Note: This is a post-review and pre-print version of the article. It may not exactly replicate the authoritative document to be published in Biological Psychology.

Recognition memory and featural similarity between concepts: the pupil's point of view

Maria Montefinese[1, *], David Vinson[1], Ettore Ambrosini[2,3,4]

[1] *Department of Experimental Psychology, University College London, 26 Bedford Way, WC1H 0AP London (United Kingdom)*

[2] *Department of Neuroscience, University of Padua, Via Giustiniani 5, 35128, Padua (Italy)*

[3] *Department of General Psychology, University of Padua, Via Venezia 8, 35131, Padua (Italy)*

[4] *Department of Neurosciences, Imaging and Clinical Sciences, University of Chieti, Via dei Vestini 33, 66100 Chieti (Italy)*

* Corresponding author:

Maria Montefinese

Department of Experimental Psychology,

University College London,

26 Bedford Way, WC1H 0AP London (United Kingdom)

email address: m.montefinese@ucl.ac.uk, maria.montefinese@gmail.com

Abstract

Differences in pupil dilation are observed for studied compared to new items in recognition memory. According to cognitive load theory, this effect reflects the greater cognitive demands of retrieving contextual information from study phase. Pupil dilation can also occur when new items conceptually related to old ones are erroneously recognized as old, but the aspects of similarity that modulate false memory and related pupil responses remain unclear. We investigated this issue by manipulating the degree of featural similarity between new (unstudied) and old (studied) concepts in an old/new recognition task. We found that new concepts with high similarity were mistakenly identified as old and had greater pupil dilation than those with low similarity, suggesting that pupil dilation reflects the strength of evidence on which recognition judgments are based and, importantly, greater locus coeruleus and prefrontal activity determined by the higher degree of retrieval monitoring involved in recognizing these items.

Keywords: pupil diameter; locus coeruleus-noradrenaline system; semantic similarity; recognition task; retrieval monitoring

1. Introduction

The size of our pupils is not constant, but rather it varies continuously as a physiological reflex in response to different factors, the most important of which is the amount of environmental light. Interestingly, it has been found that pupillary dilation –an increase of pupil diameter as compared to the baseline level- can reflect cognitive processing (Goldinger & Papesh, 2012). Indeed, pupillary dilation is modulated by a wide variety of cognitive processes such as mental arithmetic (Hess et al., 1964), working memory (Kahneman & Beatty, 1966), emotion (Bayer, Sommer, & Schacht, 2011; Bradley & Lang, 2015; Bradley, Miccoli, Escrig, & Lang, 2008; Montefinese, Costantini, Committeri, & Ambrosini, 2015), attention (Unsworth & Robison, 2016), and memory (Kafkas & Montaldi, 2011; Naber, Frassle, Rutishauser, & Einhauser, 2013). Some researchers proposed that these effects on pupillary dilation would reflect the degree of the subject's load or mental activity (i.e., the so-called 'cognitive load theory', Kahneman & Beatty, 1966). In particular, pupil dilation has been shown to be strictly related to the activity of the locus coeruleus–noradrenaline (LC-NA) system (Aston-Jones & Cohen, 2005; Joshi, Li, Kalwani, & Gold, 2016; Murphy, O'Connell, O'Sullivan, Robertson, & Balsters, 2014; Varazzani, San-Galli, Gilardeau, & Bouret, 2015), which plays an important role in cognitive processes (Bouret & Sara, 2005; Dayan & Yu, 2006). Specifically, LC BOLD activity relates to phasic changes in pupil dilation (de Gee et al., 2017; Murphy, Robertson, Balsters, & O'connell, 2011), a finding that has recently been confirmed using direct neuronal recordings from the LC (Joshi et al., 2016). The prefrontal cortex has been proposed to be involved in the pupillary responses elicited by cognitive processes (Siegle, Steinhauer, Stenger, Konecky, & Carter, 2003) due to its peculiar functional characteristics (see Steinhauer, Siegle, Condray, & Pless, 2004). Consequently, cognitively relevant pupil dilations would correspond to a task-related increase of the activity of the LC-NA system indicating an increase of cognitive demands (see review by Goldinger & Papesh, 2012).

Changes in pupil dilation have been used recently to investigate also episodic memory judgments with an old/new recognition task (Goldinger & Papesh, 2012; Heaver & Hutton, 2010, 2011; Kafkas & Montaldi, 2011). In this task, participants make old/new recognition judgments on new unstudied items and old items, which have been presented during the learning phase. Initial studies reported that when participants encountered familiar items during a recognition memory task, they showed increased pupil dilation patterns (Gardner, Beltramo, & Krkinsky, 1975; Gardner, Philp, & Radacy, 1978). These authors proposed that

pupillary dilation reflects mental effort specifically related to the mental encoding and retrieval of information from memory rather than merely the level of general, unspecific mental effort, as posited in the cognitive load theory.

Surprisingly, interest for this pupillary effect disappeared from psychophysiological research until recently, with the pivotal study of Võ et al. (2008), who found similar patterns between pupillary and ERP waveforms, reflecting mnemonic processes. In analogy to the "ERP old/new effect" (see review by Rugg & Curran, 2007), Võ et al. (2008) coined the term "pupil old/new effect" (PON effect) to indicate the pupillomotor response observed during a recognition memory task, i.e., the participants' greater pupil dilation in response to hits compared to correct rejections. They proposed that this effect reflects greater cognitive demands of recognizing an old item in comparison to rejecting a novel item, arguing that the first process requires the retrieval of qualitative contextual information about the item's presentation in the study phase, while the process leading to correct rejections does not.

While the PON effect has been replicated in many subsequent investigations (Bradley & Lang, 2015; Brocher & Graf, 2016, 2017; Evans et al., 2017; Heaver & Hutton, 2010, 2011; Hellmer, Söderlund, & Gredebäck, 2016; Kafkas & Montaldi, 2015; Montefinese, Ambrosini, Fairfield, & Mammarella, 2013c; Otero, Weekes, & Hutton, 2011; Papesh, Goldinger, & Hout, 2012), the cognitive load explanation has been questioned. Indeed, Otero et al. (2011) provided an alternative explanation of this effect (i.e., the "strength-of-memory account"), in which the magnitude of pupil dilation for old items depends on the strength of the memory trace on the basis of participants' recollection process. In other words, pupil dilation would reflect the aggregate strength of memory upon which recognition memory judgments are made (see Otero et al., 2011) rather than cognitive demands.

Recently it has been revealed that the PON effect can also occur for false recognitions as well as for veridical recognitions (Kafkas & Montaldi, 2015; Montefinese et al., 2013c; Otero et al., 2011). False recognition occurs when subjects incorrectly claim that a new item has been encountered earlier in an experiment and it is typically inferred from "old" responses to new items that are conceptually or perceptually related to previously studied items (Schacter & Slotnick, 2004). Montefinese et al. (2013c) found greater pupil dilation for false alarms (i.e., items erroneously recognized as old) than miss trials (i.e., items erroneously judged as new), suggesting that pupil dilation is related to the recognition process itself rather than its accuracy

since it seems to be caused by the participant's "old" response even when the item was not actually old. More importantly, there were faster reaction times when participants provided a correct response (i.e., hit and correct rejections) compared to when they provided an erroneous response (i.e., miss and false alarms). This is contrary to the cognitive load theory suggested by Võ et al. (2008), which would predict greater pupillary dilation (and plausibly longer reaction times) for hit and miss items because they require greater retrieval demands of study phase-related information compared to the false alarms and correct rejection items which instead don't convey this information.

However, what determines this "subjective" PON effect? Consistent with behavioural research on false memory, which disclosed how stronger relations in lexical-semantic representation between study and test concepts might yield false memories (Brainerd, Yang, Reyna, Howe, & Mills, 2008; Cann, McRae, & Katz, 2011; Montefinese, Zannino, & Ambrosini, 2015; Roediger & McDermott, 1995), sharing a particular semantic feature between verb-items (i.e., manipulability property) induced higher false recognition rates and a consequent larger pupil diameter (Montefinese et al., 2013c). However, in this study semantic relations between concepts were not operationalized and manipulated in a rigorous way. Rather, an undifferentiated measure of semantic similarity was adopted. Indeed, from those results it is unclear whether and to what degree the PON effect is driven by other semantic properties such as associative relatedness (i.e., the probability that a word in a pair is produced in response to the other in a word association task (De Deyne & Storms, 2008a, 2008b) or by lexical co-occurrence (i.e., the frequency with which a given pair of words co-occur across large text corpora (Andrews, Vigliocco, & Vinson, 2009), nor whether it is modulated by a fine-grained measure of semantic similarity between concepts, because the manipulability feature was considered in a binary way.

To better understand how semantic relations contribute to the pupillary response to false memory, here we investigated whether a quantitative, continuous measure of featural similarity between concepts derived from a feature-listing task modulates false memory in an old/new recognition task while controlling for associative relation and lexical co-occurrence. This will allow us to clarify the specific contribution of semantic similarity in modulating the pupil response to the false memory over and above that of lexical co-occurrence and especially, associative relation, which has been shown to play a critical role in predicting false memory (Gallo & Roediger, 2002; Roediger & McDermott, 1995). This is important because of the known difficulty

to empirically distinguish these measures and, consequently, to test their specific effect (McRae, Khalkhali, & Hare, 2012; Ponzetto & Strube, 2007; Spence & Owens, 1990).

In doing this, indeed, we adopted a well-defined measure of semantic relation between concept-items (Kremer & Baroni, 2011; McRae, Cree, Seidenberg, & McNorgan, 2005; Vinson & Vigliocco, 2008), which will allow us to test whether there is a fine-grained relation between memory trace and pupil dilation. This measure of featural (semantic) similarity represents the overlap of semantic features in each concept-pair, that is, the number of shared semantic features and how strong this overlap is. For instance, the concepts *motorcycle*, *scooter* and *ship* are coordinate concepts as they belong to the same superordinate category (i.e., VEHICLES) and hence, they share some features (e.g., "*has an engine*", "*used to transport*", etc.). However, *motorcycle* and *scooter*, as compared to *motorcycle* and *ship*, are closer semantic neighbors because they share many other features in addition to those related to membership in the category VEHICLES (such as, "*has handlebars*", "*has two wheels*", "*has a saddle*", "*is fast*", etc.). When the concept *motorcycle* is activated, the fact that most of its features are shared by its semantic neighbors, such as *scooter*, also causes their activation, thus increasing the probability to erroneously recognize them as old in an old/new recognition task. This assumption has been tested recently in a behavioural study, in which the likelihood of judging a concept as "old" linearly increased with its featural similarity to the studied items belonging to the same category, suggesting that meaning overlap and sharing of semantic features specifically affect recognition performance (Montefinese et al., 2015). This result is consistent with signal detection models (Stanislaw & Todorov, 1999), where old and new items represent overlapping distributions of memory trace on which participants apply a criterion to make "old" or "new" recognition judgments (Wixted, 2007).

Here, we aimed to investigate whether the pupil response to false memories may be modulated by the semantic featural overlap between concepts, while testing the contribution of associative relatedness and lexical co-occurrence. In particular, we will first analyze the trial-by-trial variations in peak pupillary response to the to-be-recognized items by using linear mixed-effects model analysis. This will allow us to assess whether pupillary response during recognition is modulated in a fine-grained and specific way by semantic similarity while controlling for the possible influence of confounding variables. Moreover, we will analyze the temporal dynamics of the effect of our experimental manipulations on pupillary response by carrying out a mass-univariate analysis with temporal cluster-based permutation tests. Finally, we will further investigate the

functional meaning and the temporal specificity of these effects on recognition-related pupillary responses by performing a principal component analysis (PCA), which will allow us to identify the different meaningful components accounting for unique variance in the pupillary response. We expect to observe greater pupil size for the false alarms to novel concepts with high featural similarity to the previously presented concepts as compared with the novel ones with low featural similarity, with a continuously increasing function as similarity increases. We also expect to identify specific components reflecting conceptually distinct recognition-related cognitive processes modulated by semantic similarity.

2. Method

2.1. Participants

We carried out a secondary analysis of the data reported in Montefinese et al. (2015) in order to reveal whether featural similarity between concepts modulates pupil responses to veridical and false memories. Twelve of the 20 participants that took part in Montefinese et al.'s study (2015) were included in the present analyses. From this previous study, we excluded the participants who had fewer than four false alarms in either high and low featural similarity conditions (see below), so to have an adequate number of trials in order to analyse pupillary responses related to false alarms. All participants were native Italian students from the University of Chieti. According to the self-report, all participants were naïve as to the purpose of the experiment, had normal or corrected-to-normal visual acuity and were right-handed. Participants provided informed consent prior to take part in the study, which was conducted in accordance with the ethical standards of the 2013 Declaration of Helsinki for human studies of the World Medical Association.

2.2. Apparatus and Stimuli

A detailed description of the apparatus and stimuli was provided in our previous report (Montefinese et al., 2015). Briefly, participants were comfortably seated in a chair in front of a 17'' LCD computer monitor (resolution: 1024 × 768 pixels) at a distance of 57 cm. Their chin and foreheads were stabilized by means of a headrest in order to reduce movement artifacts. An infrared video-based eye-tracking device (RK-826PCI pupil/corneal tracking system; ISCAN, Burlington, MA), mounted below the monitor, recorded the pupil size of the right eye at 120 Hz. Responses were recorded through two response buttons placed horizontally on a button box. The presentation of stimuli and the recording of participants' responses were controlled by customized software (see Galati et al., 2008), implemented in MATLAB (The MathWorks, Natick, MA).

Stimuli consisted of 120 Italian words denoting basic-level concepts belonging to ten categories (i.e., animals, body parts, clothes, furnishings/fittings, furniture, housing buildings, kitchenware, plants, stationary and vehicles) taken from our feature-based semantic norms for Italian (Montefinese et al., 2013b). From this total set, we created two concept groups, each containing 60 concepts (6 items for each category) to create the study and the test lists of stimuli. To create the study list, the first group of 60 concepts (Old concepts) was added to 30 filler abstract concepts and their order was pseudorandomized such that there were no more than two consecutive concepts from the same category. To create the test list, the 60 Old concepts were added to the second group of 60 (new) concepts. The latter group of new concepts was split into two subgroups of 30 concepts that had High or Low Featural Similarity (HFS and LFS, respectively; 3 concepts with HFS and 3 concepts with LFS for each category) to the Old concepts (see below). The presentation order of the concepts in the test list was pseudorandomized as for the study list, with the additional constraint that no more than two consecutive concepts from the same condition (i.e., Old, HFS, and LFS) were presented. Moreover, all old concepts were presented in the same third of the test list as at study to minimize study-test repetition lag variability (Finnigan, Humphreys, Dennis, & Geffen, 2002). The words denoting concepts were presented in black capital letters in 28-point Arial font on a gray background (RGB: 200, 200, 200) to minimize differences in the luminance during the presentation of stimuli. As words were 4 to 13 letters long, they subtended a horizontal visual angle ranging from 4.45° to 15.84° (see Montefinese et al., 2015 for further details on the Materials section).

In order to operationalize the feature similarity in a given pair of concepts, we used a well-defined measure of semantic similarity based on a feature listing task, that is, the cosine angle between two vectors representing those concepts as the corresponding feature production frequencies taken from our feature-based semantic norms (Montefinese et al., 2013b). We thus computed the semantic similarity value for each new concept as the mean cosine between the six pairs of vectors representing that new concept and each of the six old concepts belonging to the same semantic category. This value could ideally range from 0 (minimum similarity) to 1 (maximum similarity); in our new concepts it ranged from .007 to .555. Based on this featural similarity measure, we then split the six new concepts in each category so to have three LFS and three HFS concepts. Mean semantic similarity for HFS and LFS new concepts was .31 (*SD* = .14) and .14 (*SD* = .10), respectively ($t_{(58)} = 5.52$; $p < .0001$).

We matched as much as possible the Old, HFS and LFS concepts for a number of affective (valence, arousal and dominance/control) and lexical-semantic (word length, word frequency, log-transformed number and word frequency of orthographic neighbors; familiarity, typicality, imageability, concreteness, dominance, mean rank, first occurrence, lexical availability, and mean production frequency, intercorrelational density, and percentage of encyclopedic, taxonomic, functional and sensory features) variables (for a detailed description of these variables, see Fairfield, Ambrosini, Mammarella, & Montefinese, 2017; Montefinese, Ambrosini, Fairfield, & Mammarella, 2013a, 2013b, 2014b), which could affect word recognition performance (for a discussion of this topic, see e.g., Montefinese, Ciavarro, & Ambrosini, 2015; Montefinese & Vinson, 2015) (one way ANOVAs comparing Old, HFS, and LFS concepts: all $Fs_{(1,117)} < 2.21$, all $ps > .11$). Moreover, Old, HFS and LFS concepts were also controlled for the age of acquisition (one way ANOVAs comparing Old, HFS, and LFS concepts: $F_{(1,117)} = 2.65$, $p = .08$) derived from a preliminary study on an independent sample of 436 participants (363 females and 73 males; mean age = 20.75 years, $SD = 1.99$ years) who were asked to estimate the age to which they learnt a given word, in line with previous age of acquisition norms (Bird, Franklin, & Howard, 2001; Ghyselinck, De Moor, & Brysbaert, 2000; Moors et al., 2013). In particular, we asked participants to indicate the age at which they first understood the word when somebody else used it in their presence, even when they did not use the word themselves. The validity of this procedure of age of acquisition data collection has been corroborated by normative studies, which reported a significant correlation between ratings obtained in adult participants and the percentage of words known by children of various ages (De Moor, Ghyselinck, & Brysbaert, 2000; Morrison, Chappell, & Ellis, 1997).

Importantly, to investigate the impact of semantic similarity on recognition memory we controlled for the possible effect of either lexical co-occurrence or associative relatedness, by balancing HFS and LFS concepts for word textual co-occurrence and associative relatedness in free association norms. The measure of textual co-occurrence was computed as the log-transformed number of co-occurrence in a symmetrical 10-word window, calculated between each new concept and all the old concepts belonging to the same semantic category, and normalized by the orthographic frequency of the concepts in each pair, derived from "la Repubblica" corpus of Baroni et al. (2004) ($M = 25.89$ and 26.44, $SD = 2.19$ and 1.84 for LFS and HFS, respectively; $t_{(58)} = 1.05$, $p = .30$).

To quantify the latter variable, we collected free association norms on our 120 concepts in a preliminary study by using a continuous association task (De Deyne & Storms, 2008a, 2008b) in which 50 participants on average (range: 46-55) produced five associations for each of the 120 concepts. We calculated a measure of associative strength between old and new concepts as the mean percentage of participants producing either a new concept when cued with an old concept belonging to the same category, or an old concept when cued with a new concept belonging to the same category. Importantly, a two-tailed independent *t*-test showed no significant differences between LFS and HFS concepts (mean association with old items: 1% and 2.04%, *SD* = 1.51% and 3.51%, respectively; $t_{(58)} = 1.5$, $p = .14$), allowing to minimize any possible effect of this confounding variable.

Notwithstanding our efforts, however, it was not possible to perfectly match Old, LFS, and HFS for all the confounding variables, including the ones of primary theoretical interest, lexical co-occurrence and association. For this reason, we will perform control analysis accounting for the effect of these variables to provide stronger evidence of the specific, independent effect of semantic similarity in modulating recognition-related pupillary response.

2.3. Procedure

Participants' task was to make old/new recognition judgments on new unstudied concepts (HFS and LFS) and old concepts that had been presented during the study phase. During the study phase, participants viewed 90 target concepts presented one at a time in the centre of the screen for 2000 ms with an intertrial interval (a black fixation cross on a gray background) of 2000 ms. In this phase, participants were not aware of the study purpose and were asked only to read the words carefully. Subsequently, gaze position was calibrated using a standard nine-point calibration procedure (Ambrosini, Costantini, & Sinigaglia, 2011) and then participants performed a visuospatial distractor task that lasted about 10 min in order to prevent overt rehearsal of the studied concepts (Cann et al., 2011). After that, participants viewed 120 concepts (60 targets and 60 distracters) one at a time for 3000 ms at the center of the screen following the presentation of a mask (1000 ms) composed of # symbols, as the number of letters in the word, to avoid changes in luminance. During concept presentation, participants evaluated as accurately as possible whether the concept had been viewed during the study phase ("old") or whether it was presented for the first time ("new") by pressing either the right or left button on the response box. Response mapping was balanced across participants.

During the inter-trial interval (2500 ms), a black fixation cross on a gray background indicated the blinking period during which the participants were allowed (and recommended) to blink. Indeed, we asked participants to keep their heads still, to maintain fixation and to try to restrict eye blinks to the blinking phase at the end of the trial. This procedure was adopted to reduce blinking during experimental trials and to minimize the number of excluded trials.

2.4. Pupil recording

Pupil size was recorded from the right eye during the 3000 ms interval in which each item remained on the screen during the recognition test. We developed an in-house algorithm, written with Matlab (Mathworks, Natick, MA), to remove blinks as well as other minor artifacts. Blinks were identified as sudden large changes in vertical pupil diameter and were filled in by cubic spline interpolation. The percentage of interpolated samples (mean = 3.83%) was not different across experimental conditions ($F_{(2,22)}$ = 1.38, p = .272). We excluded three trials from the analysis (.21% of the total recorded trials) due to the high number of interpolated points (> 25%). Resulting pupillary data were then smoothed using an unweighted 7-point moving median filter to remove instrumental noise. Constant fluctuation in pupil size over time and inter-individual variations were controlled by computing an index that quantifies the change of pupil diameter due to the processing of the word stimuli, corrected for the baseline (pre-stimulus) pupil diameter for each trial (Pupil Dilation Ratio, PDR). This index was computed for each sample during the 3000-ms recognition period by dividing the pupil diameter by the baseline pupil diameter (i.e., the mean pupil size during the last 200-ms prior to stimulus presentation when the stimulus mask was on screen). In this manner, pupil size changes were independent of initial pupil size and comparable between participants. We also computed the peak pupil dilation (i.e., the maximum value of the PDR during the 3000-ms recognition period) as a trial-level summary measure of the evoked pupillary response to be used in linear mixed-effects model analysis (see Results).

3. Results

3.1. Behavioural data

We analyzed recognition judgments in the present sample as done in our previous study for the whole sample of participants (Montefinese et al., 2015). Briefly, we first carried out generalized linear mixed-effects model analyses (mixed-effects hierarchical logistic regressions; *lme4* package in R) on participants' responses coded as a binary variable (0 = "new", 1 = "old") (see Wright, Horry, & Skagerberg, 2009) and assessed statistical

significance of predictors by means of Wald's *z* test. Here, however, we also carried out linear mixed-effects model analyses on participants' log-transformed response times (RTs) (see e.g., Montefinese, Turco, Piccione, & Semenza, 2017) and assessed statistical significance of predictors by means of *t* tests with Satterthwaite's approximation to degrees of freedom provided by the *lmerTest* package. Twenty-two trials in which participants failed to provide a response (1.53% of the trials) were excluded from the analyses. In all of the tested models, the random part included three parameters for the random effects of Subjects and Concepts and for the by-subjects random slopes for Trial, which accounted for potential longitudinal effects of fatigue or familiarization across participants and was coded by the trial number vector zero-centered to remove the (possible) spurious correlation between the by-subjects random intercepts and slopes. Moreover, the fixed part included the parameters for the fixed effect of Intercept and Trial, as well as the parameters for the fixed effects of interest. Variables were scaled when needed to facilitate model convergence. The final models aimed to confirm the results of our previous report were determined by using the log-likelihood ratio test (for a detailed description of the procedure, see Montefinese et al., 2014) and were fitted after excluding observations with absolute standard residuals greater than 2 (always < 5% of the data). We also tested additional control models to rule out the possibility that our results were biased by the effect of lexical-semantic confounding variables (see below).

3.1.1. Recognition judgments

First, we fitted participants' responses by using a model including the fixed effect for the categorical variable of major interest, Condition, which accounts for the effect of the item true status (i.e., Old, HFS and LFS). This analysis revealed that the log odds of (erroneously) evaluating LFS and HFS concepts as ''old'' were both lower than that of (correctly) evaluating an Old concept as ''old'' (respectively, b = -1.47 and -1.17, SE = .19 and .18, z = -7.88 and -6.54, one-tailed $p < 10^{-14}$ and 10^{-10}). We next tested an additional model directly contrasting HFS and LFS concepts. This analysis revealed that the log odds of evaluating unstudied items as ''old'' was higher for HFS as compared to LFS (b = .48, SE = .27, z = 1.77, one-tailed p = .038).

These results were corroborated by the assessment of participants' recognition performance relying on signal detection theory measures. Based on participants' proportion of hits (HIT) and misses (MISS) for the Old items, as well as the proportions of false alarms (FA) and correct rejections (CR) for the HFS and

LFS unstudied items, we calculated measures for sensitivity (*d*') and decision bias (*C*) following Stanislaw and Todorov (1999). This analysis confirmed that FA rates were significantly higher for HFS than LFS items (respectively, .28 and .23, *SD* = .06 and .08; $t_{(11)}$ = 2.85, one-tailed *p* = .008, Cohen's *d* = .83). Moreover, participants recognized LFS items with significantly better sensitivity (*d*' = .85, *SD* = .34) and a significantly more conservative decision criterion (*C* = .35, *SD* = .22) as compared to HFS items (*d*' = .68, *SD* = .33; *C* = .26, *SD* = .19; both $ts_{(11)}$ = 2.97, one-tailed *p*s = .006, Cohen's *d*s = .86; see Figure 1, left panel).

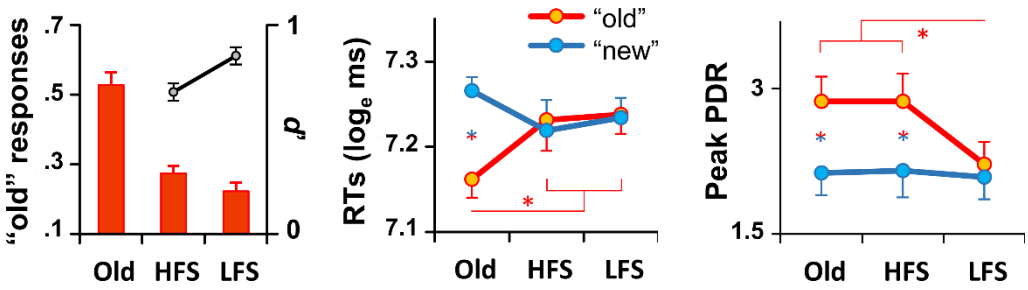

Figure 1. Descriptive results of the mixed-effects analyses.
The figure shows the unweighted mean values for the proportion of "old" responses (left panel), the RTs (middle panel), and the peak pupil dilation (right panel) as a function of Condition (Old, HFS, LFS) and Response ("old" and "new", in red and blue, respectively). The inset in the left panel shows the *d*' values for HFS and LFS conditions. Asterisks represent significant effects at the mixed-effects analyses. Error bars represent within-subject standard errors (Morey, 2008).

We then replicated our previous analysis assessing whether a more fine-grained structure of semantic similarity can influence false alarm rates for new concepts after controlling for the possible influence of confounding variables (Montefinese, Ambrosini, Fairfield, & Mammarella, 2014a). In brief, we fitted participants' responses to unstudied items with a mixed-effects model in which the fixed effect Condition was replaced with the continuous predictor semantic similarity. Moreover, the model included three parameters accounting for the effects of associative relatedness, lexical co-occurrence, and concept familiarity; note that this model was chosen to confirm our previous findings with the present reduced sample (see Montefinese et al., 2014a for details about the choice of the included variables). The analysis confirmed our previous results, revealing that the log odds of (erroneously) evaluating unstudied items as ''old'' was significantly and positively related to semantic similarity between unstudied and studied concepts (*b* = .31, *SE* = .13, *z* = 2.34, one-tailed *p* = .010) as well as to the familiarity of the unstudied concepts (*b* = 1.01, *SE* = .16, *z* = 6.22,one-

tailed $p < 10^{-9}$), but the effect of associative relatedness and lexical co-occurrence were not significant (respectively, b = -.06 and -.21, SE = .12 and .15, z = -.56 and -1.42, p = .578 and .155).

To sum up, the analyses on participants' recognition judgments not only revealed the existence of a difference in recognizing HFS as compared to LFS concepts, but also suggested that participants' recognition performance was modulated in a fine-grained way by semantic similarity between them and old concepts after controlling for the influence of associative relatedness and lexical co-occurrence, which did not reliably affect it[1].

3.1.2. Response times

We first fitted participants' RTs by using a linear mixed-effects model including parameters for the fixed effects of the categorical variable Condition and Response ("old" vs "new") and their interaction. This model revealed the significant main effect of Response factor (b = -.12, SE = .02, t = 6.29, $p < 10^{-9}$), showing that participants' RTs were faster when providing "old" responses as compared to "new" ones. Moreover, the analysis yielded a significant Condition by Response interaction, showing that participants' RTs were higher when falsely recognizing both HFS and LFS items as "old" (i.e., FA trials) as compared to when correctly recognizing Old ones (i.e., HIT trials) (respectively, b = .13 and .14, SE = .03 and .04, t = 3.64 and 3.74, both ps < .001); conversely, participants' correct "new" responses to HFS and LFS items (i.e., CR trials) were slightly faster but statistically undistinguishable from erroneous ones given to Old items (i.e., MISS trials) (both bs = -.03, SEs = .02, ts ≤ 1.26, ps ≥ .210; see Figure 1, middle panel). An additional model directly contrasting HFS and LFS concepts failed to reveal any significant differences between them (all bs ≤ .02, SEs ≥ .03, $|t|$s ≤ .74, p ≥ .464).

[1] It should be noted here that our concepts were not perfectly matched for all the remaining confounding variables, including lexical co-occurrence and association (see Apparatus and Stimuli). Therefore, in order to provide stronger evidence of the unique effect of semantic similarity in modulating participants' responses, all the reported results were confirmed by testing additional models accounting for the effect of these variables. In particular, we assessed the specific impact of semantic similarity while controlling as much as possible for the combined effect of the other confounding variables (but avoiding multicollinearity issues). We thus first applied PCA and extracted five factors accounting for more than 62% of the total variance based on inspection of eigenvalues and factorial solution. We then tested whether the inclusion of Condition or semantic similarity significantly improved the fit of models also controlling for the effect of the five PCA-derived factors. The results of the log-likelihood ratio test showed better fit for when Condition (accounting for the difference between Old, LFS, and HFS) was added to a full model already including the five PCA-derived factors ($\chi^2_{(2)} = 68.38$, $p < 10^{-14}$), while the addition of semantic similarity to the relative full model was only marginally significant ($\chi^2_{(1)} = 2.69$, p = .101).

We then carried out the same analysis described above for recognition judgments to assess the fine-grained effect of semantic similarity on RTs after controlling for the possible influence of confounding variables. Apart from a significant but paradoxically positive effect of lexical co-occurrence (b = .02, SE = .01, t = 2.02, p = .049), showing that RTs decreased as the lexical co-occurrence between unstudied and studied items increased, the analysis failed to reveal significant effects for semantic similarity (b = -.02, SE = .01, t = - 1.56, p = .123), its interaction with the Response factor (b = .01, SE = .02, t = .59, p = .555), or any other predictor (all bs ≤ .02, SEs ≥ .01, $|t|$s ≤ 1.46, p ≥ .151).[2]

3.2. Pupillary data

3.2.1. Peak pupil dilation

Pupillary data were first analyzed following the same analytical approach employed for behavioural data. Therefore, we first fitted participants' peak pupil dilation by using a linear mixed-effects model including parameters for the fixed effects of the categorical variables Condition and Response and their interaction as done for the RTs analysis. This analysis confirmed the results shown for the RTs data, revealing the significant main effect of Response factor (b = .07, SE = .01, t = 6.76, $p < 10^{-10}$). Indeed, participants' peak pupil dilation related to correct "old" responses (i.e., HIT trials) was significantly higher than that related to erroneous "new" responses (i.e., MISS trials). Moreover, the peak pupil dilation for HIT trials was significantly higher than that for false alarms to LFS items (b = -.07, SE = .02, t = -2.68, p = .008), but statistically undistinguishable from that related to false alarms to HFS (b < .01, SE = .02, t = .02, p = .842). Finally, participants' peak pupil dilation related to erroneous "new" responses to Old items (i.e., MISS trials) was statistically undistinguishable from that related to correct rejections of both HFS and LFS items (both bs ≤ .01, SEs ≥ .01, $|t|$s ≤ .43, ps ≥ .666; see Figure 1, right panel). The additional model directly contrasting HFS and LFS concepts further revealed that participants' peak pupil dilation to false alarms was significantly higher for HFS than LFS items (b = .06, SE = .02, t = 2.61, p = .009). It is important here to note that additional control analyses revealed that the effects of primary theoretical interest reported here were not biased by the timing of decision or response selection

[2] Note that additional control analyses as those described for the recognition judgments revealed that the inclusion of the Condition and Condition by Response parameters to a model including the effect of the five PCA-derived factors significantly improved the model fit ($\chi^2_{(4)}$ = 15.06, p = .005), fully confirming the reported results. On the other hand, the inclusion of the semantic similarity and semantic similarity by Response parameters to the corresponding full model was not justified ($\chi^2_{(2)}$ = .25, p = .880), again confirming the reported results.

(i.e., the RTs). Indeed, the inclusion of the parameters for the effect of Condition and its interaction with the Response factor to models including the effect of RTs significantly improved the model fit in both cases (respectively, $\chi^2_{(4)} = 13.24$, $p = .010$; $\chi^2_{(2)} = 10.70$, $p = .005$), thus providing stronger evidence of the unique effect of semantic similarity in modulating participants' peak pupil dilation over and above mere response-related or time-on-task processes.

We then assessed the fine-grained effect of semantic similarity on peak pupil dilation. Again, the effect of the Response factor was significant ($b = .05$, $SE = .01$, $t = 3.92$, $p < 10^{-4}$), confirming that peak pupil dilation was higher for "old" than "new" responses, that is, for false alarms than misses. Moreover, the analysis revealed that this effect was modulated by semantic similarity ($b = .02$, $SE = .01$, $t = 1.96$, $p = .050$), showing that the greater semantic similarity between unstudied and studied items was, the higher peak pupil dilation to false alarms was. By contrast, associative relatedness and lexical co-occurrence had no significant effect (respectively, $b = -.009$ and $.003$, $SE = .007$ and $.007$, $t = -1.31$ and $.44$, $p = .195$ and $.661$). In this case, the inclusion of the parameters for the effect of semantic similarity and its interaction with the Response factor to a model including the effect of RTs marginally improved the model fit ($\chi^2_{(2)} = 4.85$, $p = .088$).[3]

3.2.2. Pupil diameter change

We then analyzed the temporal dynamics of the participants' pupillary response during recognition of our stimuli by carrying out a massive univariate analysis followed by a non-parametric cluster-based permutation test (Maris & Oostenveld, 2007). For each time point of the participants' mean baseline-corrected pupil trace (i.e., the PDR) during the time window including the 1000-ms pre-stimulus mask presentation and the 3000-ms stimulus presentation, we assessed the effect of our experimental manipulations on the participants' pupillary response by performing a repeated measure ANOVA with Condition (Old, HFS, and LFS) and Response ("old", "new") as within-subject factors. We then corrected the results for multiple comparisons (480 tests) by carrying out a two-tailed non-parametric cluster-based permutation test based on the cluster-mass statistic (5000 permutations). With this analysis, the full time series is thus scanned blindly to

[3] Again, additional control analyses as those described for the behavioural results fully confirmed the reported results. Indeed, the inclusion of the parameters for the effect of Condition or semantic similarity (and the corresponding interactions with the Response factor) to a model including the effect of the five PCA-derived factors significantly improved the model fit in both cases (respectively, $\chi^2_{(4)} = 11.79$, $p = .019$; $\chi^2_{(2)} = 5.97$, $p = .050$), thus providing stronger evidence of the unique effect of semantic similarity in modulating participants' peak pupil dilation.

cluster together the temporally adjacent data points that exhibit a significant difference between conditions. For each statistical effect of the ANOVA, the observed cluster-level statistic (the cluster mass) is then calculated by summing the *F*-values of the data points composing each cluster. Finally, this observed cluster-mass statistic is compared with a reference null distribution generated by randomly permuting the data across the two conditions, computing the (permuted) cluster-mass statistic for each cluster in this random re-sample, and retaining the maximum cluster mass-statistic. By repeating these three steps 5000 times, a distribution of permuted cluster mass values is obtained and then used to calculate the probability of having, under the assumption that the data in the two conditions are exchangeable (i.e., not different), a cluster-mass statistic at least as extreme as the empirically observed cluster-mass statistics; in other words, the *p* value for each observed cluster. It is important to note that the false positive rate of this non-parametric cluster-based permutation test is controlled at the same alpha level used to determine statistical significance (.05). This analytical approach permits to overcome the limitations of other approaches classically used in literature, such as the use of summary measures (i.e., peak or mean) of pupil dilation and the procedures requiring some sort of averaging within time bins or across contiguous timepoints. Indeed, these approaches 1) entail arbitrary choices to decide the size of the time bins or windows, 2) may lead to severe violations of the sphericity assumption, and 3) also lead to the loss of temporal resolution.

The cluster-based permutation test revealed the significant main effect of the Condition factor in a time window ranging from 800 to 2300 ms. This result was due to greater PDR values for both Old and HFS concepts as compared to LFS ones (respectively, in a 733-1975 ms and a 958-2325 ms time window), as revealed by two post-hoc massive *t* tests followed by cluster-based permutation tests, while no significant difference emerged between PDR for Old and HFS concepts.

The analysis also revealed the significant main effect of the Response factor in a time window starting from 658 ms and lasting for all the duration of the time window of analysis. This result was due to a greater PDR for participants' "old" responses as compared to "new" ones. These two main effects were further qualified by a significant Condition by Response interaction, as detected by the cluster-based permutation test in a time window ranging from 867 to 1962 ms (see the gray shaded region in Figure 2, upper panel).

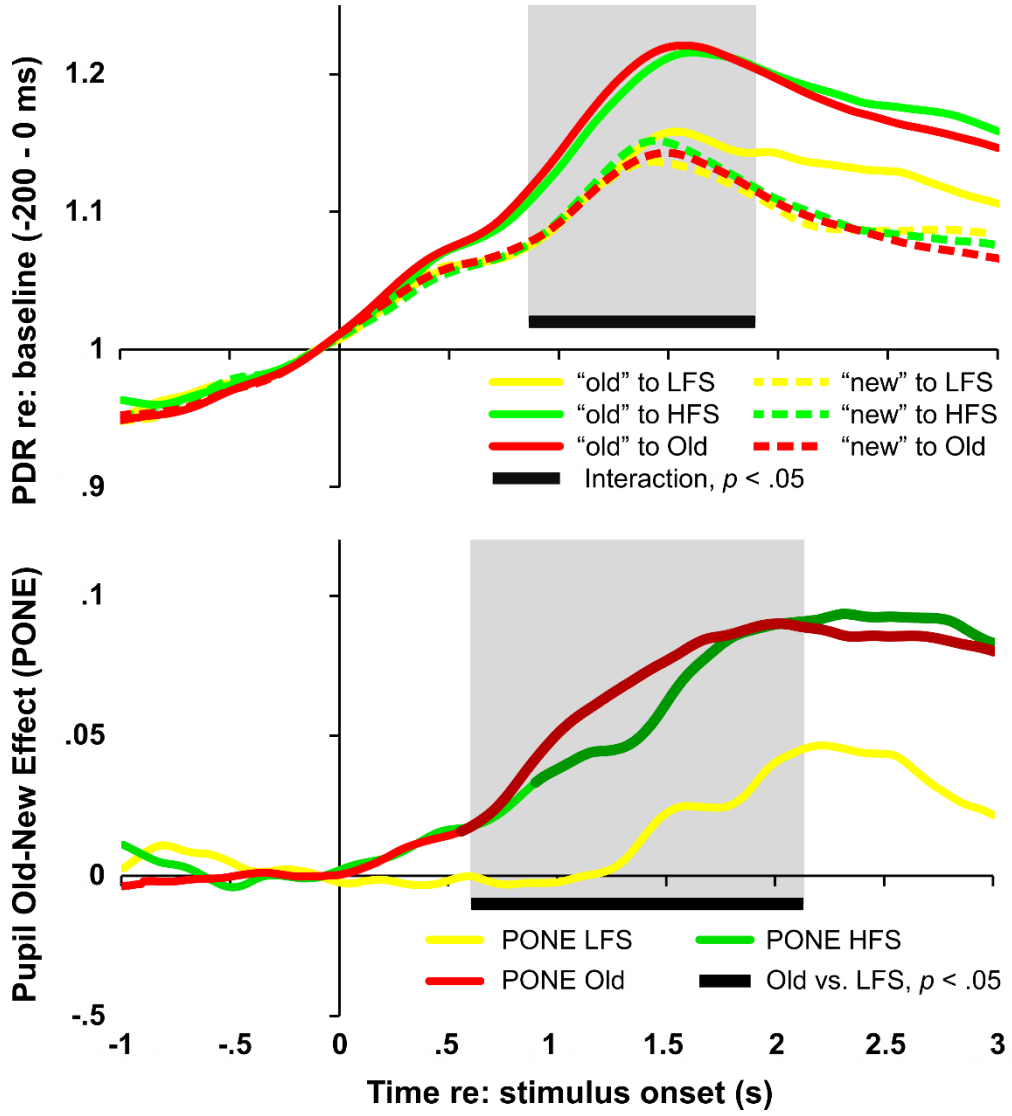

Figure 2. Results of the mass-univariate analysis, Condition by Response interaction.

The plot in the upper panel shows the time course of the PDR values as a function of Condition (Old, HFS, and LFS, in red, green, and yellow, respectively) and Response ("old" and "new", continuous and dashed lines, respectively). The gray shaded region represents the time window during which the Condition by Response interaction was significant at the cluster-based permutation test. The plot in the lower panel shows the time course of the PON effects (PONE) as a function of Condition. The darker parts of the red and green lines represent the time windows during which the corresponding PONE were significantly different from 0. The gray shaded region represents the time window during which the PONE for Old concepts was significantly greater than that for LFS ones.

We further investigated this interaction by carrying out post-hoc massive *t* tests followed by cluster-based permutation tests. This analysis revealed that the pupil old-new effect (i.e., the difference in PDR between trials with "old" and "new" responses) was significant for both Old and HFS concepts, respectively in time windows starting at 558 and 900 ms and lasting for all the duration of the time window of analysis (see

the darker part of the red and green lines in Figure 2, upper panel), but not for LFS ones. Moreover, post-hoc pairwise comparisons between the pupil old-new effect in each condition revealed that the Condition by Response interaction was due to the fact that the pupil old-new effect for Old concepts was significantly greater as compared to LFS concepts (608-2135 ms time window, see the grey shaded region in Figure 2, lower panel) but not HFS ones. However, no difference between LFS and HFS concepts has been observed.

3.2.3. Principal component analysis of PDR

We also carried out a principal components analysis (PCA) of pupillary response traces in order to investigate the time course of independent underlying factors that could help revealing the functional meaning of the differences in pupil dilation response we observed in the analyses of raw PDR data. Indeed, even if the mass-univariate analysis described above has a high temporal resolution (which allowed us to investigate the temporal profile our experimental effects), it is not able to differentiate between independent effects driven by different cognitive processes. Moreover, in our case this problem is exacerbated by the fact that the different effects we found in the mass univariate analysis had a strong temporal overlap. By contrast, the PCA identifies a small number of unique meaningful components of participants' pupillary response corresponding to different, independent cognitive processes (e.g. an early component indicates perceptual and attentional processes in response to stimulus presentation, a middle component indicates the active cognitive processing of a stimulus, a late component is related to decision processes, response selection and execution). This analysis thus allows us to infer more directly on which component (and thus cognitive process) our manipulation has an effect.

This multivariate approach allows identifying a small number of components reflecting systematic effects over many contiguous PDR time points. This analysis was performed on the PDR time courses including the 200-ms baseline and the 3000-ms stimulus presentation phases (384 time points, which were treated as dependent variables). Following previous studies using PCA on pupil data (Jainta & Baccino, 2010; Nowack, Milfont, & van der Meer, 2013; Nuthmann & van der Meer, 2005), and based on various standard criteria (i.e., > 5% of explained variance and inspection of the Scree plot and the factorial solution), we extracted three components accounting for slightly more than 90% of the total variance. We then applied a Varimax rotation to the factorial solution in order to improve it and concentrate high loadings for each factor to a specific portion of the PDR trace. As shown in Figure 3, each component is characterized by a distinct

time course of loadings throughout the stimulus presentation period. After visual inspection of the factor loadings, we renamed the components based on the time course of their loadings, so that the components (F1, F2, and F3) were ordered according to latencies to their peak loadings. The PCA thus separated the PDR trace into three time-dependent components: As shown in Figure 3, the F1 component had higher loadings from the onset of the stimulus to 475 ms and explained 18% of the total variance; the F2 one had higher loading in an intermediate time window ranging from 475 to 500 ms and explained 28.8% of the variance; the F3 component had higher loadings in the remaining portion of the PDR time course and explained the larger portion of variance (43.7%).

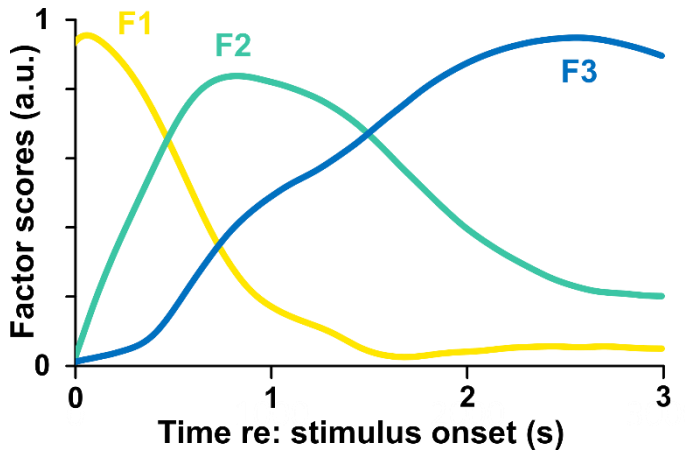

Figure 3. Results of the PCA analysis, factor loadings.
The figure shows the time course of the factor loadings for the three components identified by the PCA.

We then submitted the factor scores of every component to a Condition by Response repeated measures ANOVA. The F1 component was not significantly affected by our experimental manipulations (all Fs $\leq$ .82, ps $\geq$ .452, η^2_ps $\leq$ .07; see Figure 4, left panel). The ANOVA carried out on the F2 component revealed the main effect of the Condition factor ($F_{(2,22)}$ = 3.70, p = .041, η^2_p = .25; see Figure 4, middle panel). A Newman-Keuls's post-hoc test showed that, on average, factor scores for LFS items were significantly lower than those for Old ones (p = .042) and marginally significantly lower than those for HFS items (p = .058). The effects of Response and the Response by Condition interaction were not significant (respectively, $F_{(1,11)}$ = 1.23, p = .291, η^2_p = .10, and $F_{(2,22)}$ = 2.10, p = .146, η^2_p = .16). On the contrary, the ANOVA carried out on the F3 component revealed the significant effects of Response factor ($F_{(1,11)}$ = 25.68, p < .001, η^2_p = .70), with higher factor scores for "old" than "new" responses. This effect was further qualified by the significant Response by Condition

interaction ($F_{(2,22)}$ = 3.53, p = .047, η^2_p = .24; see Figure 4, right panel), which was explained by the fact that the old-new effect for LFS items was significantly smaller than that for HFS ones (p = .046) and marginally significantly lower than those for Old items (p = .070).

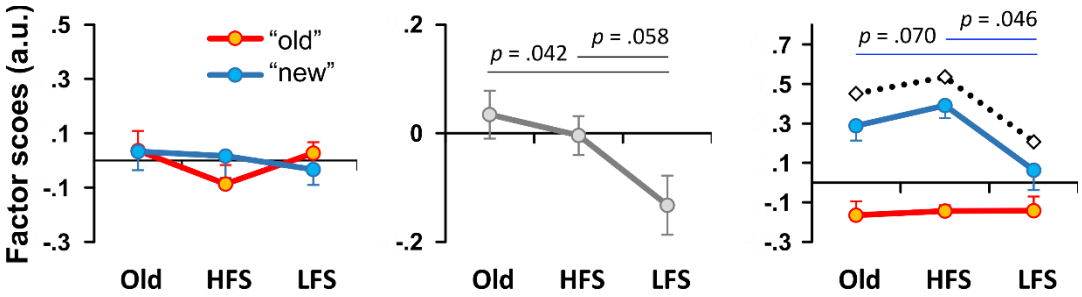

Figure 4. Results of the ANOVAs on PCA factor scores.
The figure shows the factor scores for the three components derived from the PCA analysis (F1, F2, and F3, from left to right) as a function of Condition (Old, HFS, LFS) and Response ("old" and "new", in red and blue, respectively). The left and right panels reflect the Condition by Response interaction, whereas the middle panel reflects only the main effect of Condition, which is the only significant factor in the analysis of F2. The black dotted line in the right panel represents the PON effects. P values are derived from Newman-Keuls post-hoc tests. Error bars represent within-subject standard errors (Morey, 2008).

Finally, a repeated measures ANOVA also including the component (F1, F2, and F3) as a within-subjects factor confirmed and extended the results of the previous ANOVAs. Indeed, this analysis revealed a Component by Response interaction ($F_{(2,22)}$ = 8.81, p = .002, η^2_p = .44), showing that F3 was specifically modulated by response type, with larger factor scores for "old" responses for this component as compared to all of the other conditions (all ps ≤ .007). The analysis also revealed a significant Condition by Response interaction ($F_{(2,22)}$ = 4.81, p = .019, η^2_p = .30), with significant old-new differences for Old and HFS items only (both ps = .003), and with a significant difference between "old" responses for LFS items and both HFS and Old ones (respectively, p = .002 and .004), which in turn did not differ between each other (p = .870).

4. Discussion

In this study, we tested whether featural similarity between concepts modulates false memory and recognition-related pupil responses while controlling for association and co-occurrence in an old/new recognition task. In the recognition phase, the old concepts were presented along with new concepts that either had a high or low degree of featural similarity to them. In particular, we operationalized semantic similarity as a measure of

meaning overlap derived by our feature-based norms (Montefinese et al., 2013a, 2013b): the mean cosine similarity between pairs of vectors representing a new item and each old item belonging to the same semantic category. We found that new concepts with high similarity were more often mistakenly identified as old and were related to greater pupil dilation that those with low similarity.

The behavioral results in the present sample replicated those from our previous study with the larger sample of participants, in which we showed that compared with LFS concepts, HFS concepts had significantly higher log odds of being falsely recognized as old, even after partialling out the effect of confounding variables, including associative relatedness and lexical co-occurrence, showing a fine-grained relation between featural similarity and false memories. Indeed, in line with signal detection theory, we posit that the greater the featural similarity and meaning overlap between novel and old concepts, the greater the strength of evidence and, therefore, the greater the likelihood new concepts with high similarity will be recognized as "old" as compared to the new concepts with low similarity. These results are in agreement with recent reports, showing that new items falsely recognized share semantic features with old items previously studied (Brainerd, Reyna & Ceci, 2008; Cann et al., 2011: Montefinese et al., 2013c).

Furthermore, and more important, the results of the analyses on pupillary response to recognition memory confirm that the pupil size increases in response not only to accurate memories, but also to false memories (Montefinese et al., 2013c; Otero et al., 2011), suggesting that the fact that an item had already been presented is neither sufficient nor necessary to evoke a pupil dilation.

First, we found that the pupillary response was stronger, on average, when participants gave an "old" response or not compared to when participants gave a "new" response, replicating previous studies (Montefinese et al., 2013c; Otero et al., 2011). The analysis of the PDR traces revealed that this effect was sustained, and this was confirmed by the PCA, in which we found larger factor scores for the old response compared to the new one specifically due to a relatively later, sustained component of pupillary responses (F3). This response-related sustained effect on pupillary response would thus reflect decision processes, response selection, execution and later post-processing stages, consistently with previous studies (Jainta & Baccino, 2010; Nowack et al., 2013; Nuthmann & van der Meer, 2005). This old-new response difference might also be explained, at least in part, by the retrieval success interpretation proposed to explain old-new differences in the sustained activity of right anterior prefrontal cortex commonly found in early functional

studies of recognition memory (Buckner, Koutstaal, Schacter, Wagner, & Rosen, 1998; Henson, Rugg, Shallice, & Dolan, 2000; Rugg, Fletcher, Chua, & Dolan, 1999).

Moreover, the present findings indicate that the pupil old/new effect can occur when participants provided "old" responses not only to already-presented items (i.e., the "classical" PON effect for accurate recognitions), but also to novel ones (i.e., the 'subjective' PON effect; Montefinese et al., 2013c), replicating our previous findings (Montefinese et al., 2013c; Otero et al., 2011). In our previous study, we proposed that judging an item as being "old" is "the necessary and sufficient condition to evoke a pupillary response" (p. 54, Montefinese et al., 2013c). The present results extend our previous ones by indicating that this is true to some extent, as no reliable "subjective" PON effect was found here when participants provided "old" responses to unpresented concepts that shared little semantic information with the already presented ones. This suggests that a certain level of shared information between novel and old items is needed to be reached to determine a pupil dilation in response to false memory. Therefore, judging an item as "old" is a necessary but not sufficient condition to evoke a pupillary response during recognition. Indeed, here we found a reliable "subjective" PON effect selectively for unpresented concepts that had a high semantic similarity with the already presented ones. Moreover, we also found that the second and third PCA components were modulated by the item status, showing a significant difference in factor scores between LFS concepts and Old and HFS ones. Finally, the mixed-models analysis revealed a fine-grained relation between pupil response and memory traces by showing that the peak PON effect was modulated by a quantitative measure of featural similarity between concepts. The PON effect we found seems thus to reflect the aggregate strength of evidence on which recognition judgments are based. In fact, both Old and HFS concepts generate stronger memory traces compared to the LFS ones: while correctly recognized old concepts have stronger memory traces, deriving from actually being presented during study, the incorrectly recognized new ones receive associative activation from other items (i.e., a "fake" memory trace), that in our case is stronger for the concepts with high similarity than those with low similarity. The present results thus indicate that pupil dilation can differentiate between different types of false alarms, showing how the "subjective" strength of the participants' recognition signal modulates pupil size.

The finding that the pupil responds to different false recognition judgments raises interesting questions about its functional meaning and underlying neurobiology. It has been reported that the pupil response may

reflect cortical activity determined by the LC-NA system (Aston-Jones & Cohen, 2005), and a direct correlation between pupillary response and locus coeruleus activity has also been shown during memory tests in human beings (Sterpenich et al., 2006). In particular, this system responds to salient and potentially relevant stimuli and regulates the allocation of cognitive resources (Aston-Jones & Cohen, 2005; Sara, 2009; see also Nieuwenhuis, De Geus, & Aston-Jones, 2011; Ambrosini, Vastano, Montefinese, & Ciavarro, 2013), releasing noradrenaline especially in prefrontal cortex, and plays an important role in different cognitive processes, such as working memory, decision-making, attention, memory retrieval and executive functions (Sara, 2009). It is interesting to note that there is some evidence for a right hemisphere asymmetry of the LC-NA system (Posner & Petersen, 1990), primarily from animal experiments in which frontal lesions decreased NA levels in cortex and LC (Robinson, 1979; Robinson & Coyle, 1980 see Oke, Keller, Mefford, & Adams, 1978 for evidence in human thalamus). The right-lateralization of the LC-NA system parallels those of the prefrontal cortex in recognition memory, that is well established in functional studies on memory (Cabeza, Rao, Wagner, Mayer, & Schacter, 2001; Fletcher, 1998; Henson, Rugg, Shallice, Josephs, & Dolan, 1999; Henson, Shallice, & Dolan, 1999; Schacter, Buckner, Koutstaal, Dale, & Rosen, 1997; Schacter, Curran, Galluccio, Milberg, & Bates, 1996). For example, Schacter et al. (1996) reported that a region in the dorsolateral/anterior prefrontal cortex, associated with the retrieval monitoring (Dobbins, Foley, Schacter, & Wagner, 2002; Dobbins, Rice, Wagner, & Schacter, 2003; Rugg, Fletcher, Frith, Frackowiak, & Dolan, 1996), presented greater activity during false than true recognition. Subsequent evidence (Cabeza et al., 2001; Schacter et al., 1997; Slotnick & Schacter, 2004) showed greater activation of right prefrontal cortex during false than true recognition, again suggesting a role for late-occurring verification and monitoring of the products of retrieval during episodic recognition.

Combining these notions, we propose here that during recognition, the greater pupil dilation for Old and HFS concepts, as compared to LFS ones, might reflect greater locus coeruleus and right prefrontal activity determined by the higher degree of retrieval monitoring involved in recognizing these items. This idea is supported by lower *d*' scores that we reported for the HFS than LFS concepts, and by the low accuracy with which our participants recognized Old concepts (53%). Indeed, according to the signal detection model of recognition and the pupil strength-of-memory accounts, memory strength is closer to the old-new response criterion for HFS concepts, due to the fact that **the** latter might have determined an increase in the strength of

evidence due to the greater sharing of semantic features and the meaning overlap with old concepts. This would have increased the uncertainty of the recognition process and, thus, increased monitoring requirements in order to check the relevance and validity of the retrieved information (Henson et al., 2000), which in turn, determined a greater pupil size for HFS concepts compared to the LFS ones. Interestingly, it has been posited that phasic activity of the LC-NA system driven by the threshold crossing in the task-relevant decision layer would "collapse" the different layers of task-relevant cortical networks in order to promptly couple detected targets to motor responses (Aston-Jones & Cohen, 2005). From these perspectives, we propose that our results could be explained by the activation of the LC-NA system that would have promoted the transition from monitoring the retrieved information to provide an "old" response.

One might argue that the greater the involvement of retrieval monitoring was, the greater the retrieval or cognitive effort was, and thus our results of a greater pupillary response for false memory for HFS concepts and true memory for Old ones would reflect increased cognitive load or voluntary effort required during the retrieval of episodic content (Goldinger & Papesh, 2012; Võ et al., 2008). However, if we operationalize the cognitive effort in terms of reaction times, our results rule out this hypothesis, since the pattern of results for RTs data is not consistent with it. Indeed, correct recognition of Old concepts was related to very high pupillary responses but very low RTs; moreover, the RTs for false alarms and misses to HFS were not different, while pupillary data clearly dissociated these two conditions.

Our results might also be explained by relying on an alternative explanation of the PON effect, which questioned the voluntary component of the cognitive load account. According to this study (Mill, O'Connor, & Dobbins, 2016), the pupil old/new effect would reflect involuntary orienting triggered by unexpected information, depending on the degree to which this information is unexpected. This idea is in agreement with fMRI studies assuming a role for right-lateralized bottom-up attention processes in the processing of unexpected content (Cabeza, Ciaramelli, Olson, & Moscovitch, 2008; O'Connor, Han, & Dobbins, 2010) and a role of the pupillary dilation in response to the surprise value of diagnostic information in decision-making (Preuschoff, 't Hart, & Einhäuser, 2011). This novel conceptualization of the pupil response to recognition might be consistent with our results because our participants were not aware that the task was a recognition task. Thus, it is plausible to think that the retrieved information during recognition was unexpected. As pointed by Mill et al. (2016), only the old items are able to trigger a strong orienting response since they reflect an

acontextual sense of recent encounter (familiarity) and remembrances of contextual information (recollection), in line with the dual-process models (Yonelinas, 2002). However, the current data cannot differentiate between familiarity and recollection processes underlying pupil dilation to the unexpected old concepts. Rather, we feel confident that our results mirror strength of memory trace and not (voluntary) cognitive demands due to the retrieval of contextual information of previous presentation of stimuli.

To sum up, the current data demonstrate that pupil response to false memories can be modulated by featural similarity between concepts when the effect of lexical co-occurrence and word association is controlled for. In particular, we found new words with high similarity were mistakenly identified as old, and had more pupil dilation than those with low similarity, suggesting the existence of a fine-grained relation between pupil response and the memory traces.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 702655.

References

Ambrosini, E., Costantini, M., & Sinigaglia, C. (2011). Grasping with the eyes. *Journal of Neurophysiology, 106*(3), 1437-1442. https://doi.org/10.1152/jn.00118.2011

Ambrosini, E., Vastano, R., Montefinese, M., & Ciavarro, M. (2013). Functional specificity of the locus coeruleus-norepinephrine system in the attentional networks. *Frontiers in Behavioral Neuroscience, 7*, 201. https://doi.org/10.3758/BF03203267

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review, 116*(3), 463–498. https://doi.org/10.1037/a0016261

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: Adaptive Gain and Optimal Performance. *Annual Review of Neuroscience, 28*(1), 403–450. https://doi.org/10.1146/annurev.neuro.28.061604.135709

Baroni, M., Bernardini, S., Comastri, F., Piccioni, L., Volpi, A., Aston, G., & Mazzoleni, M. (2004). Introducing the La Repubblica Corpus: A Large, Annotated, TEI(XML)-Compliant Corpus of Newspaper Italian. In *Proceedings of the 4th Edition of the Language, Resources and Evaluation Conference (LREC 2004)* (p. 5–163.).

Bayer, M., Sommer, W., & Schacht, A. (2011). Emotional words impact the mind but not the body: evidence from pupillary responses. *Psychophysiology, 48*(11), 1554–62. https://doi.org/10.1111/j.1469-8986.2011.01219.x

Bird, H., Franklin, S., & Howard, D. (2001). Age of acquisition and imageability ratings for a large set of words, including verbs and function words. *Behavior Research Methods, 33*(1), 73-79. https://doi.org/10.3758/BF03195349

Bouret, S., & Sara, S. J. (2005). Network reset: A simplified overarching theory of locus coeruleus noradrenaline function. *Trends in Neurosciences, 28*(11), 574-582. https://doi.org/10.1016/j.tins.2005.09.002

Bradley, M. M., & Lang, P. J. (2015). Memory, emotion, and pupil diameter: Repetition of natural scenes.

Psychophysiology, 52(9), 1186-1193. https://doi.org/10.1111/psyp.12442

Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology, 45*(4), 602-607. https://doi.org/10.1111/j.1469-8986.2008.00654.x

Brainerd, C. J., Yang, Y., Reyna, V. F., Howe, M. L., & Mills, B. a. (2008). Semantic processing in 'associative' false memory. *Psychonomic Bulletin & Review, 15*(6), 1035–53. https://doi.org/10.3758/PBR.15.6.1035

Brocher, A., & Graf, T. (2016). Pupil old/new effects reflect stimulus encoding and decoding in short-term memory. *Psychophysiology, 53*(12), 1823-1835. https://doi.org/10.1111/psyp.12770

Brocher, A., & Graf, T. (2017). Decision-related factors in pupil old/new effects: Attention, response execution, and false memory. *Neuropsychologia, 102*, 124-134. https://doi.org/10.1016/j.neuropsychologia.2017.06.011

Buckner, R. L., Koutstaal, W., Schacter, D. L., Wagner, A. D., & Rosen, B. R. (1998). Functional–Anatomic Study of Episodic Retrieval Using fMRI. *NeuroImage*, *7*(3), 151–162. https://doi.org/10.1006/nimg.1998.0327

Cabeza, R., Ciaramelli, E., Olson, I. R., & Moscovitch, M. (2008). The parietal cortex and episodic memory: an attentional account. *Nature Reviews Neuroscience, 9*(8), 613. https://doi.org/10.1038/nrn2459

Cabeza, R., Rao, S. M., Wagner, A. D., Mayer, A. R., & Schacter, D. L. (2001). Can medial temporal lobe regions distinguish true from false? An event-related functional MRI study of veridical and illusory recognition memory. *Proceedings of the National Academy of Sciences, 98*(8), 4805–4810. https://doi.org/10.1073/pnas.081082698

Cann, D. R., McRae, K., & Katz, A. N. (2011). False recall in the Deese-Roediger-McDermott paradigm: The roles of gist and associative strength. *The Quarterly Journal of Experimental Psychology, 64*(8), 1515–42. https://doi.org/10.1080/17470218.2011.560272

Dayan, P., & Yu, A. J. (2006). Phasic norepinephrine: a neural interrupt signal for unexpected events.

Network: Computation in Neural Systems, 17(4), 335-350. https://doi.org/10.1080/09548980601004024

De Deyne, S., & Storms, G. (2008a). Word associations: Network and semantic properties. *Behavior Research Methods, 40*(1), 213–231. https://doi.org/10.3758/BRM.40.1.213

De Deyne, S., & Storms, G. (2008b). Word associations: Norms for 1,424 Dutch words in a continuous task. *Behavior Research Methods, 40*(1), 198–205. https://doi.org/10.3758/BRM.40.1.198

de Gee, J. W., Colizoli, O., Kloosterman, N. A., Knapen, T., Nieuwenhuis, S., & Donner, T. H. (2017). Dynamic modulation of decision biases by brainstem arousal systems. *eLife, 6.* https://doi.org/10.7554/eLife.23232

De Moor, W., Ghyselinck, M., & Brysbaert, M. (2000). A validation study of the age-of-acquisition norms collected by Ghyselinck, De Moor, & Brysbaert. *Psychologica Belgica, 40*(2), 99-114.

Dobbins, I. G., Foley, H., Schacter, D. L., & Wagner, A. D. (2002). Executive Control during Episodic Retrieval. *Neuron, 35*(5), 989–996. https://doi.org/10.1016/S0896-6273(02)00858-9

Dobbins, I. G., Rice, H. J., Wagner, A. D., & Schacter, D. L. (2003). Memory orientation and success: separable neurocognitive components underlying episodic recognition. *Neuropsychologia, 41*(3), 318–333. https://doi.org/10.1016/S0028-3932(02)00164-1

Evans, S., Dowell, N. G., Tabet, N., King, S. L., Hutton, S. B., & Rusted, J. M. (2017). Disrupted neural activity patterns to novelty and effort in young adult APOE-e4 carriers performing a subsequent memory task. *Brain and Behavior, 7*(2). https://doi.org/10.1002/brb3.612

Fairfield, B., Ambrosini, E., Mammarella, N., & Montefinese, M. (2017). Affective Norms for Italian Words in Older Adults: Age Differences in Ratings of Valence, Arousal and Dominance. *PloS ONE, 12*(1). https://doi.org/10.1371/journal.pone.0169472

Finnigan, S., Humphreys, M. S., Dennis, S., & Geffen, G. (2002). ERP 'old/new' effects: memory strength and decisional factor(s). *Neuropsychologia, 40*(13), 2288–304.

Fletcher, P. (1998). The functional roles of prefrontal cortex in episodic memory. II. Retrieval. *Brain, 121*(7),

1249–1256. https://doi.org/10.1093/brain/121.7.1249

Galati, G., Committeri, G., Spitoni, G., Aprile, T., Di Russo, F., Pitzalis, S., & Pizzamiglio, L. (2008). A selective representation of the meaning of actions in the auditory mirror system. *NeuroImage, 40*(3), 1274-1286. https://doi.org/10.1016/j.neuroimage.2007.12.044

Gallo, D. a., & Roediger, I. H. L. (2002). Variability among word lists in eliciting memory illusions: evidence for associative activation and monitoring. *Journal of Memory and Language, 47*(3), 469–497. https://doi.org/10.1016/S0749-596X(02)00013-X

Gardner, R. M., Beltramo, J. S., & Krkinsky, R. (1975). Pupillary changes during encoding, storage, and retrieval of information. *Perceptual and Motor Skills, 41*(3), 951–955. https://doi.org/10.2466/pms.1975.41.3.951

Gardner, R. M., Philp, P., & Radacy, S. (1978). Pupillary changes during recall in children. *Journal of Experimental Child Psychology, 25*(1), 168–172. https://doi.org/10.1016/0022-0965(78)90046-2

Ghyselinck, M., De Moor, W., & Brysbaert, M. (2000). Age-of-acquisition ratings for 2816 Dutch four- and five-letter nouns. *Psychologica Belgica, 40*(2), 77-98.

Goldinger, S. D., & Papesh, M. H. (2012). Pupil Dilation Reflects the Creation and Retrieval of Memories. *Current Directions in Psychological Science, 21*(2), 90–95. https://doi.org/10.1177/0963721412436811

Heaver, B., & Hutton, S. B. (2010). Keeping an eye on the truth: Pupil size, recognition memory and malingering. *International Journal of Psychophysiology, 77*(3), 306–306. https://doi.org/10.1016/j.ijpsycho.2010.06.206

Heaver, B., & Hutton, S. B. (2011). Keeping an eye on the truth? Pupil size changes associated with recognition memory. *Memory, 19*(4), 398–405. https://doi.org/10.1080/09658211.2011.575788

Hellmer, K., Söderlund, H., & Gredebäck, G. (2016). The eye of the retriever: developing episodic memory mechanisms in preverbal infants assessed through pupil dilation. *Developmental Science,* 1-11. https://doi.org/10.1111/desc.12520

Henson, R. N. A., Rugg, M. D., Shallice, T., & Dolan, R. J. (2000). Confidence in Recognition Memory for Words: Dissociating Right Prefrontal Roles in Episodic Retrieval. *Journal of Cognitive Neuroscience, 2*(6), 913-923. https://doi.org/10.1162/08989290051137468

Henson, R. N. A., Rugg, M. D., Shallice, T., Josephs, O., & Dolan, R. J. (1999). Recollection and familiarity in recognition memory. *Journal of Neuroscience, 19*(10), 3962-3972.

Henson, R. N. A., Shallice, T., & Dolan, R. J. (1999). Right prefrontal cortex and episodic memory retrieval: A functional MRI test of the monitoring hypothesis. *Brain, 122*(7), 1367-1381. https://doi.org/10.1093/brain/122.7.1367

Hess, E. H., Polt, J. M., Science, S., Series, N., Mar, N., & Collins, R. L. (1964). Pupil Size in Relation to Mental Activity during Simple Problem-Solving. *Science, 143*(3611), 1190-1192.

Jainta, S., & Baccino, T. (2010). Analyzing the pupil response due to increased cognitive demand: an independent component analysis study. *International Journal of Psychophysiology*, *77*(1), 1–7. https://doi.org/10.1016/j.ijpsycho.2010.03.008

Joshi, S., Li, Y., Kalwani, R. M., & Gold, J. I. (2016). Relationships between Pupil Diameter and Neuronal Activity in the Locus Coeruleus, Colliculi, and Cingulate Cortex. *Neuron, 89*(1), 221-234. https://doi.org/10.1016/j.neuron.2015.11.028

Kafkas, A., & Montaldi, D. (2011). Recognition memory strength is predicted by pupillary responses at encoding while fixation patterns distinguish recollection from familiarity. *The Quarterly Journal of Experimental Psychology*, *64*(10), 1971–89. https://doi.org/10.1080/17470218.2011.588335

Kafkas, A., & Montaldi, D. (2015). The pupillary response discriminates between subjective and objective familiarity and novelty. *Psychophysiology, 52*(10), 1305-1316. https://doi.org/10.1111/psyp.12471

Kahneman, D., & Beatty, J. (1966). Pupil Diameter and Load on Memory. *Science, 154*(3756), 1583-1585. https://doi.org/10.1126/science.154.3756.1583

Kremer, G., & Baroni, M. (2011). A set of semantic norms for German and Italian. *Behavior Research Methods, 43*(1), 97-109. https://doi.org/10.3758/s13428-010-0028-x

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods, 164*(1), 177-190. https://doi.org/10.1016/j.jneumeth.2007.03.024

McRae, K., Cree, G., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods, 37*(4), 547-559.https://doi.org/10.3758/BRM.40.1.183

McRae, K., Khalkhali, S., & Hare, M. (2012). Semantic and associative relations in adolescents and young adults: Examining a tenuous dichotomy. *The Adolescent Brain: Learning, Reasoning, and Decision Making. 18,* 39-66. https://doi.org/10.1037/13493-002

Mill, R. D., O'Connor, A. R., & Dobbins, I. G. (2016). Pupil dilation during recognition memory: Isolating unexpected recognition from judgment uncertainty. *Cognition. 154,* 81-94. https://doi.org/10.1016/j.cognition.2016.05.018

Montefinese, M., Ambrosini, E., Fairfield, B., & Mammarella, N. (2013a). Erratum to: Semantic memory: A feature-based analysis and new norms for Italian. *Behavior Research Methods, 45*(2), 462–462. https://doi.org/10.3758/s13428-012-0291-0

Montefinese, M., Ambrosini, E., Fairfield, B., & Mammarella, N. (2013b). Semantic memory: A feature-based analysis and new norms for Italian. *Behavior Research Methods, 45*(2), 440–461. https://doi.org/10.3758/s13428-012-0263-4

Montefinese, M., Ambrosini, E., Fairfield, B., & Mammarella, N. (2013c). The 'subjective' pupil old/new effect: Is the truth plain to see? *International Journal of Psychophysiology, 89*(1), 48–56. https://doi.org/10.1016/j.ijpsycho.2013.05.001

Montefinese, M., Ambrosini, E., Fairfield, B., & Mammarella, N. (2014a). Semantic significance: a new measure of feature salience. *Memory & Cognition, 42*(3), 355–369. https://doi.org/10.3758/s13421-013-0365-y

Montefinese, M., Ambrosini, E., Fairfield, B., & Mammarella, N. (2014b). The adaptation of the Affective Norms for English Words (ANEW) for Italian. *Behavior Research Methods, 46*(3), 887–903.

https://doi.org/10.3758/s13428-013-0405-3

Montefinese, M., Ciavarro, M., & Ambrosini, E. (2015). What is the right place for atypical exemplars? Commentary: The right hemisphere contribution to semantic categorization: a TMS study. *Frontiers in Psychology, 6*(9), 1105–1109. https://doi.org/10.3389/fpsyg.2015.01349

Montefinese, M., Costantini, M., Committeri, G., & Ambrosini, E. (2015). Looking at emotions: a psychophysiological investigation on affective processing. *Neuropsychological Trends, 18,* 126.

Montefinese, M., Turco, C., Piccione, F., & Semenza, C. (2017). Causal role of the posterior parietal cortex for two-digit mental subtraction and addition: A repetitive TMS study. *NeuroImage, 155,* 72–81. https://doi.org/10.1016/j.neuroimage.2017.04.058

Montefinese, M., & Vinson, D. (2015). Can the humped animal's knee conceal its name? Commentary on: 'The roles of shared vs. distinctive conceptual features in lexical access'. *Frontiers in Psychology, 6*. https://doi.org/10.3389/fpsyg.2015.00418

Montefinese, M., Zannino, G. D., & Ambrosini, E. (2015). Semantic similarity between old and new items produces false alarms in recognition memory. *Psychological Research, 79*(5), 785–794. https://doi.org/10.1007/s00426-014-0615-z

Moors, A., De Houwer, J., Hermans, D., Wanmaker, S., van Schie, K., Van Harmelen, A.-L., … Brysbaert, M. (2013). Norms of valence, arousal, dominance, and age of acquisition for 4,300 Dutch words. *Behavior Research Methods, 45*(1), 169–177. https://doi.org/10.3758/s13428-012-0243-8

Morey, R. D. (2008). Confidence Intervals from Normalized Data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology, 4*(2), 61-64.https://doi.org/10.20982/tqmp.04.2.p061

Morrison, C. M., Chappell, T. D., & Ellis, A. W. (1997). Age of Acquisition Norms for a Large Set of Object Names and Their Relation to Adult Estimates and Other Variables. *The Quarterly Journal of Experimental Psychology: 50*(3), 528-559.https://doi.org/10.1080/027249897392017

Murphy, P. R., O'Connell, R. G., O'Sullivan, M., Robertson, I. H., & Balsters, J. H. (2014). Pupil diameter covaries with BOLD activity in human locus coeruleus. *Human Brain Mapping, 35*(8), 4140–4154.

https://doi.org/10.1002/hbm.22466

Murphy, P. R., Robertson, I. H., Balsters, J. H., & O'connell, R. G. (2011). Pupillometry and P3 index the locus coeruleus-noradrenergic arousal function in humans. *Psychophysiology, 48*(11), 1532-1543.https://doi.org/10.1111/j.1469-8986.2011.01226.x

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

Nieuwenhuis, S., De Geus, E. J., & Aston-Jones, G. (2011). The anatomical and functional relationship between the P3 and autonomic components of the orienting response. *Psychophysiology, 48*(2), 162-175. https://doi.org/10.1111/j.1469-8986.2010.01057.x

Nowack, K., Milfont, T. L., & van der Meer, E. (2013). Future versus present: Time perspective and pupillary response in a relatedness judgment task investigating temporal event knowledge. *International Journal of Psychophysiology, 87*(2), 173-182. https://doi.org/10.1016/j.ijpsycho.2012.12.006

Nuthmann, A., & van der Meer, E. (2005). Time's arrow and pupillary response. *Psychophysiology, 42*(3), 306–17. https://doi.org/10.1111/j.1469-8986.2005.00291.x

O'Connor, A. R., Han, S., & Dobbins, I. G. (2010). The Inferior Parietal Lobule and Recognition Memory: Expectancy Violation or Successful Retrieval? *Journal of Neuroscience, 30*(8), 2924-2934. https://doi.org/10.1523/JNEUROSCI.4225-09.2010

Oke, A., Keller, R., Mefford, I., & Adams, R. N. (1978). Lateralization of norepinephrine in human thalamus. *Science, 200*(4348), 1411-1413. https://doi.org/10.1126/science.663623

Otero, S. C., Weekes, B. S., & Hutton, S. B. (2011). Pupil size changes during recognition memory. *Psychophysiology, 48*(10), 1346–53. https://doi.org/10.1111/j.1469-8986.2011.01217.x

Papesh, M. H., Goldinger, S. D., & Hout, M. C. (2012). Memory strength and specificity revealed by pupillometry. *International Journal of Psychophysiology, 83*(1), 56–64.

https://doi.org/10.1016/j.ijpsycho.2011.10.002

Ponzetto, S. P., & Strube, M. (2007). Knowledge derived from wikipedia for computing semantic relatedness. *Journal of Artificial Intelligence Research, 30,* 181-212. https://doi.org/10.1.1.75.5311

Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience, 13*(1), 25-42. https://doi.org/10.1146/annurev.ne.13.030190.000325

Preuschoff, K., 't Hart, B. M., & Einhäuser, W. (2011). Pupil dilation signals surprise: Evidence for noradrenaline's role in decision making. *Frontiers in Neuroscience, 5,* 115. https://doi.org/10.3389/fnins.2011.00115

Robinson, R. (1979). Differential behavioral and biochemical effects of right and left hemispheric cerebral infarction in the rat. *Science, 205*(4407), 707–710. https://doi.org/10.1126/science.462179

Robinson, R., & Coyle, J. (1980). The differential effect of right versus left hemispheric cerebral infarction on catecholamines and behavior in the rat. *Brain Research, 188*(1), 63–78. https://doi.org/10.1016/0006-8993(80)90557-0

Roediger, H., & McDermott, K. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(4), 803. http://psycnet.apa.org/journals/xlm/21/4/803/

Rugg, M. D., & Curran, T. (2007). Event-related potentials and recognition memory. *Trends in Cognitive Sciences, 11*(6), 251–7. https://doi.org/10.1016/j.tics.2007.04.004

Rugg, M. D., Fletcher, P. C., Chua, P. M.-L., & Dolan, R. J. (1999). The Role of the Prefrontal Cortex in Recognition Memory and Memory for Source: An fMRI Study. *NeuroImage, 10*(5), 520-529.https://doi.org/10.1006/nimg.1999.0488

Rugg, M. D., Fletcher, P. C., Frith, C. D., Frackowiak, R. S. J., & Dolan, R. J. (1996). Differential activation of the prefrontal cortex in successful and unsuccessful memory retrieval. *Brain, 119*(6), 2073-2083. https://doi.org/10.1093/brain/119.6.2073

Sara, S. J. (2009). The locus coeruleus and noradrenergic modulation of cognition. *Nature Reviews Neuroscience, 10*(3), 211. https://doi.org/10.1038/nrn2573

Schacter, D. L., Buckner, R. L., Koutstaal, W., Dale, A. M., & Rosen, B. R. (1997). Late onset of anterior prefrontal activity during true and false recognition: an event-related fMRI study. *NeuroImage, 6*(4), 259-269. https://doi.org/10.1006/nimg.1997.0305

Schacter, D. L., Curran, T., Galluccio, L., Milberg, W. P., & Bates, J. F. (1996). False recognition and the right frontal lobe: A case study. *Neuropsychologia, 34*(8), 793–808. https://doi.org/10.1016/0028-3932(95)00165-4

Schacter, D. L., & Slotnick, S. D. (2004). The Cognitive Neuroscience of Memory Distortion. *Neuron, 44*(1), 149–160. https://doi.org/10.1016/j.neuron.2004.08.017

Siegle, G. J., Steinhauer, S. R., Stenger, V. A., Konecky, R., & Carter, C. S. (2003). Use of concurrent pupil dilation assessment to inform interpretation and analysis of fMRI data. *NeuroImage, 20*(1), 114–124. https://doi.org/10.1016/S1053-8119(03)00298-2

Slotnick, S. D., & Schacter, D. L. (2004). A sensory signature that distinguishes true from false memories. *Nature Neuroscience, 7*(6), 664–72. https://doi.org/10.1038/nn1252

Spence, D. P., & Owens, K. C. (1990). Lexical co-occurrence and association strength. *Journal of Psycholinguistic Research, 19*(5), 317–330. https://doi.org/10.1007/BF01074363

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, 31*(1), 137-149. https://doi.org/10.3758/BF03207704

Sterpenich, V., D'Argembeau, A., Desseilles, M., Balteau, E., Albouy, G., Vandewalle, G., … Maquet, P. (2006). The Locus Ceruleus Is Involved in the Successful Retrieval of Emotional Memories in Humans. *Journal of Neuroscience, 26*(28), 7416-7423. https://doi.org/10.1523/JNEUROSCI.1001-06.2006

Unsworth, N., & Robison, M. K. (2016). Pupillary correlates of lapses of sustained attention. *Cognitive, Affective, & Behavioral Neuroscience, 16*(4), 601-615. https://doi.org/10.3758/s13415-016-0417-4

Varazzani, C., San-Galli, A., Gilardeau, S., & Bouret, S. (2015). Noradrenaline and dopamine neurons in the reward/effort trade-off: a direct electrophysiological comparison in behaving monkeys. *The Journal of Neuroscience, 35*(20), 7866-7877. https://doi.org/10.1523/JNEUROSCI.0454-15.2015

Vinson, D. P., & Vigliocco, G. (2008). Semantic feature production norms for a large set of objects and events. *Behavior Research Methods, 40*(1), 183–190. https://doi.org/10.3758/BRM.40.1.183

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. https://doi.org/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*(1), 152–76. https://doi.org/10.1037/0033-295X.114.1.152

Wright, D. B., Horry, R., & Skagerberg, E. M. (2009). Functions for traditional and multilevel approaches to signal detection theory. *Behavior Research Methods, 41*(2), 257-267. https://doi.org/10.3758/BRM.41.2.257

Yonelinas, A. P. (2002). The Nature of Recollection and Familiarity: A Review of 30 Years of Research. *Journal of Memory and Language, 46*(3), 441–517. https://doi.org/10.1006/jmla.2002.2864