Attention, Perception, & Psychophysics
https://doi.org/10.3758/s13414-024-02888-5

The resolution of proactive interference in a novel visual working memory task: A behavioral and pupillometric study

Jamie Donenfeld[1] · Erik Blaser[1] · Zsuzsa Kaldy[1]

Accepted: 18 March 2024

Abstract

Proactive interference (PI) occurs when previously learned information impairs memory for more recently learned information. Most PI studies have employed verbal stimuli, while the role of PI in visual working memory (VWM) has had relatively little attention. In the verbal domain, Johansson and colleagues (2018) found that pupil diameter – a real-time neurophysiological index of cognitive effort – reflects the accumulation and resolution of PI. Here we use a novel, naturalistic paradigm to test the behavioral and pupillary correlates of PI resolution for *what-was-where* item-location bindings in VWM. Importantly, in our paradigm, trials (*PI* vs. *no-PI* condition) are mixed in a block, and participants are naïve to the condition until they are tested. This design sidesteps concerns about differences in encoding strategies or *generalized* effort differences between conditions. Across three experiments ($N = 122$ total) we assessed PI's effect on VWM and whether PI resolution during memory retrieval is associated with greater cognitive effort (as indexed by the phasic, task-evoked pupil response). We found strong support for PI's detrimental effect on VWM (even with our spatially distributed stimuli), but no consistent link between interference resolution and effort during memory retrieval (this, even though the pupil *was* a reliable indicator that higher-performing individuals tried harder during memory *encoding*). We speculate that when explicit strategies are minimized, and PI resolution relies primarily on implicit processing, the effect may not be sufficient to trigger a robust pupillometric response.

Keywords Proactive interference · Visual working memory · Pupillometry · Item-location memory · Cognitive effort

Introduction

Imagine that you are preparing to bake a cake and begin to gather the ingredients from the recipe. First, you take out the (white and powdery) baking soda and put it on the kitchen table, and later, you put the (white and powdery) flour on the counter next to the sink. When it is time to measure the flour, you might first try to find it on the kitchen table instead of the counter. This *proactive interference* (PI) occurs when irrelevant, previously learned information impairs memory for relevant, more recently learned information. In experimental paradigms, researchers generate PI effects by repeating to-be-remembered items or features (e.g., color, semantic relatedness) across tests. If sufficient PI is generated, then participants' performance will suffer as they inadvertently recall irrelevant information from previous tests. Underwood was the first to illustrate the PI effect in a memory task involving word lists (Underwood, 1957). An early study demonstrated PI effects in the visual modality as well, using drawings of faces as stimuli (Anderson & Paulson, 1978), although there is a recent debate about the factors that contribute to visual PI (Endress, 2022; Makovski, 2016). In the verbal modality, Johansson and colleagues (2018) found that the pupil indexes accumulation and resolution of PI, but this has yet to be investigated in the visual modality. Here, in a series of three studies using a novel task, we show that PI causes lower accuracy in visual working memory (VWM), and, using pupillometry, assess its relationship with cognitive effort.

✉ Jamie Donenfeld
jamie.donenfeld001@umb.edu

[1] Department of Psychology, University of Massachusetts Boston, 100 William T. Morrissey Blvd, Boston, MA 02125-3393, USA

Published online: 19 June 2024

Proactive interference (PI) in visual working memory (VWM)

VWM is a flexible, capacity-limited system that temporarily increases visual information availability for in-the-moment processing (Cowan, 2017). PI has been observed in many VWM studies (Cyr et al., 2017; Endress, 2022; Endress & Potter, 2014; Hartshorne, 2008; Lin & Luck, 2012; Makovski, 2016; Makovski & Jiang, 2008; Mercer & Fisher, 2022; Oberauer et al., 2017; Shoval et al., 2020; Shoval & Makovski, 2021, 2022). As with verbal stimuli, the amount of interference induced in VWM depends on the nature of the stimuli and task, including inter-item similarity (the degree of similarity among to-be-remembered stimuli), as well as more controversial issues like the effect of spatial location (see, e.g., Endress, 2022; Mercer & Fisher, 2022).

For instance, according to Shoval and colleagues (2020), change-detection paradigms with simple geometric shapes that only differ in, say, color (Hartshorne, 2008; Lin & Luck, 2012; Shoval et al., 2020) may underestimate the effect of PI in VWM. They argue that since inter-item similarity is high due to the repetition of items across sessions, this unwittingly leads to background levels of PI that may effectively wash out differences between nominal PI and no-PI conditions. In contrast, so-called "repeated-unique procedures" show larger PI effects (Shoval et al., 2020). These paradigms compare recognition performance between a large set of unique items with complex featural differences drawn from different semantic categories (low inter-item similarity) and a limited pool of repeated items (high inter-item similarity) to provide a contrast that maximizes the opportunity for PI. Endress and Potter (2014) demonstrated that VWM is both highly susceptible to PI and capacity-limited (to three to four items) when to-be-remembered stimuli – in this case, color photographs of familiar objects from the Brady and colleagues' (2008) image repository – are repeated from trial to trial, but not when stimuli are unique. Both Makovski (2016) and Shoval and colleagues (2020) replicated this result, finding small-to-moderate PI effects (PI effect sizes, the difference in performance between unique and repeated conditions: Endress & Potter (2014), $\eta^2 = 0.24$; Makovski (2016), $\eta_p{}^2 = 0.43$, and Shoval and colleagues (2020), $\eta_p{}^2 = 0.62$).

Even when using a control condition with unique items, the degree of inter-item similarity, semantic or featural, across all stimuli can affect PI. Shoval and colleagues (2020) used Brady and colleagues' (2008) repository of everyday object photographs (meaningful stimuli) to test the role of inter-item similarity within the repeated-unique paradigm. They found that meaningful, *heterogeneous* items (with relatively low inter-item similarity, i.e., from different semantic categories) caused greater PI effects than meaningful, *homogenous* stimuli (with high inter-item semantic similarity). They attribute this somewhat surprising finding to different encoding strategies: more reliance on semantic strategies in the heterogeneous case, and more subtle, featural processing in the homogeneous case. In a follow-up study, again using a selection of Brady and colleagues' (2008) stimuli, Shoval and Makovski (2022) demonstrated that heterogeneity in semantic information – not simply in visual complexity – increases PI in VWM. These findings are paralleled in the PI literature: no or low PI is observed in change-detection tasks using simple, meaningless stimuli (e.g., Hartshorne, 2008; Lin & Luck, 2012), while more moderate levels of PI have been observed in tasks using more naturalistic, meaningful stimuli (e.g., Endress & Potter, 2014).

While the level of inter-item similarity of items affects PI, so does the similarity of the to-be-remembered item and the test item. Typically, recognition memory is tested by using an exact copy of a previously seen item as a test item. In a version of the classic recent probes task (Monsell, 1978), Mercer and Fisher (2022) showed that test items (familiar photographs of everyday items from Brady and colleagues' repositories (Brady et al., 2008, 2013)) did not need to be exact copies of previously encountered items to produce PI. Test items differing in orientation or positioning, but not color, also produced a PI effect. Subsequently, Mercer and colleagues demonstrated that PI persists across various lengths of intertrial intervals (as long as 8 s) as well as during trials with distractor items, even from the same (visual) modality (Mercer et al., 2022). Thus, paradigms need not test exact copies of the to-be-remembered stimuli to induce PI, suggesting that interference-inducing memory traces may have a more general "familiarity signal" character (Mercer & Fisher, 2022).

The effect of spatial position on PI is more controversial. In addition to replicating the original findings of Endress and Potter (2014), Makovski also tested the effect of spatial position on PI in VWM (Makovski, 2016). Instead of presenting the sample items at central fixation, they presented groups of four or eight items in a circular arrangement – simultaneously in one experiment, and sequentially in another. For both spatially distributed presentation modes, they found only partial evidence of a PI effect (the only significant PI effect was for the set-size eight, simultaneously-presented condition: $\eta_p{}^2 = 0.11$), and concluded that spatial information must protect against PI in visual memory. However, Endress (2022) rebutted this claim by arguing that Makovski (2016) did not consider the ratio of total set-size to trial set-size: strong PI failed to emerge because participants had to wait too long between exposures of repeated items, making the repeated items too infrequent to produce measurable PI, and diluting the intended effect of having high inter-item similarity between repeated items.

Taken together, while the role of spatial information is still controversial, it is clear that repeated unique procedures with low inter-item semantic similarity maximize PI. Replication of the spatially distributed item-specific PI effect using a different paradigm – one that retains characteristics of PI-maximizing repeated-unique procedures but does not rely as heavily on the parameter of set size to pool size ratio – is important to understand the generalizability of PI in VWM to real-world contexts that necessarily involve items in different spatial locations (i.e., *what-was-where*?).

Mechanisms of PI resolution

Most accounts assert that PI resolution operates primarily during retrieval (Oberauer & Lin, 2023, 2017; Shoval & Makovski, 2021), though encoding also plays a role (Kliegl et al., 2015; Pastötter et al., 2011). The leading theoretical account of PI is based on *temporal distinctiveness* (Crowder, 1976; Glenberg & Swanson, 1986). According to this theory, during retrieval, when similar but no-longer-relevant representations compete with currently relevant representations, the difficulty lies in recalling the memory with the correct "timestamp" ("Did I take my medicine today or was that yesterday morning?"). The temporal distinctiveness account has been supported by multiple studies (Brown et al., 2007; Souza & Oberauer, 2015).

The neural mechanisms of proactive interference resolution have been studied for more than 25 years. In a classic functional magnetic resonance imaging (fMRI) study, Jonides et al. (1998) showed that the inferior frontal gyrus is more activated when PI is present in a verbal working memory task. More recently, it was shown that the medial temporal lobe responds differentially in a PI task when retrieval is correct versus incorrect (Öztekin et al., 2009). This provided important mechanistic evidence for the retrieval account of PI resolution. The current view is that an extended network of brain areas is involved in PI resolution, including the inferior frontal gyrus (and other areas in the frontal cortex such as the dorsolateral prefrontal cortex) and areas in the medial temporal lobe and the posterior parietal cortex (for reviews, see Hamilton et al., 2022; Kliegl & Bäuml, 2021).

Oberauer and Lin's recent *Interference Model* of WM provides a computational account of interference and its resolution (Oberauer & Lin, 2023, 2017). In that account, to-be-remembered items (e.g., colors) are bound to contexts (e.g., locations) in a two-dimensional (2D) space. Retrieval cues activate specific locations in a memory space that trigger neighboring, relevant to-be-remembered information. However, there is a degree of imprecision for each item's location and other to-be-remembered information, such that a specific retrieval cue can generate false hits that cause interference between multiple item-location bindings. The model better predicts behavioral results than other models (Oberauer & Lin, 2023, 2017).

Taken together, both behavioral and neurophysiological results support that PI manifests as a result of conflict resolution during memory retrieval. A recent *Interference Model* of WM (Oberauer & Lin, 2023, 2017) clarifies the role of context on interference levels during memory retrieval, providing a much-needed framework for predicting interference.

Pupillometric studies of PI resolution

According to Kahneman (1973), pupil diameter indexes an *intensive* aspect of attention, a kind of special case of sympathetic arousal (Bruya & Tang, 2018). Classic studies show that pupil diameter increases monotonically with increasing memory load until a critical threshold is reached, thought to correspond to the individual's working memory capacity (Kahneman & Beatty, 1966; Peavler, 1974). Similarly, pupil diameter has been shown to increase with increasing task demand until observers reach their processing limit (for a recent review, see van der Wel & van Steenbergen, 2018). It is important to note that interpreting pupil diameter changes in terms of cognitive constructs is a form of reverse inference (Poldrack, 2006); however, given the ample neurophysiological evidence, the literature has converged on accepting this as a valid causal link (Joshi & Gold, 2020; Strauch et al., 2022).

In 1975, Engle was the first to assess pupillary correlates of PI during encoding in a VWM task (Engle, 1975). Interestingly, some early studies have investigated other physiological sympathetic arousal correlates such as heart rate and skin conductance of PI (Morin et al., 1982; Wilson, 1984). To our knowledge, Johansson and colleagues (2018) were the first to explicitly test the relationship of pupil diameter, as a measure of cognitive effort, with PI *during retrieval*. They asked participants to learn lists of words from a particular semantic category and later tested the participants on free recall of the lists. This was repeated within the same semantic category for three trials, and then the fourth trial either (i) continued the same semantic category (*continuing* to induce PI) or (ii) changed the category (*releasing* the participant from PI). The authors demonstrated the classic monotonic decrease in performance across subsequent word lists as PI built up, as well as a rebound in performance when PI was released by changing the semantic category of the word list (a classic manipulation in verbal PI studies: *release from PI*; Kincaid & Wickens, 1970). Critically, the same pattern was found in the pupil diameter signal: pupil diameter increased as a function of PI accumulation and returned to baseline levels when PI was released. PCA analysis further supported a relationship between participants' ability to handle interference and more effort (higher pupil diameter) during retrieval (there was no difference in pupil diameter during encoding). They reported a behavioral PI effect size of d = 1.66 and a pupillary PI effect of d = 0.60. No studies to date have examined these processes in VWM. Our study aimed to address this gap and broaden the understanding of PI resolution across memory modalities. Critically, though, we do *not* use the classic *release from PI paradigm* described above: we use a paradigm that distributes interference continuously across both PI and NoPI conditions, so that participants cannot develop condition-specific strategies.

The current study

We used a modified Delayed Match Retrieval (DMR) four-alternative forced-choice location recognition task to test item-location binding in VWM (Kaldy et al., 2016). Experiment 1 served as a proof of concept that PI can be induced in this VWM task. In Experiments 2A and 2B (preregistered), we used this task with minor modifications to quantify PI and test whether the cognitive effort required for PI resolution can be detected using pupillometry. Finally, we combined the data from Experiments 2A and 2B and applied the same analytic framework to increase power and reconcile discrepancies between the single-study results.

DMR is a reinterpretation of the classic Delayed-Match-to-Sample paradigm and is inspired by the game *Memory* (see Fig. 1). During a trial, a set of four real-world, to-be-remembered items is shown, then hidden, and then a "sample" (that matches one of the items) is revealed. The participant is asked to indicate which of the (now hidden) items was the *match* to the sample.

Our paradigm has three important features. First, it sidesteps concerns about differential strategies during encoding. Critically, two of the four to-be-remembered items are repeated (PI items) from trial to trial, while the other two are unique (NoPI items; see Fig. 1A). Since participants do not know which of the four items they will be tested on until the cue (the "sample") appears, they are unaware *of the condition* (PI vs. NoPI) until the final test phase of the trial. Thus, they should equally encode all of the items (NoPI and PI). When participants know which condition they are in (e.g., in a blocked design), that could influence how they allocate their effort, which could be reflected in the pupil.

In "blocked" approaches, PI items are encoded during entirely separate phases than NoPI or unique items. This is true for classic *release from PI* paradigms (e.g., Johansson et al., 2018), where PI is built up *within one condition-specific block*, as opposed to across blocks of both conditions. Similarly, visual memory PI studies such as those by Endress and Potter (2014) and Shoval and colleagues (2020) also separate the buildup of PI from the unique condition (no unique images are shown during the PI encoding phase). This design results in two distinct tasks that a participant may develop differential strategies to deal with (remember repeated items; remember unique items). In our design, like recent probes tasks (see, e.g., Jonides & Nee, 2006), the buildup of PI occurs *while* unique items are presented and tested. This distributed form makes it impossible to see any differential effort deployment prior to the response phase, as PI is present in every trial. This limits the ability to detect differences in vigilance between conditions, but is also a strength as participants cannot develop condition-specific strategies that may ultimately bias their effortful engagement, providing a relatively pure measure of PI resolution during recognition.

Second, there have been many recent calls for turning toward more naturalistic tasks in cognitive science (Ibanez, 2022; Nastase et al., 2020; Sonkusare et al., 2019), including memory research (Kristjánsson & Draschkow, 2021; Maguire, 2022). Unlike PI designs that are built on the logic of change detection paradigms ("Was this item in the previous set? yes/no"), our paradigm is a four-alternative forced-choice task requiring remembering a specific item-location binding ("*Where* was *this* one?"). We believe that this type of task is more naturalistic, better capturing competencies central to everyday life: *Where did I leave my keys*? *Where was the best food source*? (Pertzov et al., 2012; Postma et al., 2008). This is consistent with the emerging view of VWM that shifts the focus from change detection-type tasks to tasks that more closely model everyday situations, such as making coffee, finding your personal items – keys, wallet, phone – before leaving the house (Kristjánsson & Draschkow, 2021). We embrace this view and are actively developing paradigms that focus on aspects of in-the-moment information processing in service of ongoing tasks (Hamilton et al., 2024; Liang et al., 2023).

Finally, we have added a verbal filler task between encoding and retrieval to minimize the opportunity for verbal rehearsal. This is a factor that is often neglected in recent PI study designs in VWM (for an exception, see Endress & Siddique, 2016).

Experiment 1

This experiment was designed to test the effect of PI in a *what-was-where* VWM task. Our first prediction was that overall accuracy would be lower in the PI condition than in the NoPI condition. Our second prediction was that accuracy would decrease across trials within each epoch to a greater extent in the PI condition compared to the NoPI condition.

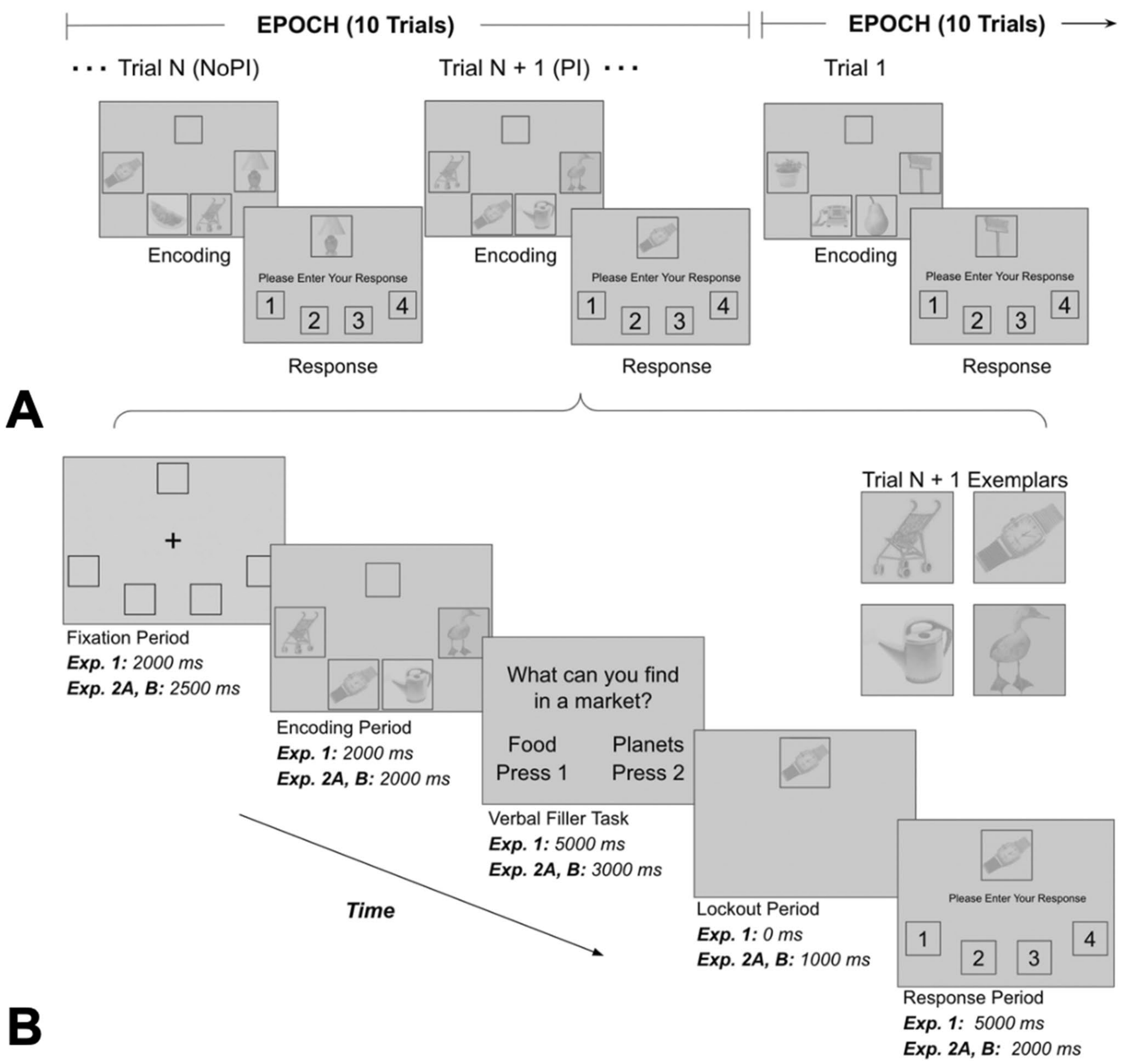

Fig. 1 **(A)** Trials within an epoch. Two images during encoding (here, the watch and the stroller) are repeated across successive trials within an epoch (but not across epochs) accompanied by two items that are always unique, with locations randomized. When a unique image is tested, the trial is a *NoPI trial* (here, trial N), and when a repeated image is tested, the trial is a *PI trial* (here, trial N+1). In subsequent epochs, the two repeating images are replaced with a new set of repeating images. **(B)** Test trial sequence. This figure shows the sequence of events of a typical trial (here, using trial *N+1* from Fig. 1A). Participants were not provided with feedback. The duration of each period depended on the experiment. The to-be-remembered set of items is shown enlarged for clarity in the inset panel. *Note:* Figures are not to scale

Methods

Participants

Forty-three participants were recruited via *Prolific*, an online research recruitment platform. Participants completed the study at home on their personal computers and entered answer choices via keypress. Nine participants were excluded (see *Exclusion criteria* below), leaving 34 participants' data for final analyses. Participants' average age was 26 years old, ranging from 20 to 34 years, and 11 identified as female. All participants were fluent in English and reported normal or corrected-to-normal vision. Participants were recruited from various countries including Poland, Chile, Portugal, Mexico, and the UK. Fifteen participants identified as White, five as Asian, eight as more than one race, and six as "other." Each participant was compensated $5 for their participation.

Materials

Images of everyday objects were selected from a collection of 2,400 unique stimuli (Brady et al., 2008). Our final set of 232 items contained only common, nameable objects with no words or letters. We used an in-house MATLAB script to convert items to grayscale and impose a common luminance histogram across all images (function *imhistmatch*), equal to the average luminance histogram of all images, to ensure that the pupillary light reflex was minimized. (Although pupil diameter was not measured in Experiment 1, we wanted to ensure that behavioral results would be comparable to subsequent experiments involving pupillometry.) Our study was built in lab.js and hosted through open-lab.online (Henninger et al., 2022; Shevchenko, 2022).

The four to-be-remembered images were arranged along a virtual arc, with the center of the target item equidistant from those image locations and 30° separating the center of each to-be-remembered image (see Fig. 1A and B). While we did not control viewing distance or device, on an average laptop screen (~35 x 20 cm) at a typical arm's-length viewing distance (~57 cm) the target scene subtended approximately 34 x 20 degrees of visual angle.

Procedure

We used a modified Delayed Match Retrieval (DMR) four-alternative forced-choice location recognition task (Kaldy et al., 2016). As described above, the condition was determined by the test item at the end of the trial: the participant was either tested on a repeated item (PI condition) or a unique item (NoPI condition; see Fig. 1A). We grouped trials into ten epochs of ten trials per epoch: five NoPI trials and five PI trials in a random sequence. Each trial consisted of four periods: *fixation* (2,000 ms), *encoding* (2,000 ms), *verbal filler task* (5,000 ms), and the *response* period (5,000 ms; see Fig. 1B). During encoding, the four to-be-remembered items were shown: two different items were repeated on each trial within a given epoch, but not across epochs (PI items), and two unique items were never repeated within or across epochs (NoPI items). The total number of PI items was 20 (two per epoch x 10 epochs), and the total number of NoPI items was 200 (20 per epoch x 10 epochs). The additional 12 stimuli noted in the *Stimuli* section were used in the practice trials. During each verbal filler task period, a question and two answer choices were presented on the screen in English, and the participant answered the question by keypress. Examples of the verbal task questions include: "*How many things are in a pair? choose: (a) 60 or (b) 2*" and "*What kind of animal is a dove? choose: (a) Bird or (b) Dinosaur.*" As a filler task, these questions were designed to be as easy to answer as possible, providing a check to ensure that participants were on-task (everyone was expected to score at or near ceiling) and acting to limit verbal rehearsal during the delay. After the 5,000-ms period for reading and answering the filler question, the actual *response* period began, with the target item appearing in the top center location. During this period, the participant indicated (via keypress: one, two, three, four, for each of the four locations) where the target had been located during encoding. The response period was a fixed length (5,000 ms) to mitigate a speed-accuracy tradeoff. Importantly, the potential for any differential encoding or other task strategies between PI and NoPI items (as might occur when PI and NoPI trials are blocked) was minimized, since PI and NoPI items were mixed within each trial.

A testing session began with instructions and three practice trials to acquaint participants with the pace of presentation and the response procedure. No performance-related feedback was offered during the practice trials and the data were not included in the final analysis. Following practice, the ten epochs of ten trials each began. The order of trial presentation within an epoch and the order of epochs across the experiment was randomized for each participant.

Exclusion criteria First, participants who had more than two missed responses out of ten possible responses in one or more epochs were removed ($n = 8$). Then, participants were excluded if their overall accuracy on the verbal filler questions was below 70% ($n = 0$). Next, trials that did not have a valid response, or where the response occurred too early (< 250 ms) were removed. After these trial-based exclusions, any participants who had accuracy lower than chance (for the NoPI condition only) were removed ($n = 1$). Overall, data from nine of the 43 online participants were excluded.

Results

Where relevant, throughout all of our reported analyses, inspection of the residuals did not reveal any substantive deviations from normality.

To test for a PI effect, we compared mean accuracy for each participant between the PI and NoPI conditions with a paired-samples t-test. We found a significant effect of *condition* (PI, NoPI) on accuracy: $t(33) = 4.33$, $p < 0.001$, Cohen's $d_z = 0.741$ (Table 1, Fig. 2). Participants' accuracy, on average, was 4.9% lower in the PI condition compared to the NoPI condition.

Table 1 Descriptive statistics and results of paired t-tests comparing mean accuracy within participants between *condition* (PI, NoPI) for all studies

	NoPI			*PI*						
Experiment	*M*	95% CI	*SD*	*M*	95% CI	*SD*	*df*	*t*	*p*	*Cohen's* d_z
1	0.673	[0.607, 0.739]	0.189	0.624	[0.561, 0.687]	0.181	33	4.318	< 0.001 ***	0.741
2A	0.617	[0.563, 0.671]	0.160	0.576	[0.522, 0.631]	0.161	35	2.664	0.0116 *	0.444
2B	0.680	[0.630, 0.730]	0.180	0.648	[0.601, 0.696]	0.171	51	2.265	0.0278 *	0.314
2A + 2B	0.654	[0.617, 0.691]	0.174	0.619	[0.583, 0.655]	0.170	87	3.420	< 0.001 ***	0.365

M mean, *SD* standard deviation

Values in square brackets indicate the 95% confidence interval around the mean. Cohen's d_z values were calculated using G*Power 3.1 using the difference between two dependent means

* $p < 0.05$, **$p < 0.01$, *** $p < 0.001$

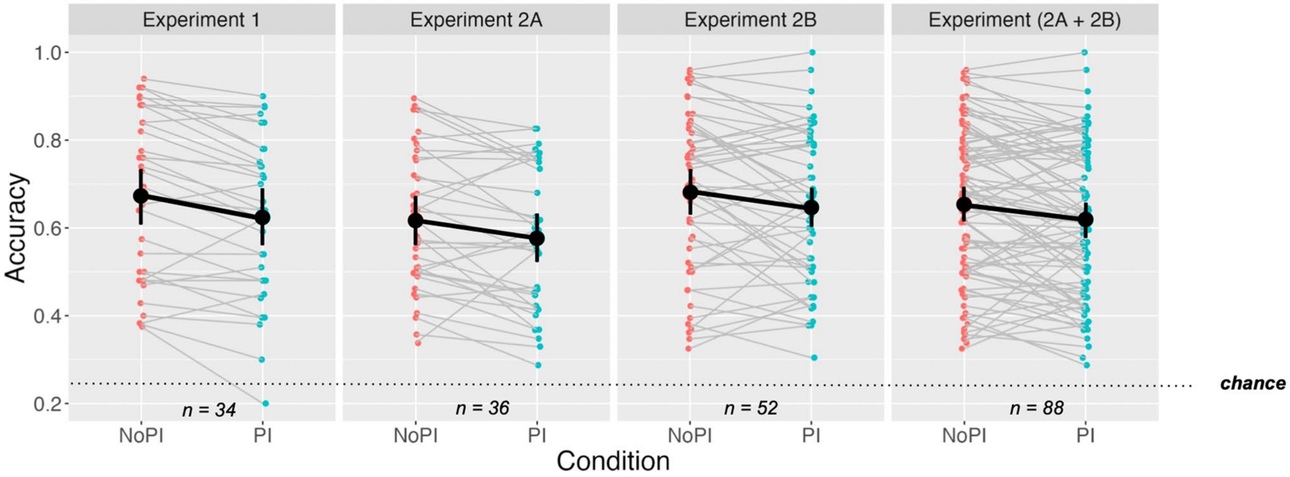

Fig. 2 The effect of *condition (PI, NoPI) o*n accuracy. Panels show accuracy in the PI versus NoPI conditions across all experiments, and reflect the data shown in Table 1. Light grey lines connect individual participants' data across condition (PI, NoPI). Large data points indicate the mean and error bars the 95% confidence interval. Chance level is shown at 25%

To test for PI accumulation within an epoch, we created an *order* factor that corresponds to whether a trial occurred in the first half of an epoch or in the second half. Performance was calculated as mean accuracy over all trials for *order* (first half, second half of the epoch), *condition* (PI, NoPI), and participant. Confirming the result of the paired t-test, the two-way repeated-measures ANOVA showed there was a significant main effect of *condition* (PI, NoPI): $F(33) = 16.5$, $p < 0.001$, $\eta^2_p = 0.333$. There was no main effect of *order* (first half, second half of the epoch): $F(33) = 2.42$, $p = 0.129$, $\eta^2_p = 0.068$. We did not find any evidence for PI accumulation across trials with the two-way repeated-measures ANOVA: there was no statistically significant interaction between the effects of *order* (first half, second half of the epoch) and *condition* (PI, NoPI) on accuracy: $F(33) = 0.117$, $p = 0.735$, $\eta^2_p = 0.004$ (Table S1, Online Supplementary Material (OSM)). A visualization of average accuracy across trials per epoch in each condition is shown in Fig. S1 (OSM).

Discussion

Using our novel DMR paradigm, we showed that spatially-distributed visual stimuli are not immune to PI: participants' accuracy was significantly lower in the PI condition compared to the NoPI condition. However, we did not observe an increase in the PI effect across the successive repetition of stimuli or a subsequent release from PI at the beginning of the new epoch that has been observed with verbal stimuli (Johansson et al., 2018; Kahneman & Beatty, 1966).

Experiment 2A

Experiment 2A used the same paradigm as Experiment 1 (with some minor modifications), and was run in our laboratory using an eye-tracker to monitor eye movements and pupil diameter. Here, we sought to replicate the results from Experiment 1 and assess whether pupil diameter indexes PI during the response period. In addition, if the pupil diameter reflects task-relevant cognitive effort to resolve PI, then we expect pupil diameter to be positively related to accuracy: participants who showed greater pupillary PI effects should also show greater behavior-based PI effects. Our final prediction was that, after categorizing all trials as correct (score = 1) or incorrect (score = 0), *success* would be related to higher pupil diameter during retrieval.

Methods

Participants

Forty participants were recruited by convenience sampling on the University of Massachusetts Boston campus. One participant was excluded due to missing data, two participants were excluded for having accuracy below chance, and one participant was removed due to low accuracy on the verbal filler task questions (see *Exclusion criteria* below). Thirty-six participants' data were included in the final analyses. Participants' average age was 20 years, ranging from 18 to 30 years, and 22 participants identified as female. All participants were fluent in English and had normal or corrected-to-normal vision. Due to an error, no race or ethnicity data were recorded. Each participant was compensated $10 cash for their participation.

Materials

Stimuli generation followed the same procedure as Experiment 1. In the lab setup, viewing distance was maintained at approximately 65 cm. The size of the images was 3.5° x 3.5°, and the radius of the annulus was approximately 15°. Relative positions of the stimuli were the same as Experiment 1, spaced at approximately 30° center-to-center (see Fig. 1).

Procedure

Trial design in Experiment 2A largely followed Experiment 1, with three minor modifications. In Experiment 2, the fixation period was increased to 2,500 ms, while the verbal filler task and response periods were reduced to 3,000 ms each. The verbal filler task was reduced to constrain the time between encoding and response in an attempt to limit verbal rehearsal further. In addition, to provide a quiet period for pupillometric measures, we introduced a response lockout period: for 1,000 ms prior to the response period, the target item was shown to the participant, but answer choices (one, two, three, and four) were occluded (see Fig. 1B). During the final 2,000 ms of the response period, the participant could enter their response. There were 103 trials in total: three practice trials followed by 100 test trials.

Tests were conducted in a moderately-lit testing room. All experimental stimuli were presented on the 23-in. display of a Tobii TX300 eye tracker running Tobii Studio (Tobii Technology, Stockholm, Sweden). The TX300 provides approximately 1° of spatial and 300-Hz temporal resolution while recording gaze and pupil diameter. During testing, participants sat approximately 65 cm from the display. Participants were instructed not to move from this position, not to interact with any devices, and not to interact with the experimenter during the testing phase. Testing began immediately following instruction and practice trials. Participants entered their responses during the verbal task and response periods using keypresses.

As in Experiment 1, participants completed three practice trials before beginning test trials. Participants were given one 2-min break after the first 50 test trials of the 100-trial block. Counterbalancing of blocks and trials was the same as in Experiment 1.

Analysis

Pupil diameter data processing To correct for any potential variability in the sampling rate, we synchronized the data to a fixed, 300-Hz timeline. The data were then preprocessed using the *pupillometryR* pipeline (Forbes, 2020): first, we smoothed the data using linear regression of the left pupil diameter against the right pupil diameter and the right pupil diameter against the left pupil diameter for every participant and every trial (Jackson & Sirois, 2009). Then we calculated the average of the left and right pupil diameters to obtain a single pupil diameter per time point. Data were then downsampled to 50 ms and filtered with a moving window (window width was set to five samples, 250 ms) median filter. We then performed linear interpolation followed by subtractive baseline correction to the mean pupil diameter across the fixation period (Mathôt et al., 2018). All further references to pupil diameter refer to these final, baseline-corrected values. (There are many analytic pipelines currently in use for pupillometry (see Mathôt et al., 2018; Sirois et al., 2023) – we made our choices based on best practices in the field.)

Exclusion criteria First, trials with more than 75% of pupil diameter data missing were removed. Then, trials were removed if there was no fixation on the central region containing all to-be-remembered items during the encoding phase. Next, we removed trials that did not have a valid

Table 2 Descriptive statistics and results of paired *t*-tests comparing mean response period (6,000–8,500 ms after trial onset) pupil diameter (mm) within participants and between *conditions (PI, NoPI)* for all studies

	NoPI			*PI*						
Experiment	*M*	95% C	*SD*	*M*	95% CI	*SD*	*df*	*t*	*p*	*Cohen's* d_z
2A	0.0895	[0.0607, 0.118]	0.0825	0.105	[0.0775, 0.133]	0.0789	35	-2.58	0.0142 *	-0.511
2B	0.0907	[0.0687, 0.113]	0.0775	0.0851	[0.133, 0.106]	0.0745	51	0.947	0.348	0.142
2A + 2B	0.0902	[0.0730, 0.107]	0.0791	0.0933	[0.0767, 0.110]	0.0765	87	-0.654	0.515	-0.0813

* $p < 0.05$, **$p < 0.01$, *** $p < 0.001$

response. After completing these trial-based exclusions, we performed participant-level exclusions. Participants who had accuracy lower than chance (for the NoPI condition only) were excluded ($n = 2$). Finally, participants were excluded if their overall accuracy on verbal filler questions was below 70% ($n = 1$).

Pupil diameter measures The phasic, task-evoked pupillary response – what we seek to assess in the current study – is fast-changing, responding to task demands as quickly as 220 ms (see Mathôt et al, 2018) with a typical window of 500–3,000 ms (see Laeng et al., 2012; Rondeel et al., 2015). (The time between subsequent response periods in our study is 7.5 s, more than adequate time for the task-evoked response to subside.) We expected that pupil diameter during retrieval would be higher in the PI condition compared to the NoPI condition. To test this, we calculated the mean pupil diameter over the last 500 ms of the lockout period plus the entire response period (2,000 ms) for each participant and each condition and compared the means between conditions (NoPI vs. PI) using a paired t-test. (In a separate exploratory analysis, we also examined the 1,000-ms period preceding each trial's response. Results based on this variable, response time tailored period, were the same as those in the main analyses, and are presented in the OSM). We also expected that the response period pupil diameter would increase across trials within an epoch in the PI condition. To test this, a two-way ANOVA was performed to assess the effect of *order* (first half, second half of the epoch) and *condition* (PI, NoPI) on response period pupil diameter. Where relevant, throughout all of our reported analyses, inspection of the residuals did not reveal any substantive deviations from normality.

Results

Accuracy Just as in Experiment 1, a paired-samples t-test comparing the mean accuracy between the PI and NoPI conditions showed a significant effect $t(35) = 2.66$, $p = 0.012$, Cohen's $d_z = 0.444$ (Table 1, Fig. 2), such that participants' accuracy was 4.1% lower in the PI compared to the NoPI condition. Confirming the results of the paired-samples t-test, a two-way repeated-measures ANOVA showed there was a significant main effect of *condition* (PI, NoPI): $F(35) = 6.72$, $p = 0.014$, $\eta^2_p = 0.161$. There was also a significant main effect of *order* (first half, second half of the epoch): $F(35) = 5.42$, $p = 0.026$, $\eta^2_p = 0.134$. Similar to Experiment 1, we did not find any evidence for PI accumulation across trials: the two-way repeated-measures ANOVA showed that the interaction between the effects of *order* (first half, second half of the epoch) and *condition* (PI, NoPI) on accuracy was not significant: $F(35) = 0.161$, $p = 0.702$, $\eta^2_p = 0.004$ (Table S2, OSM). A visualization of average accuracy versus trial number is shown in Fig. S2 (OSM).

Physiological response variable: Response period pupil diameter A paired-samples t-test comparing the mean response period pupil diameter between PI and NoPI conditions was significant: $t(35) = -2.58$, $p = 0.014$, Cohen's $d_z = -0.511$ (Table 2, Figs. 3 and 4). Participants' pupil diameter was 0.016 mm higher in the PI condition compared to the NoPI condition.[1] Confirming the results of the paired t-test, a two-way repeated-measures ANOVA showed that there was a significant main effect of *condition* (PI, NoPI): $F(35) = 5.87$, $p = 0.021$, $\eta^2_p = 0.144$. There was no significant main effect of *order* (first half, second half of the epoch): $F(35) = 0.807$, $p = 0.375$, $\eta^2_p = 0.023$. We did not find any evidence for pupillary PI accumulation across trials: the two-way repeated-measures ANOVA showed that the interaction between *order* (first half, second half of the epoch) and *condition* (PI, NoPI) on response period pupil diameter was not significant: $F(35) = 0.323$, $p = 0.573$ (Table S3, OSM).

Relationships between pupil diameter and accuracy To test whether participants who showed greater pupillary PI effects also show greater behavioral PI effects, we performed a Kendall correlation between the pupillary effect (PI response period pupil diameter – NoPI response period pupil diameter) and the behavioral PI effect (NoPI accuracy – PI accuracy) for each participant. We did not find evidence that the

[1] To mitigate the influence of potential outliers, we confirmed the results of the original analysis with a non-parametric t-test (Wilcoxon signed-rank test: W = 177, $p = 0.013$).

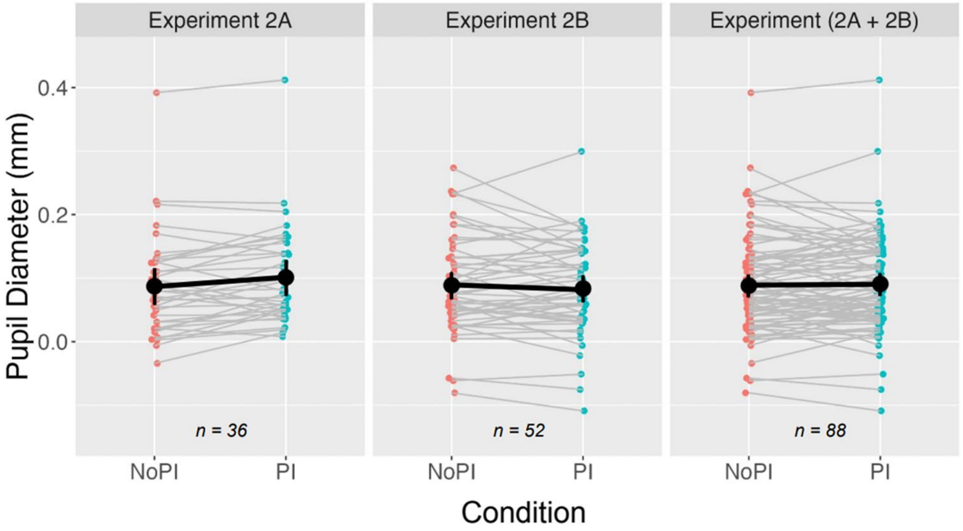

Fig. 3 The effect of *condition (PI, NoPI)* on response period pupil diameter. Panels show pupil diameter during retrieval in the PI versus NoPI conditions in Experiments 2A, 2B, and in the combined analysis (2A+2B), and reflect the data shown in Table 3. Light grey lines connect individual participants' data across condition. Large black data points indicate the mean and error bars are 95% confidence intervals

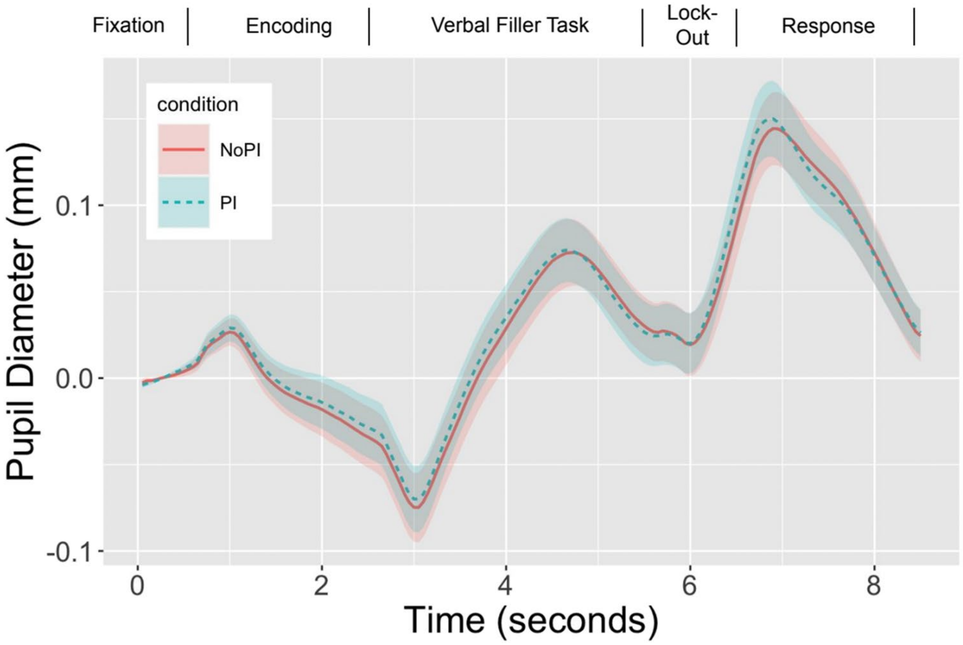

Fig. 4 Pupil diameter averaged across all trials/participants by *condition* (PI, NoPI)**.** Pupil traces in the combined Experiments 2A + 2B analysis. Data were averaged across all participants and all trials. Error ribbons indicate 95% confidence intervals

two were related: τ = -0.146, p = 0.217 (Table 3). To test whether there was a relationship between individual trial-by-trial *success (correct; incorrect)* on response period pupil diameter, we conducted a paired t-test to compare participants' average response period pupil diameter in their correct versus incorrect trials. We did not find a significant difference $t(35) = -1.14$, $p = 0.262$, Cohen's $d_z = -0.206$ (Table 4).

Table 3 Kendall correlations between participants' average pupillary and behavioral PI effects

Experiment	*Kendall's tau*	*p*
2A	-0.146	0.217
2B	-0.0935	0.328
Combined	-0.0893	0.218

Table 4 Descriptive statistics and results of paired t-tests comparing mean response period (6,000–8,500 ms after trial onset) pupil diameter (mm) within participants and between success levels *(correct, incorrect)* for all studies

	Success = incorrect			*Success = correct*						
Experiment	*M*	95% C	*SD*	*M*	95% CI	*SD*	*df*	*t*	*p*	*Cohen's* d_z
2A	0.0934	[0.0696, 0.117]	0.0694	0.103	[0.0725, 0.134]	0.0874	35	-1.14	0.262	-0.206
2B	0.105	[0.0764, 0.134]	0.101	0.0872	[0.0662, 0.108]	0.0730	51	1.92	0.0602	0.253
2A + 2B	0.100	[0.0808, 0.119]	0.0888	0.0937	[0.0764, 0.111]	0.0791	87	1.12	0.265	0.101

Discussion

Consistent with Johansson and colleagues (2018), we found that participants exerted greater effort when the task involved the resolution of interference during retrieval, but shown here, for the first time, in a VWM task. We replicated the behavioral PI effect observed in online Experiment 1 in the laboratory: accuracy was significantly lower in PI trials compared to NoPI trials, and pupil diameter during the response period was significantly higher. That said, we were not able to observe PI accumulation across trials for either the behavioral or pupil diameter outcomes. Since we *had* observed an overall behavioral effect and an overall pupil effect interference was induced, it may be that our study was not sufficiently powered to detect the time course of accumulation. Additionally, we had expected to see participants with greater interference showing greater effort during retrieval (and vice versa). However, we were not able to discern a relationship between the behavioral PI effect (NoPI accuracy – PI accuracy) and the pupillary PI effect (PI response period pupil diameter – NoPI response period pupil diameter), nor between *success* (correct, incorrect) and response period pupil diameter. Again, it may be that our study was not sufficiently powered to detect this relationship.

Experiment 2B

In Experiment 2B, our goal was to replicate Experiment 2A with a larger sample and with preregistered analyses (nearly identical to those in Experiment 2A).

Methods

Preregistration

We preregistered a replication of Experiment 2A with a larger sample size and adjustments to the number of trials. Experiment 2B was preregistered at the Center for Open Science through the Open Science Framework (Donenfeld, Blaser, & Kaldy, 2023, May 15).

Sample size rationale

Our G*Power sample size estimate (Faul et al., 2007) for the behavioral main effect ($d = 0.444$) with power = 0.8 and alpha = 0.5 was $N = 42$, and for the pupillary main effect ($d = -0.511$) it was $N = 33$. To allow for exclusions, we collected data from 60 participants.

Participants

Sixty participants were recruited using convenience sampling on the University of Massachusetts Boston campus. Seven participants were excluded for performing at or below chance, and one was removed due to low accuracy on the verbal filler task questions. Thus, a total of 52 participants were included in the final analysis. Participants' average age was 20 years, ranging from 18 to 30 years, and 35 participants identified as female. All participants were fluent in English and reported normal or corrected-to-normal vision. Twenty-four individuals identified as White and Non-Hispanic, nine individuals identified as White and Hispanic, six individuals identified as Black, eight individuals identified as Asian, and five individuals identified as belonging to more than one race. Each participant was compensated $10 cash for their participation.

Materials

Stimuli were the same as in Experiment 2A.

Procedure

The procedure followed the procedures in Experiment 2A (see Fig. 1), with one change. We had noted an increase in overall performance in Experiment 2A during the first 40 trials, so in Experiment 2B we added 40 *training trials*, followed by the usual 100 test trials. Participants were unaware that the data from the training trials would not be included in the final data analysis. Practice trials preceded the block (three practice trials preceded a question period, where participants had an opportunity to ask the experimenter questions, then three additional practice trials followed). In total, participants were given three 2-min breaks. The first break occurred after the practice trials, the second after the 40th test trial, and the third after the 80th test trial. Breaks concluded with a 5-s countdown to prepare participants for the upcoming trial.

Analysis

Our preregistered exclusions, predictions, and main statistical analyses were identical to those outlined in Experiment 2A, except the response time tailored analysis was preregistered as a potential exploratory analysis (not part of the main analysis; see OSM). The only analyses that were not listed in the preregistration were those assessing the relationship between encoding period pupil diameter and performance. Where relevant, throughout all of our analyses, inspection of the residuals did not reveal any substantive deviations from normality.

Results

Accuracy

All accuracy analyses were pre-registered. Similar to our findings in Experiments 1 and 2A, a paired-samples t-test comparing the mean accuracy between the PI and NoPI conditions was significant: $t(51) = 2.27$, $p = 0.028$, Cohen's $d_z = 0.314$ (Table 1, Fig. 2). Participants in the PI trials performed 3.2% lower than in the NoPI trials. Confirming the results of the paired t-test, the two-way repeated-measures ANOVA showed that there was a significant main effect of *condition* (PI, NoPI): $F(51) = 6.05$, $p = 0.017$, $\eta^2_p = 0.106$. There was also a significant main effect of *order* (first half, second half of the epoch), with the first half of epochs having lower average accuracy: $F(51) = 4.08$, $p = 0.049$, $\eta^2_p = 0.074$. As with Experiments 1 and 2A, we were unable to discern an accumulation of PI within epochs: the two-way repeated-measures ANOVA did not show evidence of an interaction between the effects of *order* (first half, second half of the epoch) and *condition* on accuracy: $F(51) = 3.26$, $p = 0.077$ (Table S4, OSM). A visualization of average accuracy versus trial (1–10, collapsed across epochs) is shown in Fig. S3 (OSM).

Physiological response variable: Response period pupil diameter

All retrieval-period pupil diameter analyses were pre-registered. Unlike the results of Experiment 2A, a paired-samples t-test comparing the response period pupil diameter means between PI and NoPI conditions did not show a significant effect: $t(51) = 0.947$, $p = 0.348$, Cohen's $d_z = 0.142$ (Table 2, Figs. 3 and 4). Pupil diameter in the PI trials was 0.006 mm lower than in the NoPI trials.[2] Confirming the results of the paired t-test, the two-way repeated-measures ANOVA did not show a main effect of *condition*: $F(51) = 1.26$, $p = 0.267$, $\eta^2_p = 0.024$. The main effect of *order* was also non-significant: $F(51) = 0.190$, $p = 0.665$, $\eta^2_p = 0.004$. We did not find any evidence for pupillary PI accumulation across trials in the two-way repeated-measures ANOVA: the interaction between the effects of *order* (first half, second half of the epoch) and *condition* (PI, NoPI) on response period pupil diameter was not significant: $F(51) = 0.0376$, $p = 0.847$, (Table S5, OSM).

Relationships between pupil diameter and accuracy

The result of the Kendall correlation between the pupillary effect (PI response period pupil diameter – NoPI response period pupil diameter) and the behavioral PI effect (NoPI accuracy – PI accuracy) for each participant did not show a relationship: $\tau = -0.0935$, $p = 0.328$ (Table 3). To test whether there was an effect of trial-by-trial response period pupil diameter on *success (correct, incorrect)*, we conducted a paired t-test to compare the average response period pupil diameter between each participant's set of correct trials versus their set of incorrect trials. We found that the effect of *success (correct, incorrect)* on response period pupil diameter was not statistically significant: $t(51) = 1.92$, $p = 0.0602$, Cohen's $d_z = 0.253$ (Table 4).

Discussion

In our preregistered Experiment 2B, we replicated our behavioral results from Experiments 1 and 2A, providing more evidence that PI decreases VWM performance in spatially distributed visual arrays. While we expected to find evidence of an effortful PI resolution mechanism at work during retrieval, we did not replicate our pupillary PI effect from Experiment 2A. As in Experiments 1 and 2A, there was no evidence of PI accumulation across the ten-trial epoch, in either behavioral or pupillary modalities.

Combined analysis (Experiment 2A and Experiment 2B)

Overall, while results between the studies were not contradictory, there were differences in the patterns of significance and effect sizes among the analyses. In order to address this, we combined data from studies 2A and 2B to increase power and applied the same set of analyses. Since participants in Experiments 2A and 2B were drawn

[2] To mitigate the influence of potential outliers, we confirmed the results of the original analysis with a non-parametric t-test (Wilcoxon signed-rank test: W = 753, $p = 0.563$).

from the same population and the procedures were largely identical, data from the 100 test trials (consisting of Trials 1–100 for Experiment 2A and Trials 41–140 for Experiment 2B – skipping the first 40 trials that had been preregistered as training trials) were entered into a combined analysis. There were a total of 88 participants: 36 Experiment 2A participants + 52 Experiment 2B participants = 88 combined participants; 100 trials per participant. The dataset was analyzed according to the analysis described for Experiment 2B.

Analysis

Our analytic focus for Experiments 2A and 2B has been on within-subjects assessments of pupil diameter differences in the PI versus NoPI conditions during retrieval. However, there is ample evidence that effort during *encoding* affects subsequent memory performance in WM tasks in general (Miller & Unsworth, 2021). In our paradigm, trial-by-trial fluctuations in effort during encoding were not manipulated; nevertheless, they may vary with the participant's momentary attentional state (Unsworth et al., 2018). Taking advantage of the increased sample size ($N = 88$) in this combined analysis, we assessed this "ground truth" relationship between individual differences in effort during encoding to performance by comparing the encoding pupil diameter of participants whose accuracy was above median to participants who performed below the median. Mean pupil dilation during the 2,000-ms encoding period was calculated for each *condition (PI, NoPI)* and each subject. Upper and lower median performers' pupil diameter during the encoding period were compared using an independent-samples *t*-test. Where relevant, throughout all analyses, inspection of the residuals did not reveal any substantive deviations from normality.

Results

Accuracy

The paired-samples t-test comparing mean accuracy between the PI and NoPI conditions was significant: : $t(87) = 3.42$, $p < 0.001$, Cohen's $d_z = 0.365$, consistent with our prediction (Table 1, Fig. 2). Participants performed 3.5% lower in the PI condition compared to the NoPI condition.[3] Confirming the result from the paired-samples t-test, a two-way ANOVA showed a significant main effect of *condition* (PI, NoPI): $F(87) = 12.5$, $p < 0.001$, $\eta^2_p = 0.126$. The main effect of *order* (first half, second half of the epoch) was non-significant: $F(87) = 9.41$, $p = 0.3$, $\eta^2_p = 0.098$. We did not find any evidence for behavioral PI accumulation across trials: the two-way repeated-measures ANOVA showed that the interaction between the effects of *order* (first half, second half of the epoch) and *condition* (PI, NoPI) on accuracy was not statistically significant: $F(87) = 0.961$, $p = 0.330$ (Table S6, OSM). A visualization of average accuracy versus trial number is shown in Fig. S4.

Physiological response variable: Response period pupil diameter

The paired-samples t-test comparing response period pupil diameter between PI and NoPI conditions was non-significant:: $t(87) = -0.654$, $p = 0.515$, Cohen's $d_z = -0.081$ (Table 2, Figs. 3 and 4). Pupil diameter was 0.003 mm higher in the PI condition compared to the NoPI condition.[4,5] Confirming the paired t-test result, the two-way ANOVA showed that there was no significant main effect of *condition* (PI, NoPI): $F(87) = 0.175$, $p = 0.677$, $\eta^2_p = 0.002$. There was also no significant main effect of *order* (first half, second half of the epoch): $F(87) = 0.751$, $p = 0.389$, $\eta^2_p = 0.009$.

We also did not find any evidence for pupillary PI accumulation across trials within an epoch: the two-way repeated-measures ANOVA showed that the interaction between the effects of *order* (first half, second half of the epoch) and *condition (PI, NoPI)* on response period pupil diameter was not significant: $F(87) = 0.244$, $p = 0.622$ (Table S7, OSM).

Relationships between response pupil diameter and accuracy

We did not find evidence that response period pupil diameter and their PI effect were related: $\tau = -0.0893$, $p = 0.218$ (Table 3). We also did not find evidence of an effect of *success (correct, incorrect)* on response period pupil diameter: $t(87) = 1.12$, $p = 0.265$, Cohen's $d_z = 0.101$ (Table 4).

[3] At a reviewer's request, we also conducted our core analyses on Trials 1–100 from Experiment 2A combined with Trials 1–100 from Experiment 2B (40 training trials + 60 test trials). Since we observed a practice effect in Experiment 2A, the reviewer suggested that this may be a fairer comparison. For this comparison (Trials 1–100 of Experiment 2A and Trials 1–100 of Experiment 2B), we also found a significant main effect of condition on accuracy: $t(88) = 2.44$, $p < 0.05$, consistent with our original analysis.

[4] A reviewer requested a comparison of Trials 1–100 of Experiment 2A and Trials 1–100 of Experiment 2B (see Footnote 3). Similar to our analyses in the main text, we did not find a significant main effect of *condition (PI, NoPI)* on response period pupil diameter: $t(88) = -1.59$, $p < 0.12$.

[5] To mitigate the influence of potential outliers, we confirmed the results of the original analysis with a non-parametric t-test (Wilcoxon signed-rank test; W = 1715, $p = 0.313$).

Table 5 T-test results comparing upper and lower median performers' encoding period pupil diameter in the Combined (2A + 2B) analysis

	Lower			*Upper*						
Experiment	*M*	95% CI	*SD*	*M*	95% CI	*SD*	*df*	*t*	*p*	Cohen's d
2A + 2B	-0.0109	[-0.0248, 0.0030]	0.0442	0.0108	[-0.0038, 0.0254]	0.0468	86	2.09	0.0392 *	0.475

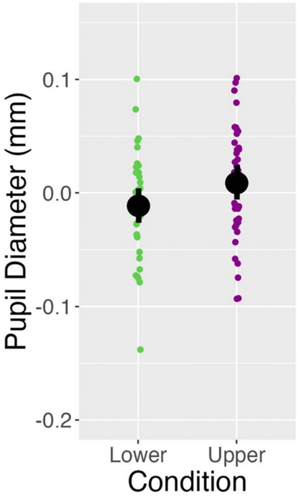

Fig. 5 Encoding period pupil diameter in the upper versus lower median performers in Experiment (2A + 2B). This figure shows upper/lower median performers' pupil diameter during the encoding period. The figure reflects the data shown in Table 5. Large data points indicate the mean and error bars the 95% confidence intervals

Relationships between encoding pupil diameter and accuracy

However, we did find that encoding period pupil diameter was higher in upper median performers compared to lower median performers: $t(86) = -2.09$, $p = 0.039$, Cohen's $d_z = 0.475$ (Table 5, Figs. 5 and 6).[6] At a reviewer's request, additional robustness checks are reported in the OSM.

Summary of results

A summary of all of our results (the observed effect sizes and significance of tests in all three studies and the combined analysis) can be found in Table 6. At a reviewer's request, we also computed adjusted likelihood ratios for our four main predictions for the combined Experiment 2A + 2B data. These are reported in the OSM.

General discussion

Proactive interference (PI) occurs when now-irrelevant memories intrude on current, task-relevant ones. Here we used a naturalistic, *what-was-where* task to test the effect of PI in visual working memory (VWM) and pupillometry to investigate the mechanism of PI resolution. Overall, we observed a robust behavioral PI effect in item-location VWM across three studies (one online, two in-lab studies, one of them preregistered; overall $N = 122$), showing that even spatially distributed stimuli are not immune to PI. However, contrary to our expectations, we did not find consistent evidence that PI induces increased cognitive effort during memory retrieval.

Our novel paradigm combines three important design elements. First, our task minimizes the possibility that different encoding strategies for unique and repeated items could influence the PI effect. Unlike previous work that used a blocked design (Endress & Potter, 2014; Makovski, 2016), in our paradigm, the condition *(PI, NoPI)* is not defined until the response period, when the participant's memory is either tested on one of the repeated items (PI condition) or one of the unique items (NoPI condition). We accomplished this by presenting both unique *and* repeated items in each trial. Second, our paradigm required participants to go beyond simple change detection, instead requiring four alternative forced-choice recognition of item-location bindings, which is consistent with a move toward tasks that better model everyday situations (Kristjánsson & Draschkow, 2021; Postma et al., 2008). Finally, we added a verbal filler task between encoding and response to protect against verbal rehearsal, which has often been neglected in PI study designs in VWM (for an exception, see Endress & Siddique, 2016). Our behavioral results are quite comparable to Makovski's (2016) existing evidence that PI affects spatial VWM with real-world, recognizable stimuli. They also expand on existing evidence that VWM is sensitive to PI in a variety of retrieval contexts and timescales.

Based on Johansson and colleagues' (2018) findings using verbal stimuli, we predicted that greater effort during retrieval would be associated with better performance and that larger pupillary effects (difference between PI and NoPI pupil diameter) would correspond to a larger behavioral PI effect (evidence of a struggle to resolve interference).

[6] To mitigate the influence of potential outliers, we confirmed the results of the original analysis with a non-parametric t-test (Mann-Whitney U test; W = 721, $p = 0.04$).

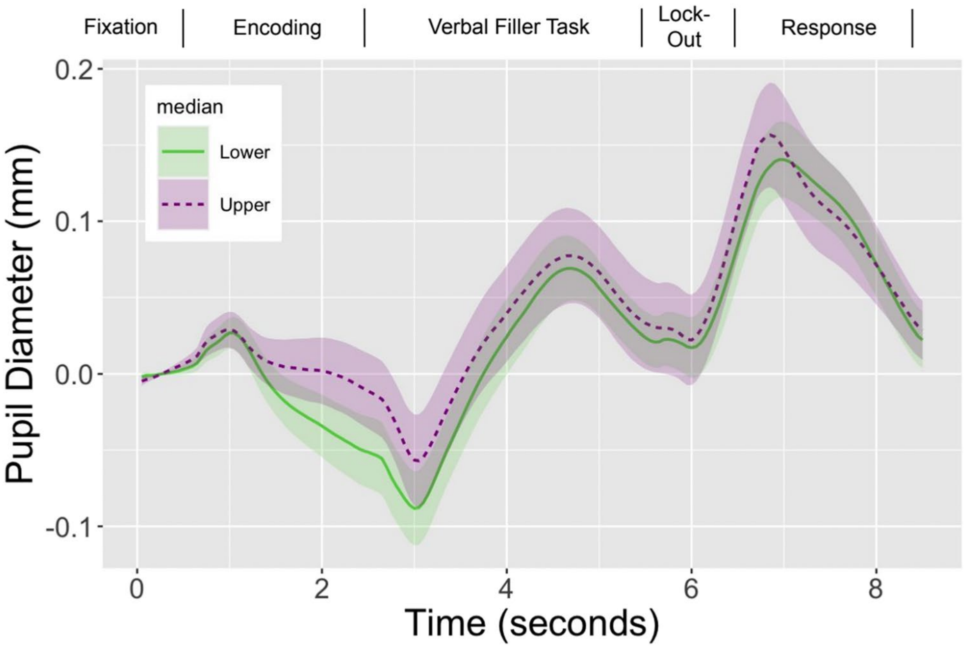

Fig. 6 Pupil diameter averaged across all trials/participants split by median performance. Baseline corrected pupil diameter traces for the combined analysis (Experiments 2A and 2B) in the upper versus lower median performers. Error ribbons indicate 95% confidence intervals

However, neither relationship was supported by our results. We speculate that the amount of interference in our task was not sufficient to trigger a substantive change in retrieval effort and/or consistently activate the mechanisms responsible for explicit interference resolution. Our mixed-condition trials, where repeated (PI) items were presented alongside unique (NoPI) items during encoding, meant that, by design, participants did not know what they would be tested on. If, instead, conditions had been blocked, there may have been a more robust apparent difference – i.e. increased *generalized* effort and/or stress in the PI block – reflected in performance and the pupil. Since our paradigm was designed to *eliminate* these generalized, potentially explicit, effort/strategy changes that might accompany a putatively "harder" condition, it better isolates any interference-resolution processes that occur during retrieval. At the same time, this removes a potential source for pupillary effects that may arise in a design with more explicit/realizable differences between conditions. In other words, the magnitude of any pupillary effect, as an index of effort toward interference resolution, will be influenced by the magnitude of the interference itself (and, as reviewed in the *Introduction*, while PI effects for verbal stimuli have been consistently large, they are more variable in the visual domain). A higher-powered study may be able to detect the present design's comparably more constrained pupillary differences between conditions.

Although we were primarily interested in pupillary correlates of interference resolution during retrieval, effort during *encoding* affects subsequent memory performance in general (Cheng et al., 2019; Miller & Unsworth, 2021), and we wanted to verify this "ground truth" relationship in our paradigm. Since participants were naïve to the condition (PI or NoPI) during encoding, we did not induce nor expect any consistent condition-dependent encoding period pupil effects. However, participants' *moment-by-moment* attentional fluctuations in engagement may nevertheless affect encoding and therefore overall performance (Unsworth et al., 2018). Indeed, in our combined (Experiments 2A + 2B) analysis, when participants were split by median performance, upper median performers had significantly larger pupils during encoding than lower median performers.

In sum, we found PI in VWM using a what-was-where task. Our paradigm was designed to eliminate generalized, potentially explicit, effort/strategy changes that might accompany a putatively "harder" PI condition, to better isolate the implicit (not strategy-driven) interference resolution processes that occur during retrieval. While we did not find consistent evidence that PI induces increased cognitive effort during memory retrieval, we observed a robust behavioral PI effect in item-location visual working memory across three studies showing that even spatially distributed stimuli are not immune to PI.

Table 6 Summary table. Effect sizes and significance of tests (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$) in our three studies and the combined Exp. 2A+2B analysis

Question	Analysis	Exp. 1	Exp. 2A	Exp. 2B (PreR)	Exp. 2A+2B	Effect size measure
		Results (EFFECT SIZES)				
Behavioral						
Does PI decrease accuracy?	Mean performance in PI and NoPI trials compared with a paired-samples t-test	0.741***	0.444*	0.314*	0.365*	d_z
Does PI accumulate over trials?	*Condition* (PI, NoPI) * *order* (first half of epoch, second half of epoch) interaction in 2*2 ANOVA	0.004	0.004	0.06	0.011	η^2_p
Pupil						
Does PI increase effort during memory ***retrieval***?	Mean pupil dilation during the response period in the PI and NPI trials compared with a paired t-test	n/a	-0.511*	0.142	-0.081	d_z
Does pupillary PI during ***retrieval*** accumulate over trials?	*Condition* (PI, NoPI) * *order* (first half of epoch, second half of epoch) interaction in 2*2 ANOVA	n/a	0.009	0.001	0.003	η^2_p
Does PI increase effort during memory retrieval? (Using a response tailored ***retrieval*** period)	Tailored response period for each individual according to their response time (1,000 ms prior to response) and compared with a paired-samples t-test	n/a	-0.261	0.358	0.030	d_z (See Suppl. Mat.)
Behavioral and Pupil						
Are the behavioral and effort (pupil diameter during ***retrieval***) effects correlated?	Kendall correlation between participants' overall behavioral PI effect (NoPI accuracy – PI accuracy) and the magnitude of the pupillary PI effect during the response period (PI pupil diameter – NoPI pupil diameter)	n/a	-0.146	-0.094	-0.089	τ
Do participants exert more effort during ***retrieval*** in correct trials?	Mean pupil diameter during the response period in the correct vs. incorrect trials compared with a paired t-test	n/a	-0.206	0.253	0.101	d_z
Do participants who perform better (in general) exert more effort during ***encoding***?	Mean pupil diameter during the encoding period in upper vs. lower median performers compared with an independent-samples t-test	n/a	–	–	0.475*	d

PreR refers to preregistered, *d* refers to Cohen's *d*, *dz* refers to Cohen's *dz* (more appropriate for within-participant comparisons)

Supplementary information The online version contains supplementary material available at https://doi.org/10.3758/s13414-024-02888-5.

Acknowledgements We would like to thank Erin Ricciarelli, Sangya Dhungana, and Tessyia Roper for their help with participant recruitment and data collection. We would also like to thank Daniel Donenfeld for substantial input in pupil preprocessing scripts.

Authors' contributions All authors contributed to the conceptualization and design of the study. JD performed the experiments, analyzed the data, and drafted the manuscript. All authors revised the manuscript, and approved the final version for submission.

Funding This work was supported by NIH #R15HD086658.

Data availability Preregistration: Experiment 2B was preregistered at the Center for Open Science through the Open Science Framework at: https://osf.io/v49nf/ [11-March-2022].

Data: The datasets generated during the current experiments are available in the Open Science Framework repository at: https://osf.io/v49nf/files/osfstorage

Code availability Pupil preprocessing and data analysis scripts are available in the Open Science Framework repository at: https://osf.io/v49nf/files/osfstorage

Declarations

Conflicts of interest/Competing interests All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

Ethics approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. Approval was granted by the Internal Review Board (IRB) of University of Massachusetts Boston: IRB study # 2020107: Visual attention and scene memory.

Consent to participate Written informed consent was obtained from all individual participants included in the study.

Consent for publication Not applicable.

Open practices statement The data and analysis scripts are available at https://osf.io/v49nf/files/osfstorage and Experiment 2B was preregistered (https://osf.io/4uvmd).

References

Anderson, J. R., & Paulson, R. (1978). Interference in memory for pictorial information. *Cognitive Psychology, 10*(2), 178–202.

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences of the United States of America, 105*(38), 14325–14329.

Brady, T. F., Konkle, T., Gill, J., Oliva, A., & Alvarez, G. A. (2013). Visual long-term memory has the same limit on Fidelity as visual working memory. *Psychological Science, 24*(6), 981–990. https://doi.org/10.1177/0956797612465439

Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review, 114*(3), 539–576.

Bruya, B., & Tang, Y.-Y. (2018). Is Attention Really Effort? Revisiting Daniel Kahneman's Influential 1973 Book Attention and Effort. *Frontiers in Psychology, 9*, 1133.

Cheng, C., Kaldy, Z., & Blaser, E. (2019). Focused attention predicts visual working memory performance in 13-month-old infants: A pupillometric study. *Developmental Cognitive Neuroscience, 36*, 100616. https://doi.org/10.1016/j.dcn.2019.100616

Cowan, N. (2017). The many faces of working memory and short-term storage. *Psychonomic Bulletin & Review, 24*(4), 1158–1170.

Crowder, M. J. (1976). Maximum likelihood estimation for dependent observations. *Journal of the Royal Statistical Society, 38*(1), 45–53.

Cyr, M., Nee, D. E., Nelson, E., Senger, T., Jonides, J., & Malapani, C. (2017). Effects of proactive interference on non-verbal working memory. *Cognitive Processing, 18*(1), 1–12.

Donenfeld, J., Blaser, E., & Kaldy, Z. (2023, May 15). *The role of effort in the resolution of proactive interference in a visual working memory task: A pupillometric study*. https://osf.io/v49nf/

Endress, A. D. (2022). Memory and Proactive Interference for spatially distributed items. *Memory & Cognition, 50*(4), 782–816.

Endress, A. D., & Potter, M. C. (2014). Large capacity temporary visual memory. *Journal of Experimental Psychology. General, 143*(2), 548–565.

Endress, A. D., & Siddique, A. (2016). The cost of proactive interference is constant across presentation conditions. *Acta Psychologica, 170*, 186–194.

Engle, R. (1975). Pupillary measurement and release from proactive inhibition. *Perceptual and Motor Skills, 41*(3), 835–842.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191.

Forbes, S. (2020). PupillometryR: An R package for preparing and analysing pupillometry data. *Journal of Open Source Software, 5*(50), 2285.

Glenberg, A. M., & Swanson, N. G. (1986). A temporal distinctiveness theory of recency and modality effects. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *12*(1), 3–15.

Hamilton, M., Roper, T., Blaser, E., & Kaldy, Z. (2024). Can't get it out of my head: Proactive interference in the visual working memory of 3- to 8-year-old children. *Developmental Psychology, 60*(3), 582–594.

Hamilton, M., Ross, A., Blaser, E., & Kaldy, Z. (2022). Proactive interference and the development of working memory. *Wiley Interdisciplinary Reviews. Cognitive Science, 13*(3), e1593.

Hartshorne, J. K. (2008). Visual working memory capacity and proactive interference. *PloS One, 3*(7), e2716.

Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., & Hilbig, B. E. (2022). Lab.js: A free, open, online study builder. *Behavior Research Methods, 54*, 556–573.

Ibanez, A. (2022). The mind's golden cage and cognition in the wild. *Trends in Cognitive Sciences, 26*(12), 1031–1034.

Jackson, I., & Sirois, S. (2009). Infant cognition: Going full factorial with pupil dilation. *Developmental Science, 12*(4), 670–679.

Johansson, R., Pärnamets, P., Bjernestedt, A., & Johansson, M. (2018). Pupil dilation tracks the dynamics of mnemonic interference resolution. *Scientific Reports, 8*(1), 4826.

Jonides, J., & Nee, D. E. (2006). Brain mechanisms of proactive interference in working memory. *Neuroscience, 139*(1), 181–193.

Jonides, J., Schumacher, E. H., Smith, E. E., Koeppe, R. A., Awh, E., Reuter-Lorenz, P. A., Marshuetz, C., & Willis, C. R. (1998). The role of parietal cortex in verbal working memory. *The Journal of Neuroscience, 18*(13), 5026–5034.

Joshi, S., & Gold, J. I. (2020). Pupil Size as a Window on Neural Substrates of Cognition. *Trends in Cognitive Sciences, 24*(6), 466–480.

Kahneman, D. (1973). *Attention and Effort*. Prentice Hall.

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science, 154*(3756), 1583–1585.

Kaldy, Z., Guillory, S. B., & Blaser, E. (2016). Delayed Match Retrieval: a novel anticipation-based visual working memory paradigm. *Developmental Science, 19*(6), 892–900.

Kincaid, J. P., & Wickens, D. D. (1970). Temporal gradient of release from proactive inhibition. *Journal of Experimental Psychology, 86*(2), 313.

Kliegl, O., & Bäuml, K.-H. T. (2021). Buildup and release from proactive interference – Cognitive and neural mechanisms. *Neuroscience & Biobehavioral Reviews, 120*, 264–278.

Kliegl, O., Pastötter, B., & Bäuml, K.-H. T. (2015). The contribution of encoding and retrieval processes to proactive interference. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *41*(6), 1778–1789.

Kristjánsson, A., & Draschkow, D. (2021). Keeping it real: Looking beyond capacity limits in visual cognition. *Attention, Perception & Psychophysics, 83*(4), 1375–1390.

Laeng, B., Sylvain, S., & Gredebäck, G. (2012). Pupillometry: A window to the preconscious? *Perspectives on Psychological Science, 7*(1), 18–27.

Liang, Y., Kaldy, Z., & Blaser, E. (2023). Young children's cost-dependent tradeoff between looking and remembering. *Journal of Vision*, *23*(9), 5766–5766.

Lin, P.-H., & Luck, S. J. (2012). Proactive interference does not meaningfully distort visual working memory capacity estimates in the canonical change detection task. *Frontiers in Psychology, 3*, 42.

Maguire, E. A. (2022). Does memory research have a realistic future? *Trends in Cognitive Sciences, 26*(12), 1043–1046.

Makovski, T. (2016). Does proactive interference play a significant role in visual working memory tasks? *Journal of Experimental Psychology*. *Learning, Memory, and Cognition, 42*(10), 1664–1672.

Makovski, T., & Jiang, Y. V. (2008). Proactive interference from items previously stored in visual working memory. *Memory & Cognition, 36*(1), 43–52.

Mathôt, S., Fabius, J., Van Heusden, E., & Van der Stigchel, S. (2018). Safe and sensible preprocessing and baseline correction of pupil-size data. *Behavior Research Methods, 50*(1), 94–106.

Mercer, T., & Fisher, L. P. (2022). Magnitude and sources of proactive interference in visual memory. *Memory*, *30*(5), 591–609.

Mercer, T., Jarvis, R.-J., Lawton, R., & Walters, F. (2022). Tracking proactive interference in visual memory. *Frontiers in Psychology, 13*, 896866.

Miller, A. L., & Unsworth, N. (2021). Attending to encode: The role of consistency and intensity of attention in learning ability. *Journal of Memory and Language, 121*, 104276.

Monsell, S. (1978). Recency, immediate recognition memory, and reaction time. *Cognitive Psychology, 10*, 465–501.

Morin, P. P., Ducharme, R., & Flash, H. (1982). Short-term memory and effects of proactive interference on heart rate. *Psychological Reports, 51*(2), 463–470.

Nastase, S. A., Goldstein, A., & Hasson, U. (2020). Keep it real: Rethinking the primacy of experimental control in cognitive neuroscience. *NeuroImage, 222*, 117254.

Oberauer, K., Awh, E., & Sutterer, D. W. (2017). The role of long-term memory in a test of visual working memory: Proactive facilitation but no proactive interference. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *43*(1), 1–22.

Oberauer, K., & Lin, H.-Y. (2023). An interference model for visual and verbal working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* https://doi.org/10.1037/xlm0001303

Oberauer, K., & Lin, H.-Y. (2017). An interference model of visual working memory. *Psychological Review, 124*(1), 21–59.

Öztekin, I., Curtis, C. E., & McElree, B. (2009). The medial temporal lobe and the left inferior prefrontal cortex jointly support interference resolution in verbal working memory. *Journal of Cognitive Neuroscience, 21*(10), 1967–1979.

Pastötter, B., Schicker, S., Niedernhuber, J., & Bäuml, K.-H. T. (2011). Retrieval during learning facilitates subsequent memory encoding. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *37*(2), 287–297.

Peavler, W. S. (1974). Pupil size, information overload, and performance differences. *Psychophysiology, 11*(5), 559–566.

Pertzov, Y., Dong, M. Y., Peich, M.-C., & Husain, M. (2012). Forgetting what was where: The fragility of object-location binding. *PloS One, 7*(10), e48214.

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences, 10*(2), 59–63.

Postma, A., Kessels, R. P. C., & van Asselen, M. (2008). How the brain remembers and forgets where things are: The neurocognition of object-location memory. *Neuroscience and Biobehavioral Reviews, 32*(8), 1339–1345.

Rondeel, E. W., van Steenbergen, H., Holland, R. W., & van Knippenberg, A. (2015). A closer look at cognitive control: Differences in resource allocation during updating, inhibition and switching as revealed by pupillometry. *Frontiers in Human Neuroscience*, *9*, 494. https://doi.org/10.3389/fnhum.2015.00494

Shevchenko, Y. (2022). Open Lab: A web application for running and sharing online Experiments. *Behavior Research Methods, 54*(6), 3118–25.

Shoval, R., Luria, R., & Makovski, T. (2020). Bridging the gap between visual temporary memory and working memory: The role of stimuli distinctiveness. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *46*(7), 1258–1269.

Shoval, R., & Makovski, T. (2021). The locus of proactive interference in visual working memory. *Journal of Experimental Psychology. Human Perception and Performance*, *47*(5), 704–715.

Shoval, R., & Makovski, T. (2022). Meaningful stimuli inflate the role of proactive interference in visual working memory. *Memory & Cognition, 50*, 1157–1168, https://doi.org/10.3758/s13421-022-01338-7

Sirois, S., Brisson, J., Blaser, E., Calignano, G., Donenfeld, J., Hepach, R., Hochmann, J.-R., et al. (2023). The pupil collaboration: A multi-lab, multi-method analysis of goal attribution in infants. *Infant Behavior & Development, 73*, 101890.

Sonkusare, S., Breakspear, M., & Guo, C. (2019). Naturalistic Stimuli in Neuroscience: Critically Acclaimed. *Trends in Cognitive Sciences, 23*(8), 699–714.

Souza, A. S., & Oberauer, K. (2015). Time-based forgetting in visual working memory reflects temporal distinctiveness, not decay. *Psychonomic Bulletin & Review, 22*(1), 156–162.

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.

Underwood, B. J. (1957). Interference and forgetting. *Psychological Review, 64*(1), 49–60.

Unsworth, N., Robison, M. K., & Miller, A. L. (2018). Pupillary correlates of fluctuations in sustained attention. *Journal of Cognitive Neuroscience, 30*(9), 1241–1253.

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.

Wilson, K. G. (1984). Psychophysiological activity and the buildup and release of proactive inhibition in short-term memory. *Psychophysiology, 21*(2), 135–142.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.