Research Article

Psychological Science
2021, Vol. 32(1) 80–95


Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0956797620958638
www.psychologicalscience.org/PS


Keeping an Eye on Effort: A Pupillometric Investigation of Effort and Effortlessness in Visual Word Recognition

Adi Shechter[1,2] and David L. Share[1,2]
[1]Department of Learning Disabilities, Faculty of Education, University of Haifa, and
[2]Edmond J. Safra Brain Research Center for the Study of Learning Disabilities, University of Haifa

Abstract
Rapid and seemingly effortless word recognition is a virtually unquestioned characteristic of skilled reading, yet the definition and operationalization of the concept of cognitive effort have proven elusive. We investigated the cognitive effort involved in oral and silent word reading using pupillometry among adults (Experiment 1, $N = 30$; Experiment 2, $N = 20$) and fourth through sixth graders (Experiment 3, $N = 30$; Experiment 4, $N = 18$). We compared multiple pupillary measures (mean, peak, and peak latency) for reading familiar words (real words) and unfamiliar letter strings (pseudowords) varying in length. Converging with the behavioral data for accuracy and response times, pupillary responses demonstrated a greater degree of cognitive effort for pseudowords compared with real words and stronger length effects for pseudowords than for real words. These findings open up new possibilities for studying the issue of effort and effortlessness in the field of word recognition and other fields of skill learning.

Keywords
effort, pupillometry, skill learning, reading, word recognition

Received 4/6/19; Revision accepted 6/25/20

One of the most distinctive characteristics of skilled reading is the sheer speed and apparent effortlessness of word recognition. Among reading researchers, there is a broad consensus that fast, near-effortless recognition of printed words (often termed *automatic* or *fluent* word reading) is crucial to successful reading development because it enables the reader to devote limited processing resources to comprehension (LaBerge & Samuels, 1974; Perfetti, 1985). For the skilled reader, the ease and speed of word recognition makes it possible for reading to become not only a means for educational and social-economic advancement but also a source of joy, as when curling up with a thrilling book. By contrast, the disabled reader's word reading is characteristically slow, error prone, and labored—a major deterrent to literacy development. Acknowledging the importance of effort in word reading, the fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association, 2013) defines *specific learning disorder* in reading as "inaccurate or slow and *effortful* [emphasis added] word reading" (p. 66).

The measurement of word-reading accuracy and speed is relatively straightforward and has consequently become universal practice in the assessment of reading skill. On the other hand, researchers have yet to reach a consensus on the definition and operationalization of the concept of effort, underscoring the ambiguity of broader constructs such as fluency and automaticity (Logan, 1997; Megherbi, Elbro, Oakhill, Segui, & New, 2018; Moors & De Houwer, 2006; Reynolds & Besner, 2006; Share, 2008; Stanovich, 1990), of which effortlessness is a core property (Kuhn, Schwanenflugel, & Meisinger, 2010; Logan, 1997).

Logan (1997) described effortlessness as the ability to perform a task with a sense of ease while performing other tasks concurrently. Although Logan's definition has been adopted by many reading researchers (e.g., Kuhn et al., 2010), the operationalization of effortlessness has generated a variety of controversial techniques such as

Corresponding Author:
David L. Share, University of Haifa, Edmond J. Safra Brain Research Center for the Study of Learning Disabilities
E-mail: dshare@edu.haifa.ac.il

the dual-task paradigm (Pashler, 1994) and the Stroop task (Labuschagne & Besner, 2015; Megherbi et al., 2018). These techniques have proven useful in capturing the outcomes of learning. However, they do not provide online, moment-to-moment insight into the dynamics of word recognition that is at the heart of current item-based models of printed-word learning (Ehri, 2014; Logan, 1988; Perry, Zorzi, & Ziegler, 2019; Share, 1995), which emphasize the micro changes in the reading process that occur as individual words are encountered during reading.

This lacuna casts a shadow over all research concerned with the development of word-reading skill, efficiency, fluency, and automaticity. In the present investigation, we took a first step toward redressing this situation by exploring the use of pupil dilation to study the critical yet neglected issue of cognitive effort in word recognition.

Statement of Relevance

The hallmark of expertise in many skill domains is the speed and apparent effortlessness of task execution. Yet mastering a skill typically starts out with slow, effortful, unskilled performance, gradually shifting with practice toward expert levels. In the case of reading, novices start off reading individual words, often letter by letter, whereas for skilled readers, reading is fast and even automatic. These characterizations come from behavioral measures of accuracy and speed, often labeled *fluency*. By contrast, direct measurement of the effort involved in reading has been largely neglected. In this investigation, we examined the effort involved in word recognition by analyzing changes in pupil dilation among skilled adult readers and elementary school children. We found that readers in each age group invested more cognitive effort in reading unfamiliar compared with familiar words, in both oral and silent reading. This approach to quantifying effort opens up new possibilities for studying allocation of effort in a range of domains of skill learning.

Pupil Dilation as a Measure of Cognitive Effort

The connection between pupil dilation and cognitive activity was noted more than 100 years ago (e.g., Mentz, 1895, cited in Kahneman, Tursky, Shapiro, & Crider, 1969). However, the revival of pupillometry as a measure of cognitive effort occurred only in the 1960s and 1970s (Beatty, 1982; Kahneman, 1973; Kahneman et al., 1969). Since then, pupil dilation has proven to be a sensitive and reliable measure of cognitive effort in a variety of domains including language, memory, decision-making, emotion, and cognitive development (Sirois & Brisson, 2014; van der Wel & van Steenbergen, 2018). While a cognitive task is performed, mental effort arouses the sympathetic system and, correspondingly, pupil diameter increases (Eckstein, Guerra-Carrillo, Singley, & Bunge, 2017). Kahneman (1973) argued that pupillometry is "the best single index" (p. 18) of effort because it captures within-task, between-task, and between-individual variation (for a review, see Beatty, 1982). Surprisingly, pupillometry has been conspicuously absent in reading research. We were able to locate only a handful of pupillometric studies of reading over the past half century since cognitive scientists rediscovered pupillometry.

Carver (1971) was probably the first to study the connection between reading and pupil size. In a study of text-reading difficulty among undergraduates, Carver found no evidence of variation in pupil dilation across difficulty levels. However, he recorded pupil size at a small number of randomly varying text locations, and the identity of specific words fixated was left uncontrolled. Moreover, by using different locations, Carver did not control for changes in gaze angle, thereby overlooking the foreshortening effect that causes the recorded pupil size to diminish as a result of rotation of the eye (Hayes & Petrov, 2016). Since Carver's work, and perhaps owing to his disappointing results, only a handful of studies have used pupillometry in reading research, typically focusing on sentence reading among highly skilled readers (e.g., Fernández, Biondi, Castro, & Agamenonni, 2016; Just & Carpenter, 1993).

Among the few studies of word recognition that used pupillometry, Kuchinke, Võ, Hofmann, and Jacobs (2007) found that peak pupil dilation in a lexical decision task was higher for low-frequency words compared with high-frequency words. Another study, conducted by Fernández et al. (2016), examined sentence processing but also looked at word length, predictability, and frequency. The results showed that mean pupil dilation was larger for longer words and smaller for more frequent and predictable words. In another study looking at the processing of single words, Mathôt, Grainger, and Strijkers (2017) found that spoken and written words conveying a sense of darkness (e.g., *night*) evoked larger pupil dilation compared with words conveying a sense of brightness (e.g., *day*). Although scarce, these studies collectively suggest that pupillometry is sensitive to the characteristics of single words, at least among skilled adult readers. To our knowledge, no study has yet addressed the issue of effort in children's word reading.

We report four experiments examining the question of effort in word reading through the lens of pupillometry among both skilled adult readers and elementary school

children. Because we were interested in reading in general, for each age group, we examined the question of cognitive effort in both oral and silent word reading.

The theoretical framework for the present study is the unfamiliar-to-familiar/novice-to-expert developmental framework outlined by Share (2008). This theory posits a fundamental and universal within-item developmental transition from unfamiliar to familiar (Share, 2008). Because every printed word is, at one point, unfamiliar, the reader must possess some means of independently deciphering novel words and morphemes. The need to identify unfamiliar printed words is crucial for the novice and expert reader alike, because a majority of words have very low frequencies and are rarely encountered in print. On the other hand, the reader must eventually be able to achieve a high degree of unitization, or "chunking," of letter strings to enable the rapid, near-effortless recognition of familiar words and morphemes via direct memory retrieval (LaBerge & Samuels, 1974; Logan, 1988, 1997; Perfetti, 1985; Perry et al., 2019).

Because our investigation ventured into largely uncharted waters, we took a measured step-by-step approach by first asking whether differences between reading familiar words (real words) and unfamiliar letter strings (pseudowords) are reflected in cognitive effort as measured by changes in pupil size. In each experiment, we predicted that pseudowords would require significantly more effort to read than real words, as indicated by greater overall pupil dilation, higher maximum (peak) dilation, and longer latencies to peak dilation. We also included the standard behavioral measures of pronunciation accuracy and response latencies, anticipating slower responses and lower accuracy for pseudowords. In addition, we examined the length effect, which is widely regarded as reflecting the serial letter-by-letter processing typical of pseudowords. We predicted a familiarity-by-length interaction, in which length effects on both behavioral and pupillometric measures would be greater for pseudowords than real words, as reported previously for response times by Weekes (1997).

Experiment 1: Oral Reading (University Students)

Method

Participants. Because it was not possible to rely on prior research to determine the required sample size, we conducted a power analysis (using G*Power Version 3; Faul, Erdfelder, Lang, & Buchner, 2007) with power set at .80, a relatively conservative alpha of .01, and an intermediate effect size (f) of .25. The analysis indicated that a sample size of 33 participants was necessary. Participants were 34 students from the University of Haifa who had no reported past or present reading difficulties or attention deficits, whose mother tongue is Hebrew, and who had normal or corrected-to-normal vision. Four participants were excluded because they did not provide a minimum of 20 valid responses in each of the four conditions (i.e., at least 50% correct responses with no more than 20% missing pupil data). Considering the large effect sizes of our observed findings, we saw no reason to run additional participants; hence, the final sample contained 30 participants (24 female; age: M = 27.5 years, SD = 5.85) who had 9.3% missing trials on average. Each student signed an informed consent form prior to the experiment and received course credit or a monetary payment of 40 shekels (around $11) for participation.[1]

Design. The experiment had a fully within-subjects 2 × 2 design with two levels of familiarity (unfamiliar letter strings [pseudowords] vs. familiar real words) and two lengths (three letters vs. five letters). Each of the four conditions contained 40 random items (i.e., 160 target stimuli). The inclusion of an additional 80 fillers yielded a total of 240 trials. These were divided into four blocks, each block containing 20 pseudowords (10 of each length), 20 real words (10 of each length), and 20 fillers. Each stimulus appeared only once during the experiment. Yoked pairs of target stimuli (a real word and its matched pseudoword) were separated by an intervening block (Blocks 1 and 3 or 2 and 4).

Stimuli.

Target stimuli. The examination of reading by pupillometry poses a number of challenges regarding potential confounds with luminance because the pupil's response to light is larger than the pupil's cognitive response (Granholm & Steinhauer, 2004). We maintained identical luminance levels across conditions by creating yoked pairs of target stimuli (i.e., pairs of real words and pseudowords). We first compiled a list of common real words. For each length (three and five letters), 75 pointed (fully vocalized) words were selected from two academic frequency-based word lists in Hebrew (Balgur, 1968; Mahelman, Rozen, & Shaked, 1960). Our aim was to include words that would be familiar not only to adults but also to children. Only high-frequency words were included, covering various parts of speech. The available corpuses (Balgur, 1968; Mahelman et al., 1960), however, have two shortcomings: They are old and possibly outdated; furthermore, they may not reflect printed word frequencies. To validate the frequency of candidate items, we asked 17 teachers currently teaching fourth- to sixth-grade classes to respond to an online questionnaire containing two separate lists of words: three-letter words and five-letter words. Each

list contained 100 words: the 75 high-frequency candidate words and another 25 rare words (i.e., low-frequency words, selected from the lists of Mahelman et al., 1960, and Balgur, 1968). Using a five-point Likert-type scale, teachers were asked to evaluate, "How many times a student in 4th-6th grade would have seen the printed word?" Response options were *not at all* (1), *several times* (2), *dozens of times* (3), *hundreds of times* (4), and *thousands of times* (5). The mean frequency rating was then calculated for each item.

Next, for each candidate (high-frequency) real word, we created a matched pseudoword by scrambling the letters while preserving the vowel diacritics (e.g., שֶׁלֶג [ʃɛlɛg], the word for *snow*, became לֶשֶׁג [lɛʃɛg]). Because Hebrew contains five letters that have a word-final form (ך, ם, ן, ף, ץ), their position was preserved as well. Pairs were dropped from the final list if all possible combinations of letters in the real word produced only real words or an illegal morphophonological pattern (e.g., the word עִיר [ʔiʁ], the word for *city*, could not be successfully scrambled). The final list contained 40 word–pseudoword pairs. There was no significant difference between the familiarity ratings of the three-letter target (real) words (M = 4.15, SD = 0.28) and the five-letter words (M = 4.25, SD = 0.20), $t(39) = -1.71$, $p = .10$. These target stimuli subtended a visual angle of 1.11° to 1.61° for height and 2.21° to 4.82° for width from a viewing distance of 57 cm.

Filler words. Twenty filler words, representing a variety of parts of speech and length (i.e., two to eight letters), were added to each block to provide a more ecologically valid range of word frequencies and minimize possible strategic artifacts that can arise when the set of experimental stimuli includes a large proportion of pseudoword stimuli. From the viewing distance of 57 cm, these stimuli subtended a visual angle of 1.00° to 1.61° for height and 1.31° to 5.82° for width. All stimuli were centered and presented in white text (RGB values = 255, 255, 255) on a gray background (RGB values = 128, 128, 128).

Procedure. The data were collected in a dimly illuminated sound-reduced room at the Edmond J. Safra Brain Research Center for the Study of Learning Disabilities at the University of Haifa. Participants were asked to read aloud all letter strings (words, pseudowords, and fillers), which were presented one at a time on a computer screen. Each block began with an instruction screen, and the participant was asked to read the displayed word aloud. Participants were informed that the printed word would disappear automatically. Two practice trials were then presented. After calibration and validation, a drift correction was displayed and the block began.

Figure 1 illustrates the procedure. Each trial commenced with a central fixation cross presented for 500 ms, followed by a gray fixation screen. The fixation screen presented a string of Xs—the same as the number of characters in the upcoming string—to avoid luminance confounding. This fixation screen appeared for 1,000 ms and was followed immediately by the stimulus word. Because pupil-size changes are characterized by much slower responses than typical behavioral measures such as reaction times (Partala & Surakka, 2003), stimuli remained on the screen for 3,300 ms. The trial ended with a blank screen displayed for 1,500 ms. Pronunciation onset latencies were recorded by a voice key. Each pronunciation was also audio-recorded. During the task, however, errors were manually documented by the tester, who sat behind the participant in front of the host computer.

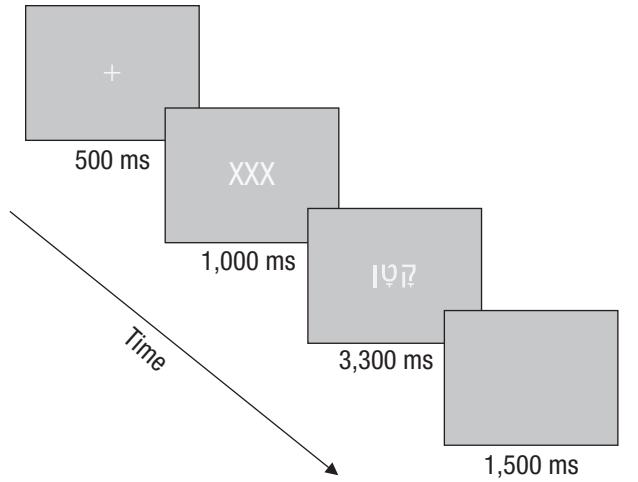

Fig. 1. Example trial sequence in Experiment 1. After viewing a string of Xs (presented to avoid luminance confounding), participants saw a stimulus consisting of a real word or a pseudoword, which they were asked to read aloud.

Apparatus. Pupillometry data were recorded with an EyeLink 1000 Plus (SR Research, Kanata, Ontario, Canada), a video-based eye tracker with a sampling rate of 1,000 Hz. The experimental materials were presented using the EyeLink's Experiment Builder software. Participants wore headphones (HS-11V stereo headphones with microphone, SilverLine, China), placed their chin on a chin rest, and adjusted the microphone to their mouth. Next, participants were asked to pronounce a sample word (the word שָׁלוֹם [ʃɑlom], which means *hello*) to ensure there was sufficient freedom for opening the jaw. Participants' eyes were 57 cm away from a 24-in. LCD monitor (XL24II monitor, BenQ, Taipei, Taiwan; Quadro K620 graphics card, NVIDIA, Santa Clara, CA) with 1,024- × 768-pixel resolution and a refresh rate of 60 Hz. The threshold level for the voice key was defined as 0.1 audio level. To ensure reliable pupil-size data, we preceded each block with calibration and validation. In addition, the display

stimuli were located in the center of the screen. Participants were instructed to look at the center of the screen during the entire session without shifting their gaze position. In addition, to avoid extreme luminance changes, we presented the same white text (RGB values = 255, 255, 255) on a gray background (RGB values = 128, 128, 128) on all screens.

Statistical analysis.

Pupil-data analysis. We used the divisive baseline-correction method (percentage of relative change = 100 × pupil size/baseline) for analyzing changes in pupil size (e.g., Binda, Pereverzeva, & Murray, 2014). Raw pupil data were analyzed using CHAP software (Hershman, Henik, & Cohen, 2019). For each trial, z scores were first calculated, and then trials in which z scores exceeded 2.5 were classified as outliers and omitted from further analyses. Next, for each participant, we excluded trials with 20% of missing pupil values and above. Responses to filler words as well as incorrect and missing responses were excluded from the analysis. In addition, the yoked item for any deleted item was also deleted from the analyses to maintain identical luminance levels across conditions. Eyeblinks were detected using Hershman, Henik, and Cohen's (2018) algorithm, and missing values were replaced with a linear interpolation. Time courses were aligned with stimulus onset and divided by the baseline (the average pupil size 200 ms before stimulus onset).

Rather than rely on a single dependent measure, we included several common parameters of pupillary responses with the aim of obtaining converging findings across multiple measures. Consequently, we used average pupil dilation as an overall index of the amount of cognitive effort invested in reading each item (Kahneman et al., 1969), peak (maximum) pupil dilation as an indicator of the maximum invested effort, and peak latency (the time elapsed from stimulus onset to peak dilation) as a reflection of processing speed (Zekveld, Kramer, & Festen, 2011).

Response time analyses. Only when both members of yoked pairs were pronounced correctly were their naming latencies included in the analysis. Response times greater than 2 standard deviations above or below the participant mean were excluded. Finally, for each of the four experimental conditions, response times were averaged within participants.

Accuracy analyses. For each participant, we calculated the percentage of target stimuli pronounced correctly in each of the four experimental conditions.

Results

Pupil dilation. Pupillary data were submitted to a 2 (familiar vs. unfamiliar) × 2 (three letters vs. five letters) repeated measures analysis of variance (ANOVA) using a time window from stimulus onset to 4,300 ms (1,000 ms after stimulus offset). Figure 2 displays the average proportional changes of pupillary responses in the four conditions in Experiment 1.

Mean relative changes in pupil size. The mean relative changes in pupil size over this time window revealed a significant main effect for word familiarity, $F(1, 29) = 47.81$, $p < .001$, $\eta_p^2 = .62$, observed power (OP) = 1.00. As predicted, relative changes in pupil size were significantly larger for pseudowords (M = 11.07%, SD = 7.83) than for real words (M = 6.38%, SD = 5.31), consistent with the hypothesis that a greater degree of cognitive effort is involved in reading unfamiliar letter strings. There was also a main effect for length, $F(1, 29) = 24.82$, $p < .001$, $\eta_p^2 = .46$, OP = 1.00, indicating larger dilation for five-letter strings (M = 10.06%, SD = 6.74) than for three-letter strings (M = 7.39%, SD = 6.46). That is, as measured by pupil size, more cognitive effort appears to be invested in pronouncing longer letter strings. As predicted, we also found a significant interaction between familiarity and length, $F(1, 29) = 13.38$, $p = .001$, $\eta_p^2 = .32$, OP = .94. Follow-up t tests revealed a significant length effect for pseudowords, with larger relative changes for five-letter strings (M = 13.17%, SD = 8.68) compared with three-letter strings (M = 8.97%, SD = 7.52), $t(29) = -5.19$, $p < .001$. A (smaller) length effect was also observed for real words, with larger relative changes for five-letter strings (M = 6.95%, SD = 5.06) than for three-letter strings (M = 5.81%, SD = 5.90), $t(29) = -2.20$, $p = .04$.

Peak dilation. A two-way repeated measures ANOVA of peak pupil dilation also revealed significant main effects for both word familiarity, $F(1, 29) = 58.80$, $p < .001$, $\eta_p^2 = .67$, OP = 1.00 (pseudowords: M = 20.61%, SD = 11.28; real words: M = 13.87%, SD = 7.78), and length, $F(1, 29) = 24.22$, $p < .001$, $\eta_p^2 = .46$, OP = 1.00 (five-letter strings: M = 19.02%, SD = 10.08; three-letter strings: M = 15.46%, SD = 9.09). On this measure, too, the predicted interaction between familiarity and length was significant, $F(1, 29) = 12.26$, $p = .002$, $\eta_p^2 = .30$, OP = .92, confirming higher peaks for five-letter pseudowords (M = 23.28%, SD = 12.41) compared with three-letter pseudowords (M = 17.94%, SD = 10.81), $t(29) = -5.10$, $p < .001$, as well as higher peaks for five-letter real words (M = 14.76%, SD = 7.99) compared with three-letter real words (M = 12.98%, SD = 8.05), $t(29) = -2.61$, $p = .01$.

Latency to peak dilation. The third pupillary measure, latency to peak dilation, also revealed a significant main effect for word familiarity, $F(1, 29) = 29.51$, $p < .001$, $\eta_p^2 = .50$, OP = 1.00 (pseudowords: M = 2,399 ms, SD = 362; real words: M = 2,004 ms, SD = 409), but not for length, $F(1, 29) = 1.33$, $p = .26$, $\eta_p^2 = .04$, OP = .20 (five-letter strings: M = 2,282 ms, SD = 464; three-letter strings: M = 2,122 ms,

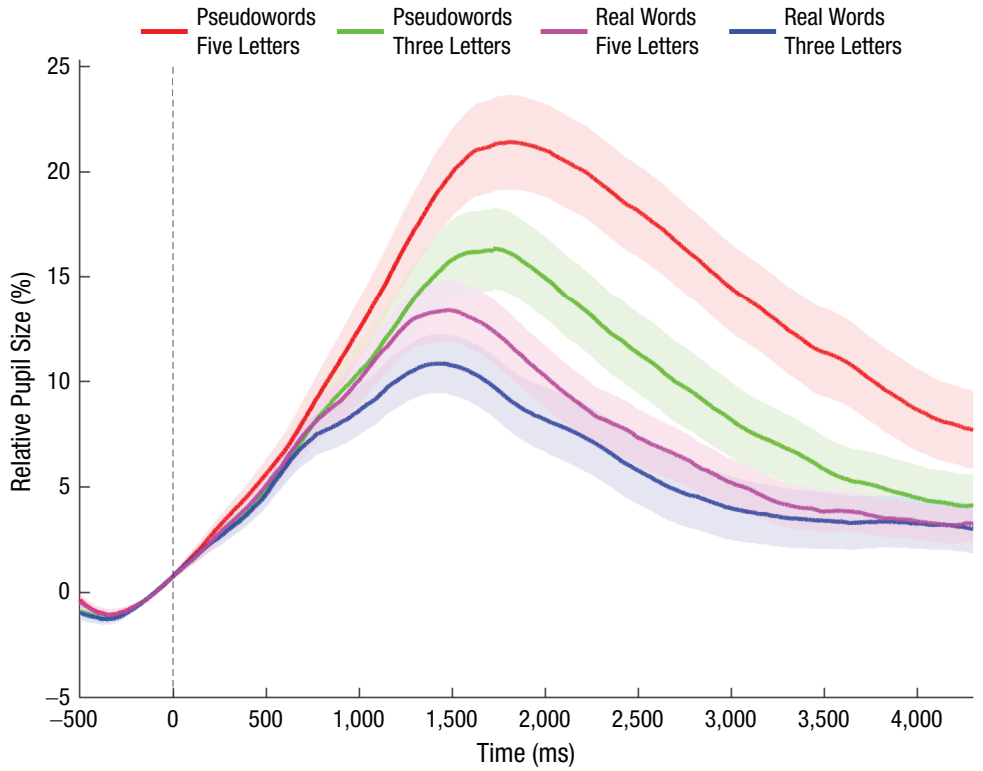

Fig. 2. Relative changes in pupil size for the four conditions in Experiment 1, from stimulus onset (Time 0, the dashed vertical line) to 1,000 ms after stimulus offset. The shaded areas depict standard errors of the mean.

SD = 541). The interaction between familiarity and length, although in the predicted direction, failed to reach significance, $F(1, 29) = 2.35$, $p = .14$, $\eta_p^2 = .08$, OP = .32 (pseudowords—three letters: *M* = 2,276 ms, *SD* = 501, five letters: *M* = 2,522 ms, *SD* = 505; real words—three letters: *M* = 1,967 ms, *SD* = 695, five letters: *M* = 2,042 ms, *SD* = 528).[2]

Behavioral data. Both times and accuracy for yoked pairs pronounced correctly were analyzed with a two-way repeated measures ANOVA with familiarity (familiar vs. unfamiliar) and length (three letters vs. five letters) as within-subjects factors.

Pronunciation onset latencies. For pronunciation onset latencies, the main effect for word familiarity was significant, $F(1, 29) = 108.70$, $p < .001$, $\eta_p^2 = .79$, OP = 1.00, with faster responses for real words than for pseudowords (Table 1). A main effect for length was also observed, $F(1, 29) = 27.83$, $p < .001$, $\eta_p^2 = .49$, OP = 1.00, indicating faster responses for three-letter strings compared with five-letter strings. The interaction between familiarity and length was also significant, $F(1, 29) = 28.83$, $p < .001$, $\eta_p^2 = .50$, OP = 1.00. A significant length effect was present for pseudowords, with faster response times for three-letter strings compared with five-letter strings, $t(29) = -6.63$, $p < .001$, but no length effect was found for real words, $t(29) = -0.93$, $p = .36$.

Pronunciation accuracy. For pronunciation accuracy, a main effect for familiarity was observed, $F(1, 29) = 40.27$, $p < .001$, $\eta_p^2 = .58$, OP = 1.00, with superior accuracy for real words compared with pseudowords. A main effect for length was also evident, $F(1, 29) = 12.04$, $p = .002$, $\eta_p^2 = .29$, OP = .92, with greater accuracy for three-letter strings compared with five-letter strings. We also found a significant interaction between familiarity and length, $F(1, 29) = 19.24$, $p < .001$, $\eta_p^2 = .40$, OP = .99. For real words, accuracy was close to ceiling levels of performance for both lengths, $t(29) = -2.04$, $p = .05$. For pseudowords, accuracy was significantly higher for the shorter items, $t(29) = 4.06$, $p < .001$ (Table 1).

Summary

In summary, our results clearly showed that pupillary responses are indeed sensitive to the familiarity and length of individual letter strings. As anticipated, unfamiliar letter strings appear to require greater cognitive effort, as indicated by multiple measures of pupil dilation. Both overall pupil-size changes and peak-dilation

Table 1. Mean Pronunciation Onset Latencies and Accuracy in Oral Word Reading Among University Students ($N = 30$)

	Onset latency (ms)			Accuracy (%)		
Word type	Three letters	Five letters	*M*	Three letters	Five letters	*M*
Pseudowords	1,013 (207)	1,119 (245)	1,066 (222)	95.8 (4.0)	91.3 (8.3)	93.5 (5.8)
Real words	829 (134)	840 (150)	834 (138)	99.3 (1.5)	99.9 (0.5)	99.6 (0.8)
M	921 (165)	980 (189)		97.5 (2.6)	95.6 (4.3)	

Note: Standard deviations are given in parentheses.

analyses (but not peak latency) confirmed greater length effects for pseudowords compared with real words, indicating that reading longer letter strings, especially longer pseudowords, demands additional mental effort. Furthermore, the pupillary data were largely in accordance with the behavioral predictions regarding lower accuracy and slower pronunciation latencies for pseudowords, longer strings, and their interaction. Interestingly, the attenuated length effect for real words was statistically significant on both mean dilation and peak dilation but not significant on pronunciation accuracy and speed, hinting that pupillometric measures may be more sensitive to word-level variables than traditional speed and accuracy measures.

To ensure that our findings were not simply task specific, resulting perhaps from pupillary responses for speech output (i.e., articulatory demands), we conducted a follow-up study (Experiment 2) in which we traced pupillary responses during silent word reading. This study used a novel variant of the delayed naming task that we call the "silent-then-oral-reading" procedure.

Experiment 2: Silent-Then-Oral Reading (University Students)

Method

Participants. Twenty-three students from the University of Haifa participated in this experiment. On the basis of an a priori power analysis (G*Power) using the reported effect size from Experiment 1, we estimated that a required sample size of 18 participants would be necessary to achieve an effect size (f) of .35 (power = .80, $\alpha = .01$). We collected data from five more participants in anticipation of possible dropouts or equipment failure. All observers were native Hebrew speakers with no reported past or present reading difficulties or attention deficits and with normal or corrected-to-normal visual acuity. After the exclusion of three participants, the final sample contained 20 participants (16 female; age: M = 26.5 years, SD = 4.76) who had 5.9% missing trials on average. Of the excluded participants, one did not reach the criterion of 20 valid trials per condition (i.e., 50% correct responses with no more than 20% missing pupil data). Another participant reported feeling unwell and coughed frequently during data recording, and an additional participant was omitted because of equipment failure. Written informed consent was obtained from all participants before the experiment, and each student received course credit or a monetary payment of 40 shekels (around $11) for participation.

Design and procedure. The design, stimuli, and apparatus were the same as in Experiment 1. The procedure was similar to that in Experiment 1, with some exceptions. Each trial commenced, as in Experiment 1, with a central cross presented for 500 ms, followed by a gray fixation screen with a string of Xs for 1,000 ms, followed immediately by the stimulus item. Here, instead of reading aloud, participants were asked to read the displayed stimulus silently and press a response button after completing a single reading. The stimulus disappeared when the key was pressed or 4,000 ms after stimulus onset in the case of a missing response. Next, a blank screen was presented for 1,500 ms (following Hershman & Henik, 2019). This was followed by the simultaneous presentation of a 300-ms auditory tone (beep) and the reappearance of the letter string. At this point, participants were asked to read the stimulus aloud to allow the tester to document reading accuracy, on the assumption that the accuracy of oral reading at the second appearance of the stimulus would, in the vast majority of cases, reflect the accuracy of the immediately preceding silent reading. The trial ended with a blank screen displayed for 1,500 ms. Figure 3 illustrates the procedure.

Statistical analysis. Pupil-data analyses included only yoked pairs of target stimuli that were pronounced correctly. We omitted responses to filler words as well as incorrect responses. Because we were interested in silent word reading, we excluded items that were pronounced aloud inaccurately as well as items with missing response times for key presses indicating completion of the silent reading. Because response times varied from trial to trial for each participant, we created a mean score for each

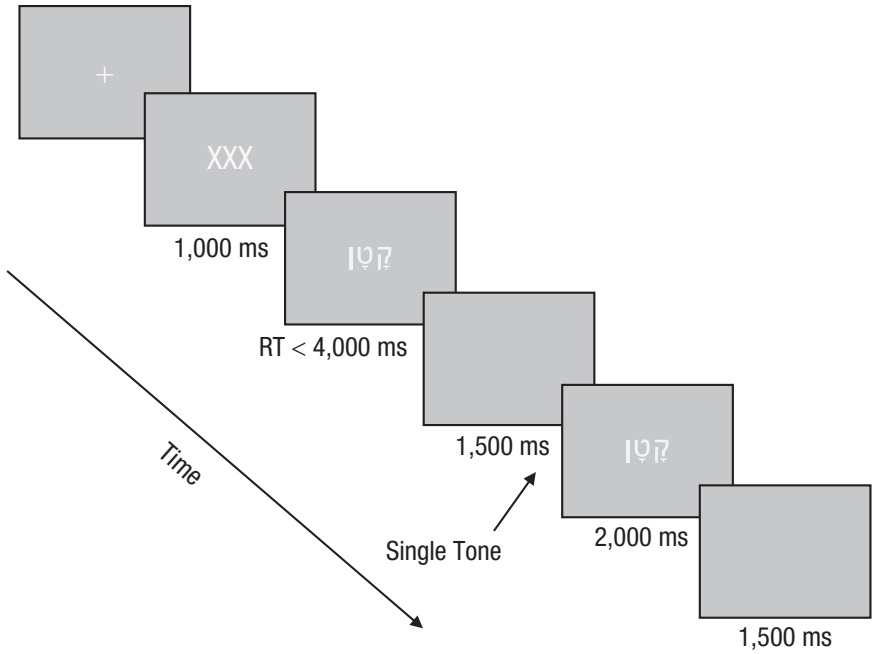

Fig. 3. Example trial sequence in Experiment 2. After viewing a string of Xs, participants saw a stimulus consisting of a real word or a pseudoword, and they had to press a response key to indicate that they had silently read the stimulus a single time. Following a 1,500-ms interval, the stimulus was presented again, together with an auditory tone (beep). Participants then read the stimulus aloud. RT = response time.

condition based on each individual per-trial time window from stimulus onset to tone. Response time and accuracy analyses were the same as in Experiment 1.

Results

Pupil dilation. A fully within-subjects 2 × 2 ANOVA with two levels of familiarity (familiar vs. unfamiliar) and two lengths (three letters vs. five letters) was conducted using a time window from stimulus onset to auditory tone (1,500 ms after silent reading). Figure 4 presents the average pupillary responses in the four conditions in Experiment 2.

Mean relative changes in pupil size. For relative changes in pupil size, we obtained a main effect of word familiarity, $F(1, 19) = 4.80$, $p = .04$, $\eta_p^2 = .20$, OP = .55 (pseudowords: $M = 5.28\%$, $SD = 4.34$; real words: $M = 4.13\%$, $SD = 3.00$), but not for length, $F(1, 19) = 2.67$, $p = .12$, $\eta_p^2 = .12$, OP = .34 (five-letter strings: $M = 4.97\%$, $SD = 3.89$; three-letter strings: $M = 4.44\%$, $SD = 3.35$). We also confirmed the predicted familiarity-by-length interaction, $F(1, 19) = 10.57$, $p = .004$, $\eta_p^2 = .36$, OP = .87, indicating a length effect for pseudowords (five-letter strings: $M = 5.93\%$, $SD = 5.06$; three-letter strings: $M = 4.62\%$, $SD = 3.90$), $t(19) = -2.48$, $p = .02$, but not for real words (five-letter strings: $M = 4.01\%$, $SD = 3.08$; three-letter strings: $M = 4.26\%$, $SD = 2.98$), $t(19) = 1.13$, $p = .27$.

Peak dilation. For peak dilation, the main effect of familiarity was again significant, $F(1, 19) = 9.89$, $p = .005$, $\eta_p^2 = .34$, OP = .85 (pseudowords: $M = 15.60\%$, $SD = 6.98$; real words: $M = 13.33\%$, $SD = 4.87$), as was length, $F(1, 19) = 9.27$, $p = .007$, $\eta_p^2 = .33$, OP = .82 (five-letter strings: $M = 15.20\%$, $SD = 6.44$; three-letter strings: $M = 13.74\%$, $SD = 5.32$). The interaction between familiarity and length was also significant, $F(1, 19) = 13.55$, $p = .002$, $\eta_p^2 = .42$, OP = .94, because of higher peaks for five-letter pseudowords ($M = 17.05\%$, $SD = 8.17$) compared with three-letter pseudowords ($M = 14.16\%$, $SD = 6.11$), $t(19) = -3.58$, $p = .002$, but length was not significant in the case of real words (five letters: $M = 13.35\%$, $SD = 5.14$; three letters: $M = 13.31\%$, $SD = 4.73$), $t(19) = -0.10$, $p = .92$.[3]

Latency to peak dilation. For latency to peak dilation, all effects were, once again, significant: a main effect for word familiarity, $F(1, 19) = 36.28$, $p < .001$, $\eta_p^2 = .66$, OP = 1.00 (pseudowords: $M = 1{,}813$ ms, $SD = 488$; real words: $M = 1{,}423$ ms, $SD = 329$); a main effect for length, $F(1, 19) = 49.00$, $p < .001$, $\eta_p^2 = .72$, OP = 1.00 (five-letter strings: $M = 1{,}738$ ms, $SD = 433$; three-letter strings: $M = 1{,}499$ ms, $SD = 357$); and an interaction between familiarity and length, $F(1, 19) = 30.20$, $p < .001$, $\eta_p^2 = .61$, OP = 1.00. As before, the interaction derived from the fact that, among pseudowords, slower peak dilations were observed for five-letter pseudowords ($M = 2{,}029$ ms, $SD = 575$) compared with three-letter pseudowords

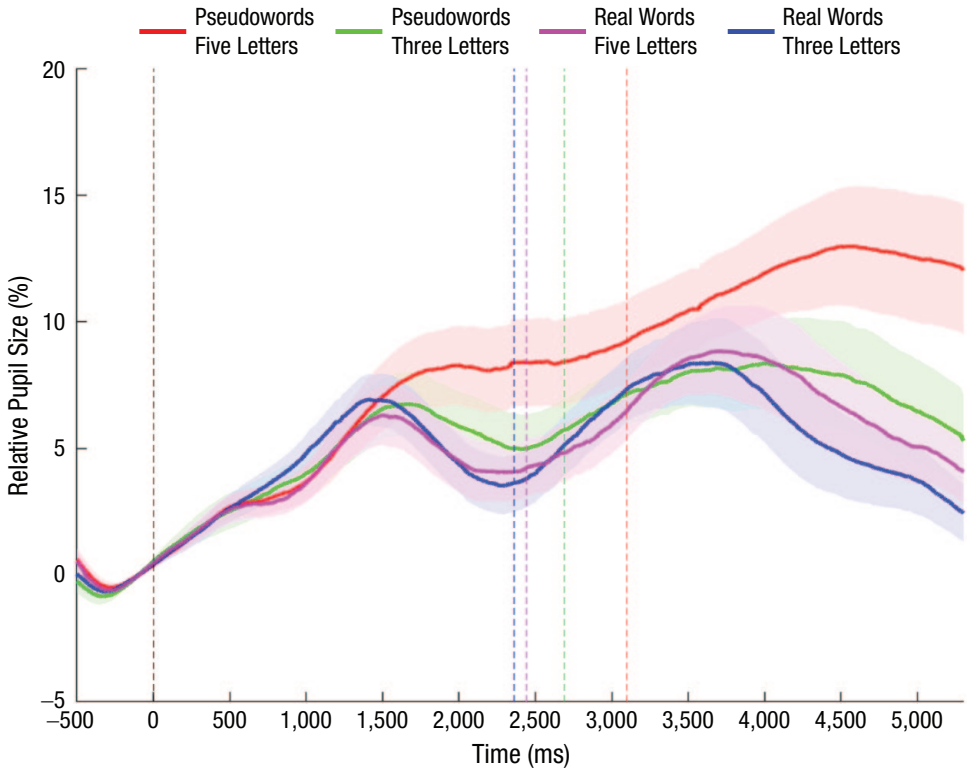

Fig. 4. Relative changes in pupil size for the four conditions in Experiment 2 from stimulus onset (Time 0; the black, dashed vertical line) to trial offset. Each colored, dashed vertical line marks 1,500 ms after the mean response time for the given condition. For each condition, the time course from stimulus onset to the colored vertical line represents the silent-reading mode. The time course from the colored vertical line to trial offset represents the oral-reading mode. The shaded areas depict standard errors of the mean.

(M = 1,598 ms, SD = 426), $t(19) = -7.07$, $p < .001$, with no significant difference between three-letter and five-letter real words (five letters: M = 1,446 ms, SD = 362; three letters: M = 1,400 ms, SD = 309), $t(19) = -1.42$, $p = .17$.

Behavioral data. Both response times and accuracy for yoked pairs pronounced correctly were again analyzed using a two-way repeated measures ANOVA with familiarity (familiar vs. unfamiliar) and length (three letters vs. five letters) as within-subjects factors.

Response times. For response times, we observed main effects for word familiarity, $F(1, 19) = 63.68$, $p < .001$, $\eta_p^2 = .77$, OP = 1.00, and length, $F(1, 19) = 61.10$, $p < .001$, $\eta_p^2 = .76$, OP = 1.00, in addition to a significant familiarity-by-length interaction, $F(1, 19) = 47.62$, $p < .001$, $\eta_p^2 = .72$, OP = 1.00. Length effects were found for both pseudowords, $t(19) = -7.80$, $p < .001$, and real words, $t(19) = -4.74$, $p < .001$ (Table 2).

Pronunciation accuracy. For pronunciation accuracy, the main effect for familiarity was significant, $F(1, 19) = 18.12$, $p < .001$, $\eta_p^2 = .49$, OP = .98, but the main effect for length was not, $F(1, 19) = 0.30$, $p = .59$, $\eta_p^2 = .02$, OP = .08. The interaction between familiarity and length was nonsignificant as well, $F(1, 19) = 0.30$, $p = .59$, $\eta_p^2 = .02$, OP = .08 (Table 2).

Summary

In summary, Experiment 2 confirmed that silent word reading of unfamiliar letter strings (pseudowords) indeed demands more effort than silent reading of familiar (real) words, as indicated by each of the pupillometric measures (mean, peak dilation, and peak latency). Furthermore, the multiple pupillometric analyses clearly pointed to length effects only for pseudowords, emphasizing the greater mental effort invested in reading longer pseudowords silently. Thus, Experiments 1 and 2 together supply clear evidence that pupillary responses are sensitive to word familiarity and its interaction with length in both oral and silent reading among skilled adult readers.

Experiments 3 and 4 replicated these findings with elementary school children.

Table 2. Mean Response Times and (Estimated) Accuracy in Silent Word Reading Among University Students ($N = 20$)

	Response time (ms)			Accuracy (%)		
Word type	Three letters	Five letters	*M*	Three letters	Five letters	*M*
Pseudowords	1,154 (422)	1,546 (596)	1,350 (504)	97.3 (3.5)	96.8 (4.0)	97.0 (3.2)
Real words	835 (298)	921 (344)	878 (319)	100 (0.0)	100 (0.0)	100 (0.0)
M	995 (353)	1,233 (454)		98.6 (1.7)	98.4 (2.0)	

Note: Standard deviations are given in parentheses.

Experiment 3: Oral Reading (Fourth to Sixth Graders)

Method

Participants. A pool of 38 children in the upper elementary grades (fourth to sixth grades) was recruited for this experiment. Four children reported attentional difficulties and were excluded. The other 34 participants, all native Hebrew speakers, reported no past or present reading difficulties or attentional deficits and had normal or corrected-to-normal vision. The data from four participants were excluded because they did not reach a minimum of 20 valid trials in each of the four conditions (i.e., 50% correct responses with no more than 20% of missing pupil values). The final sample numbered 30 participants (17 female; age: $M = 10.42$ years, $SD = 1.02$) who had 20.1% missing trials on average. Of these participants, nine were fourth graders, 10 were fifth graders, and 11 were sixth graders. Each student and his or her parent signed a voluntary informed consent form prior to the experiment and received a small gift for participation.

Design and procedure. As in Experiment 1, this experiment had a fully within-subjects 2 × 2 design with two levels of familiarity (familiar vs. unfamiliar) and two lengths (three letters vs. five letters). However, to accommodate the younger age group, the current experiment included slightly fewer trials: 200 (instead of 240) likewise divided into four blocks. Each block contained 50 items, 10 fillers (instead of 20 for the adults), and the same 20 pseudowords (10 of each length) and 20 real words (10 of each length).

The procedure and apparatus were the same as in Experiment 1, with the exception that the duration of stimulus presentation was longer (4,700 ms) to accommodate this younger sample (Fig. 5). As in Experiment 1, target stimuli were yoked pairs (a real word and its matched pseudoword). Filler words were randomly selected from the fillers in Experiment 1. Data were analyzed as in Experiment 1.

Results

Pupil dilation. Pupillary data were analyzed over a time window from stimulus onset to 5,700 ms (1,000 ms after stimulus offset). Changes in pupil dilation for the four conditions are depicted in Figure 6.

Mean relative changes in pupil size. Analyses of relative changes in pupil size replicated the pattern of adult data in Experiment 1. We found a significant main effect for word familiarity, $F(1, 29) = 7.96$, $p = .01$, $\eta_p^2 = .22$, OP = .78 (pseudowords: $M = 9.41\%$, $SD = 5.70$; real words: $M = 7.72\%$, $SD = 4.60$); a main effect for length, $F(1, 29) = 14.98$, $p = .001$, $\eta_p^2 = .34$, OP = .96 (five-letter strings: $M = 9.39\%$, $SD = 5.31$; three-letter strings: $M = 7.74\%$, $SD = 4.82$); and a significant familiarity-by-length interaction, $F(1, 29) = 5.06$, $p = .03$, $\eta_p^2 = .15$, OP = .59. Follow-up t tests confirmed a significant length effect for pseudowords (five-letter strings: $M = 10.83\%$, $SD = 6.95$; three-letter strings: $M = 8.00\%$, $SD = 4.85$), $t(29) = -4.18$, $p < .001$, but not for real

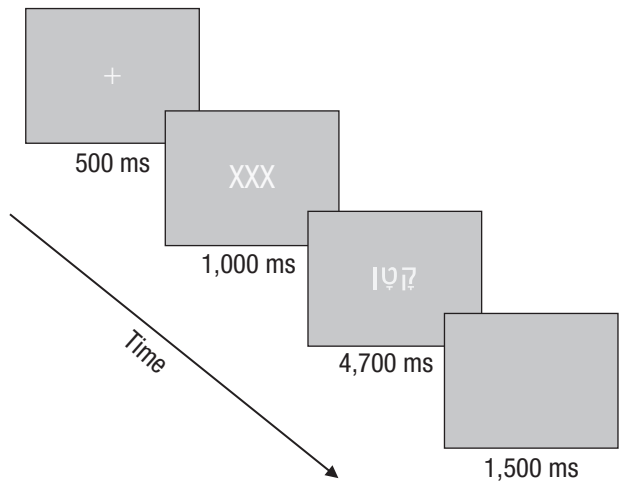

Fig. 5. Example trial sequence in Experiment 3. After viewing a string of Xs, participants saw a stimulus consisting of a real word or a pseudoword, which they were asked to read aloud. The experiment differed from Experiment 1 in that the time allocated for participants' response was longer.

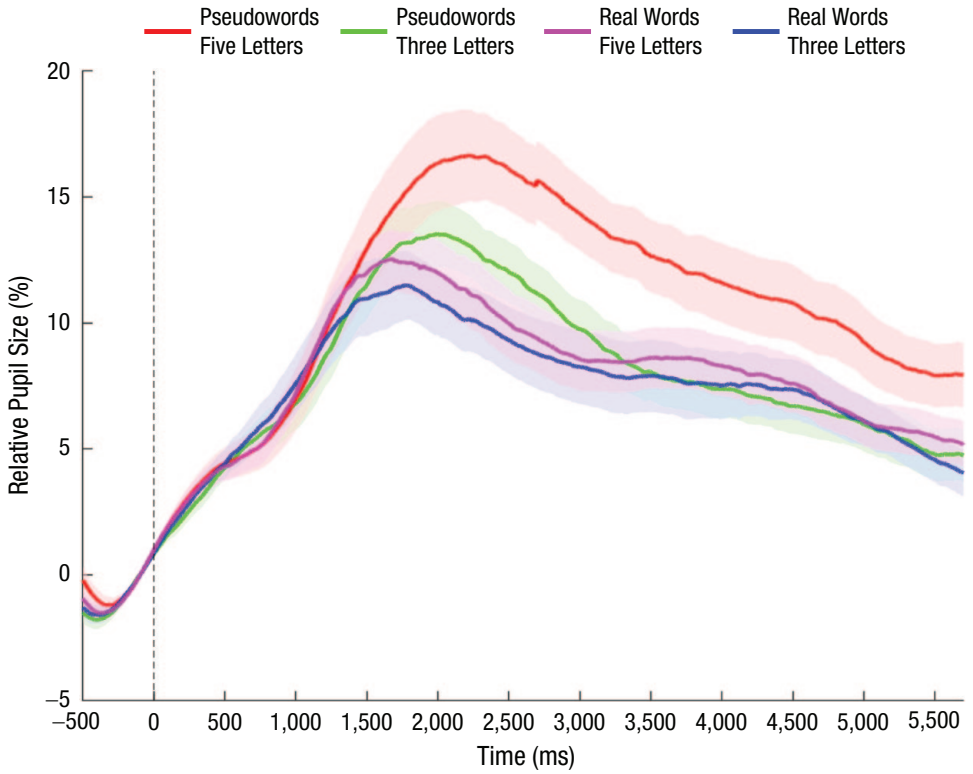

Fig. 6. Relative changes in pupil size for the four conditions in Experiment 3 from stimulus onset (Time 0, the dashed vertical line) to 1,000 ms after stimulus offset. The shaded areas depict standard errors of the mean.

words (three letters: $M = 7.48\%$, $SD = 5.28$; five letters: $M = 7.96\%$, $SD = 4.65$), $t(29) = -0.71$, $p = .48$.

Peak dilation. For peak dilation, too, ANOVAs indicated significant main effects for both word familiarity, $F(1, 29) = 9.39$, $p = .005$, $\eta_p^2 = .25$, OP = .84 (pseudowords: $M = 17.49\%$, $SD = 8.87$; real words: $M = 14.95\%$, $SD = 7.34$), and length, $F(1, 29) = 15.78$, $p < .001$, $\eta_p^2 = .35$, OP = .97 (five-letter strings: $M = 17.27\%$, $SD = 8.22$; three-letter strings: $M = 15.17\%$, $SD = 7.72$), as well as a significant familiarity-by-length interaction, $F(1, 29) = 5.79$, $p = .02$, $\eta_p^2 = .17$, OP = .64. As before, this interaction resulted from a significant length effect for pseudowords (five-letter strings: $M = 19.30\%$, $SD = 10.40$; three-letter strings: $M = 15.68\%$, $SD = 7.69$), $t(29) = -4.52$, $p < .001$, but not for real words (three letters: $M = 14.67\%$, $SD = 8.29$; five letters: $M = 15.24\%$, $SD = 7.04$), $t(29) = -0.67$, $p = .51$.

Latency to peak dilation. In ANOVAs examining latency to peak dilation, the main effect for word familiarity was again significant, $F(1, 29) = 12.85$, $p = .001$, $\eta_p^2 = .31$, OP = .93 (pseudowords: $M = 2{,}376$ ms, $SD = 772$; real words: $M = 1{,}994$ ms, $SD = 706$), as was the main effect for length, $F(1, 29) = 5.08$, $p = .03$, $\eta_p^2 = .15$, OP = .59 (five-letter strings: $M = 2{,}316$ ms, $SD = 819$; three-letter strings: $M = 2{,}054$ ms, $SD = 675$). The interaction between familiarity and length, although in the predicted direction, was not significant, $F(1, 29) = 0.81$, $p = .38$, $\eta_p^2 = .03$, OP = .14 (pseudowords—three letters: $M = 2{,}175$ ms, $SD = 776$, five letters: $M = 2{,}576$ ms, $SD = 876$; real words—three letters: $M = 1{,}933$ ms, $SD = 939$, five letters: $M = 2{,}055$ ms, $SD = 1{,}028$).

Behavioral data.

Pronunciation onset latencies. Analyses of pronunciation onset latencies revealed a main effect for word familiarity, $F(1, 29) = 63.57$, $p < .001$, $\eta_p^2 = .69$, OP = 1.00; a main effect for length, $F(1, 29) = 14.69$, $p < .001$, $\eta_p^2 = .34$, OP = .96; and a significant familiarity-by-length interaction, $F(1, 29) = 7.26$, $p = .01$, $\eta_p^2 = .20$, OP = .74. As expected, follow-up tests confirmed a significant length effect for pseudowords, $t(29) = -4.54$, $p < .001$, but not for real words, $t(29) = -1.17$, $p = .25$ (Table 3).

Pronunciation accuracy. Accuracy analyses also produced a main effect for familiarity, $F(1, 29) = 69.67$, $p < .001$, $\eta_p^2 = .71$, OP = 1.00; a main effect for length, $F(1, 29) = 18.84$, $p < .001$, $\eta_p^2 = .39$, OP = .99; and a significant familiarity-by-length interaction, $F(1, 29) = 63.44$, $p < .001$, $\eta_p^2 = .69$, OP = 1.00, as well as a length effect for both pseudowords, $t(29) = 6.59$, $p < .001$, and real words, $t(29) = -2.68$, $p = .01$ (Table 3).

Table 3. Mean Pronunciation Onset Latencies and Accuracy in Oral Word Reading Among Fourth to Sixth Graders ($N = 30$)

	Onset latency (ms)			Accuracy (%)		
Word type	Three letters	Five letters	*M*	Three letters	Five letters	*M*
Pseudowords	1,362 (358)	1,465 (387)	1,413 (367)	88.5 (12.4)	78.8 (11.7)	83.6 (11.4)
Real words	1,114 (268)	1,139 (261)	1,127 (258)	97.4 (4.7)	99.3 (1.7)	98.4 (2.9)
M	1,238 (303)	1,302 (307)		93.0 (7.9)	89.0 (6.3)	

Note: Standard deviations are given in parentheses.

Summary

In summary, the results of this experiment essentially replicated Experiment 1 and extended the results obtained with skilled adult readers to developing readers. More cognitive resources were required to read unfamiliar letter strings compared with familiar (real word) strings, as reflected in multiple pupillary measures (mean relative changes, peak dilation, and peak latency), consistent with lower accuracy and slower pronunciation times. We also observed greater length costs for pseudowords compared with real words not only on behavioral measures but also on pupillometric measures of mean overall changes and peak dilation.

To our knowledge, this is the first study to successfully apply pupillometric methods to single-word reading in children. Here, too, we conducted a follow-up silent-then-oral-reading study among another sample of fourth- to sixth-grade children (Experiment 4) to confirm that our findings are generalizable to word reading as such, as opposed to being a task-specific effect possibly reflecting the articulatory-motor demands of vocalization.

Experiment 4: Silent-Then-Oral Reading (Fourth to Sixth Graders)

Method

Based on the reported effect size from Experiment 3, an a priori power analysis (G*Power) estimated that a sample size of 18 participants would be necessary to achieve an effect size (f) of .35 (power = .80, $\alpha = .01$). A new sample of 21 fourth to sixth graders was recruited for this experiment; they were all native Hebrew speakers, had no reported reading or attentional deficits, and had normal or corrected-to-normal vision. Three children did not reach the criterion of 20 valid trials per condition (i.e., 50% correct responses with no more than 20% missing pupil data), leaving a final sample of 18 participants (11 female; age: $M = 9.97$ years, $SD = 0.90$) who had 16.22% missing trials on average. Of these participants eight were fourth graders, four were fifth graders, and six were sixth graders. Each student and his or her parent signed a voluntary informed consent form prior to the experiment and received a small gift for participation.

The design, stimuli, and apparatus were the same as in Experiment 3. The procedure and statistical analyses were the same as in Experiment 2.

Results

Pupil dilation. A within-subjects 2 × 2 design with two levels of familiarity (familiar vs. unfamiliar) and two lengths (three letters vs. five letters) was conducted using a time window from stimulus onset to the auditory tone (1,500 ms after the participant's key press indicating completion of silent reading) to examine pupil dilation. Figure 7 presents the average pupillary responses in the four conditions in Experiment 4.

Mean relative changes in pupil size. Analyses of relative changes in pupil size replicated the silent-reading results for adults in Experiment 3. First, we found a significant main effect for word familiarity, $F(1, 17) = 4.93$, $p = .04$, $\eta_p^2 = .23$, OP = .55 (pseudowords: $M = 6.30\%$, $SD = 3.95$; real words: $M = 5.31\%$, $SD = 3.56$). In addition, the overall main effect for length approached significance, $F(1, 17) = 3.47$, $p = .08$, $\eta_p^2 = .17$, OP = .42 (five-letter strings: $M = 6.16\%$, $SD = 3.99$; three-letter strings: $M = 5.46\%$, $SD = 3.44$), whereas the familiarity-by-length interaction was significant, $F(1, 17) = 5.03$, $p = .04$, $\eta_p^2 = .23$, OP = .56, with a significant length effect for pseudowords (five-letter strings: $M = 7.17\%$, $SD = 4.73$; three-letter strings: $M = 5.43\%$, $SD = 3.50$), $t(17) = -2.73$, $p = .01$, but not real words (three letters: $M = 5.48\%$, $SD = 3.63$; five letters: $M = 5.14\%$, $SD = 3.89$), $t(17) = 0.62$, $p = .55$.

Peak dilation. For the measure of peak dilation, we also obtained significant main effects for both word familiarity, $F(1, 17) = 8.15$, $p = .01$, $\eta_p^2 = .32$, OP = .77 (pseudowords: $M = 18.22\%$, $SD = 6.07$; real words: $M = 16.22\%$, $SD = 4.54$), and length, $F(1, 17) = 6.33$, $p = .02$, $\eta_p^2 = .27$,

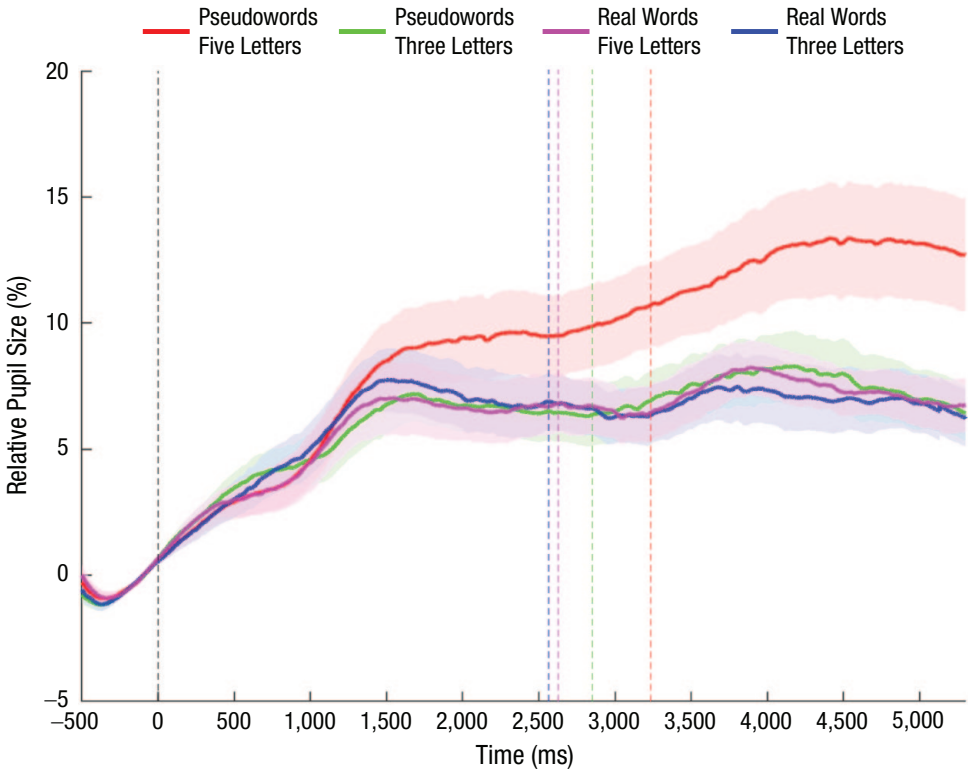

Fig. 7. Relative changes in pupil size for the four conditions in Experiment 4, from stimulus onset (Time 0; the black, dashed vertical line) to trial offset. Each colored, dashed vertical line marks 1,500 ms after the mean response time (completion of silent reading) for the given condition. For each condition, the time course from stimulus onset to the colored vertical line represents the silent-reading mode. The time course from the colored vertical line to the trial offset represents the oral-reading mode. The shaded areas depict standard errors of the mean.

OP = .66 (five-letter strings: M = 18.00%, SD = 5.85; three-letter strings: M = 16.44%, SD = 4.67), as well as a significant familiarity-by-length interaction, $F(1, 17) = 6.65$, $p = .02$, $\eta_p^2 = .28$, OP = .68. A length effect was once again found for pseudowords (five-letter strings: M = 19.88%, SD = 7.87; three-letter strings: M = 16.57%, SD = 4.85), $t(17) = -2.91$, $p = .01$, but not for real words (three letters: M = 16.31%, SD = 4.88; five letters: M = 16.12%, SD = 4.55), $t(17) = 0.30$, $p = .77$ (see Note 2).

Latency to peak dilation. Results for latency to peak dilation were consistent with the outcomes for mean dilation and peak dilation, showing a significant main effect for word familiarity, $F(1, 17) = 31.17$, $p < .001$, $\eta_p^2 = .65$, OP = 1.00 (pseudowords: M = 2,032 ms, SD = 517; real words: M = 1,618 ms, SD = 319); a significant main effect for length, $F(1, 17) = 13.90$, $p = .002$, $\eta_p^2 = .45$, OP = .94 (five-letter strings: M = 1,928 ms, SD = 438; three-letter strings: M = 1,722 ms, SD = 394); and a significant interaction between familiarity and length, $F(1, 17) = 17.15$, $p = .001$, $\eta_p^2 = .50$, OP = .97. A length effect was found for pseudowords (five-letter strings: M = 2,236 ms, SD = 621; three-letter strings: M = 1,828 ms, SD = 459), $t(17) = -4.92$, $p < .001$, but not for real words (three letters: M = 1,617 ms, SD = 372; five letters: M = 1,620 ms, SD = 319), $t(17) = -0.06$, $p = .96$.

Behavioral data.

Response times. The ANOVA on response times yielded a main effect for word familiarity, $F(1, 17) = 44.18$, $p < .001$, $\eta_p^2 = .72$, OP = 1.00; a main effect for length, $F(1, 17) = 45.96$, $p < .001$, $\eta_p^2 = .73$, OP = 1.00; and a significant familiarity-by-length interaction, $F(1, 17) = 40.64$, $p < .001$, $\eta_p^2 = .71$, OP = 1.00. Follow-up tests confirmed a significant length effect for pseudowords, $t(17) = -7.64$, $p < .001$, and (marginally) for real words, $t(17) = -2.10$, $p = .05$ (Table 4).

Pronunciation accuracy. For pronunciation accuracy, there was again a main effect for familiarity, $F(1, 17) = 13.87$, $p = .002$, $\eta_p^2 = .45$, OP = .94; a main effect for length, $F(1, 17) = 6.24$, $p = .02$, $\eta_p^2 = .27$, OP = .65; and a significant familiarity-by-length interaction, $F(1, 17) = 13.95$, $p = .002$, $\eta_p^2 = .45$, OP = .94, indicating larger

Table 4. Mean Response Times and Accuracy in Silent Word Reading Among Fourth to Sixth Graders ($N = 18$)

	Response time (ms)			Accuracy (%)		
Word type	Three letters	Five letters	*M*	Three letters	Five letters	*M*
Pseudowords	1,317 (526)	1,705 (604)	1,511 (556)	92.1 (12.6)	86.5 (13.5)	89.3 (12.5)
Real words	1,038 (308)	1,102 (354)	1,070 (325)	98.9 (2.1)	100 (0.0)	99.4 (1.1)
M	1,178 (411)	1,404 (465)		95.5 (7.3)	93.3 (6.8)	

Note: Standard deviations are given in parentheses.

length effects for pseudowords, $t(17) = 3.25$, $p = .005$, than for real words, $t(17) = -2.20$, $p = .04$ (Table 4).

Summary

In summary, replicating the silent-reading results of Experiment 2 with university students, Experiment 4 confirmed that readers in the fourth to sixth grades also invested more cognitive effort in silently reading pseudowords compared with real words, as indicated by multiple pupillary measures (mean relative changes, peak dilation, and peak latency). Furthermore, we consistently obtained length effects for pseudowords but not for real words on multiple pupillometric measures. Together, the results of Experiments 3 and 4 support the contention that changes in pupil size are a reliable and sensitive index of the cognitive effort invested in both oral and silent reading among developing readers.

Discussion

For more than a century, the notion of mental effort, and effortlessness in particular, has been a common denominator in the psychological literature on skill learning in general and visual word recognition in particular. Like the broader, multifaceted constructs of automaticity and fluency of which it is a defining property (Kuhn et al., 2010; Logan, 1997), word-reading effortlessness or near effortlessness has long been regarded as a distinctive feature of skilled reading. The obverse case of unskilled or impaired reading is typically defined as inaccurate or slow and effortful (American Psychiatric Association, 2013). Yet despite the continued popularity and intuitive appeal of these bread-and-butter concepts, their definition and operationalization have proven surprisingly elusive. In the present investigation, we set out to redress this gap in our knowledge by exploring the applicability of pupillometry as a direct measure of the cognitive effort involved in word reading.

Our findings provided clear evidence that pupillary responses are sensitive to the cognitive effort involved in single-word reading not only among skilled readers (Fernández et al., 2016; Kuchinke et al., 2007; Mathôt et al., 2017) but also among school-age readers in both oral- and silent-reading modes. The data from four experiments were near unanimous in showing that readers, both young and old, are not only slower and less accurate but also allocate more cognitive resources when reading unfamiliar letter strings (i.e., pseudowords) compared with familiar (real) words. Furthermore, our study also examined the length effect—widely regarded as reflecting reliance on the serial, letter-by-letter processing typical of unfamiliar letter strings. We predicted and repeatedly confirmed a significant familiarity-by-length interaction; length effects on behavioral and pupillometric measures were consistently stronger for pseudowords than for real words. These findings corroborate the widespread assumption that reading via a sequential process of letter-to-sound translation and synthesis indeed demands more cognitive resources than reading via direct memory-retrieval mechanisms (e.g., Ehri, 2014; LaBerge & Samuels, 1974; Logan, 1988, 1997; Share, 1995, 2008). This observation, moreover, merges the study of reading with the study of human skill learning in general (e.g., Anderson, 1981; Logan, 1988). Common to almost all skill learning is a transition from slow, effortful, step-by-step, unskilled performance to rapid, near-effortless, one-step, or "unitized" skilled performance.

If replicated,[4] our findings have the potential to open up new avenues of research capable of providing a deeper understanding of the ubiquitous but troublesome concepts of fluency and automaticity. Pupillometry may offer reading researchers a more sensitive moment-by-moment glimpse into the dynamics of word recognition (including developmental, interindividual, and intraindividual variation) that goes beyond the standard measures of skill growth such as reading accuracy and rate or some combination of these two (such as words correctly read per minute). And because learning to read is a paradigmatic case of skill learning, pupillometry has potentially far-reaching applications to a wide variety of domains of skill learning.

We acknowledge that our study is only a first sortie into uncharted waters. This essentially pretheoretical

investigation, nonetheless, raises a host of questions for future work. What is the nature of the association between pupil dilation and standard measures of reading proficiency, and how does this vary across and within levels of reading ability? When does a novel printed word become a familiar unitized orthographic pattern in the course of repeated exposures, and how does this relate to the shape of the effortful-to-(near)-effortless trajectory? Is the learning function monotonic, is it discontinuous with a critical threshold, or does it follow the well-known reaction time power law (Logan, 1988)? Does the disabled reader's word reading remain forever effortful? What exactly is "effort" in the brain? These are just some of the many questions that lie ahead.

Transparency

Action Editor: Rebecca Treiman
Editor: D. Stephen Lindsay
Author Contributions

D. L. Share conceived the idea for this study. A. Shechter developed the experimental design and materials, implemented the experiments, analyzed the results, and wrote the first draft of the manuscript. Both authors revised and approved the final manuscript for submission.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This research was supported by the Ministry of Science & Technology, Israel; by the Israel Science Foundation (Grant No. 1002/20 to D. L. Share); and by the Edmond J. Safra Brain Research Center for the Study of Learning Disabilities.

Open Practices

Data for all experiments, including the two pilot experiments, have been made publicly available on OSF at https://osf.io/hk4yq/. The design and analysis plans for the experiments were not preregistered.

ORCID iDs

Adi Shechter https://orcid.org/0000-0001-8863-9981
David L. Share https://orcid.org/0000-0001-9737-572X

Acknowledgments

We thank Sam Hutton, Stav Magalnik, and Amir Yair for their assistance in designing the experiments. We thank Ronen Hershman, Stuart Steinhauer, Noga Cohen, and Amit Yashar for their valuable comments. We also thank Tami Katzir for her support in this work. Finally, we are grateful to the children, the parents, and the students who participated in this study.

Notes

1. Ethical approval for all four experiments was obtained by the Ethical Committee of the Faculty of Education of the University of Haifa (No. 18/427).

2. Item (in addition to subject) analyses are standard practice in the cognitive-behavioral literature. However, the situation is different for psychophysiological research such as pupillometry, as discussed in Kelbsch et al.'s (2019) review article "Standards in Pupillography." Nonetheless, to allay concerns, we ran item analyses for all three pupillometry measures in each of the four experiments. The outcomes of those analyses produced a near-identical pattern of results.

3. The astute reader will notice that the reported numbers do not match the observed peaks in Figure 4. This is because the actual numbers and the graphical presentation were calculated in different ways. In the statistical analyses of the data, we determined peak dilation as the maximum relative change in each individual trial during the time interval from stimulus onset (Time 0) to the tone (the colored vertical lines), which varied from item to item. The graph in Figure 4 presents an average value at each time point, which is necessarily lower than the true individual (per trial) values reported in the text.

4. The present experiments were preceded by two smaller-scale pilot experiments, one with skilled adult readers ($N = 8$) and one for fourth to sixth graders ($N = 8$). Statistically, and numerically, the two sets of findings for the oral reading tasks were almost identical. The data for all experiments are available on OSF (https://osf.io/hk4yq/).

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.

Anderson, J. R. (1981). *Cognitive skills and their acquisition*. Hillsdale, NJ: Erlbaum.

Balgur, R. (1968). *List of basic words for school*. Tel Aviv, Israel: Otsar Hamoreh.

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

Binda, P., Pereverzeva, M., & Murray, S. O. (2014). Pupil size reflects the focus of feature-based attention. *Journal of Neurophysiology*, *112*, 3046–3052. doi:10.1152/jn.00502.2014

Carver, R. P. (1971). Pupil dilation and its relationship to information processing during reading and listening. *Journal of Applied Psychology*, *55*, 126–134. doi:10.1037/h0030664

Eckstein, M. K., Guerra-Carrillo, B., Singley, A. T. M., & Bunge, S. A. (2017). Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development? *Developmental Cognitive Neuroscience*, *25*, 69–91.

Ehri, L. (2014). Orthographic mapping in the acquisition of sight word reading, spelling memory, and vocabulary learning. *Scientific Studies of Reading*, *18*, 5–21. doi:10.1080/10888438.2013.819356

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191.

Fernández, G., Biondi, J., Castro, S., & Agamenonni, O. (2016). Pupil size behavior during online processing of

sentences. *Journal of Integrative Neuroscience*, *15*, 485–496. doi:10.1142/S0219635216500266

Granholm, E., & Steinhauer, S. R. (2004). Pupillometric measures of cognitive and emotional processes. *International Journal of Psychophysiology*, *52*, 1–6. doi:10.1016/j.ijpsycho.2003.12.001

Hayes, T. R., & Petrov, A. A. (2016). Mapping and correcting the influence of gaze position on pupil size measurements. *Behavior Research Methods*, *48*, 510–527.

Hershman, R., & Henik, A. (2019). Dissociation between reaction time and pupil dilation in the Stroop task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*, 1899–1909. doi:10.1037/xlm0000690

Hershman, R., Henik, A., & Cohen, N. (2018). A novel blink detection method based on pupillometry noise. *Behavior Research Methods*, *50*, 107–114. doi:10.3758/s13428-017-1008-1

Hershman, R., Henik, A., & Cohen, N. (2019). CHAP: Open-source software for processing and analyzing pupillometry data. *Behavior Research Methods*, *51*, 1059–1074. doi:10.3758/s13428-018-01190-1

Just, M. A., & Carpenter, P. A. (1993). The intensity dimension of thought: Pupillometric indices of sentence processing. *Canadian Journal of Experimental Psychology*, *47*, 310–339. doi:10.1037/h0078820

Kahneman, D. (1973). *Attention and effort*. Upper Saddle River, NJ: Prentice Hall.

Kahneman, D., Tursky, B., Shapiro, D., & Crider, A. (1969). Pupillary, heart rate, and skin resistance changes during a mental task. *Journal of Experimental Psychology*, *79*, 164–167. doi:10.1037/h0026952

Kelbsch, C., Strasser, T., Chen, Y., Feigl, B., Gamlin, P. D., Kardon, R., . . . Wilhelm, B. J. (2019). Standards in pupillography. *Frontiers in Neurology*, *10*, Article 129. doi:10.3389/fneur.2019.00129

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

Kuhn, M. R., Schwanenflugel, P. J., & Meisinger, E. B. (2010). Aligning theory and assessment of reading fluency: Automaticity, prosody, and definitions of fluency. *Reading Research Quarterly*, *45*, 230–251. doi:10.1598/RRQ.45.2.4

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, *6*, 293–323. doi:10.1016/0010-0285(74)90015-2

Labuschagne, E. M., & Besner, D. (2015). Automaticity revisited: When print doesn't activate semantics. *Frontiers in Psychology*, *6*, Article 117. doi:10.3389/fpsyg.2015.00117

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*, 492–527. doi:10.1037/0033-295X.95.4.492

Logan, G. D. (1997). Automaticity and reading: Perspectives from the instance theory of automatization. *Reading & Writing Quarterly: Overcoming Learning Difficulties*, *13*, 123–146. doi:10.1080/1057356970130203

Mahelman, I., Rozen, H. B., & Shaked, Y. (1960). *List of basic words for Hebrew literacy instruction*. Jerusalem, Israel: Department of Education and Culture, World Zionist Organization.

Mathôt, S., Grainger, J., & Strijkers, K. (2017). Pupillary responses to words that convey a sense of brightness or darkness. *Psychological Science*, *28*, 1116–1124. doi:10.1177/0956797617702699

Megherbi, H., Elbro, C., Oakhill, J., Segui, J., & New, B. (2018). The emergence of automaticity in reading: Effects of orthographic depth and word decoding ability on an adjusted Stroop measure. *Journal of Experimental Child Psychology*, *166*, 652–663. doi:10.1016/j.jecp.2017.09.016

Moors, A., & De Houwer, J. (2006). Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin*, *132*, 297–326. doi:10.1037/0033-2909.132.2.297

Partala, T., & Surakka, V. (2003). Pupil size variation as an indication of affective processing. *International Journal of Human-Computer Studies*, *59*, 185–198.

Pashler, H. (1994). Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin*, *116*, 220–244. doi:10.1037/0033-2909.116.2.220

Perfetti, C. A. (1985). *Reading ability*. Oxford, England: Oxford University Press.

Perry, C., Zorzi, M., & Ziegler, J. C. (2019). Understanding dyslexia through personalized large-scale computational models. *Psychological Science*, *30*, 386–395.

Reynolds, M., & Besner, D. (2006). Reading aloud is not automatic: Processing capacity is required to generate a phonological code from print. *Journal of Experimental Psychology: Human Perception and Performance*, *32*, 1303–1323.

Share, D. L. (1995). Phonological recoding and self-teaching: Sine qua non of reading acquisition. *Cognition*, *55*, 151–218. doi:10.1016/0010-0277(94)00645-2

Share, D. L. (2008). On the Anglocentricities of current reading research and practice: The perils of overreliance on an "outlier" orthography. *Psychological Bulletin*, *134*, 584–615. doi:10.1037/0033-2909.134.4.584

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

Stanovich, K. E. (1990). Concepts in developmental theories of reading skill: Cognitive resources, automaticity, and modularity. *Developmental Review*, *10*, 72–100. doi:10.1016/0273-2297(90)90005-O

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*, 2005–2015. doi:10.3758/s13423-018-1432-y

Weekes, B. S. (1997). Differential effects of number of letters on word and nonword naming latency. *The Quarterly Journal of Experimental Psychology Section A*, *50*, 439–456.

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*, 498–510. doi:10.1097/AUD.0b013e31820512bb