Universiteit Leiden
The Netherlands

In the Eye of the Beholder: The Relationship Between Pupil Diameter and Recognition

Wessels, Pauline

Citation

Wessels, P. (2022). *In the Eye of the Beholder: The Relationship Between Pupil Diameter and Recognition*.

Version:	Not Applicable (or Unknown)
License:	License to inclusion and publication of a Bachelor or Master thesis in the Leiden University Student Repository
Downloaded from:	https://hdl.handle.net/1887/3439624

Note: To cite this publication please use the final published version (if applicable).

In the Eye of the Beholder: The Relationship Between Pupil Diameter and Recognition

Pauline Wessels (s2088320)
Research Master Psychology (Cognitive Neuroscience)
Faculty of Social Sciences, Leiden University
Supervisor: Dr. Samarth Varma
Second reader: Dr. David Vogelsang

20-06-2022

Word count: 9000

Abstract

Interest in assessing memory processes using pupillometry has recently increased. The pupillary old/new effect is well-supported and several explanations are proposed. However, studies relating pupil diameter (PD) to word frequency and confidence provide conflicting results. Moreover, few replications have been performed despite the recent increase in awareness of them. Thus, we aimed to replicate several findings of Papesh et al. (2012) relating the PD to memory accuracy, word status, word frequency, and confidence.

Twenty students from Leiden University were presented with high- and low-frequency English words and non-words during the study phase, after which a 3-minute break followed. During the test phase, participants were again presented with the old words or with new words. They needed to make old/new judgements and indicate their confidence during recognition. PD was measured during both study and test phase.

Unlike previous findings, there was no difference in PD between old and new items, nor did we find a difference in PD between word frequencies. PD was larger for high-confidence decisions compared to medium- and low-confidence, and PD was also larger for hits compared to misses.

Thus, our results suggest that the PD is related to memory accuracy and confidence, replicating two findings of Papesh et al. (2012), both in size and direction. Additionally, the results seem to support the *strength of memory effect* explanation associated with PD changes. Moreover, the effect of confidence on PD may reflect a subjective experience of memory strength, favoring the view that sees recognition memory as one process and on a continuum, and not consisting of the two separate processes of familiarity and recognition. The PD did not seem to be related to word frequency and word status, thereby not replicating these results from Papesh et al. These null effects can be attributed to small sample size and improper manipulation of word frequency as the encoding stimuli were not in the native language of the participants.

Keywords: pupillary old/new effect, pupillometry, recognition, word frequency, confidence.

Layman's Abstract

Interest in examining the memory process of recognition by measuring the pupil diameter (PD) with an eye tracker has recently increased. The most well-supported effect is the pupillary old/new effect. Previously presented words (old words) are related to a larger PD compared to words that have not been presented previously (new words). Several explanations of this effect have been proposed. The PD has also been related to word frequency (how often a word is used in daily life) and confidence, but results are conflicting. Thus, we aimed to replicate several findings of Papesh et al. (2012) relating the PD to memory accuracy (hit or miss), word status (old or new), word frequency, and confidence.

Twenty students from Leiden University were presented with high- and -low-frequency English words and non-words during the study phase, after which a 3-minute break followed. During the test phase, participants were again presented with the words or with new words. They needed to make old/new judgements and indicate their confidence in their judgment. PD was measured during both study and test phase.

Surprisingly, there was no difference in PD between old and new items, nor was there a difference in PD between word frequencies. PD was larger for high-confidence decisions compared to medium- and low-confidence, and PD was also larger for hits compared to misses.

Thus, our results suggest that the PD is related to memory accuracy and confidence, replicating two findings of Papesh et al. (2012). Additionally, the results seem to support the strength of memory explanation. Moreover, the effect of confidence on PD may reflect an subjective experience of memory strength, favoring the view that sees recognition memory as one process and not consisting of the two separate processes of familiarity and recognition. The PD was not related to word frequency and word status, thereby not replicating these results from Papesh et al. These null effects can be attributed to small sample size and improper manipulation of word frequency as the encoding stimuli were not in the native language of the participants.

In the Eye of the Beholder: The Relationship Between Pupil Diameter and Recognition.

‘Beauty is in the eye of the beholder’ is a common expression in the English language. Although it is meant as a figure of speech, it can also be taken quite literally as the pupil expands when looking at someone you find attractive (e.g., Aboyoun & Dabbs, 1998). Not only does the pupil respond to arousal, light, and proximity (Mathôt, 2018), pupil size can also be used as a marker of several cognitive functions, like attention. More recently, studies have been attempting to link pupil size to memory processes. In this paper, we examined the relationship between pupil diameter and the memory process of recognition. Specifically, we aimed to replicate several findings of Papesh et al. (2012) by looking at the relationship between the pupil diameter and word status (old or new), word frequency, accuracy (hit or miss), and confidence levels. By replicating these findings, we aim to add to the growing body of knowledge regarding the relationship between the pupil and memory. Moreover, by performing a replication we hope to address the current replication crisis in psychological research and to improve the quality of research.

Locus Coeruleus, Norepinephrine, and the Pupil

The size of the pupil diameter (PD) has been known to change in response to changes in arousal, illumination, or changes in proximity of an object (Mathôt, 2018). More specifically, the pupil dilates when arousal increases, illumination decreases, or when an object is far away. The pupil constricts when arousal decreases, illumination increases, or when an object is close by. Constriction of the pupil is controlled by the parasympathetic nervous system. Once this nervous system is activated, the sphincter muscle that is located around the iris contracts and the pupil constricts. Moreover, constriction is mediated by the Edinger-Westphal complex (Papesh & Goldinger, 2015). When this nucleus is inhibited, the sphincter muscles relax, and the pupil can dilate again. Dilation, on the other hand, is controlled by both the activation of the sympathetic nervous system and the inhibition of the parasympathetic nervous system (Mathôt, 2018). Once the sympathetic nervous system is activated, the dilator muscles in the eye contract and the pupil dilates. This dilation pathway is mediated via the hypothalamus and the locus coeruleus (LC). More specifically, activity of the LC can inhibit the Edinger-Westphal complex. As mentioned earlier, this allows the pupil to dilate. This is why dilation of the pupil is controlled by both the activation of the sympathetic nervous system and the inhibition of the parasympathetic nervous system.

The LC is located in the pons of the brainstem, and it is the only source of the neurotransmitter norepinephrine (NE) in the brain (Aston-Jones & Cohen, 2005). This NE is distributed to all areas of the brain via projections from the LC. The pattern of release of NE depends on the mode of activity of the LC of which there are two: tonic and phasic (Benarroch, 2009). Tonic activity is displayed in response to emotional arousal or changes in illumination, and it is characterized by a sustained and regular firing pattern, whereas phasic activity is related to specific cognitive events, and it is characterized by bursts of activity in response to these cognitive events. These bursts of activity are called the task-evoked pupillary reflexes (TEPRs). Tonic changes in pupil size are generally much larger compared to phasic changes in pupil size (Papesh & Goldinger, 2015).

This LC-NE system has been suggested to play a role in the memory processes of encoding (Cohen Hoffing & Seitz, 2015), consolidation, and retrieval of memories (Demberg, 2013; Laeng et al., 2012). Specifically, when LC neurons are activated by stimuli (i.e. phasic activity), NE is released and, via the projections from the LC to the hippocampus, leads to long-term potentiation (LTP) in the dentate gyrus and CA3 in the hippocampus at β-receptors (Sara, 2009). Moreover, it seems that dopamine is released together with NE, contributing to LTP as well (Yamasaki & Takeuchi, 2017). This process of LTP facilitates the encoding of stimuli. Memory consolidation and retrieval are facilitated by the interaction between NE and other neurotransmitters, and stress hormones in the amygdala and hippocampus (Benarroch, 2009).

Since the LC-NE system is responsible for the dilation and constriction of the pupil and has been linked to memory as well, it was suggested that the pupil could be used as an index of memory. Pupillometry is a method that has been used to study memory processes. In pupillometry, an eye tracker is used to measure the pupil of the subject while they are participating in an experiment. Pupillometry is non-invasive, cheap, and since the pupillary response is a reflex, it cannot be manipulated by the participant. Although pupillometry has been used since the 1960's to study cognitive events (e.g., Kahneman & Beatty, 1966), it was nearly abandoned in the 1970's to 1990's. Only more recently has its popularity increased again and has it been used to study memory processes (Papesh & Goldinger, 2015). Most of these studies focused on capacity-limited processes in immediate, short-term, or working memory, while the studies focusing on long-term memory have been limited. However, the studies that did focus on long-term memory have found interesting, albeit sometimes conflicting, results. We discuss some of these results below.

The Pupillary Old/New Effect

Cognitive Load

The PD has been related to memory in several different ways. The pupillary old/new effect is well established. In one of the first studies that investigated human memory using the

PD, participants were presented with words during the study phase, and they had to judge whether words that were presented in the test phase were old or new (Võ et al., 2008). In the test phase, the PD was larger during words that were correctly recognized as old (hits) compared to words that were correctly recognized as new (correct rejections). Võ et al. (2008) argued that the differences in PD were due to differences in cognitive load as recollecting an old word, recovering specific contextual information that is associated with that word, and evaluating whether it is old or new would be more effortful than only deciding a word was new.

Several other explanations of the pupillary old/new effect have been suggested since. For example, Otero et al. (2011) proposed that instead of cognitive load the pupillary old/new effect reflects the strength of the memory traces. They argued that if the cognitive load hypothesis was correct and the PD indeed reflected different levels of cognitive load during recollection, that items that were encoded well during study should be associated with a smaller PD during test as recollection would be less effortful compared to items that were poorly encoded. However, deeply encoded items were related to a larger PD during test than poorly encoded items. Poorly encoded items were related to larger PD compared to new items. This suggested that PD changes during recognition reflect more than just differences in cognitive load. Moreover, assuming that the pupillary old/new effect indeed reflects the strength of memory traces, Otero et al. hypothesized that the pupillary old/new effect should persist in new items that are semantically close to old items. Previous research had shown that if new items are semantically or categorically related to the old items, that these new items are more often wrongfully judged as old. Otero et al. found exactly what they expected. Thus, the findings of Otero et al. could not be explained by the cognitive load hypothesis. Instead, they hypothesized that the pupillary old/new effect reflects the strength of the memory traces.

Several studies have supported this hypothesis. For example, in a within-subjects design, participants performed a memory task in which they had to make old/new judgments and indicate their confidence in their judgements as normal, feign amnesia, or respond ‘new’ to every trial (Heaver & Hutton, 2011). The second condition represented a high cognitive effort condition whereas the latter condition represented a low cognitive effort condition. The pupillary old/new effect was present in every condition. Again, this suggests that the pupillary old/new effect is not entirely dependent on cognitive load as there would have been a difference in the effect between the latter two conditions that differed in the cognitive load required. A similar conclusion can be drawn from Experiments 4 and 5 in the study by Brocher and Graf (2016). Instead of making old/new judgments, participants needed to indicate whether the presented stimuli were words or non-words. Brocher and Graf hypothesized that this would lead to weaker encoding as participants were never explicitly instructed to remember the stimuli. Weaker encoding in turn would lead to weaker memory traces and this would lessen the pupillary old/new effect if the strength of memory traces would indeed explain the pupillary old/new effect. They found that the pupillary old/new effect was indeed reduced and more so for non-words than for words (Experiment 4). The old/new effect disappeared entirely when participants needed to focus on responding fast (Experiment 5).

Additional opposition of the cognitive load hypothesis comes from the study of Kafkas and Montaldi (2015). In two experiments, participants performed an old/new recognition task where they had to answer on a 5-point scale (*new*, *weakly familiar*, *moderate familiar*, *strong familiar*, and *recollected*), representing a more cognitively demanding task (Experiment 1), or simply indicate *yes* or *no,* representing a lower cognitively demanding task (Experiment 2). If the cognitive load hypothesis was true, there would be a difference in the pupillary old/new effect between the two experiments as they differed in cognitive load. Additionally, using a within-subjects design, the focus was either on the familiar stimuli or on the new stimuli. This ensured that both the familiar and the novel stimuli were seen as targets. Consequently, the enlarged PD that is associated with familiar stimuli in the pupillary old/new effect could not be caused by the fact that familiar stimuli received more attention because they were seen as targets. This would cause a larger PD compared to new stimuli as those were not seen as targets and thus received less attention. Experiment 1 showed that the familiar stimuli were related to a larger PD during recognition compared to new stimuli. Thus, in contrast to what Võ et al. (2008) argued, familiar-based recognition is able to produce an enlarged PD even when recollection is absent. Moreover, since performance in the familiar-focused and novelty-focused conditions in both Experiment 1 and Experiment 2 was similar, it is unlikely that a difference in difficulty in identifying familiar or new stimuli, a difference in task difficulty, or targetness is responsible for the pupillary old/new effect. Instead, Kafkas and Montaldi (2015) proposed that different underlying processing mechanisms are used for making familiarity and novelty decisions and that this is reflected by the different pupil responses for old and new stimuli. Support for this proposal comes from an fMRI study that showed that two different, but partially overlapping brain networks, compute the familiarity and novelty signals when those stimuli are detected (Kafkas & Montaldi, 2014).

Additionally, Kafkas and Montaldi (2015) investigated whether the pupillary old/new effect was affected more by the objective or the subjective old/new status of a stimulus. The objective old/new status refers to whether the stimulus is truly old or new, whereas the subjective old/new status refers to whether the participant perceives the stimulus to be old or new independent of whether it is truly old or new. The subjective status can be congruent to the objective status (hits or correct rejections) or it can be incongruent (false alarms or misses). In line with previous research (Montefinese et al., 2013; Otero et al., 2011), misses and false alarms were related to a larger PD compared to correct rejections. Furthermore, using the extent of constriction of the pupil instead of the absolute PD like previous research, Naber et al. (2013) found a stronger pupil constriction for both objectively and subjectively new stimuli during recognition. They concluded that pupil constriction signals novelty. Thus, these studies suggest that the PD is influenced by both objective and subjective oldness or newness of a stimulus. Objective and subjective oldness of a stimulus is associated with a larger PD, whereas objective and subjective newness is associated with a constricted or a smaller PD.

To recap, the pupillary old/new effect is a well-established phenomenon. Several explanations have been proposed but the most consistent explanation is the strength of memory traces stating that a stronger memory trace is associated with a larger PD during recognition. Moreover, the pupillary old/new effect seems to be present for both objective and subjective old- and newness of the stimuli.

The Pupil and Word Frequency, Retrieval Success, and Confidence

Word Frequency

Besides the pupillary old/new effect, several other effects of the PD in response to stimuli have been established. For example, word frequency has been shown to affect the PD (Haro et al., 2017; Kuchinke et al., 2007; Papesh & Goldinger, 2011; Papesh et al., 2012). The frequency of a word refers to how often it is used in daily life. Papesh et al. (2012) showed that during encoding, non-words (NW) were associated with a larger PD compared to high-frequency (HF) and low-frequency (LF) words, but only in trials that lead to hits during recognition. During recognition, NW were associated with the largest PD, followed by LF and HF words between which there was no statistically significant difference. In contrast, Papesh and Goldinger (2011) did find a difference in PD response to HF and LF words. The PD dilated more during recognition than during encoding for both HF and LF words, but this effect was stronger for the LF words. Additionally, in the study by Kuchinke et al. (2007), participants performed a lexical decision task in which they were presented with words varying in their emotional valence and word frequency. The pupil showed a higher peak dilation for HF words compared to LF words. Since Kuchinke et al. did not investigate memory and used a different study design compared to Papesh et al., we cannot directly compare the results of these studies and therefore we cannot conclude whether the results of these studies are in line with each other. We can, however, conclude that both studies show that the pupil responds to word frequency.

Retrieval Success

Moreover, the PD seems to be able to predict which items will be remembered during recognition based on the size of the PD during encoding (Naber et al., 2013; Papesh et al., 2012). Papesh et al. (2012) found that higher cognitive effort during the encoding of stimuli, as reflected by a larger PD, predicted the accuracy in a recognition test, as the stimuli that were attended to more during encoding were more likely to be remembered during recognition. Similarly, in the study of Kucewicz et al. (2018), words that were later recalled correctly were associated with a larger PD during encoding compared to later forgotten words. Moreover, this study used free recall, suggesting that the PD during encoding is able to predict correctly remembered stimuli both during forced recognition and free recall. However, in contrast to Papesh et al. and Kucewicz et al., Naber et al. (2013) found opposite results: A stronger pupil constriction during encoding was associated with remembered compared to forgotten stimuli. This difference in results might be explained by differences is study design as Papesh et al. and Kucewicz et al. used auditorily presented and visually presented verbal stimuli, respectively, whereas Naber et al. used visually presented natural scenes as stimuli. Auditory stimuli elicit larger pupillary responses compared to visual stimuli (Klinger et al., 2011), which could explain the difference in results between Papesh et al. and Naber et al. However, it cannot explain the difference between Naber et al. and Kucewicz et al. as they both visually presented their stimuli. It might be that the type of stimuli (natural scenes versus words, respectively) could explain the difference in results.

Confidence

Lastly, the PD appears to be related to memory strength as reflected by the level of confidence of participants. During encoding, the PD was larger for items that were later remembered correctly with high confidence compared to items that were remembered correctly with lower confidence (Papesh et al., 2012). During recognition, the PD was larger for high-confidence decisions compared to low-confidence decisions. Heaver and Hutton (2011) had similar findings: During recognition, the pupil dilation ratio (PDR) was higher for high-confidence hits compared to low-confidence hits. Moreover, the PDR was higher for high-confidence hits compared to high-confidence correct rejections. However, several studies found conflicting results (Kafkas & Montaldi, 2011; Naber et al., 2013; Suzuki et al., 2018). For example, Kafkas and Montaldi (2011) concluded that a smaller PD during recognition was related to higher reported memory strength, and Naber et al. (2013) reported that at the highest level of confidence the pupil constricted more strongly for new items compared to old items. Thus, Papesh et al. (2012) and Heaver and Hutton found that a larger pupil both at encoding and recognition was associated with higher confidence during recognition, whereas Kafkas and Montaldi (2011) and Naber et al. associated a smaller or constricted pupil during recognition with higher confidence. Again, the difference in results could be explained by the differences in study design. Papesh et al. and Heaver and Hutton explicitly instructed their participants to memorize the stimuli, whereas Kafkas and Montaldi (2011) used incidental encoding. Incidental encoding eliminates TEPRs as participants do not actively attempt to remember stimuli. Moreover, incidental encoding recruits different neural processes compared to intentional encoding (Kapur et al., 1996). Thus, the difference in results between Papesh et al. and Heaver and Hutton, and Kafkas and Montaldi (2011) could be explained by using intentional versus incidental encoding. Differences between the former two studies and Naber et al. might be more difficult to explain as, like Papesh et al. and Heaver and Hutton, they used intentional encoding. However, Naber et al. used images of natural scenes as stimuli, whereas Papesh et al. and Heaver and Hutton used words as stimuli. As mentioned earlier, auditory stimuli elicit larger pupillary responses than visual stimuli (Klinger et al., 2011), which explains the difference between Naber et al. and Papesh et al. However, it does not explain the differences between Naber et al. and Heaver and Hutton as both studies visually presented the stimuli. It might be possible that the different types of stimuli (natural scenes versus nouns, respectively) could explain the differences in results.

To recap, the size of the pupil changes in response to word frequency, retrieval success, and confidence during recognition. However, for the latter two, findings remain inconsistent as to whether remembered items and high memory strength are associated with a large or a small pupil. Some conflicting findings can be explained by differences in study design, whereas others still require an explanation.

The Current Study

Given that the interest to link the pupil to long-term memory has been quite recent, the number of studies investigating this field are relatively limited. Studies with the main goal to replicate are even more scarce. The current replication crisis in psychological research, in which studies replicate findings with weaker evidence than the original study or findings are not replicated at all (Open Science Collaboration, 2015), has highlighted the need for replications in order to improve the quality of research and its conclusions. Thus, given the conflicting findings discussed previously and given the lack of replication studies in this field, we intended to replicate several findings from Papesh et al. (2012) in order to improve the quality of research in this field. Moreover, we aimed to contribute to this growing body of knowledge that links the PD to long-term memory. We focused on recognition and the effects of word frequency, confidence, word status (old or new), and accuracy (hits or miss) on the PD. Our study design was based on the design of Papesh et al. as we aimed to replicate several of their findings. Participants performed a memory task in which English words, both HF and LF words and NW, were presented during the study phase. Then, a 3-minute break followed after which the words from the study phase and new words were presented to the participants. They then needed to make an old/new judgement and indicate their confidence in their decision. The PD was measured during the study and test phases.

Our research questions were as follows:

1. Is the presentation of NW associated with a larger PD compared to HF and LF words during recognition?
2. Is there a difference in PD between hits and misses during recognition?
3. Is there a difference in PD when processing old items versus new items during recognition?
4. Is a larger PD during recognition related to higher confidence?

Based on the findings of Papesh et al. (2012) and previous research, we formulated the following hypotheses:

1. We expect a difference in PD between NW, HF, and LF words during recognition. Specifically, the PD of NW will be larger compared to HF and LF words.
2. We expect a difference in PD between hits and misses during recognition. Specifically, the PD of hits will be larger than the PD of misses.
3. We expect a difference in PD between old and new stimuli during recognition. Specifically, the PD of old stimuli will be larger than the PD of new items.
4. We expect a difference in PD depending on confidence level. Specifically, a larger PD will be associated with high confidence compared to low and medium confidence.

In this study, we focused on recognition instead of on both encoding and recognition like Papesh et al. (2012). We hope that in the future this fundamental knowledge linking the PD to recognition can be used in practical settings, like education. Studies have already started to investigate the usefulness of pupillometry in assessing the accuracy of identification in a lineup by eyewitnesses (Elphick et al., 2020), and the results are promising.

Methods

Participants

Twenty-six participants between the ages of 18 and 30 were recruited via SONA and via personal connections of the researcher. Participants were students from Leiden University and were of varying nationalities and studies. They were given a monetary reward of 7 Euros or 2 study credits for their participation, or participation was voluntary. Individuals with memory impairments, hearing impairments, impaired vision (corrected-to-normal vision is sufficient), or insufficient mastery of the English language (a native speaker, a high school degree in the Netherlands, or a bachelor or master's degree at a Dutch university was deemed sufficient) were excluded. This study was approved by the Research Ethics Committee of Leiden University (2021-06-09.S VARMA-V1-3293).

One participant was excluded as they did not make any responses, and five participants were excluded after inspection of the PD data as large sections of the PD were not measured during the study phase, contained too much noise, or had a too low number of valid samples. Thus, 20 participants remained for the analyses.

Materials

The auditory stimuli were individually recorded by a non-native male without a strong or discernable accent. 80 non-words (e.g., *flazick*), 40 high-frequency words (e.g., *also*), and 40 low-frequency words (e.g., *anvil*) were recorded. The verbal stimuli matched the ones used in Papesh et al. (2012; see Appendix).

The experiment was presented on a computer screen and the auditory stimuli were presented via Sennheiser HD 202 headphones. E-prime 3 software was used to run the experiment and record responses. A Tobii X3-120 eye tracker with a sampling frequency of 120 Hz was used to record the PD during the study and test phases. The lighting was kept at a constant, dim level to not interfere with the PD. A mouse was used to select the response during the test phase.

Lastly, informed consent forms, payment forms, and pens were present for the participants.

Procedure

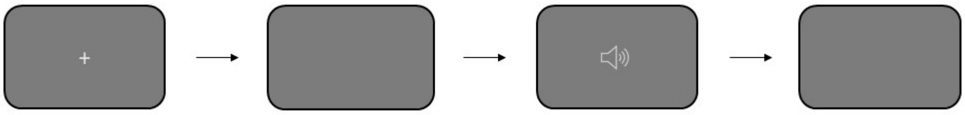
Similar to Papesh et al. (2012), the experiment consisted of a study phase, followed by a 3-minute break, and lastly a test phase. PD was measured during both study and test phases.

Upon arrival at the lab, participants were given verbal instructions by the researcher after which they signed the informed consent form. Then, participants took place behind the computer screen and eye tracker located in a cubicle and put on the headphones that they kept on throughout the entire experiment. Next, the eye tracker was calibrated by having the participants follow a red, moving dot across the screen which fixated at five different positions. Calibration was performed again if there were missed fixations or if the fixations were outside the acceptable range.

During the study phase, participants were presented with 40 NW, 20 HF, and 20 LF words in random order. They were randomly selected from the entire list of stimuli. All screens in the trial procedure (see Figure 1) had the same dark grey background to not influence the PD. First, a fixation screen with a lighter grey plus sign was presented for 1s, followed by a pre-stimulus screen of 1s, an empty screen presenting the auditory stimulus for the duration of the stimulus, and lastly a post-stimulus screen of 1s.

Figure 1.

Schematic Representation of an Encoding Trial.



Note. The trial procedure of a single encoding trial. First, a fixation screen with a dark grey background and a lighter grey plus sign was presented for 1s, followed by a pre-stimulus screen of 1s, an empty screen presenting the auditory stimulus for the duration of the stimulus, and lastly a post-stimulus screen of 1s.

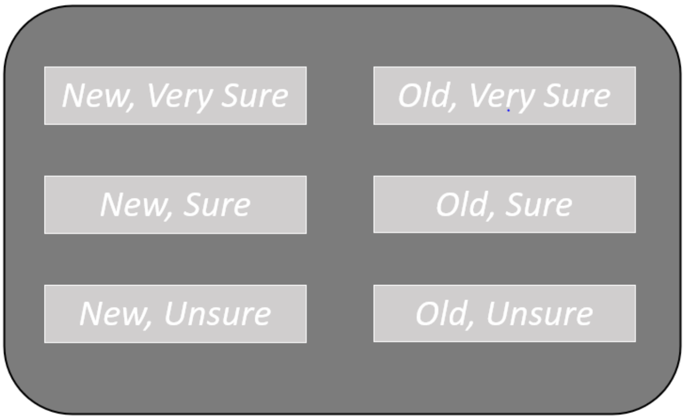
Next, a 3-minute break followed during which participants relaxed. During the test phase, participants were presented with 80 NW, 40 HF, and 40 LF words in random order. Half of each word frequency had been presented during the study phase (old) and the other half of each word frequency was presented for the first time (new). Participants had to judge whether the stimulus was old or new and indicate their confidence in their judgment. These old/new judgments and confidence ratings were made on a 6-point scale (ranging from *very sure new* to *unsure old*). Participants indicated their decision by clicking the corresponding button on screen using the mouse. This ensured that participants kept their gaze aimed at the screen during recognition.

Similar to the trial procedure during the study phase, all screens in the trial procedure in the test phase had the same dark grey background. First, a fixation screen with a lighter grey plus sign was presented for 1s, followed by a pre-stimulus screen of 1s, an empty screen presenting the stimulus for the duration of the stimulus, a post-stimulus screen of 1s, and lastly a screen of max. 3s containing the six response buttons (see Figure 2).

At the end of the study, participants were debriefed and compensated. The study took approximately 45 minutes.

Figure 2.

Schematic Representation of the Response Screen of a Recognition Trial.



Note. The last screen in a recognition trial, containing the six response buttons. The screen had a duration of max. 3s. The other screens presented previously in the recognition trial are exactly the same as the screens presented in an encoding trial procedure.

Covid-19 Protocol

Since data collection took place during the Covid-19 pandemic, Leiden University devised and provided research protocols to ensure safety of the participants and researchers. The protocols involved cleaning all surfaces before and after participants, wearing masks, ventilating the room for at least 15 minutes after participants, and a symptom checklist before the study took place for every participant and researcher. Protocols changed during the data collection depending on the current measures of the Dutch government.

Analyses

Assumption checks for normality, independence, and sphericity were performed for all statistical tests.

Recognition Accuracy

D' scores were calculated to assess recognition accuracy. These were compared across word frequency (HF, LF, and NW) using a RM ANOVA. Paired samples *t*-tests were used to compare the means of the three word frequencies if the RM ANOVA was statistically significant.

Pupillometry Data

Preprocessing. Pupillometry data was preprocessed using the PhysioDataToolbox (Sjak-Shie, 2021). After importing the data files, we applied the pupil diameter module analyzer to remove blinks and interpolate the data to fill gaps (default settings). Epochs were created for every combination of word frequency, word status, and response, for both fixations and stimuli. The durations for fixations were the same length as the fixation screen, and the duration of the stimuli were the same length as the stimuli plus 1s. Epochs were also created for trials to which no responses were made. This resulted in 50 epochs in total. After the analyzer was applied to the data set from each participant, the preprocessed data sets could be exported. Next, in Excel the eye with the highest number of valid samples was selected per participant and used for analyses. Then, the PD was baseline corrected by subtracting the average PD of the fixation epoch from the maximum PD of the stimulus epoch plus 1s (peak PD) per trial during recognition. Trials to which no responses were made or in which either the PD for the fixation epoch or the stimulus epoch was missing were removed, as the baseline-corrected PD would be incorrect. Next, for each trial the level of confidence (either high, medium, or low) and the accuracy (hit, miss, false alarm, or correct rejection) were coded based on the response on the 6-point scale and the status of the word (old or new). Finally, pivot tables were used to calculate the average peak PD for high, medium, and low confidence, and for HF_Old, LF_Old, NW_Old, HF_New, LF_New, NW_New, hits, and misses for each participant in order to be able to perform the statistical test in SPSS.

Recognition Trials. We performed a 2 (word status: old or new) x 3 (word frequency: HF, LF, or NW) RM ANOVA to assess whether PD differs for old and new words during recognition (RQ3), to assess whether PD differs per word frequency during recognition (RQ1), and to assess whether there is an interaction between these factors. If there was a statistically significant interaction, paired sampled *t*-tests were used to further investigate this interaction. Moreover, we performed a RM ANOVA with confidence (high, medium, or low) as a factor to assess whether PD differs depending on confidence during recognition (RQ4). Lastly, we performed a RM ANOVA with accuracy (hit or miss) as factor to assess whether there is a difference in PD between hits and misses during recognition (RQ2). The Bonferroni method was used to correct for multiple comparisons.

Results

Recognition Accuracy

The data met the assumptions of independence, sphericity, and normality. Based on the histogram, the distribution of *D'* for LF words was slightly skewed to the right due to one outlier. However, this was not to an extent that we cannot assume normality (p = .200).

The mean *D'* was 1.38 (SD = 0.69). There was no statistical difference in recognition accuracy between the three word types, $F(2, 38) = 0.34$, $p = .711$, $\eta_p^2 = 0.02$. Means for HF and LF words and NW can be found in Table 1.

Table 1

Means and Standard Deviations (SD) for D' and Pupil Diameter (PD) in mm per Word Type

Word Frequency	*D'*		PD	
	M	*SD*	*M*	*SD*
HF	1.45	0.90	0.92	0.37
LF	1.31	0.72	0.91	0.36
NW	1.38	0.35	0.88	0.36

Pupillometry Data

The data met the assumptions of independence and sphericity unless mentioned otherwise. Based on the histograms, the distributions of the PD for high, medium, and low confidence decisions seemed to be slightly skewed to the right. So were the distributions of the PD for LF_Old, NW_Old, LF_New, NW_New, and Miss variables. However, based on the Shapiro-Wilk tests, only NW_Old deviated significantly from normality (p = .026). The other variables were normally distributed (p > .05). PD for the variables Hit, HF_Old, and HF_New were normally distributed (p > .05).

Test Trials

The assumption of sphericity was not met for both the main effects and the interaction, therefore the Greenhouse-Geisser correction was used for the 2 x 3 RM ANOVA between word status and word type. Testing our first hypothesis, there was no statistical difference in PD between the three word types during recognition, $F(1.53, 29.12) = 2.35$, $p = .124$, $\eta_p^2 = 0.11$ (see Figure 3A). Means for HF and LF words and NW can be found in Table 1. Furthermore, testing our third hypothesis, there was no statistical difference in PD between old (M = 0.92, SD = 0.35) and new items (M = 0.88, SD = 0.38), $F(1, 19) = 2.56$, $p = .126$, η_p^2

= 0.12. Lastly, the interaction between word status and word type on PD was not statistically significant, $F(2, 26.87) = 0.34$, $p = .639$, $\eta_p^2 = 0.02$.

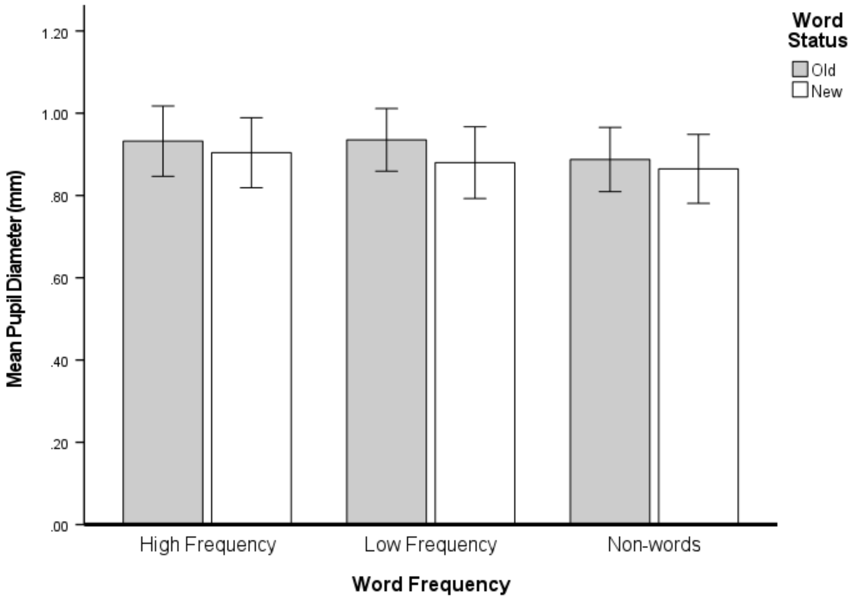
Testing our second hypothesis, there was a statistical difference in PD between hits and misses, $F(1, 19) = 6.44$, $p = .020$, $\eta_p^2 = 0.25$. This effect was large in magnitude. Hits (M = 0.95, SD = 0.36) had a larger PD compared to misses (M = 0.89, SD = 0.36; see Figure 3B).

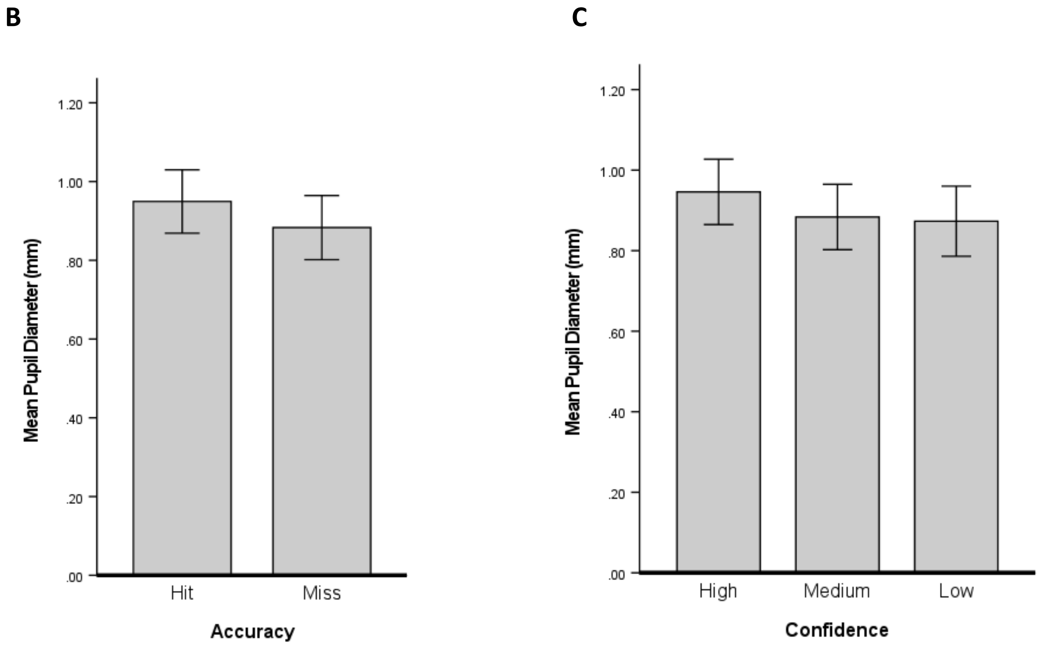
Lastly, testing our fourth hypothesis, the PD during recognition significantly differed depending on confidence level, $F(2, 38) = 14.12$, $p < .001$, $\eta_p^2 = 0.43$. This effect was large in magnitude. Further pairwise comparisons indicated that high-confidence decisions (M = 0.95, SD = 0.36) were associated with a larger PD compared to medium-confidence (M = 0.88, SD = 0.36, p = .002) and low-confidence decisions (M = 0.87, SD = 0.39, p = .002), whereas there was no statistical difference in PD between medium-confidence and low-confidence decisions (p > .05; see Figure 3C).

Figure 3.

Baseline-Corrected Pupil Diameters by Word Frequency and Word Status (A), Accuracy (B), and Confidence (C).

A





Note. Error bars represent standard error of the mean.

Discussion

We aimed to replicate several findings from Papesh et al. (2012) to contribute to the growing body of knowledge regarding the relationship between the PD and memory. Papesh et al. found that the PD was related to word status, word frequency, accuracy, and confidence of the participants. However, previous research has provided conflicting results, and few replications have been performed in this field despite the recent increase in awareness of the importance of replications in psychological research (Open Science Collaboration, 2015). In the current study, participants performed a memory task in which they were presented with English HF and LF words and NW during the study phase. During the test phase, they made old/new judgments and indicated their confidence in their judgment on a 6-point scale. The PD was measured by an eye tracker during both the study and test phase. However, for our analysis we focused only on the PD during the test phase. Based on Papesh et al., we expected the PD to be larger for NW compared to HF and LF words, the PD to be larger for hits compared to misses, the PD to be larger for old compared to new stimuli, and lastly, we expected the PD to be larger for high-confidence decisions compared to low- and medium-confidence decisions. The results showed that while some hypotheses are confirmed, others are not. Specifically, confirming our hypotheses, the PD was larger for hits compared to misses. Moreover, high-confidence decisions were related to a larger PD compared to medium- and low-confidence decisions. Surprisingly, and in contrast to what we expected, the

PD did not differ between old and new items. Thus, we did not replicate the pupillary old/new effect. Additionally, the PD did not differ between NW and HF and LF words. Therefore, based on our results, the PD seems to be related to accuracy and confidence level, but not to word status or word frequency. We will now dive further into these results.

Pupillary Old/New Effect

Most surprisingly, we did not replicate the pupillary old/new effect. The PD did not statistically differ between old and new items. This was very striking as almost, if not all, studies examining this effect do find it. We identified four possible reasons why we did not find the old/new effect. First, studies differ in which trials they include in the analysis of the old/new effect. Some studies (e.g., Võ et al., 2008) only use hits and correct rejections for the analysis of the old/new effect, whereas others (e.g., Heaver & Hutton, 2011) include all old and new trials regardless of the accuracy of the responses of the participants. However, all these studies did find the effect, therefore this cannot be the reason for the differences in results between our study and previous research.

Second, our sample size is quite small, which can lead to low power, and this in turn increases chances of a false negative or, in other words, a Type II error. However, our sample size is similar to the sample sizes of other studies that did find the effect (e.g., Brocher & Graf, 2016; Montefinese et al., 2013; Võ et al., 2008), suggesting that our sample size might not be problematic. Additionally, we performed a post-hoc power analysis to inspect our power. It showed that the power for the main effect of word status was .33. This value is much lower than the ideal value of .80 for power, suggesting that our small sample size might have led to an increased chance of a false negative. However, post-hoc power analyses are not a reliable measure of true power as the post-hoc power is completely dependent on the *p*-value, and therefore non-significant *p*-values will always produce low power (Hoenig & Heisey, 2001). Since we cannot be sure of our true power, we cannot establish whether a small sample size did or did not pose a problem for our power. Thus, a small sample size leading to low power might be a viable reason why we did not find the pupillary old/new effect. This would also mean that the other tests performed in this study are underpowered and we might have to interpret them with caution.

A third possible reason why we did not find the pupillary old/new effect is high task difficulty. If the experiment was too difficult, cognitive load would have been high in every condition. Consequently, the PD would have been large in every condition and significant differences in PD between conditions would have been unlikely. However, the overall *D*' was 1.38, suggesting that the experiment was not too difficult as accuracy was high.

A final reason explaining the lack of pupillary old/new effect is an interaction between confidence levels and word status. In Naber et al.’s (2013) study, an old/new paradigm was used and participants indicated their confidence in their decisions. Naber et al. found that the old/new effect was present overall, but when they inspected the effect at different levels of confidence, it only remained at high confidence levels. Thus, it might be that in our data the pupillary old/new effect is present at high confidence levels but not at medium and low confidence levels. These latter two levels could mask the old/new effect when one does not discriminate between confidence levels. In sum, several explanations as to why we did not find the pupillary old/new effect are presented, but the only viable explanations are low power due to a small sample size and an interaction between confidence levels and word status.

Since we did not replicate the pupillary old/new effect, we cannot state whether our study supports or rejects the cognitive load explanation of this effect. Our results seem to suggest that recognizing an old word and deciding it is old is not more or less effortful than seeing a new word and deciding it is new, as there was no difference in PD between old and new stimuli. In contrast, previous studies that supported the cognitive load hypothesis (e.g., Võ et al., 2008) suggested that recollecting an old word, recovering specific contextual information that is associated with that word, and then evaluating whether it is old or new would be more effortful than merely deciding a word was new. More effort would cause increased activation of the LC, which in turn would release more NE, and cause an increase in PD. However, since our study was not specifically designed to test the cognitive load hypothesis, we can only provide weak evidence of its refutal.

Word Frequency

Regarding our first hypothesis, we did not find a difference in PD between HF and LF words and NW, suggesting that lexicality and word frequency do not influence the PD. In contrast, previous research showed that NW were associated with a larger PD compared to HF and LF words (Papesh et al., 2012), and that HF words were associated with a larger PD than LF words (Kuchinke et al., 2007; Papesh & Goldinger, 2011). A possible explanation for the difference in findings might be that in the previous studies, participants were presented with words in their native language, whereas in the present study participants were of varying nationalities but were all presented with English words. Thus, it might be that the difference between HF and LF words was not large enough as participants might not use English enough on a daily basis for there to be differences in the frequency of words for them. Additionally, it is also possible that the LF words were not known to the participants, therefore they would have a similar effect on the PD as NW would have. Continuing this line of reasoning, we would indeed not expect a difference in PD between HF and LF words and NW if these stimuli do not differ in frequency for the participant. Moreover, Brysbaert, Mandera, and Keuleers (2017) suggested that the effect of word frequency (high-frequency words are processed more efficiently than low-frequency words) depends on the individual and their own exposure to language. This supports the idea that our participants had a different exposure to the English language compared to natives, and that this causes a difference in word frequency and in turn its effect on PD. Future studies are therefore recommended to use stimuli in the native language of the participants to replicate this effect. The current study could have been improved by asking the participants whether they indeed experienced the frequencies as we expected. Then, we could have pooled the participants based on their subjective experience of frequency. If the effect of subjective experience of word frequency on PD is similar to the subjective experience of word status in the subjective old/new effect (e.g., Montefinese et al., 2013), we might have seen a difference in PD between the three word frequencies.

Besides the limitations of a small sample size and the use of non-native speakers, noisy data could have contributed to both non-significant results. We could have improved the preprocessing of the pupil size data in order to reduce noise in the data. Instead of using the standard settings of the pupil diameter analyzer in the PhysioDataToolbox, we could have adjusted these settings to make the data even less noisy. Moreover, we could have manually removed the outliers that remained after applying the analyzer in the data of each participant to reduce noise even more.

Accuracy

Regarding our second hypothesis, we found a difference in PD between hits and misses. In line with Papesh et al. (2012), the PD of hits was larger than the PD of misses suggesting that the PD is able to distinguish between remembered and forgotten old items. Besides hits and misses, Montefinese et al. (2013) examined the PD of false alarms and correct rejections as well. They found that the PD of hits was larger than the PD of false alarms, and that the PD of false alarms was larger than the PD of both correct rejections and misses. Interestingly, they did not find a significant pupil response for misses at all, concluding that presentation of a stimulus is not necessary to elicit a pupillary response and that the only prerequisite for an increase in PD is that the individual perceives that stimulus to be old, regardless of whether it is truly old or not. We did not include the PD of false alarms and correct rejections in our analyses and therefore we cannot conclude whether we found a subjective old/new effect like Montefinese et al. We can conclude that our finding that the PD of hits is larger compared to misses is in line with Montefinese et al.'s finding regarding the PD of hits and misses. Moreover, we found a similar effect size for the effect of hits and misses on the PD ($\eta_p^2 = 0.25$) that was found in both Montefinese et al. ($\eta_p^2 = 0.22$) and in Papesh et al. ($\eta_p^2 = 0.21$). Thus, we replicated both the direction and the size of the effect of hits and misses on the PD.

Moreover, this result seems to support the strength of memory traces hypothesis as the PD seems to be able to distinguish between forgotten items and remembered items. As both hits and misses are old items but hits are remembered while misses are not, this suggests that hits are encoded better, and thus have stronger memory traces, compared to misses. Given that we found a difference in PD between hits and misses, this seems to suggest that the PD reflects the strength of memory traces as better encoded words, and thus words with stronger memory traces, are related to a larger PD compared to words that are less well encoded. We could have provided stronger evidence for this hypothesis if we had related the PD during encoding to later forgotten and remembered words. If the PD during encoding for later remembered words would have been larger than the PD during encoding for later forgotten words, this would have suggested that words that were attended to more during encoding, and therefore formed stronger memory traces, are more likely to be remembered compared to words with less strong memory traces.

Confidence

Regarding our fourth and final hypothesis, we found an effect of confidence on PD. Again in line with Papesh et al. (2012), high-confidence decisions were related to a larger PD compared to medium- and low-confidence decisions. Moreover, we found an effect size that is similar to Papesh et al.'s effect size ($\eta_p^2 = 0.43$ and $\eta_p^2 = 0.57$, respectively). It is not surprising that our effect size is smaller in magnitude compared to the effect size of Papesh et al. as most replications find a smaller effect size than the original study (Open Science Collaboration, 2015). When comparing our result to other studies, it is relatively easy to compare it to the study of Papesh et al., but it is more difficult to compare our result with the results of Naber et al. (2013) as their study only analyzed confidence in combination with word status. Thus, they did not look at the sole effect of confidence on PD like this study and Papesh et al. did. Naber et al. found that although there was a pupillary old/new effect present overall, it only remained at high levels of confidence when they looked at different levels of confidence. Thus, although we cannot compare our results to the results of Naber et al. directly, all studies do suggest that confidence levels influence the PD.

Furthermore, Heaver and Hutton (2011) proposed that confidence might reflect a subjective experience of the strength of a memory. It has been suggested that, instead of consisting of the two processes of recognition and familiarity and being an all-or-nothing process, recognition is based on the strength of a memory and varies along a continuum (Wixted, 2007). If the strength of a memory passes the threshold, it is declared to be old. Assuming that Heaver and Hutton's proposal is true and given that the PD is assumed to reflect the ongoing memory processes, high confidence is expected to be related to a larger PD compared to low confidence. This is exactly what we and Papesh et al. (2012) found. Moreover, Heaver and Hutton found a larger PDR for high-confidence hits than for high-confidence correct rejections. Like Naber et al. (2013), the old/new effect stayed significant when they examined high-confidence trials only. Together with the finding that there were no differences in PD between correct rejections and false alarms despite participants being significantly more confident in correct rejections, Heaver and Hutton concluded that the difference in PD between old and new items was not just a difference in confidence, although confidence levels do contribute to the pupil size. As mentioned earlier, it is difficult to directly compare our results to the results of Heaver and Hutton or Naber et al. as they both examined confidence in combination with word status. However, their studies do suggest that only relating confidence to PD like we and Papesh et al. did, does not tell the whole story. We can conclude, though, that our results support the hypothesis proposed by Heaver and Hutton since we found a larger PD to be associated with higher confidence.

To conclude, in this replication study we aimed to contribute to the growing body of knowledge regarding the relationship between the pupil and recognition. Surprisingly, we did not replicate the pupillary old/new effect, nor did we find a difference in PD between HF and LF words and NW. We did, however, find a difference in PD between hits and misses, and we found an effect of confidence on PD, replicating these effects of previous research both in direction and size. Moreover, our results also favor the strength of memory hypothesis over the cognitive load hypothesis. Lastly, our results support the hypothesis proposed by Heaver and Hutton (2011) stating that the effect of confidence on PD may reflect the subjective experience of memory strength, favoring the view that sees recognition memory as one process and on a continuum instead of consisting of the two separate processes of familiarity and recognition. Based on the limitations of this study, we recommend future studies to perform an a priori power analysis to ensure that the study has sufficient power, and to use verbal stimuli that are in the native language of the participants.

Appendix

Stimuli

High Frequency Words	Low Frequency Words	Non-words	
Also	Anvil	Mazz	Borse
Basis	Binder	Flazick	Lexel
Big	Blame	Infloss	Zeat
Boy	Bleed	Wurve	Squeet
Car	Boar	Sarlin	Ashwan
Care	Brood	Breen	Corple
Church	Burglar	Preck	Meegon
Day	Calf	Freem	Forch
Door	Chose	Tupe	Lapek
Else	Clove	Tramet	Remond
End	Coop	Greele	Yole
Face	Cork	Sagad	Ostrem
Fact	Fake	Goip	Sorneg
Feet	Fool	Hinsup	Rebook
Fire	Glean	Hesting	Nork
Force	Glove	Neep	Blukin
Girl	Grapes	Hine	Chark
Good	Haze	Erbow	Brant
Hand	Heal	Manuge	Daver
Head	Locker	Zolite	Loash
Heard	Moot	Vorgo	Reast
Help	Nail	Swoke	Dorve
High	Propel	Puxil	Roaken
Hope	Repeal	Fegole	Floak
House	Rouge	Humax	Kosspow
Level	Slate	Gurst	Vour
Like	Sneak	Bilark	Bawn
Made	Stamp	Modge	Plitch
Man	Starch	Vasult	Tink
Paper	Stove	Yertan	Rotail

Real	Thief	Lactain	Skave
Simple	Thumb	Rensor	Yolash
Still	Tulip	Seck	Duforst
Stood	Wade	Blemin	Sleam
Strong	Wallet	Natch	Yusock
Table	Weld	Plaret	Yince
Took	Wolf	Verm	Gisto
Water	Worm	Subar	Behick
Wife	Yolk	Glane	Murch
Woman	Yore	Serp	Redent

References

Aboyoun, D. C., & Dabbs, J. N. (1998). The Hess pupil dilation findings: Sex or novelty? *Social Behavior and Personality*, *26*, 415-419. doi: 10.2224/ sbp.1998.26.4.415

Aston-Jones, G., & Cohen, J. (2005). An integrative theory of the locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annual Review of Neuroscience*, *28*, 403-450. doi: 10.1146/annurev.neuro.28.061604.135709

Benarroch, E. E. (2009). The locus ceruleus norepinephrine system: Functional organization and potential clinical significance. *Neurology*, *73*, 1699-1704. doi: 10.1212/WNL.0b013e3181c2937c

Brocher, A., & Graf, T. (2016). Pupil old/new effects reflect stimulus encoding and decoding in short-term memory. *Psychophysiology*, *53*, 1823-1835. doi: 10.1111/psyp.12770

Brysbaert, M., Mandera, P., & Keuleers, E. (2017). The word frequency effect in word processing: an updated review. *Current Directions in Psychological Science*, *27*, 45-50. doi:10.1177/0963721417727521

Cohen Hoffing, R., & Seitz, A. R. (2015). Pupillometry as a glimpse into the neurochemical basis of human memory encoding. *Journal of Cognitive Neuroscience*, *27*, 765-774. doi: 10.1162/jocn_a_00749

Demberg, V. (2013). Pupillometry: the index of cognitive activity in a dual-task study. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *35*, 2154-2159. Retrieved from https://escholarship.org/uc/item/4vf5w6bn

Elphick, C. E. J., Pike, G. E., & Hole, G. J. (2020). You can believe our eyes: measuring implicit recognition in a lineup with pupillometry. *Psychology, Crime & Law*, *26*, 67-92. doi: 10.1080/1068316X.2019.1634196

Haro, J., Guasch, M., Vallès, B., & Ferré, P. (2017). Is pupillary response a reliable index of word recognition? Evidence from a delayed lexical decision task. *Behavior Research Methods*, *49*, 1930-1938. doi: 10.3758/s13428-016-0835-9

Heaver, B., & Hutton, S. (2011). Keeping an eye on the truth? Pupil size changes associated with recognition memory. *Memory*, *19*, 398-405. doi: 10.1080/09658211.2011.575788

Hoenig, J. M., & Heisey, D. M. (2011). The abuse of power: the pervasive fallacy of power calculation for data analysis. *The American Statistician*, *55*, 19-24. doi: 10.1198/000313001300339897

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,* 1971-1989. doi: 10.1080/17470218.2011.588335

Kafkas, A., & Montaldi, D. (2014). Two separate, but interacting, neural systems for familiarity and novelty detection: a dual-route mechanism. *Hippocampus, 24,* 516-527. doi: 10.1002/hipo.22241

Kafkas, A., & Montaldi, D. (2015). The pupillary response discriminates between subjective and objective familiarity and novelty. *Psychophysiology, 52,* 1305-1316. doi: 10.1111/psyp.12471

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science, 154,* 1583-1585. doi: 10.1126/science.154.3756.1583

Kapur, S., Tulving, E., Cabeza, R., McIntosh, A. R., Houle, S., & Craik, F. I. M. (1996). The neural correlates of intentional learning of verbal materials: a PET study. *Cognitive Brain Research*, *4*, 243-249. doi: 10.1016/S0926-6410(96)00058-4

Klinger, J., Tversky, B., & Hanrahan, P. (2011). Effects of visual and verbal presentation on cognitive load in vigilance, memory, and arithmetic tasks. *Psychophysiology*, *48*, 323-332. doi: 10.1111/j.1469-8986.2010.01069.x

Kucewicz, M. T., Dolezal, J., Kremen, V., Berry, B. M., Miller, L. R., Magee, A. L., Fabian, V., & Worrell, G. A. (2018). Pupil size reflects successful encoding and recall of human memory in humans. *Scientific Reports, 8,* 1-7. doi: 10.1038/s41598-018-23197-6

Kuchinke, L., Võ, M. L.-H., Hofmann, M., & Jacobs, A. M. (2007). Pupillary responses during lexical decisions vary with word frequency but not emotional valence. *International Journal of Psychophysiology*, *65*, 132-140. doi: 10.1016/j.ijpsycho.2007.04.004

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry: a window to the preconscious? *Perspectives on Psychological Science*, *7*, 18-27. doi: 10.1177/1745691611427305

Mathôt, S. (2018). Pupillometry: psychology, physiology, and function. *Journal of Cognition*, *1*, 1-23. doi: 10.5334/joc.18

Montefinese, M., Ambrosino, E., Fairfield, B., & Mammarella, N. (2013). The “subjective” pupil old/new effect: Is the truth plain to see?. *International Journal of Psychophysiology*, *89,* 48-56. doi: 10.1016/j.ijpsycho.2013.05.001

Naber, M., Frässle, S., Rutihauser, U., & Einhäuser, W. (2013). Pupil size signals novelty and predicts later retrieval success for declarative memories of natural scenes. *Journal of Vision*, *13,* 1-20. doi: 10.1167/13.2.11

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349,* 1-8. doi: 10.1126/science.aac4716

Otero, S. C., Weekes, B. S., & Hutton, S. B. (2011). Pupil size changes during recognition memory. *Psychophysiology, 48,* 1346-1353. doi: 10.1111/j.1469-8986.2011.01217.x

Papesh, M. H., & Goldinger, S. D. (2011). Your effort is showing! Pupil dilation reveals memory heuristics. In Higham, P., Leboe, J. (Eds), Constructions of Remembering and Metacognition. Palgrace Macmillan, pp. 215-224.

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

Papesh, M. H., & Goldinger, S. D. (2015). Pupillometry and memory: external signals of metacognitive control. In G. H. E. Gendolla, M. Tops, & S. Koole (Eds.), *Handbook of biobehavioral approaches to self-regulation* (pp. 125-139). New York, USA: Springer New York. doi: 10.1007/978-1-4939-1236-0_9

Porter, G., Troscianko, T., & Gilchrist, I. D. (2007). Effort during visual search and counting: insights from pupillometry. *Quarterly Journal of Experimental Psychology, 60,* 2011-229. doi: 10.1080/17470210600673818

Sara, S. J. (2009). The locus coeruleus and noradrenergic modulation of cognition. *Nature Reviews*, *10*, 211-223. doi: 10.1038/nrn2573

Suzuki, Y., Minami, T., & Nakauchi, S. (2018). Association between pupil dilation and implicit processing prior to object recognition via insight. *Scientific Reports*, *8,* 1-10. doi: 10.1038/s41598-018-25207-z

Sjak-Shie, E. E. (2021). PhysioData Toolbox (Version 0.6.1) [Computer software]. Retrieved from https://PhysioDataToolbox.leidenuniv.nl.

Võ, M. L.-H., 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*, 130-140. doi: 10.1111/j.1469-8986.2007.00606.x

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, *114,* 152-176. doi: 10.1037/0033-295X.114.1.152

Yamasaki, M., & Takeuchi, T. (2017). Locus coeruleus and dopamine-dependent memory consolidation. *Neural Plasticity, 2017,* 1-15. doi: 10.1155/2017/8602690