64, 95 % CI [.40, .78].

As this is the final experiment in the series, we undertook an additional correlation model that combined all four experiments (combined N = 188 for Aloud – Silent and N = 143 for Control - Silent). Here, the pupil-behavioral correlation was observed to be credible for the production effect, r = .31, 95 % CI [.17, .46], but not for the control effect, r = −.02, 95 % CI [−.19, .14]. Also of note, the pupillary production effect correlation was apparently driven more by the late time window in Experiments 3 and 4 (reflecting word + response or response alone) than by the early time window (reflecting instruction or instruction + word + holding period). In the combined data, despite a credible positive association between memory performance and pupil size for the aloud condition, r = .27, 95 % CI [.12, .41], no similar trend was observed for the silent, r = .04, 95 % CI [−.13, .20], or control, r = −0.07, 95 % CI [−.21, .08] conditions; however, for the control condition there was a credible positive association observed for the early time window, r = .20, 95 % CI [.01, .39], and a negative association observed for the late time window, r = −.19, 95 % CI [−.37, .01].

8.3. Discussion

The post-cuing procedure in Experiment 4 permitted separate evaluation of the processes involved in encoding the word from those involved in the productive act itself. Separate deflections were observed for each, with aloud and control trials producing similar waveforms (here aloud was slightly, but non-significantly, higher than control throughout) both of which were significantly greater in magnitude compared to the waveform for silent trials. For the silent trials, pupil size initially increased before gradually declining to baseline levels, with a modest peak surrounding the onset of the 'Go' signal (where participants said the word silently in their head). For aloud and control trials, pupil size declined more gradually during the 'holding' period and peaked to a much greater degree upon 'Go' signal onset.

It is noteworthy that the aloud/control > silent pattern in pupil dilation emerged even during the instruction and word onset phases and was sustained throughout the holding period because at this point all conditions had the same processing demands: reading the stimulus, alerting to the condition, and awaiting response. Increased pupil size in the aloud and control conditions compared to the silent condition suggests that participants were preparing to speak even early in the trial and remained at a heightened level of engagement until the response was made. This notion lends support to the hypothesis that there is a preparatory component to the production effect that may reflect differences in attention related to preparing an overt and target-relevant response. It also aligns with earlier work arguing that knowing that an overt response must soon be made has consequences for participant motivation and behavior (e.g., Forrin et al., 2019). This might also explain the imagined production benefit reported by Jamieson and Spear (2014).

The fact that aloud and control trials did not differ with respect to pupil size in our primary model was unexpected. Of course, participants do speak in both cases. Given the additional processing involved in preparing a unique verbal response on aloud trials, however, we had predicted some difference especially given that we observed such differences in Experiment 2. Nonetheless, in the present case, speaking in either condition appeared to require roughly equal attention, with the check response readily "pre-loaded" as soon as the instruction appeared. Whereas reading a unique word benefited memory—potentially by forcing participants to engage specifically with that word—repeating "check" had no such benefit, likely because speaking in this case diverted processing away from the target word. Silent trials once again demonstrated an initial peak at word and instruction onset followed by a decline to baseline, which could reflect disengagement. A second peak occurred following 'Go' signal onset, reflecting momentary re-engagement, but the magnitude was much smaller for silent trials than for aloud or control trials.

In the present case, unlike in our prior experiments, the pupillary production effect was not credibly associated with the behavioral production effect. However, when the Bayesian priors for this analysis were instead mildly informed by the earlier studies, the effect did become credible, r = .22, 95 % CI [.01, .43]. Critically, the pupillometric-

behavioral production effect correlation was credible when the results of the four experiments were combined.

9. General discussion

Production has been presented as a simple encoding technique requiring little knowledge or practice but nevertheless affording a substantial mnemonic benefit. However, effective use requires understanding its underlying processes. Past research has focused largely on whether distinctiveness adequately explains the mnemonic benefit of production (see MacLeod & Bodner, 2017, for a summary); other processes, notably enhanced attention or encoding that could occur in addition to the productive act itself, have been less examined despite having been considered in early studies (e.g., MacDonald & MacLeod, 1998).

9.1. The pupillometric production effect

Across four experiments, we adopted a novel psychophysiological approach using changes in pupil size to unobtrusively quantify processing demands during production and to relate those demands to the behavioral production effect. We consistently observed a robust behavioral production effect (aloud > silent). We also observed an unexpected but consistent control effect (control < silent): Repeating an unrelated control word impaired memory for the target word on that trial. In our pupillary data, reading a word aloud and repeating a control word were each associated with a positive pupil deflection tied to word or instruction onset. Aloud trials tended to demonstrate larger, later peaks than control trials (except for Experiment 4). Reading silently was instead invariably associated with a dilation peak surrounding initial encoding of the word followed by a decline to or below baseline level, with only slight recovery in Experiment 4 when a "Go" signal indicated that the response should be made (i.e., saying the word silently in their head).

In each experiment, the pupillary production effect (aloud > silent) was predictive of the behavioral production effect, including in time windows preceding word onset (although in Experiment 4 this correlation was not credible without priors informed by the preceding experiments). Yet a comparable analysis of the negative control effect tended to produce no such brain-behavior relation, despite the two pupillary effects themselves being correlated. The findings are broadly compatible with classic distinctiveness-based accounts, but they also suggest a role for other processes.

Making any response encourages greater engagement with the task, as demonstrated by both aloud trials and control trials exhibiting greater pupil sizes than silent trials throughout, including periods where there was no distinctive information to encode (e.g., the instruction period of Experiment 3) and when task demands were putatively matched across trials (e.g., the "holding" period of Experiment 4). These differences reflect an increase in pupil size for aloud trials and control trials accompanied by a decrease for silent trials.

Throughout this study we have interpreted greater pupil dilation as reflecting increased attention during aloud and control encoding trials. Beyond attention, however, there is a sizeable literature suggesting that pupil dilation may instead index processes such as cognitive effort (van der Wel & van Steenbergen, 2018) or working memory (Unsworth & Robison, 2015). While these mechanisms are likely at play in any cognitive study, we reason that in the present case pupil dilation is unlikely to be driven by them. Cognitive effort would seem to be an unlikely contributor in our experiments given that we observed similar pupil deflections for the aloud and control conditions: Surely repeating the same word each time is less effortful than reading a new word. The design of our experiments also helped to rule out working memory as crucial: In all experiments (except Experiment 3), the condition assignment and target word were presented in unison such that nothing had to be held in working memory before a response was made. As a result of our systematic progressions of experimental design, these alternative factors are unlikely to be the main drivers of the pupillometric responses that we observed.

Consistent across all experiments, pupil size was not monotonically related to later memory performance. Put simply, there was a clear dissociation between pupil dilation and encoding efficacy. Beginning in Experiment 2 and carrying on through Experiments 3 and 4, we observed that pupil dilation was comparably large in both the aloud and control conditions, yet only the aloud condition yielded a substantial memory benefit. Indeed, the control condition always led to performance worse than silent reading. This could reflect cue overload (Watkins & Watkins, 1975) in that the control trials involve two words—the presented target word and the word "check"—whereas the silent trials involve only the presented target word. More generally, the repeated word "check" could simply be interfering with the presented target word on that trial.

In addition, in Experiments 2 and 3, pupil size in the aloud condition correlated positively with memory performance whereas pupil size in the control condition did not. We argue that increased attentional engagement—which we see pupil dilation as indexing—is by itself not sufficient for successful memory encoding. Rather, what matters is *where* attention is directed. In the control condition, when participants prepare to speak their attention is directed to a repeated, non-informative response ("check") rather than to the target word, yielding pupil dilation without a related memory benefit. We interpret pupil dilation as indexing a heightened state of attention preceding and during speaking—regardless of what is spoken. It is only in the aloud condition, however, that this attention is beneficially directed toward the target word.

This interpretation is consistent with previous work showing that pupil dilation reflects time pressure, arousal, and task urgency rather than encoding efficacy (Gross & Dobbins, 2021; Lloyd & Nieuwenhuis, 2024; Murphy et al., 2016; Robison et al., 2022; Unsworth & Miller, 2020). In addition, the delayed naming literature has used a pre-cueing paradigm like our Experiment 4 and demonstrated that pupil size reflects greater attention beginning with pre-speech planning and carrying through to post-lexical processing, matching the waveforms observed here as well (Goldinger et al., 1997; Papesh & Goldinger, 2012). Therefore, recognizing the distinction between greater pupil dilation and encoding efficacy per se will help prevent misinterpretation in future studies that use pupillometry to index learning and memory.

9.2. Theoretical interpretation

Putting this into the context of contemporary theory, the distinctiveness account (e.g., Conway & Gathercole, 1987; MacLeod et al., 2010) might interpret Experiments 1 and 2 as reflecting cognitive demands associated with distinctive encoding of the study word, preparing the unique response, and binding sensorimotor features (i.e., the production record) to the study episode. These features could then be used to drive memory benefits as proposed by the distinctiveness heuristic account (e.g., Dodson & Schacter, 2001)—where these features are used strategically to guide memory—or as proposed by the relative distinctiveness account (e.g., Jamieson et al., 2016)—where the presence of these features interacts with implicit retrieval processes to give rise to the effect. Either account likewise easily accommodates the finding that repeating a control word aloud produces a similar pupil deflection but of lesser magnitude, and each account would have (correctly) predicted that this deflection on control trials would not predict later memory. This perspective also aligns with Hassall et al.'s (2016) interpretation of the P300 in their ERP study, which was viewed as an indicator of distinctive encoding.

A strict interpretation of the distinctiveness account has greater difficulty, however, explaining the precipitous drop in pupil size for silent trials, as well as the aloud/control > silent pattern observed in pupil size during the instruction period of Experiment 3 and the holding

period of Experiment 4. Concerning the former, the most obvious interpretation is that once participants have read a word in the silent condition, that trial is completed and they become disengaged. This perspective is supported by findings that participants report paying less attention to silent words than to aloud words (e.g., Fawcett & Ozubko, 2016), that they tend to mind-wander more during silent reading (e.g., Varao Sousa et al., 2013), and that pupillary constriction is typical when mind-wandering (e.g., Unsworth & Robison, 2018). Further, the only consistent finding in our exploratory moving window analysis was that increased pupil size during the early portion of the silent trials in each experiment was predictive of later memory for those silent words, suggesting that remaining attentive early in the trial (as evidenced by larger pupil sizes) would predict better memory for silent words, in turn resulting in a smaller production effect.[4]

The pupillometric control effect is another piece of evidence suggesting that simply forcing engagement is insufficient to induce a memory benefit for words studied in the affected trials (see MacLeod, 1975, for an analogous finding in the context of the release from proactive interference paradigm). If it were the case that the behavioral production effect arose because participants were alert during aloud trials and disengaged during silent trials, a memory benefit would be expected for control trials where they were also alerted. However, if anything, memory was worse for words in the control condition than for words in the silent condition. This aligns with Bailey et al. (2021) who observed similar activity in aloud and control trials with no association between activation on control trials and later recognition, just as we saw in our correlation analyses.

Participants were attentive on control trials while focusing on making the "check" response, but this appears to have undermined further processing of the target word. Therefore, it is not engagement alone but the features with which participants engage that is critical. For example, Bailey et al. (2021) observed greater activation of motor and auditory cortices at test for aloud and control words than for silent words, suggesting that participants were potentially re-activating the production record in both conditions but, because on control trials the record did not contain discriminative information between episodic events, it did not benefit memory performance for the target words. Rather than simply encouraging general attentiveness as in control trials, reading aloud forces attention to the target word and, in so doing, encourages distinctive processing liable to improve memory performance.

9.3. Implications for encoding techniques

To summarize, the production effect probably arises due to a combination of processes that vary depending on methodological factors. This is, of course, not unique to production. Fawcett et al. (2022) identified similar sentiments having emerged within the enactment effect (Russ et al., 2003) and the generation effect (Rosner et al., 2013) literatures as well. and Fawcett and Ozubko (2016, see also Ozubko et al., 2012) drew similar conclusions and proposed a dual-process account of the production effect, attributing the effect of production on recollection to potential distinctive encoding and the effect of production on familiarity to enhanced encoding owing to attentional or motivational factors. This was also used to explain why the production effect is usually (but not always, see Whitridge et al., 2024) smaller in between-subjects designs. MacLeod and Bodner (2017) concluded that the smaller between-subjects effect was due to strengthening—which can also be seen as greater attention—and that the larger within-subject effect was due to this strengthening plus a larger contribution of distinctiveness (see also, Fawcett & Ozubko, 2016).

One possible framing would be that reading a word aloud serves as a desirable difficulty, orienting attention to the target word and encouraging encoding of all features necessary to produce that word. In doing so, however, participants are also driven to stay alert during the trial. Remaining focused on the word may engage additional processing, including enhanced processing of conceptual information (e.g., Fawcett et al., 2022; Lu et al., 2025). Further, this state of heightened engagement could facilitate binding of the encoded features into the study episode. This latter component also aligns with the recent finding that, contrary to past expectation (MacLeod et al., 2010), the production effect improves at least certain forms of implicit memory (Lu et al., 2025; Mama, 2025).

In this way, the production effect emerges not through distinctive encoding or attention alone, but rather through the interplay of the two, with production (and its preparatory requirements) serving to force attentional engagement, the effect of which is then at least partially mediated by distinctive encoding, including but not limited to sensorimotor processes. Such a framework might also explain why control trials exhibit a similar pupillary signature to aloud trials without the mnemonic benefit: Although attention is engaged, processing is oriented *away* from the target word, discouraging distinctive processing of the word's features.

One implication of this view would be that the mechanisms implied by our pupillometric analyses—including response preparation, execution, etc.—are representative of the mechanisms present during a typical task. These mechanisms also align well with phenomenological reports by participants (e.g., Fawcett & Ozubko, 2016). It is important to keep in mind that this article has focused exclusively on study phase processes because we were interested in leveraging the unobtrusive nature of pupillometry to evaluate the roles of attention and distinctive encoding in the emergence of the effect. This focus on the study phase does not address potential test phase mechanisms, including whether a distinctiveness heuristic is used by participants and whether production is used as a contextual cue (possibly facilitated by sensorimotor reinstatement) to facilitate memory retrieval (see Wakeham-Lewis et al., 2022; Whitridge et al., 2024; Zhou & MacLeod, 2022). These test-phase processes are worthy of their own future investigation.

10. Conclusion

Reading words aloud renders them more memorable than words that are read silently. The current series of experiments provides evidence compatible with distinctiveness as a central process in the production effect while suggesting that additional processes are involved. In particular, the finding of a pupillary response *prior* to the presentation of target information supports earlier findings that attention likely plays a role in the production effect. We propose that production leads to heightened attention, which is a crucial antecedent or facilitator of distinctive encoding. The current study therefore refines the distinctiveness versus strength discussion (see MacLeod & Bodner, 2017) in the production effect literature by positioning attention as a prerequisite for effective encoding during production. Future studies should further examine whether attention plays a critical or supportive role in improving memory via production. Furthermore, the current studies highlight pupillometry as a useful metric for examining processes underlying the production effect. Given its simple implementation yet powerful efficacy as a mnemonic technique, the production effect stands as a critical method worthy of investigation focused on both mechanistic and applied perspectives.

[4] These findings would also appear to align with the "lazy reading" hypothesis (cf. Begg & Snider, 1987) deriving from studies of the generation effect which would propose that silent words are processed to a lesser extent than aloud words. However, the lazy reading hypothesis has been challenged several times by the fact that forcing participants to initially engage elaboratively with the word via generation or semantic analysis (MacLeod et al., 2010, Experiments 7 and 8; see also Forrin et al., 2014) or by imagery (Forrin et al., 2014) prior to production does not interact with the production effect.

Author note

The * denotes that the first (JMF) and second (BRTR) listed authors share first authorship of this work equally. This research was supported by a Natural Sciences and Engineering Research Council (NSERC) of Canada postdoctoral scholarship to BRTR, and by NSERC Discovery Grants A7459 to CMM and RGPIN-2017-05250 to JMF. Parts of this work were presented at the annual meetings of the Canadian Society for Brain, Behavior, and Cognitive Science (in 2018, 2019, 2020, and 2025), and at the annual meeting of the Psychonomic Society in 2019. Experiment 2 and Experiments 3 and 4 previously served as JCT's directed studies project and as HVW's MSc thesis, respectively. The authors have no conflicts of interest to declare.

CRediT authorship contribution statement

Jonathan M. Fawcett: Writing – review & editing, Writing – original draft, Visualization, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Brady R.T. Roberts:** Writing – review & editing, Writing – original draft, Software, Resources, Project administration, Methodology, Investigation, Conceptualization. **Hannah V. Willoughby:** Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation. **Jenny C. Tiller:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Conceptualization. **Kathleen L. Hourihan:** Writing – review & editing, Writing – original draft, Supervision. **Colin M. MacLeod:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2025.106326.

Data availability

Data are available via the links provided in the manuscript.

References

Bailey, L. M., Bodner, G. E., Matheson, H. E., Stewart, B. M., Roddick, K., O'Neil, K., … Fawcett, J. M. (2021). Neural correlates of the production effect: An fMRI study. *Brain and Cognition, 152*. https://doi.org/10.1016/j.bandc.2021.105757. Article 105757.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., … Treiman, R. (2007). The english lexicon project. *Behavior Research Methods, 39*(3), 445–459.

Begg, I., & Snider, A. (1987). The generation effect: Evidence for generalized inhibition. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 13*, 553–563. https://doi.org/10.1037/0278-7393.13.4.553

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. https://doi.org/10.7551/mitpress/4561.001.0001.

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., Jamieson, R. K., Cormack, D. T., McDonald, D. T., & 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. (2012). Reassessing the basis of the production effect in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(6), 1711–1719. https://doi.org/10.1037/a0028466

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

Bürkner, P.-C. (2021). Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software, 100*(5), 1–54. https://doi.org/10.18637/jss.v100.i05

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

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology, 33A*(4), 497–505. https://doi.org/10.1080/14640748108400805

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

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*(6), 671–684. https://doi.org/10.1016/S0022-5371(72)80001-X

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 and Review, 8*(1), 155–161. https://doi.org/10.3758/BF03196152

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

Fabiani, M., & Donchin, E. (1995). Encoding processes and memory organization: A model of the von Restorff effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(1), 224–240. https://doi.org/10.1037/0278-7393.21.1.224

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

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

Fawcett, J. M., 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., Jonker, T. R., & MacLeod, C. M. (2014). Production improves memory equivalently following elaborative vs. non-elaborative processing. *Memory, 22*, 470–480. https://doi.org/10.1080/09658211.2013.798417

Forrin, N. D., & MacLeod, C. M. (2016). Order information is used to guide recall of long lists: Further support 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., & 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*. https://doi.org/10.1016/j.jml.2019.104050. Article 104050.

Franklin, M. S., Broadway, J. M., Mrazek, M. D., Smallwood, J., & Schooler, J. W. (2013). Window to the wandering mind: Pupillometry of spontaneous thought while reading. *Quarterly Journal of Experimental Psychology, 66*(12), 2289–2294. https://doi.org/10.1080/17470218.2013.858170

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

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

Geller, J., Still, M. L., & Morris, A. L. (2016). Eyes wide open: Pupil size as a proxy for inhibition in the masked-priming paradigm. *Memory & Cognition, 44*(4), 554–654. https://doi.org/10.3758/s13421-015-0577-4

Goldinger, S. D., Azuma, T., Abramson, M., & Jain, P. (1997). Open wide and say "Blah!": Attentional dynamics of delayed naming. *Journal of Memory and Language, 37*(2), 190–216. https://doi.org/10.1006/jmla.1997.2518

Grandchamp, R., Braboszcz, C., & Delorme, A. (2014). Occulometric variations during mind wandering. *Frontiers in Psychology, 5*. https://doi.org/10.3389/fpsyg.2014.00031. Article 31.

Gross, M. P., & Dobbins, I. G. (2021). Pupil dilation during memory encoding reflects time pressure rather than depth of processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 47*(2), 264–281. https://doi.org/10.1037/xlm0000818

Hassall, C. D., Quinlan, C. K., Turk, D. J., Taylor, T. L., & Krigolson, O. E. (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

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

Hoeks, B., & Levelt, W. J. M. (1993). Pupillary dilation as a measure of attention: A quantitative system analysis. *Behavior Research Methods, Instruments, & Computers, 25*(1), 16–26. https://doi.org/10.3758/BF03204445

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

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. https://doi.org/10.1093/acprof:oso/9780195169669.001.0001.

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

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

Kafkas, A., & Montaldi, D. (2011). Recognition memory strength is predicted by pupillary responses at encoding while fixation patterns distinguish recollection from familiarity. *Quarterly Journal of Experimental Psychology, 64*(10), 1971–1989. https://doi.org/10.1080/17470218.2011.588335

Kamp, S.-M., Forester, G. R., Murphy, A. R., Brumback, T., & Donchin, E. (2012). Testing a distinctiveness explanation of the primacy effect in free recall using event-related potentials. In , *Vol. 34. Proceedings of the annual meeting of the cognitive science society* (pp. 539–544). Retrieved from https://escholarship.org/uc/item/7rk363fw.

Keene, P. A., deBettencourt, M. T., Awh, E., & Vogel, E. K. (2022). Pupillometry signatures of sustained attention and working memory. *Attention, Perception & Psychophysics, 84*(8), 2472–2482. https://doi.org/10.3758/s13414-022-02557-5

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). https://doi.org/10.1016/j.jml.2021.104299. Article 104299.

Kučera, H., & Francis, W. (1967). *Computational analysis of present day American English.* Providence, RI: Brown University Press.

Lloyd, B., & Nieuwenhuis, S. (2024). The effect of reward-induced arousal on the success and precision of episodic memory retrieval. *Scientific Reports, 14*(1). https://doi.org/10.1038/s41598-024-52486-6

Lu, X., Wang, J., & MacLeod, C. M. (2025). Production increases both true and false recognition. *Journal of Memory and Language, 140*, 104584. https://doi.org/10.1016/j.jml.2024.104584

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/S0001-6918(97)00047-4

MacLeod, C. M. (1975). Release from proactive interference: Insufficiency of an attentional account. *American Journal of Psychology, 88*(3), 459–465. https://doi.org/10.2307/1421776

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

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

Mama, Y. (2025). The production effect in implicit memory: Mixed evidence from process dissociation procedure, lexical decision, word-stem completion, and category exemplar generation tests. *Experimental Psychology, 71*(5), 298–311. https://doi.org/10.1027/1618-3169/a000633

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. (2018a). 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

Mama, Y., & Icht, M. (2018b). 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

Murphy, P. R., Boonstra, E., & Nieuwenhuis, S. (2016). Global gain modulation generates time-dependent urgency during perceptual choice in humans. *Nature Communications, 7*(1). https://doi.org/10.1038/ncomms13526

Murray, D. J. (1965). Vocalization-at-presentation, auditory presentation and immediate recall. *Nature, 207*(5000), 1011–1012. https://doi.org/10.1038/2071011a0

Naber, M., Frässle, S., Rutishauser, U., & Einhäuser, W. (2013). Pupil size signals novelty and predicts later retrieval success for declarative memories of natural scenes. *Journal of Vision, 13*(2), 11. https://doi.org/10.1167/13.2.11

Nakamura, R., Nouchi, R., Yagi, A., Yamaya, N., Ota, M., Ishigooka, M., & Kawashima, R. (2023). Neural representation of a one-week delay in remembering information after production and self-generated elaboration encoding strategy. *Acta Psychologica, 240*, Article 104051. https://doi.org/10.1016/j.actpsy.2023.104051

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

Ozubko, J. D., Major, J., & MacLeod, C. M. (2013). Remembered study mode: Support for the distinctiveness account of the production effect. *Memory, 22*(5), 509–524. https://doi.org/10.1080/09658211.2013.800554

Papesh, M. H., & Goldinger, S. D. (2012). Pupil-BLAH-metry: Cognitive effort in speech planning reflected by pupil dilation. *Attention, Perception, & Psychophysics, 74*(4), 754–765. https://doi.org/10.3758/s13414-011-0263-y

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

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., … Lindeløv, J. K. (2019). PsychoPy2: Experiments in behaviour made easy. *Behavior Research Methods, 51*(1), 195–203. https://doi.org/10.3758/s13428-018-011-y

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.* Vienna, Austria: R Foundation for Statistical Computing. https://www.r-project.org/.

Robison, M. K., Trost, J. M., Schor, D., Gibson, B. S., & Healey, M. K. (2022). Pupillary correlates of individual differences in long-term memory. *Psychonomic Bulletin & Review, 29*(4), 1355–1366. https://doi.org/10.3758/s13423-022-02081-5

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives in Psychological Science, 1*(3), 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x

Rosner, Z. A., Elman, J. A., & Shimamura, A. P. (2013). The generation effect: Activating broad neural circuits during memory encoding. *Cortex, 49*(7), 1901–1909. https://doi.org/10.1016/j.cortex.2012.09.009

Russ, M. O., Mack, W., Grama, C.-R., Lanfermann, H., & Knopf, M. (2003). Enactment effect in memory: Evidence concerning the function of the supramarginal gyrus. *Experimental Brain Research, 149*(4), 497–504. https://doi.org/10.1007/s00221-003-1398-4

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*(1). https://doi.org/10.1016/j.jml.2021.104219. Article 104219.

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

Smallwood, J., Brown, K. S., Tipper, C., Giesbrecht, B., Franklin, M. S., Mrazek, M. D., … Schooler, J. W. (2011). Pupillometric evidence for the decoupling of attention from perceptual input during offline thought. *PLoS One, 6*(3), Article e18298. https://doi.org/10.1371/journal.pone.0018298

Smilek, D., Carriere, J. S. A., & Cheyne, J. A. (2010). Out of mind, out of sight: Eye blinking as indicator and embodiment of mind wandering. *Psychological Science, 21*(6), 786–789. https://doi.org/10.1177/0956797610368063

SR Research. (2010). *Eyelink 1000 [Apparatus and software]*. Mississauga, Ontario, Canada: SR Research.

SR Research. (2020). *Experiment builder (version 2.2.299) [computer software]*. Mississauga, Ontario, Canada: SR Research.

Sterpenich, V., D'Argembeau, A., Desseilles, M., Balteau, E., Albouy, G., Vandewalle, G., Degueldre, C., Luxen, A., Collette, F., & Maquet, P. (2006). The locus coeruleus is involved in the successful retrieval of emotional memories in humans. *Journal of Neuroscience, 26*(28), 7416–7423. https://doi.org/10.1523/JNEUROSCI.1001-06.2006

Strauch, C., Wang, C.-A., Einhäuser, W., Van der Stigchel, S., & Naber, M. (2022). Pupillometry as an integrated readout of distinct attentional networks. *Trends in Neurosciences, 45*(8), 635–647. https://doi.org/10.1016/j.tins.2022.05.003

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

Unsworth, N., & Miller, A. L. (2020). Encoding dynamics in free recall: Examining attention allocation with pupillometry. *Memory & Cognition, 49*(1), 90–111. https://doi.org/10.3758/s13421-020-01077-7

Unsworth, N., & Robison, M. K. (2015). Individual differences in the allocation of attention to items in working memory: Evidence from pupillometry. *Psychonomic Bulletin & Review, 22*(3), 757–765. https://doi.org/10.3758/s13423-014-0747-6

Unsworth, N., & Robison, M. K. (2018). Tracking arousal state and mind wandering with pupillometry. *Cognitive, Affective, & Behavioral Neuroscience, 18*(4), 638–664. https://doi.org/10.3758/s13415-018-0594-4

Varao Sousa, T. L., Carriere, J. S. A., & Smilek, D. (2013). The way we encounter reading material influences how frequently we mind wander. *Frontiers in Psychology, 4*. https://doi.org/10.3389/fpsyg.2013.00892. Article 892.

Verney, S. P., Granholm, E., & Marshall, S. P. (2004). Pupillary responses on the visual backward masking task reflect general cognitive ability. *International Journal of Psychophysiology, 52*(1), 23–36. https://doi.org/10.1016/j.ijpsycho.2003.12.003

Võ, M. L., Jacobs, A. M., Kuchinke, L., Hofmann, M., Conrad, M., Schacht, A., & Hutzler, F. (2008). The coupling of emotion and cognition in the eye: Introducing the pupil old/new effect. *Psychophysiology, 45*(1), 130–140. https://doi.org/10.1111/j.1469-8986.2007.00606.x

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

Watkins, O. C., & Watkins, M. J. (1975). Buildup of proactive inhibition as a cue-overload effect. *Journal of Experimental Psychology: Human Learning and Memory, 1*(4), 442–452. https://doi.org/10.1037/0278-7393.1.4.442

van der Wel, P., & van Steenbergen, H. (2018). Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic Bulletin & Review, 25*(6), 2005–2015. https://doi.org/10.3758/s13423-018-1432-y

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

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

Yonelinas, A. P. (2001). Consciousness, control and confidence: The three cs of recognition memory. *Journal of Experimental Psychology: General, 130*, 361–379. https://doi.org/10.1037/0096-3445.130.3.361

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

Zhang, B., Meng, Z., Li, Q., Chen, A., & Bodner, G. E. (2023). EEG-based univariate and multivariate analyses reveal that multiple processes contribute to the production effect in recognition. *Cortex, 165*, 57–69. https://doi.org/10.1016/j.cortex.2023.04.006

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