TITLE:

Is pupillary response a reliable index of word recognition? Evidence from a delayed lexical decision task

AUTHORS:

Juan Haro[1], Marc Guasch[1], Blanca Vallès[1], Pilar Ferré[1]

[1]Research Center for Behavior Assessment (CRAMC) and Department of Psychology. Universitat Rovira i Virgili. Tarragona. Spain.

RUNNING HEAD:

Pupillary response and word recognition

CORRESPONDING AUTHOR:

Juan Haro

Research Center for Behavior Assessment (CRAMC) and Department of Psychology.

Universitat Rovira i Virgili.

Crta. de Valls s/n, Campus Sescelades, 43007, Tarragona. Spain.

E-mail: juan.haro@urv.cat

Telephone Number: +34-977-558567

Abstract

Previous word recognition studies have shown that pupillary response is sensitive to word frequency. However, such pupillary effect may be due to the process of planning or executing a response, instead of being an index of word processing. With the aim of exploring this possibility, we recorded the pupillary response in two experiments involving a lexical decision task (LDT). In the first experiment, participants completed a standard LDT, whereas in the second, participants performed a delayed LDT. The delay in the response allowed us to compare pupil dilation with and without the response execution component. Results showed that pupillary response was modulated by word frequency both in the standard and the delayed LDT. This finding supports the reliability of using pupillometry for word recognition research. However, our results also suggest that tasks that do not require a response during pupil recording lead to clearer and stronger effects.

Keywords

Visual word recognition; Pupillometry; Lexical decision task; Delayed lexical decision task; Word frequency effect

The study of the pupillary response has interested psychologists for many years (see Laeng, Sirois, & Gredeback, 2012, for an overview). This measure has several interesting properties for the study of cognitive phenomena. Among them, the pupillary response can provide information about the time-course of the cognitive phenomenon under study. It is also sensitive to processes that are only partially activated but never pass the threshold for eliciting overt behavior or for reaching consciousness (Laeng et al., 2012). Another remarkable property is that it is not affected by participant's strategies, as it is quite difficult to be controlled voluntarily. Due to these advantages, the pupillary response has been widely used in cognitive psychology to study a large variety of different cognitive processes, including attention allocation (Karatekin, Couperus, & Marcus, 2004), face perception (Goldinger, He, & Papesh, 2009), arithmetics (Klingner, Tversky, & Hanrahan, 2011), or working memory (Heitz, Schrock, Payne, & Engle, 2008), among others.

Following this line of inquiry, the aim of the present work was to investigate whether the pupillary response can be reliably applied to word recognition research, a field in which it has been scarcely used. Word recognition is a complex process. In order to understand it, psycholinguistic research has focused on identifying the variables that influence the processing of words. Among these variables, word frequency is considered the best predictor of word recognition (e.g., Keuleers, Diependaele, & Brysbaert, 2010). The frequency of a word is usually measured as the number of occurrences of a word in a given language, and this measure has been shown to predict performance in several experimental tasks, such as lexical decision (Rubenstein, Garfield, & Millikan, 1970), naming (Forster & Chambers, 1973), or perceptual identification (Manelis, 1977). A consistent finding from these tasks is that high-frequency words (e.g., *people*) are processed faster and more accurately than low-frequency words (e.g., *nuance*). An

account of this processing advantage proposes that high-frequency words have a higher resting activation level than low-frequency words, thus less activation would be needed in order to recognize a high-frequency word than a low-frequency word (e.g., McClelland & Rumelhart, 1981).

To our knowledge, three word recognition studies have examined to date whether pupillary response is modulated by word frequency (Kuchinke, Võ, Hofmann, & Jacobs, 2007; Schmidtke, 2014). In these studies, the recorded measures of the pupillary response were: a) the point of maximum pupil dilation in response to the presentation of a stimulus (i.e., peak dilation), and b) the point in time where this peak dilation is achieved (i.e., peak latency). It is assumed that the larger the dilation or the longer the latency, the higher the cognitive effort or the processing difficulty. As for the results, Kuchinke et al. (2007) found that participants exhibited larger pupil peak dilations to low-frequency words than to high-frequency words in a lexical decision task (LDT). In the same line, Schmidtke (2014) observed that low-frequency words elicited later pupil peak dilations than high-frequency words during a spoken recognition task. Finally, Papesh and Goldinger (2012) showed that peak dilations for low-frequency words were larger than peak dilations for high-frequency words before, during, and after naming responses. Importantly, the results of these studies showed a convergence between behavioral and physiological data, as slower response times were associated with larger pupil peak sizes (Kuchinke et al., 2007; Papesh and Goldinger, 2012) or delayed pupil peak latencies (Schmidtke, 2014).

However, there is the possibility that these findings may be biased by an experimental confounding factor, since in these studies participants were not only asked to recognize a word, but also to execute a response. In the work of Kuchinke et al. (2007), participants performed a LDT. In each trial, participants were presented with a

string of letters that could be a word or a nonword (e.g., *tapem*). Then, participants were required to press one mouse button if the string was a word, and to press another button if the string was a nonword. In the other study (Schmidtke, 2014), participants had to match spoken words to pictures (i.e., visual-world paradigm). Each trial started with the presentation of four images while participants heard "Click on the [target word]", and the trial ended when participants clicked with the mouse on one of the images. Finally, Papesh and Goldinger (2012) examined if word frequency effects are restricted to early processes of perception and lexical access, or if these effects continue into postaccess processes. Indeed, such study was aimed to assess word frequency effects during speech planning and speech execution, instead of how word frequency influences the first stages of word processing. This represents a significant difference with respect to Kuchinke et al. (2007) and Schmidkte (2014). Thus, as the main goal of the present study was to examine word frequency effects during early word processing stages, in the following we will focus only in Kuchinke et al. and Schmidkte studies.

Taking into account the characteristics of such experimental tasks, it could be argued that the modulation of the pupillary response observed in these studies may be reflecting not only differences in the processing of low- and high-frequency words, but also the preparation and execution of a response. This would be in agreement with evidences that the pupillary response is affected by planning and executing a motor response (e.g., Hupé, Lamirel, & Lorenceau, 2009; Moresi et al., 2008). Indeed, Moresi et al. (2008) found that the difficulty of response preparation during a finger-cuing task (Miller, 1982) was correlated with pupil size, as more difficult cues elicited larger pupil dilations during response preparation and execution. In another study, Hupé et al. (2009) recorded pupil changes while subjects continuously reported changes in the perception of visual ambiguous stimuli, observing that 70% of pupil dilation could be accounted

for by the motor response. A possible explanation for these findings is that the pupillary response is closely linked with the activity of the locus coeruleus (see Laeng et al., 2012, for a review), a subcortical structure involved in a large variety of processes, including task-related decision processes and the execution of behavioral responses. Thus, any change in the activity of the locus coeruleus due to planning or executing a response may have an effect in the pupillary response (Hupé et al., 2009).

Given the above, the aim of the present study was to test whether the modulation of the pupillary response by word frequency found by Kuchinke et al. (2007) and Schmidtke (2014) could be due to a confounding effect of response execution, or if it rather reflects a genuine effect on word processing. We believe that addressing this issue is important for two main reasons: 1) to determine if the study of the pupillary response can be reliably applied to word recognition research, and 2) to help identifying the proper methodological requirements for the study of pupillary response in word recognition research. To this aim, we conducted two experiments in which we manipulated the requirements of the experimental task. In Experiment 1a, we examined the effects of word frequency on the pupillary response during a standard LDT. The objective of this first experiment was to replicate the word frequency effect reported in previous studies by using a task that requires planning and executing a response (Kuchinke et al., 2007; Schmidtke, 2014). In Experiment 1b, we explored word frequency effects on the pupillary response in a delayed LDT. In contrast to Experiment 1a, the delayed LDT allowed us to observe changes in the pupil size, avoiding any influence of participant's response, thus providing a purer measure of the modulation of the pupillary response during word processing.

Method for Experiments 1a and 1b

The study consisted of two experiments in which word frequency (low-,

medium-, and high-frequency) was manipulated within participants. In addition, task (standard LDT and delayed LDT) was manipulated between participants. In Experiment 1a, participants completed a standard LDT, whereas in Experiment 1b, participants completed a delayed LDT. We have combined both experiments and report them as one.

Participants

Sixty Spanish speakers were recruited for the study. Half of them (21 women and 9 men; mean age = 20.63, *SD* = 3.18) participated in Experiment 1a, and the other half (28 women and 2 men; mean age = 19.70, *SD* = 2.52) participated in Experiment 1b. All of them were students from the Universitat Rovira i Virgili (Tarragona, Spain) who received academic credits for their contribution. They had either normal or corrected-to-normal vision and reported no history of major visual impairments.

Materials

The stimulus set of both experiments included 75 Spanish words. Stimuli were divided into three conditions according to word frequency: 25 low-frequency words (less than 10 occurrences per million; e.g., *bautizo*, "baptism"), 25 medium-frequency words (between 10 and 30 occurrences per million; e.g., *industria*, "industry"), and 25 high-frequency words (more than 30 occurrences per million; e.g., *anillo*, "ring"). In addition to word frequency occurrences per million, we ensured that conditions differed also in log frequency, lemma frequency and log lemma frequency (all *p*s < .001).

Experimental conditions were matched for word length, number of syllables, number of neighbors, number of higher frequency neighbors, mean Levenshtein distance of the 20 closest words (old20), bigram frequency and trigram frequency (all *p*s > .10). Furthermore, concreteness, imageability, context availability, arousal, and emotional valence ratings were equivalent across conditions (all *p*s >.10). Familiarity, age-of-acquisition, and contextual diversity ratings could not be matched across

conditions (all *p*s < .01) because of the high correlation between them and word frequency. These variables were obtained from different sources. Word frequency occurrences per million, log frequency, lemma frequency, log lemma frequency, old20, number of neighbors, number of higher frequency neighbors, bigram frequency, trigram frequency, contextual diversity, and log contextual diversity were obtained from the EsPal subtitles corpus (Duchon, Perea, Sebastián, Martí, & Carreiras, 2013). On the other hand, familiarity, concreteness, imageability, context availability, arousal, and emotional valence ratings were taken from Guasch, Ferré, and Fraga (2015), and age-of-acquisition values were obtained from the database of Alonso, Fernandez, and Díez (2014). Full details of the experimental items are shown in Table 1.

Table 1

Characteristics of the stimuli used in the experiments (standard deviations are shown in parentheses)

	FRE	LEM	CTD	FAM	AoA	LNG	SYL	CON	IMA	CTA	VAL	ARO	OLD	NEI	NHF	BFQ	TFQ
Low-frequency	3.6	5.5	1.6	5.0	8.1	7.3	3.1	4.7	4.9	5.2	5.4	5.3	2.0	2.1	0.2	4992.5	674.2
	(2.3)	(3.5)	(0.9)	(0.4)	(1.0)	(1.5)	(0.7)	(1.1)	(1.4)	(0.7)	(1.8)	(1.4)	(0.5)	(2.3)	(0.7)	(3619.1)	(1025.4)
Medium-frequency	18.8	43.2	7.1	5.7	6.5	7.3	2.8	4.7	5.3	5.4	5.3	5.0	1.9	4.0	0.3	6060.2	890.2
	(6.1)	(78.3)	(2.1)	(0.6)	(1.8)	(1.4)	(0.6)	(1.1)	(1.4)	(0.5)	(1.2)	(1.2)	(0.5)	(5.5)	(0.6)	(3659.6)	(572.5)
High-frequency	54.3	97.5	18.0	6.5	5.4	6.8	2.8	4.6	4.8	5.3	5.2	4.9	1.9	2.7	0.1	6638.1	1294.8
	(18.2)	(75.3)	(5.1)	(0.3)	(1.5)	(1.7)	(0.7)	(1.3)	(1.8)	(0.7)	(1.7)	(1.1)	(0.4)	(2.3)	(0.3)	(4375.4)	(1337.0)

Note. FRE = word frequency per million; LEM = lemma frequency per million; CTD = contextual diversity; FAM = familiarity; AoA = age-of-acquisition; LNG = word length; SYL = number of syllables; CON = concreteness; IMA = imageability; CTA = context availability; VAL = emotional valence; ARO = arousal; OLD = old20; NEI = number of substitution neighbors; NHF = number of higher frequency substitution neighbors; BFQ = mean bigram frequency; TFQ = mean trigram frequency.

Additionally, we created a set of 75 pronounceable nonwords that were legal in Spanish by using the Wuggy pseudoword generator (Keuleers & Brysbaert, 2010). They were matched to the experimental stimuli in subsyllabic structure and transition frequencies. Finally, six words and six nonwords were selected as practice stimuli and were presented before the experimental trials.

Procedure

The procedure for both experiments was identical except for the task employed. Participants were tested individually in a medium-illuminated room. They were seated with their head on a chinrest with forehead support. Chinrest was adjusted for each participant in order to stabilize their head and keep a constant distance of 60 cm between their eyes and the monitor (a 19'' computer screen set to a resolution of 1024x768 pixels).

Right eye's pupil diameter and position were continuously recorded at a sampling rate of 1000 Hz, using an EyeLink 1000 eye tracker. This eye tracker measures pupil diameter in arbitrary units (range: 400 – 16,000 units). It can measure pupil diameter with a resolution of 0.2% of diameter (e.g., a resolution of 0.01 mm for a 5 mm pupil) and has a spatial resolution of 0.01º Root Mean Square.

Stimuli were presented using the Experiment Builder software. All stimuli were drawn in black lower-case characters (font type Arial, 24 pixels) in the center of a gray background screen (RGB 150). In addition to pupillary data, behavioral measures (RTs and response accuracy) were also recorded for each stimulus during the experimental task. Both types of data were recorded with the Experiment Builder software.

In Experiment 1a, participants completed a standard LDT. Each trial started with the presentation of a fixation cross ("+") in the center of the screen for 1000 ms. After that, the fixation cross was replaced by a letter string representing a Spanish word or a

nonword. Participants were instructed to press with the right hand either the mouse button labeled as “YES” (left button) or “NO” (right button), as quickly and accurately as they could, indicating whether or not the letter string was a Spanish word. The letter string remained on the screen for 2000 ms or until a response was made, and it was followed by a new fixation cross for 1200 ms. Preceding each trial, a self-paced display was presented in which participants were allowed to blink. When they were ready to start a new trial, participants had to fixate their gaze on a circle located at the center of the screen and then press the space bar of the keyboard.

In Experiment 1b, participants were asked to perform a delayed LDT. In this case, each trial started with the presentation of a fixation cross (“+”) in the center of the screen for 1000 ms. Then, the fixation cross was replaced by a letter string displaying a Spanish word or a nonword. It remained on the screen for 500 ms, and was followed by a new fixation cross for 1500 ms. After that, a question appeared in the screen asking participants to indicate if the letter string was a Spanish word. Participants responded by pressing with the right hand the mouse button labeled as either “YES” (left button) or “NO” (right button). If participants responded before the question, a “Too quick” feedback message was displayed. The question remained on the screen for 2000 ms or until a response was made.

The stimuli were presented in a different randomized order for each participant. There were 150 experimental trials, with 12 preceding practice trials. The experimental trials were divided into two blocks. Between blocks, participants were allowed to take a short break. At the beginning of the experiment and after the break, a calibration routine was performed.

Data cleaning and selection

Two types of data were registered in both experiments: behavioral data (RTs and

errors) and pupillary response (peak dilations and peak latencies). The process of data cleaning and selection was nearly identical in both experiments. In Experiment 1a, trials with error responses or non-responses (4.29%), trials with RTs below 300 ms or over 1500 ms (1.64%), and trials with RTs greater/lower than 2 standard deviations above/below the participant's mean (4.27%), were excluded from all analyses. As a whole, 459 trials (10.2% of the total) were rejected. None of the participants was rejected from the analyses due to the number of errors committed, as accuracy was very high (between 0% and 10.67% of errors, M = 4.29%, SD = 2.44%). On the other hand, in Experiment 1b, we first removed trials with incorrect responses and non-responses (2.91%), and trials with RTs greater/lower than 2 standard deviations above/below the participant's mean (5.30%). In this experiment it was important to remove also anticipation responses, that is, trials where the participants responded before the question appeared (1.53%). As a whole, 428 trials (9.74% of the total) were rejected. None of the participants was rejected from analyses by the number of errors committed (including anticipations). The accuracy ranged from 0.67% to 10.67%, (M = 4.44%, SD = 2.35%).

All pupillary data were processed using a Python script. First, samples with saccades or eye blinks were removed. We extended the rejection area with 25 samples on both sides for saccades and 50 samples for blinks to exclude pre- and post-artifacts (Van Rijn, Dalenberg, Borst, & Sprenger, 2012). Missing samples were filled in by linear interpolation, a procedure used in similar studies (e.g., Kuchinke et al., 2007). Finally, pupillary data were smoothed with a five-point moving-average smoothing filter.

Relevant pupillary variables were computed, on a trial-by-trial basis, from the time window comprised between 200 ms before target onset and 1500 ms after target

onset. A baseline pupil diameter was defined by averaging the pupil diameter during the 200 ms preceding the target onset (while the fixation cross was displayed). Next, pupil peak dilation and pupil peak latency were calculated. Peak dilation was computed as the difference between the participant's baseline pupil diameter and the maximum pupil diameter from the target onset to 1500 ms. To allow the comparison between participants, peak dilation was converted into relative dilation expressed as a proportional difference (in percentage of change) from the baseline. Peak latency was defined as the time elapsed from the target onset to the peak dilation. Trials for which the baseline diameter was higher than the peak dilation (6.33% in Experiment 1a and 14.57% in Experiment 1b) were removed, following the procedure used by Schmidtke (2014). In addition, we removed 4 trials from Experiment 1a (0.09% of the total) with more than 50% of missing samples. In sum, after data cleaning, 4181 data points were submitted to analyses (2136 from Experiment 1a, and 2045 from Experiment 1b).

Data analysis

Analyses were performed with R using the lme4 package (Bates et al., 2014). The effect of each dependent variable (reaction times, peak latencies and peak dilations) was analyzed separately using linear mixed-effect models (e.g., Baayen, 2008; Baayen, Davidson, & Bates, 2008). In each analysis, word frequency (low-, medium-, and high-frequency), task (standard LDT and delayed LDT) and the word frequency x task interaction were included as fixed effects, and participants and words as random effects (adjusting for the intercept). We first fitted linear mixed-effects models to the data. Then, outliers 2.5 SD below and above the model residuals mean were removed from the dataset (e.g., Baayen, 2008; Tremblay & Tucker, 2011), and models were refitted to the trimmed data. Of note, less than 3% of data points were removed after applying this trimming procedure. The significance of fixed effects was determined using log-

likelihood ratio tests (R function anova). Namely, we evaluated the contribution of each fixed effect and the interaction by comparing a model that included the effect of interest to one that did not include such effect. The P-values for pairwise comparisons were estimated by the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2014), which relies on t-distributions with degrees of freedom derived by the Satterthwaite approximation. Of note, we did not perform accuracy analyses due to the low number of error responses.

Results

Reaction times

Average of response times for each condition and task are presented in Figure 1 (averaged over individual trials).

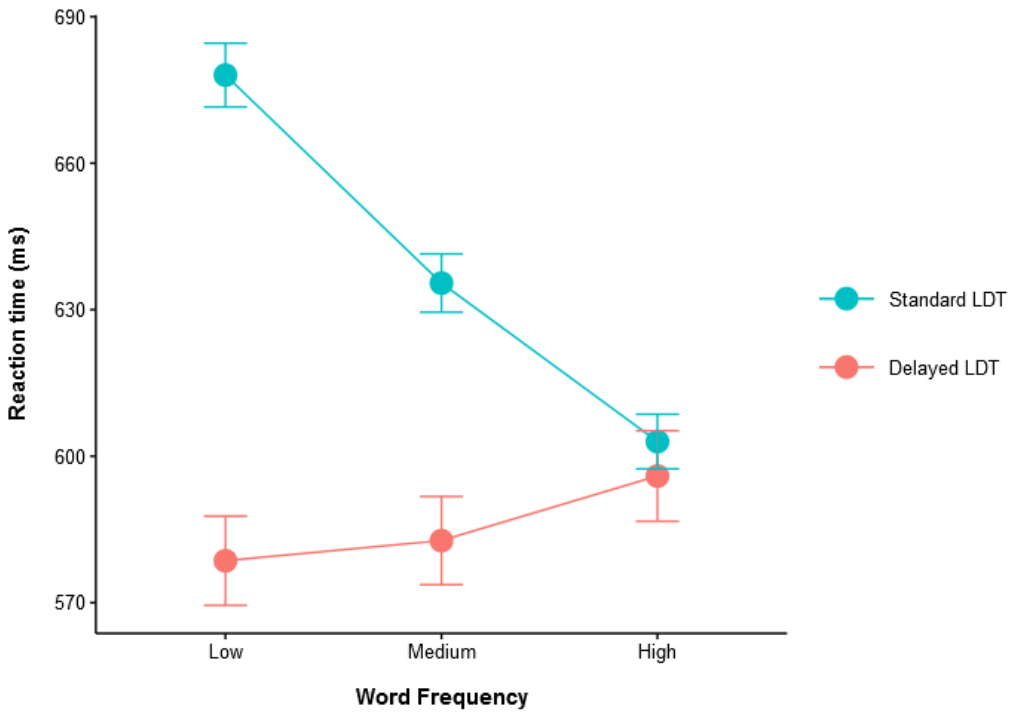

Figure 1. Average of reaction times (in ms) per word frequency for each task. Error bars represent the standard error of the mean.

There was a significant effect of word frequency on RTs, χ^2 (2) = 17.82, p

< .001. Low-frequency words were responded slower with respect to medium-frequency words, ß = -21.16, SE = 7.25, t = -2.92, p = .005, and to high-frequency words, ß = -31.46, SE = 7.25, t = -4.34, p < .001. No differences were observed between medium-frequency words and high-frequency words, ß = -10.29, SE = 7.19, t = -1.43, p = .16. In addition, average RTs did not differ between tasks, χ^2 (1) = 2.82, p = .09. The interaction of word frequency and task reached significance, χ^2 (2) = 55.5, p < .001. Word frequency had a significant effect on RTs in the standard LDT, χ^2 (2) = 36.51, p < .001, but not in the delayed LDT, χ^2 (2) = 2.17, p = .34. In the standard LDT, low-frequency words were responded slower than medium-frequency words, ß = -46.74, SE = 11.83, t = -3.95, p < .001, and high-frequency words, ß = -79.34, SE = 11.81, t = -6.72, p < .001. Additionally, medium-frequency words were responded slower than high-frequency words, ß = -32.60, SE = 11.75, t = -2.78, p = .007. In sum, behavioral results showed the expected word frequency effect in the standard LDT: participants responded faster to high-frequency words than to low-frequency words. Furthermore, the effect showed a clear linear trend.

Peak latency

The figure 2 shows the mean of peak latencies for each condition and task, averaged over individual trials. We first ensured that baseline pupil diameter was equivalent across conditions in both tasks: LDT, F (2, 58) = 0.76, p = .47, and delayed LDT, F (2, 58) = 0.41, p = .66.

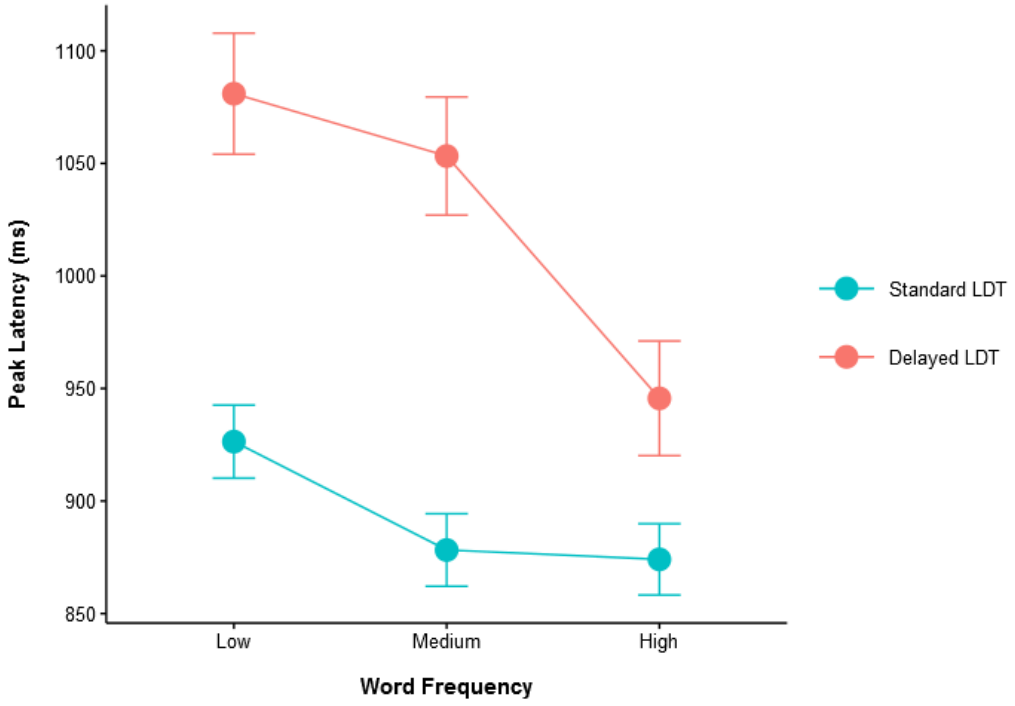

Figure 2. Average of peak latencies (in ms) per word frequency for each task. Error bars represent the standard error of the mean.

There was a main effect of word frequency on peak latency, χ^2 (2) = 16.39, p < .001. Peak latencies for low-frequency words were marginally slower in comparison to medium-frequency words, ß = -41.74, SE = 22.44, t = -1.86, p = .07, and significantly slower with respect to high-frequency words, ß = -94.36, SE = 22.43, t = -4.21, p < .001. In addition, peak latencies for medium-frequency words were slower than peak latencies for high-frequency words, ß = -52.61, SE = 22.21, t = -2.37, p = .02. There was also a significant effect of the task, as peak latencies were slower in the delayed LDT with respect to the standard LDT, ß = -129.17, SE = 48.51, t = -2.66, p = .01. The interaction between word frequency and task was significant, χ^2 (2) = 8.61, p = .01. Word frequency modulated peak latencies in both tasks: standard LDT, χ^2 (2) = 7.29, p = .03, and delayed LDT, χ^2 (2) = 14.21, p < .001, but the pattern of results differed

between tasks. In the standard LDT, differences were found between low-frequency words and medium-frequency words, ß = -50.94, SE = 21.78, t = -2.34, p = .02, and between low-frequency words and high-frequency words, ß = -52.19, SE = 21.69, t = -2.41, p = .02. Conversely, in the delayed LDT, differences were observed between high-frequency words and medium frequency words, ß = -111.32, SE = 38.14, t = -2.92, p = .004, and between high-frequency words and low-frequency words, ß = -142.36, SE = 38.45, t = -3.70, p < .001. In any case, it should be noted that pupil dilation needed more time to reach its peak when responding to low-frequency words than when responding to high-frequency words in both tasks.

Peak dilation

Peak dilations for each condition and task, averaged over individual trials, are presented in Figure 3. In addition, grand average of peak dilations during a trial are shown in Figure 4 (standard LDT) and in Figure 5 (delayed LDT).

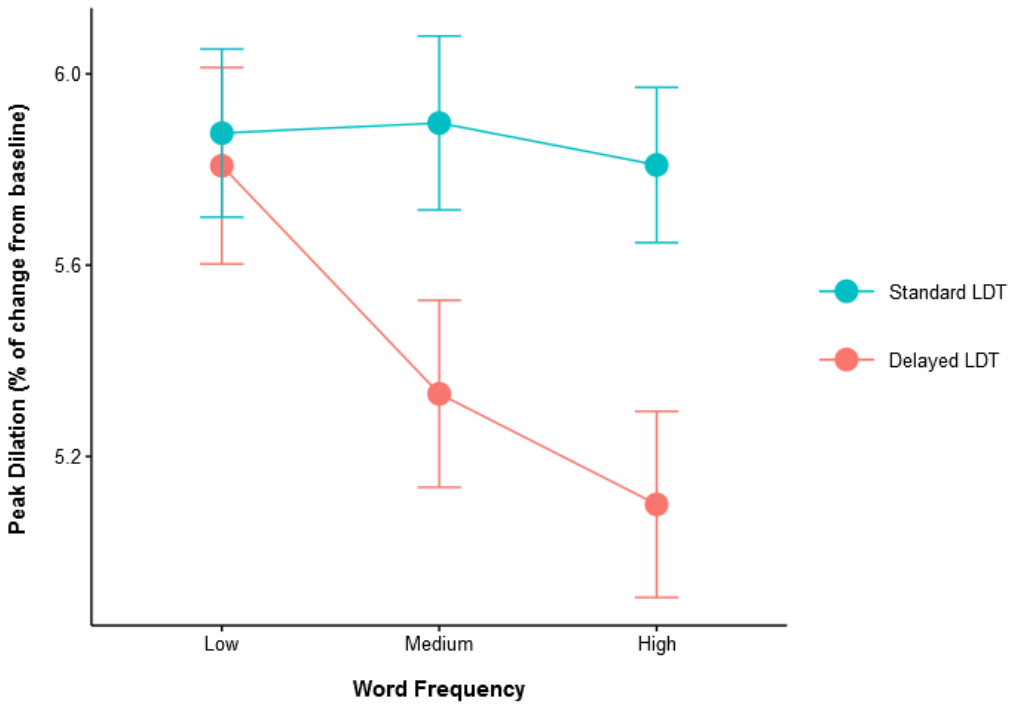

Figure 3. Average of peak dilations (in percentage of change from baseline) per word frequency for each task. Error bars represent the standard error of the mean.

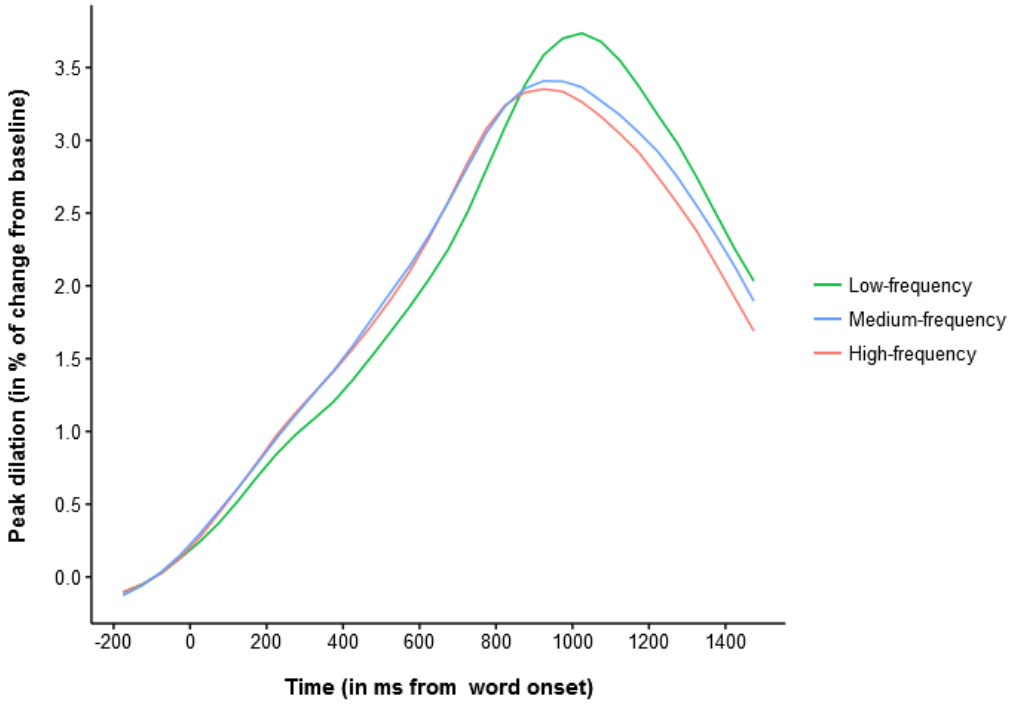

Figure 4. Grand average of peak dilations during a trial for each type of word in the standard LDT. Note that these values do not correspond to the average of peak latencies (Figure 2) and the average of peak dilations (Figure 3), given that peak dilations occurred at different times for different trials.

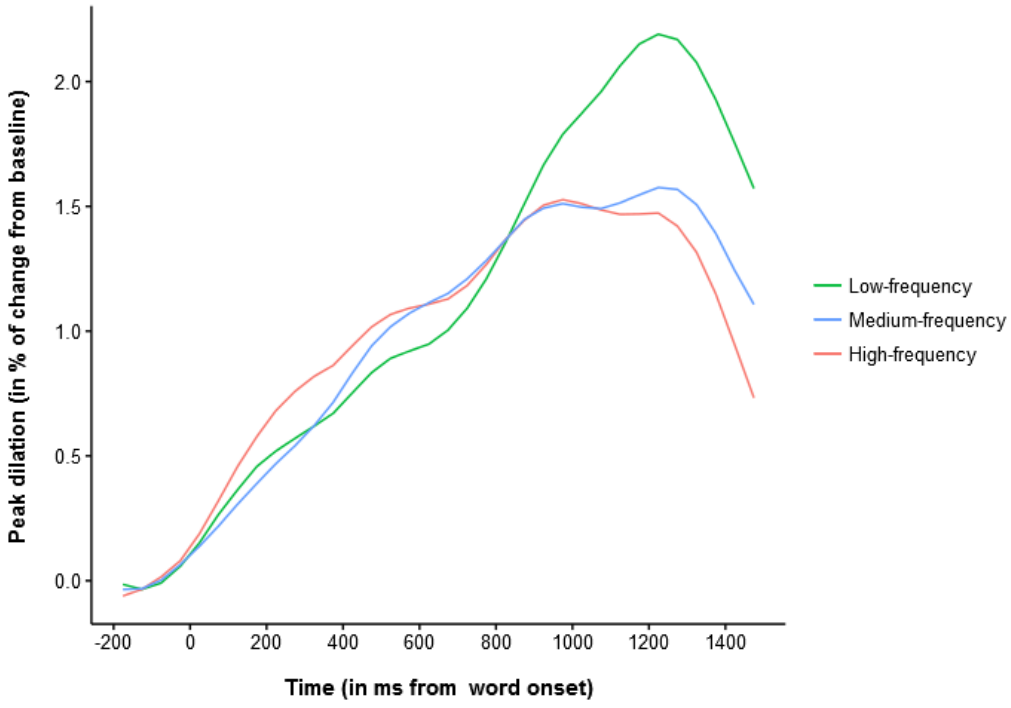

Figure 5. Grand average of peak dilations during a trial for each type of word in the delayed LDT. Note that these values do not correspond to the average of peak latencies (Figure 2) and the average of peak dilations (Figure 3), given that peak dilations occurred at different times for different trials.

The main effect of word frequency on peak dilation was not significant, χ^2 (2) = 4.71, p = .09. Likewise, the main effect of task did not reach significance, χ^2 (1) = 2.18, p = .14. Importantly, there was a significant interaction between word frequency and task, χ^2 (2) = 7.66, p = .02. Peak dilations were modulated by word frequency in the delayed LDT, χ^2 (2) = 11.08, p = .004, but not in the standard LDT, χ^2 (2) = 0.44, p = .80. In the delayed LDT, peak dilations were higher for low-frequency words in comparison to medium-frequency words, ß = -0.49, SE = 0.22, t = -2.18, p = .03, and to high-frequency words, ß = -0.75, SE = 0.22, t = -3.36, p = .001. Medium frequency words and high-frequency words elicited similar peak dilations, ß = -0.26, SE = 0.22, t = -1.19, p = .24. This interaction between word frequency and task suggests that pupillary response might have been affected by task characteristics. The similarities and dissimilarities between the results of the experiments 1a and 1b will be discussed in the next section.

General discussion

The present study aimed to test whether previous reports of a word frequency effect on the pupillary response in word recognition studies may be due to response planning or execution, rather than to word processing per se. To address this issue, we recorded pupillary responses to words differing in lexical frequency (low-, medium-, and high-frequency) during two LDT experiments. In the first one, participants performed a standard LDT, that is, they were asked to respond, as fast as possible, if a

string of letters corresponded to a word in Spanish or not. In contrast, in the second experiment participants performed a delayed LDT. In this task, they were not required to execute a response while the word was presented, allowing us to record pupillary response avoiding any potential effect of response planning or execution on such measure.

In both experiments we observed a modulation of the pupillary response produced by word frequency, as low-frequency words were associated with delayed pupil peaks in comparison to high-frequency words. Thus, a similar pattern of results was obtained although in Experiment 1a a response was required while the word was presented, whereas in Experiment 1b it was not. Accordingly, the findings of the present study are in line with those of Kuchinke et al. (2007) and Schmidtke (2014). Importantly, they suggest that the word frequency effect on the pupillary response reported in these studies was not due to a confounding effect produced by response execution. The word frequency effect on pupillary response may suggest that pupil size reflects the amount of activation needed for a word to reach the recognition threshold: an earlier and lower pupil peak would indicate that less activation is needed for a word to be recognized (Schmidtke, 2014).

Of note, although there was a significant effect of word frequency in peak latency in both experiments, pupil peak dilation was affected by word frequency only in the delayed LDT. This difference between tasks may indicate that when no immediate response is required, and so there is no time pressure, participants could perform a deeper processing of the stimulus in comparison to when an immediate response is required. This is also supported by the fact that pupil peak latencies were larger in the delayed LDT. Consequently, this deeper processing in the delayed LDT would allow participants to dedicate more time and resources to stimulus processing, leading to

clearer and larger differences between low- and high frequency words than in a task in which participants are urged to respond as fast as they can. This would be in in line with the results of Stone and Van Orden (1993), who found that the word frequency effect in a LDT was larger when difficult nonwords were included. The cause of this increased frequency effect might be that participants were compelled to analyze the stimuli in more depth to distinguish between words and nonwords. Thus, in that study as well as in the present study, the more time and resources devoted to stimulus processing, the larger and clearer the frequency effects. Accordingly, this would explain why peak dilation differences between low- and high-frequency words were observed in the delayed LDT, but not the standard LDT.

Taking all the above into account, we consider that there are at least two advantages of using a delayed response task, or even a task requiring no response at all, when using pupillometry to study word recognition. First of all, it avoids any potential influence of response execution during word processing. In this way, the pupillary response represents a purer measure of word processing than behavioral responses (e.g., RTs or percentage of errors), given that the latter do not allow us to separate the processing and response components in LDT. Consequently, the analysis of the pupillary response gives us the opportunity to test experimental hypotheses concerning word processing that would not be possible to test by recording behavioral responses. The second advantage of using a delayed response task is that it may lead to clearer and stronger experimental effects by allowing participants to perform a deeper stimulus processing.

In sum, the present study provides evidence of the reliability of pupillometry for word recognition research. We found that pupillary response was affected by word frequency when participants performed a LDT, either delayed or not. Thus, we can be

confident that the reported word frequency effect is not due to planning or executing a response. On the other hand, we have argued that using a task that allows isolating word processing from response execution may be more suitable for pupillometry research.

ACKNOWLEDGEMENTS:

This research was funded by the Spanish Ministry of Economy and Competitiveness (PCIN-2015-165-C02-02 and PSI2015-63525-P) and by a grant of the University Rovira i Virgili (2014PFR-URV-B2-37).

References

Alonso, M. A., Fernandez, A., & Díez, E. (2014). Subjective age-of-acquisition norms for 7,039 Spanish words. *Behavior Research Methods*, *47*(1), 268–274. doi:10.3758/s13428-014-0454-2

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*, 403–450. doi:10.1146/annurev.neuro.28.061604.135709

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412.

Bates D., Maechler M., Bolker B., Walker S. (2014). lme4: Linear Mixed-Effects Models Using Eigen and S4. R package version 1, 1–10. Available online at: http://CRAN.R-project.org/package=lme4

Duchon, A., Perea, M., Sebastián, N., Martí, M. A., & Carreiras, M. (2013). EsPal: one-stop shopping for Spanish word properties. *Behavior Research Methods*, *45*(4), 1246–58. doi:10.3758/s13428-013-0326-1

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior*, *12*(6), 627–635. doi:10.1016/S0022-5371(73)80042-8

Goldinger, S. D., He, Y., & Papesh, M. H. (2009). Deficits in cross-race face learning: insights from eye movements and pupillometry. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *35*(5), 1105–22. doi:10.1037/a0016548

Guasch, M., Ferré, P., & Fraga, I. (2015). Spanish norms for Affective and Lexico-Semantic variables for 1,400 words. *Behavior Research Methods*. Advance online publication. doi:10.3758/s13428-015-0684-y

Heitz, R. P., Schrock, J. C., Payne, T. W., & Engle, R. W. (2008). Effects of incentive on working memory capacity: behavioral and pupillometric data. *Psychophysiology*, *45*(1), 119–129. doi:10.1111/j.1469-8986.2007.00605.x

Hupé, J. M., Lamirel, C., & Lorenceau, J. (2009). Pupil dynamics during bistable motion perception. *Journal of vision, 9*(7), 1-19. doi: 10.1167/9.7.10.

Karatekin, C., Couperus, J. W., & Marcus, D. J. (2004). Attention allocation in the dual-task paradigm as measured through behavioral and psychophysiological responses. *Psychophysiology*, *41*(2), 175–185. doi:10.1111/j.1469-8986.2004.00147.x

Keuleers, E., & Brysbaert, M. (2010). Wuggy: a multilingual pseudoword generator. *Behavior Research Methods*, *42*(3), 627–633. doi:10.3758/BRM.42.3.627

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono- and disyllabic words and nonwords. *Frontiers in Psychology, 1*, 1–15. doi:10.3389/fpsyg.2010.00174

Klingner, J., Tversky, B., & Hanrahan, P. (2011). Effects of visual and verbal presentation on cognitive load in vigilance, memory, and arithmetic tasks. *Psychophysiology*, *48*(3), 323–332. doi:10.1111/j.1469-8986.2010.01069.x

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*(2), 132–140. doi:10.1016/j.ijpsycho.2007.04.004

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2014). lmerTest: Tests for random and fixed effects for linear mixed effect models: R package version 2.0-6. http://CRAN.R-project.org/package=lmerTest

Laeng, B., Sirois, S., & Gredeback, G. (2012). Pupillometry: A Window to the Preconscious? *Perspectives on Psychological Science*, *7*(1), 18–27. doi:10.1177/1745691611427305

Manelis, L. (1977). Frequency and meaningfulness in tachistoscopic word perception. *The American Journal of Psychology*, *90*(2), 269–280. doi:10.2307/1422049

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological review*, *88*(5), 375–407. doi:10.1037/0033-295X.88.5.375

Miller, J. (1982). Discrete versus continuous stage models of human information processing: In search of partial output. *Journal of Experimental Psychology: Human Perception and Performance, 8*(2), 273–296. doi:10.1037/0096-1523.8.2.273

Moresi, S., Adam, J. J., Rijcken, J., Van Gerven, P. W., Kuipers, H., & Jolles, J. (2008). Pupil dilation in response preparation. *International Journal of Psychophysiology*, *67*(2), 124–130. doi:10.1016/j.ijpsycho.2007.10.011

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

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

Rubenstein, H., Garfield, L., & Millikan, J. A. (1970). Homographic entries in the internal lexicon. *Journal of verbal learning and verbal behavior, 9*(5), 487–494. doi:10.1016/S0022-5371(70)80091-3

Schmidtke, J. (2014). Second language experience modulates word retrieval effort in bilinguals: Evidence from pupillometry. *Frontiers in Psychology*, *5*, 137. doi:10.3389/fpsyg.2014.00137

Stone, G. O., & Van Orden, G. C. (1993). Strategic control of processing in word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(4), 744–774. doi:10.1037/0096-1523.19.4.744

Tremblay, A., & Tucker, B. V. (2011). The effects of N-gram probabilistic measures on the recognition and production of four-word sequences. *The Mental Lexicon, 6*(2), 302–324.

Van der Molen, M. W., Boomsma, D. I., Jennings, J. R., & Nieuwboer, R. T. (1989). Does the heart know what the eye sees? A cardiac/pupillometric analysis of motor preparation and response execution. *Psychophysiology*, *26*(1), 70–80. doi:10.1111/j.1469-8986.1989.tb03134.x

Van Rijn, H., Dalenberg, J. R., Borst, J. P., & Sprenger, S. A. (2012). Pupil Dilation Co-Varies with Memory Strength of Individual Traces in a Delayed Response Paired-Associate Task. *PLoS ONE*, *7*(12). doi:10.1371/journal.pone.0051134