Behav Res (2017) 49:1930–1938
DOI 10.3758/s13428-016-0835-9

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

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

Published online: 6 December 2016

Abstract Previous word recognition studies have shown that the pupillary response is sensitive to a word's frequency. However, such a pupillary effect may be due to the process of executing a response, instead of being an index of word processing. With the aim of exploring this possibility, we recorded the pupillary responses in two experiments involving a lexical decision task (LDT). In the first experiment, participants completed a standard LDT, whereas in the second they performed a delayed LDT. The delay in the response allowed us to compare pupil dilations with and without the response execution component. The results showed that pupillary response was modulated by word frequency in both the standard and the delayed LDT. This finding supports the reliability of using pupillometry for word recognition research. Importantly, 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 processes. 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, never passing the threshold for eliciting overt behavior or for reaching consciousness (Laeng et al., 2012). Another remarkable property is that it is not affected by a participant's strategies, because it is quite difficult to control voluntarily. Due to these advantages, the pupillary response has been widely used in cognitive psychology to study a large variety of cognitive processes, including attention allocation (Karatekin, Couperus, & Marcus, 2004), face perception (Goldinger, He, & Papesh, 2009), arithmetic (Klingner, Tversky, & Hanrahan, 2011), and 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, psycholinguists have focused on identifying the variables that influence the processing of words. One of the most relevant variables is word frequency, which is usually measured as the number of occurrences of a word in a given language. This measure has been shown to predict performance in several experimental tasks, such as lexical decision (Rubenstein, Garfield, & Millikan, 1970), naming (Forster & Chambers, 1973), and perceptual identification (Manelis, 1977). In fact, word frequency is considered the best predictor of word recognition (e.g., Keuleers, Diependaele, & Brysbaert, 2010). A consistent finding from those 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 do low-frequency words, and thus that less activation is needed to recognize a high-frequency word than

✉ Juan Haro
juan.haro@urv.cat

[1] Research Center for Behavior Assessment (CRAMC) and Department of Psychology, Universitat Rovira i Virgili, Crta. de Valls s/n, Campus Sescelades, 43007 Tarragona, Spain

to recognize a low-frequency word (e.g., McClelland & Rumelhart, 1981).

To our knowledge, only two word recognition studies to date have examined 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., the peak dilation), and (b) the point in time at which this peak dilation is achieved (i.e., the 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. found that participants exhibited larger pupil peak dilations to low-frequency than to high-frequency words in a lexical decision task (LDT). Along the same lines, Schmidtke observed that low-frequency words elicited later pupil peak dilations (i.e., larger peak latencies) than did high-frequency words during a spoken word recognition task. Another study focused on word production, conducted by Papesh and Goldinger (2012), also investigated the effects of word frequency. These authors aimed to assess whether frequency effects in naming were allocated in the speech-planning stage or in the response execution stage. Toward that end, they used a delayed-naming procedure and included some trials in which participants did not have to produce any response. Consistent with the word recognition studies, Papesh and Goldinger found that the peak dilations for low-frequency words were larger than the peak dilations for high-frequency words before, during, and after naming responses. The procedure used by these authors was very interesting, since it allowed them to separate distinct components in the naming task and to examine the effects of the relevant variables in each component. A similar approach might be used with tasks tapping the early stages of word recognition. We did this in the present study, in which we focused on the LDT and tested whether the modulation of pupillary response by frequency is a reliable index of word-processing effort.

As we explained above, both Kuchinke et al. (2007) and Schmidtke (2014) found a modulation of pupil response by word frequency. Furthermore, both studies showed a convergence between behavioral and physiological data, since slower reaction times (RTs) were associated with larger pupil peak amplitudes (Kuchinke et al., 2007) or delayed pupil peak latencies (Schmidtke, 2014). However, there is the possibility that these findings may been biased by an experimental confounding factor, since in both studies participants were not only asked to recognize a word, but also to execute a response. In the work of Kuchinke et al., participants performed an LDT. In each trial, they 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., a 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.

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 differences not only in the processing of low- and high-frequency words, but also in the execution of a response. This would be in agreement with evidence 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. found that the difficulty of response preparation during a finger-cuing task (Miller, 1982) was correlated with pupil size, with more-difficult cues eliciting larger pupil dilations during response preparation and execution. In another study, Hupé et al. recorded pupil responses while their participants 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 on the pupillary response (Hupé et al., 2009).

Given the results 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 whether it rather reflects a genuine effect on word processing. We believe that addressing this issue is important for two main reasons: (1) to determine whether the study of pupillary responses can be reliably applied to word recognition research, and (2) to help identify the proper methodological requirements for the use of the pupillary response in this field of research. To that end, we conducted two experiments in which we examined the effects of word frequency during an LDT, by recording both the pupillary response and behavioral measures. The experimental stimuli of these two experiments were the same; namely, we compared low-, medium-, and high-frequency words. The differences between the two experiments were found in the requirements of the experimental tasks. In Experiment 1a, we used 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 required executing a response (Kuchinke et al., 2007; Schmidtke, 2014). On the other hand, 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 response execution, and thus providing a purer measure of the modulation of the pupillary response during word processing. Due to the similarities between the two experiments, we report them together.

Method

Participants

Sixty Spanish speakers were recruited for the study. Half of them (21 women and nine men; mean age = 20.63, *SD* = 3.18) participated in Experiment 1a, and the other half (28 women and two 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 sets of both experiments included 75 Spanish words. The 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 the conditions also differed in log frequency, lemma frequency, and log lemma frequency (all *p*s < .001).

The 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, trigram frequency, concreteness, imageability, context availability, arousal, and emotional valence ratings (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 correlations 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, the familiarity, concreteness, imageability, context availability, arousal, and emotional valence ratings were taken from Guasch, Ferré, and Fraga (2016), and the 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.

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 procedures for both experiments were 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. The 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-in. computer screen set to a resolution of 1,024 × 768 pixels).

The right eye's pupil diameter and position were continuously recorded at a sampling rate of 1000 Hz, using an EyeLink 1000 eyetracker. This eyetracker measures pupil diameter in arbitrary units (range: 400–16,000 units). It can measure pupil diameter with a resolution of 0.2% of the diameter (e.g., a resolution of 0.01 mm for a 5-mm pupil) and has a spatial resolution of 0.01° root mean square.

Stimuli were presented using the Experiment Builder software. All stimuli were drawn in black lowercase characters (font type Arial, 24 pixels) in the center of a gray background screen (RGB 150). In addition to the 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 1,000 ms. After that, the fixation cross was replaced by a letter string representing a Spanish word or nonword. Participants were instructed to press with their right hand either the mouse button labeled "YES" (left button) or the one labeled "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 until a response was made or a timeout (2,000 ms), and it was followed by a new fixation cross for 1,200 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.

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 (2.3)	5.5 (3.5)	1.6 (0.9)	5.0 (0.4)	8.1 (1.0)	7.3 (1.5)	3.1 (0.7)	4.7 (1.1)	4.9 (1.4)	5.2 (0.7)	5.4 (1.8)	5.3 (1.4)	2.0 (0.5)	2.1 (2.3)	0.2 (0.7)	4,992.5 (3,619.1)	674.2 (1,025.4)
Medium frequency	18.8 (6.1)	43.2 (78.3)	7.1 (2.1)	5.7 (0.6)	6.5 (1.8)	7.3 (1.4)	2.8 (0.6)	4.7 (1.1)	5.3 (1.4)	5.4 (0.5)	5.3 (1.2)	5.0 (1.2)	1.9 (0.5)	4.0 (5.5)	0.3 (0.6)	6,060.2 (3,659.6)	890.2 (572.5)
High frequency	54.3 (18.2)	97.5 (75.3)	18.0 (5.1)	6.5 (0.3)	5.4 (1.5)	6.8 (1.7)	2.8 (0.7)	4.6 (1.3)	4.8 (1.8)	5.3 (0.7)	5.2 (1.7)	4.9 (1.1)	1.9 (0.4)	2.7 (2.3)	0.1 (0.3)	6,638.1 (4,375.4)	1,294.8 (1,337.0)

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

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 1,000 ms. Then, the fixation cross was replaced by a letter string displaying a Spanish word or nonword. The letter string remained on the screen for 500 ms and was followed by a new fixation cross for 1,500 ms. After that, a question appeared on the screen asking participants to indicate whether the letter string was a Spanish word. They responded by pressing with their right hand either the mouse button labeled “YES” (left button) or the one labeled “NO” (right button). If participants responded before the question, a “Too quick” feedback message was displayed. The question remained on the screen for 2,000 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 responses (peak dilations and peak latencies). Of note, whereas the RTs in the standard task were measured from the onset of the presentation of the letter string, in the delayed task they were measured from the onset of the probe question until the participant's response. Furthermore, the processes of data cleaning and selection were nearly identical in both experiments. In Experiment 1a, trials with incorrect responses or nonresponses (4.29%), trials with RTs below 300 ms or over 1,500 ms (1.64%), and trials with RTs greater than two standard deviations above or 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, because accuracy was very high (between 0% and 10.67% errors, *M* = 4.29%, *SD* = 2.44%). In Experiment 1b, as in Experiment 1a, we first removed trials with incorrect responses and nonresponses (2.91%) and trials with RTs greater than two standard deviations above or below the participant's mean (5.30%). In this experiment, it was important to also remove anticipation responses—that is, trials on which 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 removed from the analyses by the numbers of errors committed (including anticipations). In this experiment, the accuracy ranged from 0.67% to 10.67% (*M* = 4.44%, *SD* = 2.35%), which were very similar values to those from Experiment 1a.

All pupillary data were processed using a Python script. First, samples with saccades or eye blinks were removed. We extended the rejection area by 25 samples on both sides for saccades, and by 50 samples on either side for blinks, to exclude pre- and postblink 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, the 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 extending between 200 ms before target onset and 1,500 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, the 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 1,500 ms. To allow for comparisons between participants, the peak dilation was converted into a relative dilation, expressed as a proportional difference (in percentage of change) from the baseline. The 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 Exp. 1a and 14.57% in Exp. 1b) were removed, following the procedure used by Schmidtke (2014). In addition, we removed four trials from Experiment 1a (0.09% of the total) due to more than 50% of missing samples. In sum, after the data cleaning, 4,181 data points were submitted to the analyses (2,136 from Exp. 1a, and 2,045 from Exp. 1b).

Data analysis

Analyses were performed with R using the lme4 package (Bates, Maechler, Bolker, & Walker, 2014). The effect of each dependent variable (RTs, peak latencies, and peak dilations) was analyzed separately using linear mixed-effects 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 × 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*s below and above the model residuals' mean were removed from the dataset (e.g., Baayen, 2008; Tremblay & Tucker, 2011), and the models were refitted to the trimmed data. Of note, fewer than 3% of the data points were removed after applying this trimming procedure. The significance of the fixed effects was determined using log-likelihood ratio tests (R function anova). Namely, we evaluated the contributions of each fixed effect and the interaction by comparing a model that included the effect of interest to one that did not include such an effect. The *p* values for pairwise comparisons were estimated using the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2014), which relies on *t* distributions with degrees of freedom derived by the Satterthwaite approximation. We would like to note that we did not perform accuracy analyses, due to the low number of error responses.

Results

Reaction times

Averages of the RTs for each condition and task are presented in Fig. 1 (averaged over individual trials).

There was a significant effect of word frequency on RTs, $\chi^2(2) = 17.82$, $p < .001$. Low-frequency words were responded to more slowly with respect to either medium-frequency words, $\beta = 21.16$, $SE = 7.25$, $t = 2.92$, $p = .005$, or high-frequency words, $\beta = 31.46$, $SE = 7.25$, $t = 4.34$, $p < .001$. No differences were observed between medium-frequency and high-frequency words, $\beta = 10.29$, $SE = 7.19$, $t = 1.43$, $p = .16$. In addition, the average RTs did not differ between tasks, $\chi^2(1) = 2.82$, $p = .09$. The interaction of word frequency and task reached significance, $\chi^2(2) = 55.5$, $p < .001$. Word frequency had a significant effect on RTs in the standard LDT, $\chi^2(2) = 36.51$, $p < .001$, but not in the delayed LDT, $\chi^2(2) = 2.17$, $p = .34$. In the standard LDT, low-frequency words were responded to more slowly than medium-frequency words, $\beta = 46.74$, $SE = 11.83$, $t = 3.95$, $p < .001$, and high-frequency words, β= 79.34, $SE = 11.81$, $t = 6.72$, $p < .001$. Additionally, medium-frequency words were

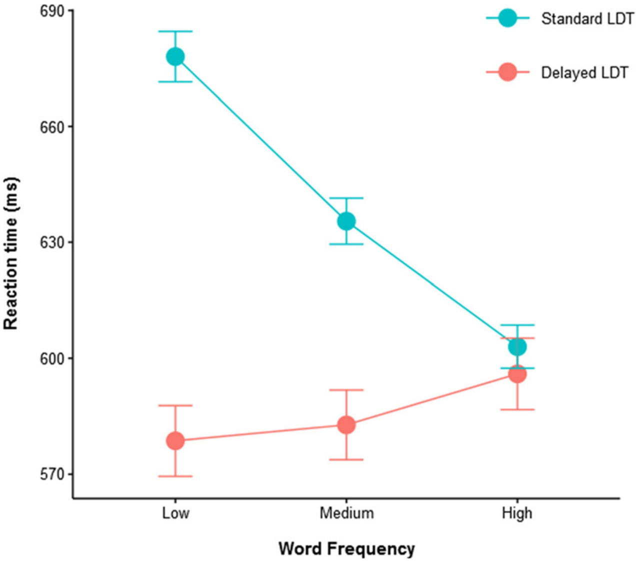

Fig. 1 Average reaction times (in milliseconds) per word frequency for each task. Error bars represent the standard errors of the means

responded to more slowly than high-frequency words, β = 32.60, SE = 11.75, t = 2.78, p = .007.

Peak latency

Figure 2 shows the means of the peak latencies for each condition and task, averaged over individual trials. We first ensured that the baseline pupil diameters were equivalent across conditions in both tasks: standard LDT, $F(2, 58) = 0.76$, p = .47, and delayed LDT, $F(2, 58) = 0.41$, p = .66.

There was a main effect of word frequency on peak latencies, $\chi^2(2) = 16.39$, $p < .001$. The peak latencies for low-frequency words were marginally larger in comparison to those for medium-frequency words, $\beta = 41.74$, $SE = 22.44$, $t = 1.86$, $p = .07$, and significantly larger than those for high-frequency words, $\beta = 94.36$, $SE = 22.43$, $t = 4.21$, $p < .001$. In addition, the peak latencies for medium-frequency words were larger than the peak latencies for high-frequency words, $\beta = 52.61$, $SE = 22.21$, $t = 2.37$, $p = .02$. We also observed a significant effect of task, in which peak latencies were larger in the delayed than in the standard LDT, $\beta = 129.17$, $SE = 48.51$, $t = 2.66$, $p = .01$. The interaction between word frequency and task was significant, $\chi^2(2) = 8.61$, $p = .01$. Word frequency modulated the peak latencies in both tasks—the standard LDT, $\chi^2(2) = 7.29$, $p = .03$, and the delayed LDT, $\chi^2(2) = 14.21$, $p < .001$—but the patterns of results differed between the two tasks. In the standard LDT, differences were found between low-frequency and medium-frequency words, $\beta = 50.94$, $SE = 21.78$, $t = 2.34$, $p = .02$, and between low-frequency and high-frequency words, $\beta = 52.19$, $SE = 21.69$, $t = 2.41$, $p = .02$. Conversely, in the delayed LDT, the differences were observed between high-frequency and medium-frequency words, $\beta = 111.32$, $SE = 38.14$, $t = 2.92$, $p = .004$, and between high-frequency and low-frequency words, $\beta = 142.36$, $SE = 38.45$, $t = 3.70$, $p < .001$.

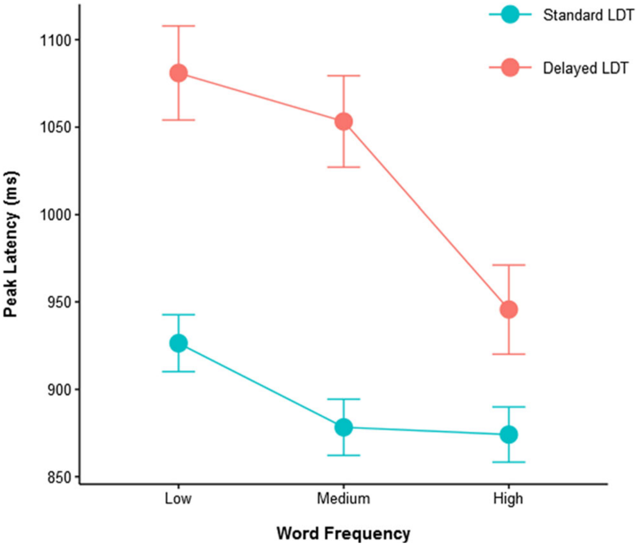

Fig. 2 Average peak latencies (in milliseconds) per word frequency for each task. Error bars represent the standard errors of the means

Peak dilation

Peak dilations for each condition and task, averaged over individual trials, are shown in Fig. 3. In addition, grand averages of the peak dilations during trials are shown in Fig. 4 (standard LDT) and in Fig. 5 (delayed LDT).

The main effect of word frequency on peak dilations was not significant, $\chi^2(2) = 4.71$, $p = .09$. Likewise, the main effect of task did not reach significance, $\chi^2(1) = 2.18$, $p = .14$. Importantly, there was a significant interaction between word frequency and task, $\chi^2(2) = 7.66$, $p = .02$: Peak dilations were modulated by word frequency in the delayed LDT, $\chi^2(2) = 11.08$, $p = .004$, but not in the standard LDT, $\chi^2(2) = 0.44$, $p = .80$. In the delayed LDT, peak dilations were higher for low-frequency words in comparison to both medium-frequency words, $\beta = 0.49$, $SE = 0.22$, $t = 2.18$, $p = .03$, and high-frequency words, $\beta = 0.75$, $SE = 0.22$, $t = 3.36$, $p = .001$. Medium-frequency and high-frequency words elicited similar peak dilations, $\beta = 0.26$, $SE = 0.22$, $t = 1.19$, $p = .24$.

Discussion

In the present study, we aimed to test whether previous reports of a word frequency effect on the pupillary response in word recognition studies may have been due to response 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) during two LDT experiments. In the first experiment, participants performed a standard LDT; that is, they were asked to respond, as quickly as

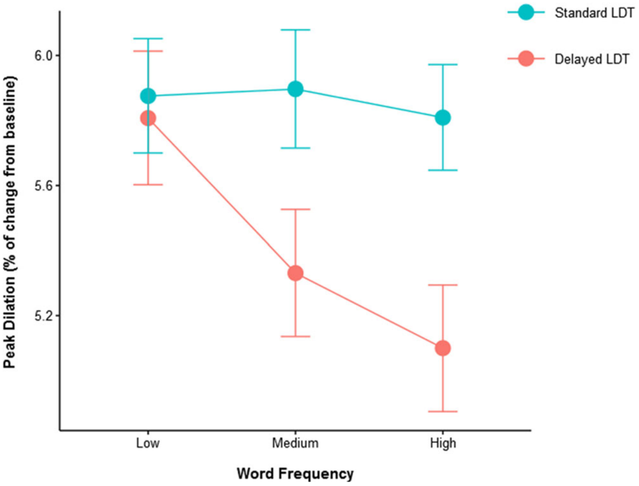

Fig. 3 Average peak dilations (in percentages of change from baseline) per word frequency for each task. Error bars represent the standard errors of the means

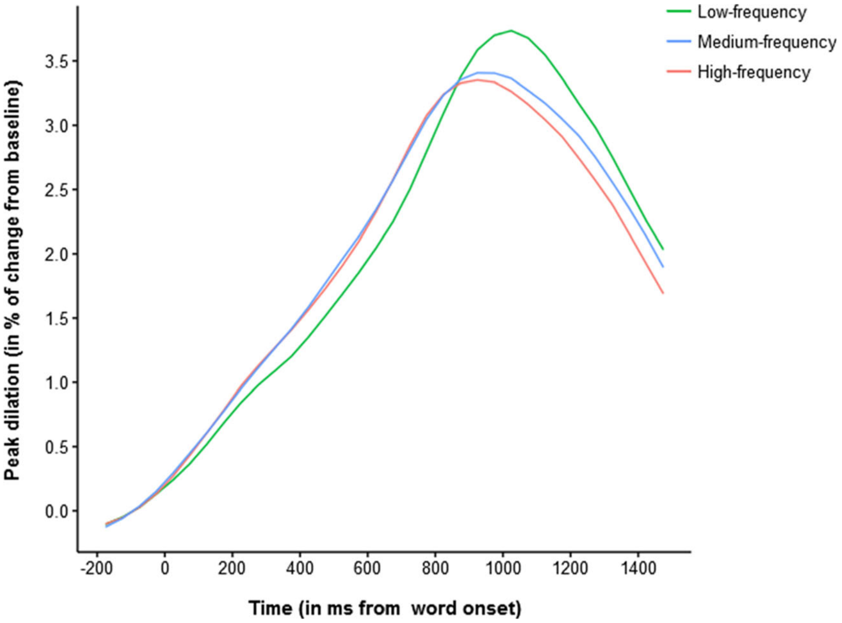

Fig. 4 Grand averages of peak dilations during the trials for each type of word in the standard lexical decision task. Note that these values do not correspond to either the average peak latencies (Fig. 2) or the average peak dilations (Fig. 3), given that peak dilations occurred at different times for different trials

possible, with whether or not a string of letters corresponded to a word in Spanish. 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 execution on such a measure.

The results showed the expected behavioral word frequency effect in the LDT, with participants responding faster to high-frequency than to low-frequency words in the standard LDT. More importantly, we observed that pupillary response was modulated by word frequency in both experiments. Peak latencies were larger for low-frequency than for high-frequency words when participants performed a standard LDT. Furthermore, we found larger peak latencies as well as larger peak dilations for low-frequency with respect to high-frequency words when pupillary responses were registered during a delayed LDT. Thus, a similar pattern of results was obtained, although in the standard LDT a response was

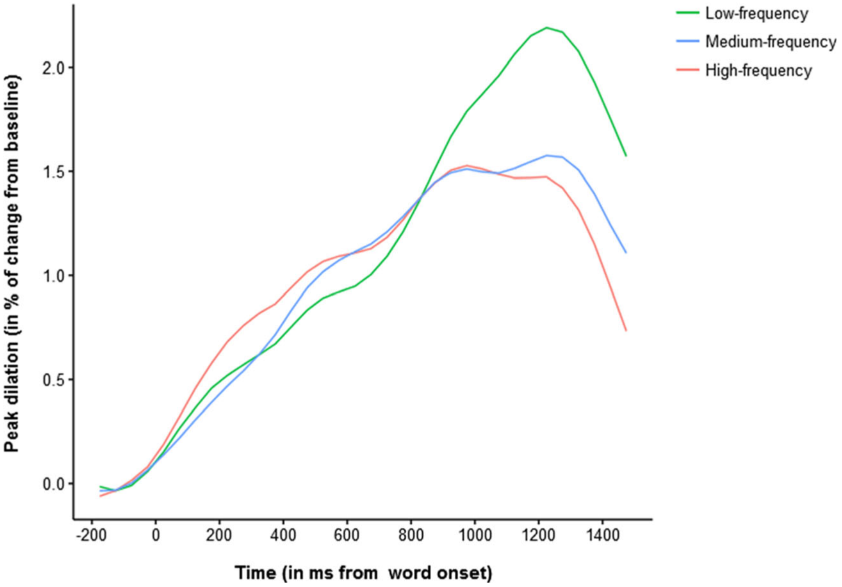

Fig. 5 Grand averages of peak dilations during the trials for each type of word in the delayed lexical decision task. Note that these values do not correspond to the average peak latencies (Fig. 2) or the average peak dilations (Fig. 3), given that peak dilations occurred at different times for different trials

required while the word was being presented, whereas in the delayed LDT it was not. These findings are in line with previous reports showing an influence of word frequency on the pupil peak latency (Schmidtke, 2014) and pupil peak dilation (Kuchinke et al., 2007). Accordingly, our results suggest that the word frequency effect on the pupillary response found in those studies was not due to a confounding with response execution. Rather, this effect seems to be an index of the difficulty of word processing. Thus, pupil size may reflect the amount of activation needed for a word to reach the recognition threshold: An earlier and lower pupil peak would indicate that less activation was needed for a word to be recognized (Schmidtke, 2014).

Nevertheless, we should mention some discrepancies between our results and those observed in previous studies, as well as between those obtained by using different tasks in the present study. First, the null effect on peak dilations found in the standard LDT is partially at odds with the results of Kuchinke et al. (2007), who reported that word frequency modulated peak dilation during an LDT. We believe that this inconsistency may be due at least partially to some particularities of the stimulus selections. For instance, Kuchinke et al. only included two levels of word frequency (low and high), whereas the distribution of word frequencies in our experiment was wider (low, medium, and high). Importantly, Schmidtke (2014) was also unable to find a word frequency effect on peak dilations, and the distribution of word frequencies in their experiment was also wider than the one used by Kuchinke et al. Another relevant difference between our study and the work of Kuchinke et al. is that those authors not only manipulated word frequency, but also the emotional valence of words. Although we do not know the extent to which the comparison of two extreme conditions of frequency (high vs. low) or the orthogonal manipulation of frequency and emotional valence may have affected their pattern of results, what is clear is that these methodological differences make the work of Kuchinke not entirely comparable to ours.

Whichever is the cause of the divergences between studies, we consider that the fact that word frequency only affected peak dilations in the delayed LDT is an interesting result. It may indicate that when no immediate response is required, so there is no time pressure, participants could perform deeper processing of the stimulus than when an immediate response is required. This is also supported by the finding 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 were urged to respond as quickly as they could. This would be in line with the results of Stone and Van Orden (1993), who found that the word frequency effect in an 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 one, the more time and resources that were devoted to stimulus processing, the larger and clearer the frequency effects.

Taking all of the above into consideration, in our opinion 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, this technique avoids any potential influences of response execution during word processing. In this way, the pupillary response represents a purer measure of word processing than do behavioral responses (e.g., RTs or percentages of errors), given that the latter do not allow us to separate the processing and response components in the LDT. Thus, the analysis of the pupillary response with the delayed task gives us the opportunity to test experimental hypotheses concerning word processing that would not be possible to test by recording behavioral responses alone or by using the standard task. 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 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 an LDT, either delayed or not delayed. Thus, we can be confident that the reported word frequency effect is not due to 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.

Author note This research was funded by the Spanish Ministry of Economy and Competitiveness (Grant No. PSI2015-63525-P) and by a grant of the University Rovira i Virgili (Grant No. 2015PFR-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,* 268–274. doi:10.3758/s13428-014-0454-2

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. doi:10.1016/j.jml.2007.12.005

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4 (R package version 1.1-10). Retrieved from 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,* 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,* 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,* 1105–22. doi:10.1037/a0016548

Guasch, M., Ferré, P., & Fraga, I. (2016). Spanish norms for affective and lexico-semantic variables for 1,400 words. *Behavior Research Methods, 48,* 1358–1369. 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,* 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), 10:1–19. 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,* 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,* 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*(174), 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,* 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,* 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). Retrieved from 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,* 18–27. doi:10.1177/1745691611427305

Manelis, L. (1977). Frequency and meaningfulness in tachistoscopic word perception. *American Journal of Psychology, 90,* 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,* 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,* 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,* 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,* 754–765. doi:10.3758/s13414-011-0263-y

Rubenstein, H., Garfield, L., & Millikan, J. A. (1970). Homographic entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior, 9,* 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,* 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. *Mental Lexicon, 6,* 302–324. doi:10.1075/ml.6.2.04tre

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,* e51134. doi:10.1371/journal.pone.0051134