COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 48 (2024) e13446

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13446

Pupils Dilate More to Harder Vocabulary Words than Easier Ones

Ishanti Gangopadhyay,[a] Daniel Fulford,[b,c] Kathleen Corriveau,[d] Jessica Mow,[b] Pearl Han Li,[e] Sudha Arunachalam[f]

[a]*Department of Speech, Language and Hearing Sciences, Indiana University*
[b]*Department of Occupational Therapy, Boston University*
[c]*Department of Psychological and Brain Sciences, Boston University*
[d]*Department of Counseling Psychology and Applied Human Development, Boston University*
[e]*Department of Psychology and Neuroscience, Duke University*
[f]*Department of Communicative Sciences and Disorders, New York University*

Received 29 September 2023; received in revised form 27 March 2024; accepted 8 April 2024

Abstract

Understanding cognitive effort expended during assessments is essential to improving efficiency, accuracy, and accessibility within these assessments. Pupil dilation is commonly used as a psychophysiological measure of cognitive effort, yet research on its relationship with effort expended specifically during language processing is limited. The present study adds to and expands on this literature by investigating the relationships among pupil dilation, trial difficulty, and accuracy during a vocabulary test. Participants ($n = 63$, $M_{age} = 19.25$) completed a subset of trials from the Peabody Picture Vocabulary Test while seated at an eye-tracker monitor. During each trial, four colored images were presented on the monitor while a word was presented via audio recording. Participants verbally indicated which image they thought represented the target word. Words were categorized into Easy, Medium, and Hard difficulty. Pupil dilation during the Medium and Hard trials was significantly greater than during the Easy trials, though the Medium and Hard trials did not significantly differ from each other. Pupil dilation in comparison to trial accuracy presented a more complex pattern, with comparisons between accurate and inaccurate trials differing depending on the timing of the stimulus presentation. These results present further evidence that pupil dilation increases with cognitive effort associated with vocabulary tests, providing insights that could help refine vocabulary assessments and other related tests of language processing.

Keywords: Vocabulary; Task-evoked Pupillary response; Eye-tracking

Correspondence should be sent to Ishanti Gangopadhyay, 2631 E. Discovery Pkwy, Rm. C2015, Bloomington, IN 47408 USA. E-mail: ishgang@iu.edu

1. Introduction

Standard assessments of a variety of cognitive functions, including memory and language, involve measuring accuracy on probes. For example, in a widely used measure of receptive vocabulary, the Peabody Picture Vocabulary Test (PPVT; Dunn & Dunn, 2007), participants view an array of four pictures and hear a word that matches one of them; they are asked to identify which picture matches the word. In these assessments, the difficulty of the words increases as participants proceed through the task, allowing the investigator to identify an individual score. As the task becomes more difficult for the participant, accuracy naturally decreases. However, for trials on which the participant gives an accurate response, it is impossible to determine whether the task was very difficult and effortful or whether they knew the answer easily. Likewise, for an inaccurate response, it is unknown whether the participant did not know the word or whether they were just not engaged. Evaluating the point at which the task becomes very effortful for the participant would help in the design of adaptive assessments (e.g., Sharma, Papamitsiou, Olsen, & Giannakos, 2020) by tailoring the difficulty of questions to the individual's ability level.

One objective indicator of cognitive effort is pupil dilation (e.g., Beatty, 1982; van der Wel & van Steenbergen, 2018). The pupils dilate for several different reasons, including not only changes in the brightness of ambient lighting, but also internal participant experiences such as emotional arousal (e.g., Bradley, Miccoli, Escrig, & Lang, 2008; Partala & Surakka, 2003) and task-related processing difficulty (e.g., Ahern & Beatty, 1979; Alnæs et al., 2014; Beatty, 1982). Assessing cognitive effort from pupil dilation, or pupillometry, is valuable because one cannot consciously manipulate pupil size or inhibit pupillary responses to stimuli (e.g., Loewenfeld, 1993). Pupillometry is a well-established method for assessing cognitive load and has been applied specifically in language processing tasks (e.g., Chapman & Hallowell, 2015; Demberg & Sayeed, 2016; McLaughlin et al., 2022; Tromp, Hagoort, & Meyer, 2016; Zekveld, Kramer, & Festen, 2011). These studies take advantage of the fact that changes in pupil diameter can be time-locked to stimuli that elicit different cognitive demands, where larger pupil dilations indicate greater cognitive loads.

In the current study, we apply this approach to vocabulary knowledge. We ask whether pupil dilation is linked to the difficulty of the words and to participants' response accuracy. This line of inquiry is significant because by characterizing pupillary responses in vocabulary tasks with varying difficulty, we hope to be able to create more adaptive and personalized assessments. Such tailored assessments could ensure that the assessment windows accurately match each participant's ability, thereby enhancing the precision of the assessment outcomes. Additionally, pupil dilation can serve as an indicator of cognitive fatigue, which can affect the accuracy of the assessment. Therefore, pupillometry can reflect how engaged a participant is in a language task and allow the assessor to decide whether the assessment results are likely to be trustworthy.

There is limited evidence showing that pupil dilation is related to the difficulty of language assessments (Chapman & Hallowell, 2015; Ledoux et al., 2016). Chapman and Hallowell (2015) presented adults with a single image and a matching word (non-matching words were also included as filler trials); there was no accompanying task. They found greater pupil

dilation during the presentation of harder words as compared to easier words for typically developing adult participants (and a similar pattern for patients with aphasia). This finding suggests that participants expend cognitive effort on retrieving vocabulary words even when not prompted with a particular task. Nevertheless, because participants were only presented with a single picture at a time and were not asked to label the referent, it is unknown whether such pupil dilation was related to accurate word knowledge.

In another study, Ledoux et al. (2016) compared eye-gaze measures typically used in language processing studies (related to saccades and fixations) with pupillometry and event-related potential data. Their goal was to explore variability in these measures when participants were presented with familiar versus unfamiliar words. As in the current study, the authors showed participants an array of four pictures and provided an auditory recording of a word that matched one of the pictures. They found that pupil dilation (relative to baseline pupil size) was greater for words that were expected to be unknown (more difficult words) than words that were expected to be known (easier words).

We aimed to replicate and extend this finding by comparing pupil dilation not only by item difficulty but also by accuracy, exploring whether there are different patterns on trials in which participants respond accurately as compared to inaccurately. We used trials from the PPVT as the vocabulary measure. Accordingly, our study (like that of Ledoux et al., 2016) included different images across trials. The variability in the PPVT images raises a potential confound, namely, that differences in pupillary response across trials are related to processing the images rather than processing the words. In other words, it is possible that across the different trials, variability in the images (e.g., shape, complexity, color) may also affect pupil dilations. For instance, the first few trials that depict referents even preschoolers are familiar with (e.g., toys) tend to have more colorful images than the later trials. To control for potential differences in image properties across the trials with varying difficulty, we included two timing conditions (see below) where the participant either heard the word 1 s after the picture presentation or 4 s after the picture presentation. This timing difference allowed us to see whether effects of image properties on pupil dilation occur only when participants first see the images, or whether effects of word difficulty obtain *over and above* these image-related differences, after the word is heard. We also included image brightness as a factor in the analyses to provide an additional control for differences in image properties.

2. Method

2.1. Participants

The final sample included 63 undergraduate students at Boston University in the United States. Participants received course credit for participation. Of the 63 participants, 10 did not complete the demographics form. Of the remaining 53 participants, 34 were female and 19 were male. Ages ranged from 18 to 24 years ($M = 19.25$, $SD = 1.27$). Most participants (22) identified as White; 19 as Asian, five as multiracial, five as Black, and two participants did not report. Four participants identified themselves as Hispanic/Latino (of any race). Informed

consent was obtained following procedures approved by Boston University's Institutional Review Board. Participants were randomly assigned to one of two conditions in a between-subjects design: a 1-s condition ($n = 33$) and a 4-s condition ($n = 30$). The conditions were identical except for the timing of when the target word was heard relative to trial onset (see below).

2.2. Materials

The PPVT is a widely used, validated and norm-referenced measure of receptive vocabulary for American English, which is designed for children ages 2.5 years through adulthood. Each of the 228 trials on the 4th edition of the PPVT (Dunn & Dunn, 2007) used in the current study consists of four colored images arranged in quadrants; only one image matches the target word. Target words belong to different content areas (e.g., animals, tools, vegetables, body parts, etc.) and parts of speech (e.g., nouns, verbs, adjectives). The test was designed to increase progressively in difficulty from trials 1 to 228, with the first several trials being appropriate for very young children, and the last several unfamiliar to many adults (the norms show a mean raw score of 206.2 words for the best-performing age group—adults aged 41–50). The PPVT-4 includes two forms (A and B) that have different words to allow for retesting; we selected a subset of 108 words from the PPVT-4 Form A. We selected 36 Easy words, 36 Medium words, and 36 Hard words based on where they appear in the assessment. For the Easy category, we selected words 1–36 (e.g., *spoon*, *turtle*, *fence*); preschool-aged children are expected to know these words (as indicated by the fact that when administering the assessment to a 5-year-old, you would begin the assessment at item 37 to bypass these initial "too easy" words). For the Medium difficulty category, we selected words from the middle of the assessment, 121–156 (e.g., *puzzled*, *tusk*, *hedge*); when administering the assessment to adults 19 years of age and older, you would begin the assessment just after these, at item 157. For the Hard category, we selected the 36 words that appear at the end of the assessment, 193–228 (e.g., *trajectory*, *cupola*, *supine*), which includes words that many adults would not know.

Two additional items from relatively early in the assessment, words 37 and 38, were used as practice trials. These occurred as the first two trials in the procedure. The remaining trials were arranged in a fixed order for each participant, pseudorandomized such that no more than two words of the same difficulty occurred in a row. The auditory stimuli consisted of the target words, recorded by a native English speaker in an audio booth (as single words in isolation, e.g., "flower"). The recordings were synchronized with the visual stimuli according to the timeline in Fig. 1.

2.3. Apparatus

Participants viewed the stimuli on a Tobii T60XL eye-tracker monitor, which samples at 60 frames/s. They sat in a comfortable stationary chair with their eyes positioned within the range of 60–75 cm from the monitor. Ambient lighting was kept constant across all participants to control for luminance.

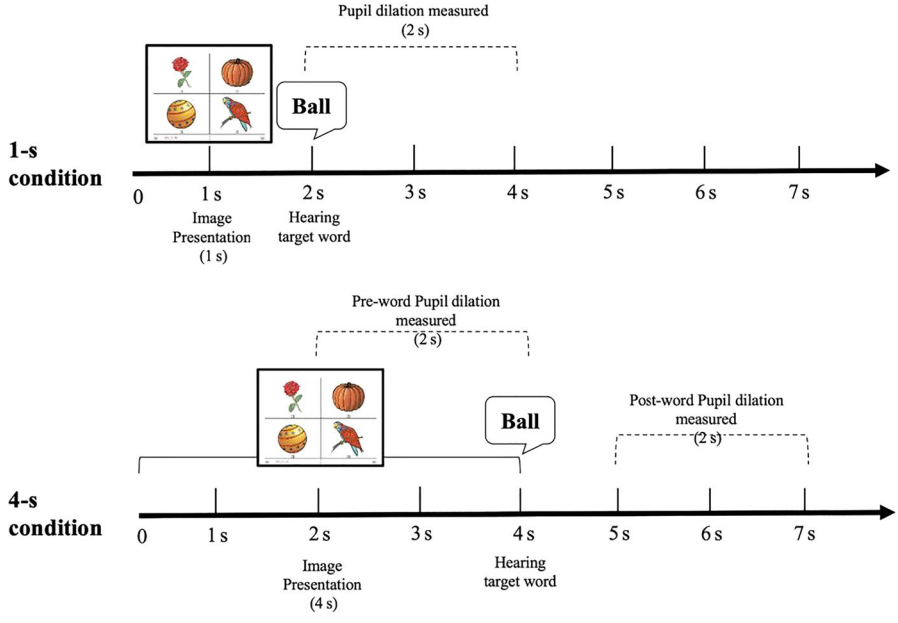

Fig. 1. Timeline of the Peabody Picture Vocabulary Test task in the 1-s and 4-s conditions.

2.4. Design and procedure

Participants sat in front of the monitor and underwent a standard 5-point calibration. They then participated in a digit span task and the PPVT task and then two unrelated tasks and a battery of self-report questionnaires in a fixed order. Only the PPVT task was included in the current analysis.

For the PPVT task, participants were assigned to either a 1-s or a 4-s condition (see Fig. 1). Because the images differed across trials, and because image properties are known to affect pupil size (e.g., Sirois & Brisson, 2014), we included two conditions with different timing to ensure that any effects observed were due to word difficulty and not due to visual properties of the images. (We also measured the average brightness on each trial in Adobe Photoshop and included this as a factor in analyses.)

In the 1-s condition, the display on each trial (i.e., four images arranged in quadrants) was presented on the screen for 1 s prior to the audio presentation of the target word. In the 4-s condition, the image was presented on the screen for 4 s prior to the audio presentation of the target word.

To measure participants' accuracy at identifying the target image on each trial, they were invited to provide a verbal response from 1 to 4 indicating the picture that matched the label (see Fig. 1). Responses were recorded by an experimenter.

2.5. Analysis plan

Our primary measure was pupil diameter (in millimeters, unstandardized) as reported by Tobii Studio software (version 3.1.6, Tobii Technology). We analyzed data from the right

eye only. Data preprocessing involved determining the usability of the right pupil data. We removed: (1) all saccades and "unclassified" data points so that we were left only with fixations, (2) all data points with a validity code of 2 or higher in the right eye (which included blinks) or a pupil dilation of "NA," and (3) all data points for which the participant was looking at coordinates outside the screen.

We then aggregated pupil diameter over a baseline period and a response period. We followed standard analysis procedures (e.g., Ledoux et al., 2016; McLaughlin & Van Engen, 2020). All analyses were conducted in R version 4.1.2 (R Core Team, 2022) and the lme4 package version 1.1.28 (Bates, Mächler, Bolker, & Walker, 2015); the lmerTest package version 3.1.3 (Kuznetsova, Brockhoff, & Christensen, 2017) provided *p*-values. A secondary dependent measure was Accuracy on each trial as indicated by the participants' verbal indication of the image that matched the previously heard label.

Next, we corrected pupil dilations in the analysis window for baseline dilation. For the 1-s condition, we used a baseline window of 500–999 ms[1]. We averaged across this baseline period for each trial for each participant to yield a single baseline value per trial. The test window was 2000–4000 ms. To correct for baseline, we subtracted this baseline value from the pupil dilation at each time point for each participant and trial.

For the 4-s condition, we evaluated the data in two ways. First, to assess the possibility that any effects of Trial Difficulty observed in the 1-s condition were due to luminance or other differences in the images, we repeated the same process as for the 1-s condition, using a baseline of 500–999 ms and a test window of 2000–4000 ms. Differences between trials of different difficulties in this window could not be due to the difficulty of the word, as the word had not yet been uttered.

Second, we asked whether there were effects of Trial Difficulty in response to the word by using a baseline of 2000–3999 ms and a test window of 5000–7000 ms. The choice of this longer baseline period, which was identical to the test period used for the 1-s condition but before the word was heard in the 4-s condition, allowed us to correct for baseline differences related to luminance or image differences that accumulated as participants had time to inspect the pictures.[2]

To analyze the data statistically, we used growth curve analysis (Mirman, 2017). Baseline-corrected pupil diameters were aggregated into time bins of 20 ms. Standard orthogonal polynomial time variables of linear time, quadratic time, and cubic time were included in the growth curve model. We tested multiple base models to determine which time terms were contributing significantly, using an ANOVA to identify the best-fitting model. All time terms that contributed to significantly better fit were included as fixed effects in subsequent analyses, along with random effects of participant and item. (We initially included a full random effects structure with random slopes for the orthogonal polynomial time terms, but these models did not converge.) To examine the effects of Trial Difficulty, we ran our first model with Easy trials as the reference level to ascertain if Medium and Hard trials differed from it and then a second model with Medium trials as the reference level to look for differences between Medium and Hard trials.

Finally, to look for the effects of trial accuracy, we repeated the analyses, now including whether participants were accurate or not (Trial Accuracy) as a fixed effect but removing

the interaction with time terms from the model so that the Accuracy main effect would be interpretable on its own. Given the high rates of accuracy on Easy and Medium trials, we only conducted these analyses with Hard Difficulty words.

3. Results

3.1. Accuracy

Due to experimenter error, six participants did not have accuracy data recorded; these participants were excluded resulting in a sample of 58 adults for the analyses involving accuracy. Mean participant accuracy was 85 words (standard deviation = 7 words), ranging from 64 to 100. As expected, accuracy varied by trial difficulty. On Easy trials, participants were nearly 100% accurate; there were only three inaccurate responses from two participants (in the 4-s condition). On Medium Difficulty trials, accuracy was 93% (1-s condition, 93% accurate; 4-s condition, 93% accurate). On Hard trials, accuracy was 45% (1-s condition, 43% accurate; 4-s condition, 47% accurate). Because of the high accuracy on Easy and Medium Difficulty trials, we conducted analyses of pupil dilation by accuracy only on Hard trials (see below).

3.2. 1-second condition

Model comparison using ANOVA indicated that the model including linear, quadratic, and cubic time terms was the best fitting. We first evaluated this model with the Easy condition as the reference level. We only report parameters of interest in the text, which relate to the main effects of Trial Difficulty, Accuracy, and image brightness; full models are available in the Supplementary Materials. The parameter of interest was Trial Difficulty (Easy, compared to Medium; Easy; compared to Hard), and there were significant main effects of both (Medium $\beta = 0.058$, $p < .001$; Hard $\beta = 0.062$, $p < .001$; Supplementary Table S1). We re-ran the model including image brightness; this model yielded a nonsignificant effect of brightness ($\beta = -0.0034$, $p = .074$), and crucially, the effects of Trial Difficulty remained unchanged (Medium $\beta = 0.050$, $p < .001$; Hard $\beta = 0.060$, $p < .001$; Supplementary Table S2). See Fig. 2 (raw pupil data available in Supplementary Fig. S1).

We then repeated the analysis with Medium as the reference level to look for differences between the Medium and Hard conditions. We found no significant difference between them (Hard $\beta = 0.0039$, $p < .77$; Supplementary Table S3).

With the subset of participants for whom we had accuracy data, we removed Easy and Medium trials and sum-coded Accuracy (inaccurate as -0.5, accurate as 0.5). We removed the time terms from this analysis so that the main effect of Accuracy would be interpretable. The results indicated a significant main effect of Accuracy ($\beta = 0.0083$, $p < .001$) on pupil dilation, with trials on which participants responded inaccurately showing slightly greater dilation ($m = 0.29$) than those on which they were accurate ($m = 0.28$; Supplementary Table S4). See Fig. 3 (raw pupil data in Supplementary Fig. S2).

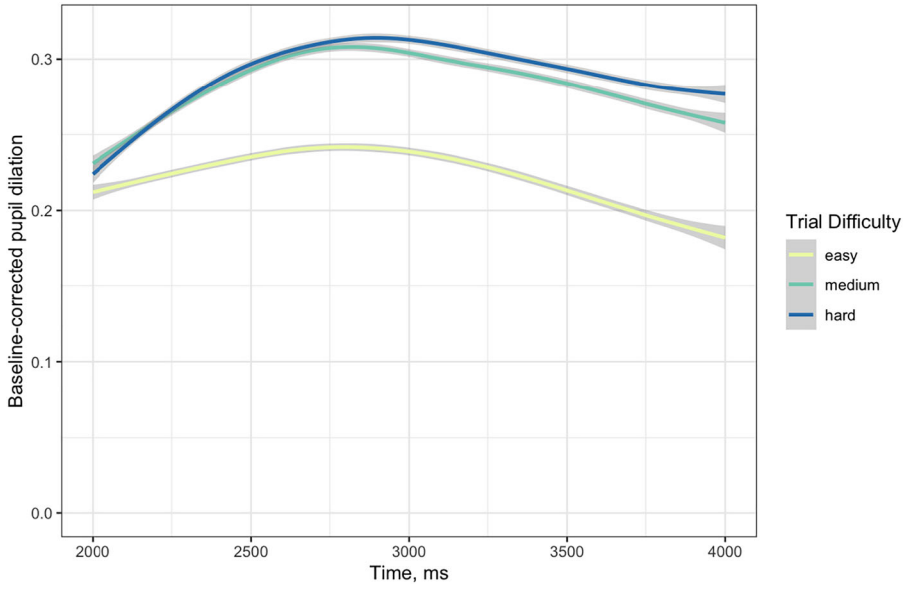

Fig. 2. Baseline-corrected pupil dilation from 2000 to 4000 ms for the 1-s condition by trial difficulty. Lines are loess-smoothed.

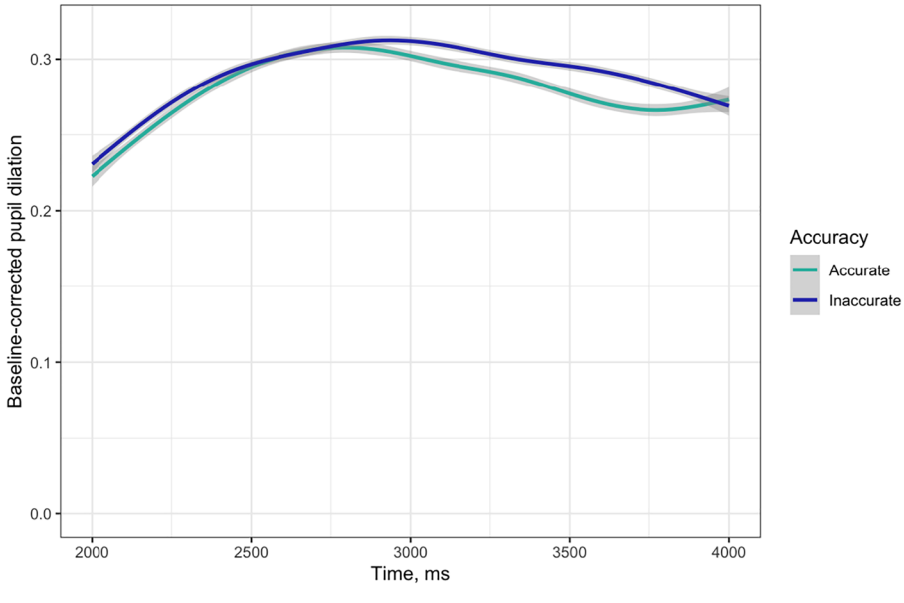

Fig. 3. Baseline-corrected pupil dilation from 2000 to 4000 ms for Hard trials in the 1-s condition by Accuracy of participant response. Lines are loess-smoothed.

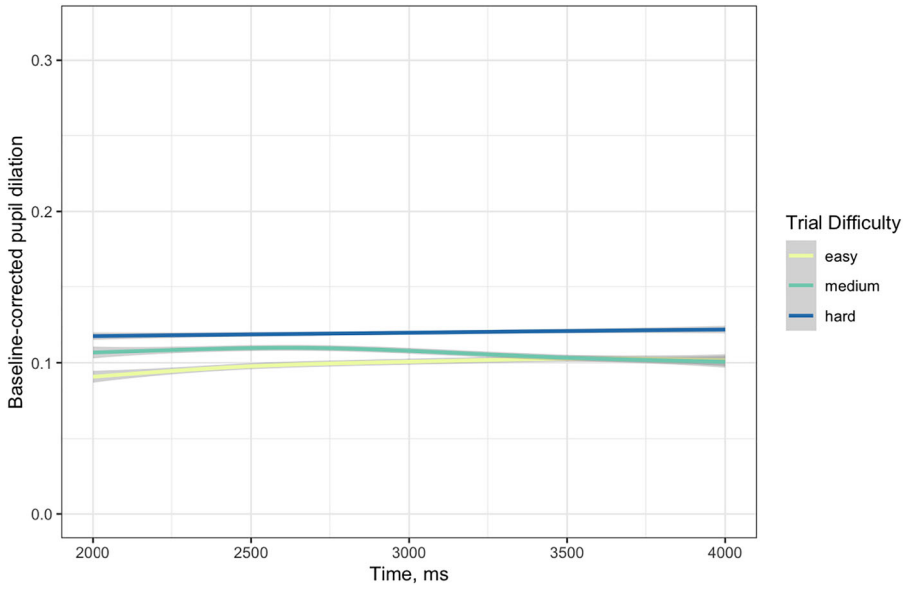

Fig. 4. Baseline-corrected pupil dilation from 2000 to 4000 ms for the 4-s condition (before the word was uttered) by trial difficulty. Lines are loess-smoothed.

3.3. 4-second condition: Pre-word comparison

We first analyzed the 4-s condition data as we did for the 1-s condition, despite that the word had not yet been heard during the 2000–4000 ms test window. Model comparison revealed that the model with no time terms as fixed effects was best fitting. The model with Easy as the reference level for Trial Difficulty revealed no significant main effects for either Medium trials ($\beta = 0.0046$, $p = .69$) or Hard trials ($\beta = 0.019$, $p = .10$; Supplementary Table S5). We again re-ran the model including brightness; there was no significant effect of brightness ($\beta = -0.028$, $p = .081$; Supplementary Table S6). The same pattern occurred with Medium as the reference level; there was no effect for Hard trials ($\beta = 0.014$, $p = .22$; Supplementary Table S7). See Fig. 4 (raw pupil data in Supplementary Fig. S3).

3.4. 4-second condition: Post-word comparison

We then repeated the analyses for the 4-s condition using baseline and test windows that reflected when the auditory stimuli occurred; we used a baseline of 2000–3999 ms and a test window of 5000–7000 ms. Model comparison revealed that, as for the 1-s condition, the best-fitting model included linear, quadratic, and cubic time terms. The model with Easy as the reference level for Trial Difficulty revealed significant main effects for both Medium trials ($\beta = 0.052$, $p < .001$) and Hard trials ($\beta = 0.063$, $p < .001$; Supplementary Table S8). We re-ran this model including brightness; there was no significant main effect of brightness ($\beta = -0.0017$, $p = .32$), and we continued to find significant main effects of Trial Difficulty (Medium $\beta = 0.047$, $p < .001$; Hard $\beta = 0.062$, $p < .001$; Supplementary Table S9). With Medium as the reference level, we found no significant difference for Hard trials ($\beta = 0.011$,

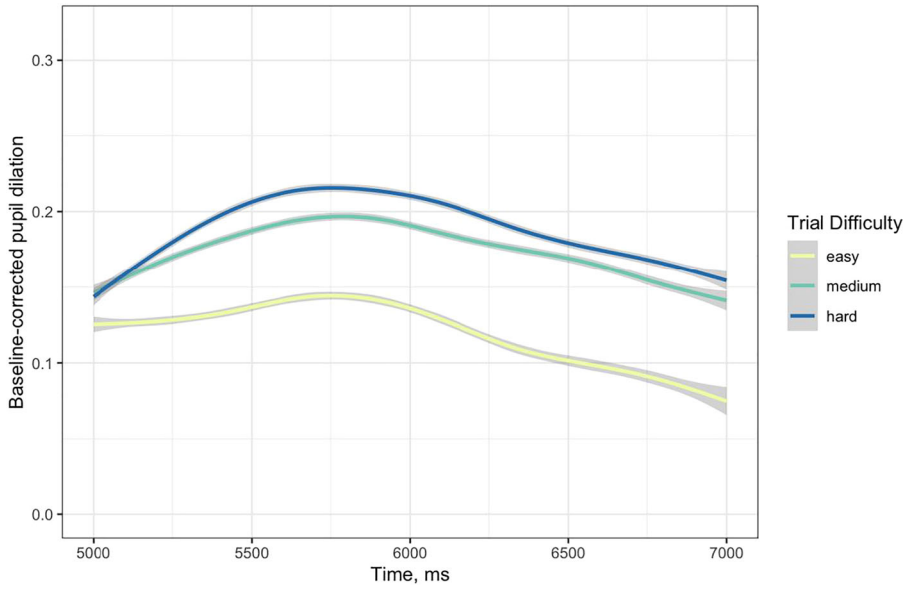

Fig. 5. Baseline-corrected pupil dilation from 5000 to 7000 ms for the 4-s condition (after the word was uttered) by trial difficulty. Lines are loess-smoothed.

$p = .35$; see Fig. 5; Supplementary Table S10 and raw pupil data in Supplementary Fig. S4). Factoring in accuracy on Hard trials for the subset of participants for whom we had accuracy data, we found a significant main effect of Accuracy ($\beta = 0.015$, $p < .001$; see Fig. 6 and raw pupil data in Supplementary Fig. S5), with greater pupil dilation on accurate trials ($m = 0.21$) than inaccurate trials ($m = 0.18$; Supplementary Table S11).

To further probe this unexpected result, that of greater pupil dilation on accurate trials than on inaccurate trials in the 4-s post-word comparison, we ran post hoc analyses with the Medium trials in the 4-s condition (we did not do so for the Easy trials because accuracy was almost 100%). Results revealed a nonsignificant effect of Accuracy ($p = .08$). However, there was a trend toward greater pupil dilation for accurate trials ($M = 0.18$) than for inaccurate trials ($M = 0.16$). A follow-up analysis on the Medium trials in the 1-s condition revealed that inaccurate trials ($M = 0.34$) had significantly greater pupil dilation ($p < .001$) than accurate trials ($M = 0.28$).

4. Conclusion

Our results show a clear effect of word difficulty on pupil dilation in a standard vocabulary assessment, even when we controlled for image properties including display timing and image brightness. Specifically, Easy trials elicited significantly smaller pupil dilation than Medium and Hard trials; there were no differences in dilation between Medium and Hard trials. Thus, we replicate and extend previous work showing similar findings (Chapman & Hallowell, 2015; Ledoux et al., 2016).

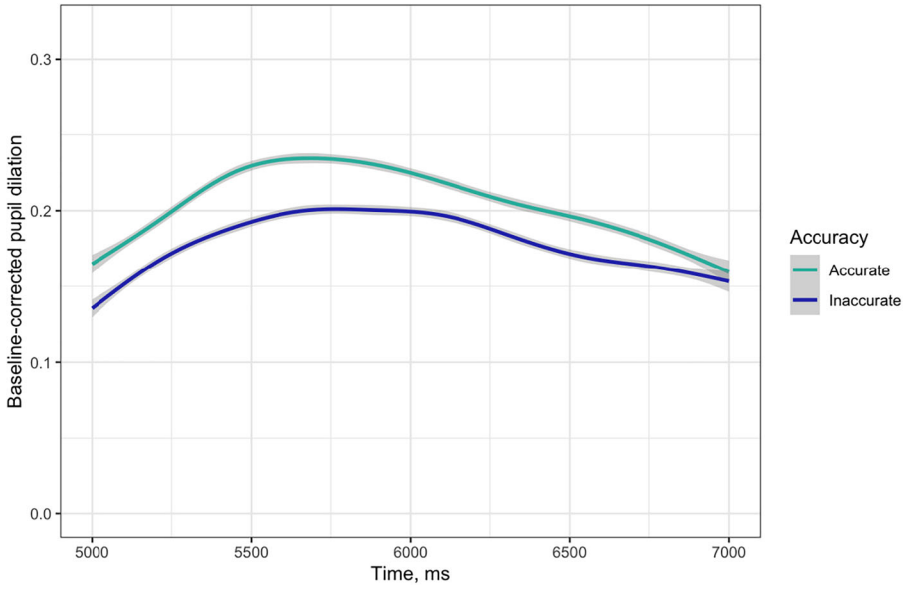

Fig. 6. Baseline-corrected pupil dilation from 5000 to 7000 ms for Hard trials in the 4-s condition by Accuracy of participant response. Lines are loess-smoothed.

We further examined accuracy on Hard trials only, on which participants were accurate slightly less than half the time, to evaluate whether participants displayed more or less effort on trials on which they succeeded. However, the results are not straightforward, because the 1-s and 4-s conditions show different patterns. In the 1-s condition, Hard trials resulted in slightly greater pupil dilation when participants were inaccurate than accurate, but in the 4-s condition, Hard trials resulted in greater pupil dilation when participants were accurate than inaccurate.

Why would we see a condition difference in the impact of accuracy on pupil dilation? One possibility is that in the 4-s condition, participants have a longer time to view the images prior to the onset of the word. They may have time to label all of the images and imagine what the upcoming word might be. This may impact their effort level: Specifically, participants may be more likely to recognize that they cannot recall a label that is consistent with the image displayed, and so reduce their effort, leading to less effort on trials on which they are inaccurate than trials on which they are accurate. This explanation is conceivable given that our post hoc analyses on the Medium trials yielded a similar pattern of results in the 4-s condition (i.e., greater pupil dilation on accurate trials) but not in the 1-s condition (i.e., greater pupil dilation on inaccurate trials). Thus, with additional time given, participants may have indeed reduced their effort when they could not identify the pictures and anticipate the target word.

A limitation of the current study is that we used participants' accuracy on each trial as a factor in the analyses, but with a one-in-four chance of guessing the correct answer, it could be that participants did not know some of the words for which they gave correct answers.

However, the opposite can also be true, wherein, participants knew the words but gave incorrect answers. This is particularly true in the Hard trials, where participants demonstrated low accuracy. It could be the case that although participants correctly knew the definition of the vocabulary item, they were unable to select the correct visual representation, thus resulting in an incorrect response. It is possible that as words became more difficult, their visual depictions got more challenging to identify. Additional work will be needed to explore associations between accuracy and effort and add additional measures beyond accuracy (e.g., participant confidence in their answer).

This line of research has the potential to contribute to enhanced vocabulary assessments. As suggested in Ledoux et al. (2016) and Coderre, Chernenok, Bosley, Gordon, and Ledoux (2019), we believe it is important to develop eye-gaze measures that can improve assessments of vocabulary. Eye-gaze assessments can both be more accessible, in offering a low task-demand way to indicate responses (as opposed to responding verbally or with a pointing gesture), as well as more insightful into cognitive processing (such as effort) in addition to simply indicating whether participants know the word. Moreover, such investigations could lead to more adaptive assessments (Sharma et al., 2020). For example, by taking into account how difficult an item was for a test-taker rather than only whether the response was accurate (which might be by chance), we may be able to terminate the assessment earlier, maximizing efficiency and avoiding frustration.

As a first step in these directions, the current study provides crucial foundational information by demonstrating that the difficulty of a vocabulary word predictably affects pupil dilation, with harder words eliciting greater cognitive effort. Moreover, the timing structure of the task has differential effects on the effort they expend on difficult words and in turn whether participants ultimately guess accurately or inaccurately.

Acknowledgments

All deidentified data and analysis code are available via the Open Science Framework at https://osf.io/64gv8/?view_only=fe09c8a4010b406d9e9e6f4b59a31e09. The authors would like to thank Ellinor Hull for her support during the initial stages of the project, and Sophia Aburida and Jourdan Parent for their assistance with data collection.

Conflict of Interest Statement

Author DF reports consulting fees from Click Therapeutics and Boehringer Ingelheim.

Notes

1 The duration of the baseline period used in prior studies varies widely, but we selected a period that has been used in prior work that also involved language processing tasks (e.g., McLaughlin et al., 2022) and that, crucially, does not overlap with the auditory stimulus.

2 Thanks to an anonymous reviewer comment, we repeated the analyses using a 500-ms baseline from 3500 to 3999 ms for the 4-s condition; the patterns were identical. See the Supplementary Materials (Tables S12–S15).

References

Ahern, S., & Beatty, J. (1979). Pupillary responses during information processing vary with Scholastic Aptitude Test scores. *Science*, *205*(4412), 1289–1292. https://doi.org/10.1126/science.472746

Alnæs, D., Sneve, M. H., Espeseth, T., Endestad, T., van de Pavert, S. H., & Laeng, B. (2014). Pupil size signals mental effort deployed during multiple object tracking and predicts brain activity in the dorsal attention network and the locus coeruleus. *Journal of Vision*, *14*(4), 1. https://doi.org/10.1167/14.4.1

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, *91*(2), 276–292. https://doi.org/10.1037/0033-2909.91.2.276

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. http://doi.org/10.1111/j.1469-8986.2008.00654.x

Chapman, L. R., & Hallowell, B. (2015). A novel pupillometric method for indexing word difficulty in individuals with and without aphasia. *Journal of Speech, Language, and Hearing Research: JSLHR*, *58*(5), 1508–1520. https://doi.org/10.1044/2015_JSLHR-L-14-0287

Coderre, E. L., Chernenok, M., Bosley, L., Gordon, B., & Ledoux, K. (2019). Implicit measures of receptive vocabulary knowledge in individuals with level 3 autism. *Cognitive and Behavioral Neurology*, *32*(2), 95–119. https://doi.org/10.1097/WNN.0000000000000194

Demberg, V., & Sayeed, A. (2016). The frequency of rapid pupil dilations as a measure of linguistic processing difficulty. *PLOS ONE*, *11*(1), e0146194. https://doi.org/10.1371/journal.pone.0146194

Dunn, L. M., & Dunn, D. M. (2007). *Peabody Picture Vocabulary Test* (4th ed.). San Antonio, TX: Pearson/PsychCorp.

Ledoux, K., Coderre, E., Bosley, L., Buz, E., Gangopadhyay, I., & Gordon, B. (2016). The concurrent use of three implicit measures (eye movements, pupillometry, and event-related potentials) to assess receptive vocabulary knowledge in normal adults. *Behavior Research Methods*, *48*(1), 285–305. https://doi.org/10.3758/s13428-015-0571-6

Loewenfeld, I. E. (1993). *The pupil: Anatomy, Physiology, and clinical applications*. Ames, IA: Iowa State University Press.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

McLaughlin, D. J., & Van Engen, K. J. (2020). Task-evoked pupil response for accurately recognized accented speech. *The Journal of the Acoustical Society of America*, *147*(2), EL151–EL156.

McLaughlin, D. J., Zink, M. E., Gaunt, L., Spehar, B., Van Engen, K. J., Sommers, M. S., & Peelle, J. E. (2022). Pupillometry reveals cognitive demands of lexical competition during spoken word recognition in young and older adults. *Psychonomic Bulletin & Review*, *29*(1), 268–280. https://doi.org/10.3758/s13423-021-01991-0

Mirman, D. (2017). *Growth curve analysis and visualization using R*. Chapman and Hall/CRC.

Partala, T., & Surakka, V. (2003). Pupil size variation as an indication of affective processing. *International Journal of Human-Computer Studies*, *59*(1-2), 185–198. http://doi.org/10.1016/S1071-5819(03)00017-X

R Core Team. (2022). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Sharma, K., Papamitsiou, Z., Olsen, J. K., & Giannakos, M. (2020). Predicting learners' effortful behaviour in adaptive assessment using multimodal data. Proceedings of the Tenth International Conference on Learning Analytics & Knowledge, Frankfurt, Germany (pp. 480–489). https://doi.org/10.1145/3375462.3375498

Sirois, S., & Brisson, J. (2014). Pupillometry. *Wiley Interdisciplinary Reviews*: *Cognitive Science*, *5*(6), 679–692. https://doi.org/10.1002/wcs.1323

Tromp, J., Hagoort, P., & Meyer, A. (2016). Pupillometry reveals increased pupil size during indirect request comprehension. *The Quarterly Journal of Experimental Psychology*, *69*(6), 1093–1108. https://doi.org/10.1080/17470218.2015.1065282

van der Wel, P., & van Steenbergen, H. (2018). Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic Bulletin & Review*, *25*(6), 2005–2015. https://doi.org/10.3758/s13423-018-1432-y

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2011). Cognitive load during speech perception in noise: The influence of age, hearing loss, and cognition on the pupil response. *Ear and Hearing*, *32*(4), 498–510. https://doi.org/10.1097/AUD.0b013e31820512bb

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supporting Information
Supporting Information