Behavior Research Methods
https://doi.org/10.3758/s13428-018-1134-4

The human task-evoked pupillary response function is linear: Implications for baseline response scaling in pupillometry

Jamie Reilly[1,2] · Alexandra Kelly[1,2] · Seung Hwan Kim[3] · Savannah Jett[4] · Bonnie Zuckerman[1,2]

Abstract
The human task-evoked pupillary response provides a sensitive physiological index of the intensity and online resource demands of numerous cognitive processes (e.g., memory retrieval, problem solving, or target detection). Cognitive pupillometry is a well-established technique that relies upon precise measurement of these subtle response functions. Baseline variability of pupil diameter is a complex artifact that typically necessitates mathematical correction. A methodological paradox within pupillometry is that linear and nonlinear forms of baseline scaling both remain accepted baseline correction techniques, despite yielding highly disparate results. The task-evoked pupillary response (TEPR) could potentially scale nonlinearly, similar to autonomic functions such as heart rate, in which the amplitude of an evoked response diminishes as the baseline rises. Alternatively, the TEPR could scale similarly to the cortical hemodynamic response, as a linear function that is independent of its baseline. However, the TEPR cannot scale both linearly and nonlinearly. Our aim was to adjudicate between linear and nonlinear scaling of human TEPR. We manipulated baseline pupil size by modulating the illuminance in the testing room as participants heard abrupt pure-tone transitions (Exp. 1) or visually monitored word lists (Exp. 2). Phasic pupillary responses scaled according to a linear function across all lighting (dark, mid, bright) and task (tones, words) conditions, demonstrating that the TEPR is independent of its baseline amplitude. We discuss methodological implications and identify a need to reevaluate past pupillometry studies.

Keywords Pupillometry · Executive functioning · Arousal · Cognitive load · Psychophysics

Introduction

The human pupil dilates and constricts in response to variations in luminance, effectively gating light to the retina. Pupillary dilation is also evoked by many other factors that modulate sympathetic and/or parasympathetic nervous system functioning, including acoustic startle, pleasant tastes, fatigue, habituation, sexual arousal, fear, and imagined light (Laeng & Sulutvedt, 2014; Loewenfeld & Lowenstein, 1993; Peysakhovich, Vachon, & Dehais, 2017; Tryon, 1975; Zavagno, Tommasi, & Laeng, 2017). One particular source of pupillary dilation has been the subject of intense interest since Hess and Polt (1964) demonstrated that the pupil dilates in tandem with cognitive processing demands during verbal problem solving. Beatty (1982) referred to this phenomenon of phasic pupil dilation time-locked to a specific cognitive event as the *task-evoked pupillary response* (hereafter abbreviated TEPR).

Phasic arousal is associated with up-regulation of the sympathetic nervous system in the context of task engagement. In contrast, modulation of tonic arousal is typically linked to the relatively slower and sustained buildup of processing resources (Aston-Jones & Cohen, 2005). The amplitude of the TEPR is typically an order of magnitude smaller than luminance-induced movements, with dilation peaking at approximately 0.1 mm above the tonic baseline (Beatty & Lucero-Wagoner, 2000; Peysakhovich et al., 2017). Cognitive pupillometry involves the precise measurement of these subtle phasic response functions, typically via sensitive ocular imaging modalities such as infrared eye tracking.

Variations in both luminance and tonic arousal are associated with rising and falling baselines in pupillary diameter (Loewenfeld & Lowenstein, 1993; Papesh & Goldinger, 2015; Peysakhovich, Causse, Scannella, & Dehais, 2015).

✉ Jamie Reilly
reillyj@temple.edu

[1] Eleanor M. Saffran Center for Cognitive Neuroscience, Temple University, Philadelphia, PA, USA

[2] Department of Communication Sciences and Disorders, Temple University, Philadelphia, PA, USA

[3] Department of Slavic & Eastern Languages, Boston College, Boston, MA, USA

[4] Department of Linguistics and Cognitive Science, Pomona College, Claremont, CA, USA

Published online: 27 September 2018

Thus, pupillary data constitute a nonstationary time series in which putative "events" are initiated at different baseline pupil diameters. These rising and/or falling baselines typically necessitate some form of mathematical correction in order to permit direct contrasts of nested events.

Despite the relative maturity of pupillometry as a psychophysical measurement technique, the field has yet to empirically validate and/or reach methodological consensus on a definitive baseline correction procedure (but see Mathôt, Fabius, Van Heusden, & Van der Stigchel, 2018).[1] Nevertheless, an informal set of practices has emerged. Many researchers collect brief intervals of neutral baseline data during periods of rest or habituation immediately preceding each TEPR. The pupillary response is then derived by computing change scores during the window following the event relative to the prestimulus baseline. Baseline correction facilitates group-level contrasts by standardizing all change scores as relative, initiating from 0 mm.

Baseline correction within pupillometry typically assumes one of two forms. *Subtractive correction* models the TEPR as a linear function, with all evoked responses independent of their starting point. A linear response pattern for a target detection, for example, would be exemplified by similar amplitudes of pupil dilation whether the listener was hyperaroused, with an elevated tonic baseline pupil diameter, or in a resting state, with correspondingly lower baseline pupil diameter. In contrast, nonlinear response scaling techniques (e.g., proportional, logarithmic, or power) model the evoked response as progressively diminishing at elevated baseline levels. Consequently, a proportionally scaled difference in pupillary peak amplitude from 6.0 to 6.5 mm represents an 8.3% change, whereas the same absolute magnitude difference (0.5 mm) initiated at a different baseline (3.0 to 3.5 mm) would reflect a 16.7% difference.

The most common nonlinear baseline correction procedure in cognitive pupillometry involves *proportional (or divisive) correction*. Proportional correction typically involves deriving change scores for each observed event relative to its baseline amplitude via the following formula, where i is the respective time bin and b is the mean of the baseline interval preceding the specific event:

$$\%\text{pupillary change} = \frac{\mu_i - \mu_b}{\mu_b}.$$

Numerous pupillometry studies—past and present—have employed this nonlinear correction technique (Duñabeitia & Costa, 2015; Graham, Hoover, Ceballos, & Komogortsev, 2011; Hayashi, Someya, & Fukuba, 2010; Hess & Polt, 1960, 1964; Iqbal, Zheng, & Bailey, 2004; Janisse, 1974; Kankipati, Girkin, & Gamlin, 2011; Mathôt, Grainger, & Strijkers, 2017; Spitschan, Jain, Brainard, & Aguirre, 2014; Weiss, Trehub, Schellenberg, & Habashi, 2016).

Many physiological responses scale nonlinearly. Skin conductance, heart rate, and hearing thresholds all scale as predicted by the law of initial values, which describes a function in which the magnitude of an evoked response progressively dampens with rising amplitude of the baseline (Lacey, 1956; Wilder, 1958). Other processes, such as the cortical hemodynamic response function (HRF), appear to scale linearly. The determination of an appropriate response function to characterize the TEPR is an empirical question that remains largely unresolved. As a consequence, two parallel lines of baseline response modeling have co-evolved. If linear and nonlinear approaches converged upon the same result, there would be little cause for concern; however, linear and nonlinear scaling often yield highly disparate results from the same time series. Because the pupil cannot scale both linearly and nonlinearly under the same conditions, one evolutionary line of pupil response modeling is vulnerable to modeling error, though the question of which remains unanswered.

Perturbation of baseline pupil amplitude

Homogeneity between two TEPRs elicited at different baseline pupillary diameters would support a linear physiological scaling process. Experimental confirmation of this effect would require manipulation of the baseline pupil diameter. There are numerous methods of effecting such movements, including manipulations of tonic arousal and illuminance. Variations in the tonic baseline are typically small, slow, and unpredictable, relative to the more robust, immediate, and reproducible pupil movements evoked by altering ambient light levels (Beatty & Lucero-Wagoner, 2000; Peysakhovich et al., 2015). Luminance-induced pupil movements primarily involve the modulation of parasympathetic pathways, inducing in most people a much larger dynamic range of constriction/dilation from 2.0 to 9.0 mm (Loewenfeld & Lowenstein, 1993; Wang & Munoz, 2015).

Luminance by task demand interactions

In a traditional pupillometry experiment, the researcher maintains tight control over luminance while manipulating a specific cognitive variable. Here we conducted the reverse manipulation, holding cognitive task demands constant while manipulating luminance. The advantage of this technique is that if all other cognitive variables are controlled, then any differences in evoked responses at different luminance levels must reflect the initial state of the pupil (Bradshaw, 1969;

[1] Numerous disciplines have adopted their own techniques to cope with variable baselines. In fMRI BOLD signal processing, for example, software packages such as AFNI account for the variable baselines associated with signal drift by using linear and least-squares detrending (Cox, 1996).

Peysakhovich et al., 2015; Pfleging, Fekety, Schmidt, & Kun, 2016; Steinhauer, Siegle, Condray, & Pless, 2004; Xu, Wang, Chen, & Choi, 2011).

Steinhauer et al. (2004) measured pupil amplitudes in dim versus moderate ambient light as participants performed serial-7 mental subtraction (i.e., repeated subtraction by 7, initiated from a random seed) relative to a less-demanding continuous calculation task (i.e., adding 1, initiated from a random seed). The primary aim of this investigation was to dissociate parasympathetic from sympathetic contributions to the TEPR, leveraging the manipulation of illuminance as a robust means of isolating these relative contributions. That is, a cognitively demanding TEPR elicited in darkness reflects predominantly sympathetic dilation, whereas similar cognitive demands elicited in bright light include both parasympathetic and sympathetic responses. Participants demonstrated a task-by-illuminance interaction, such that the average pupil dilation over approximately 1 min of mental calculation was highest for the more demanding task ("subtract 7") when participants completed the calculations in bright light. These results suggest that unique nervous system processes influence the response properties of the TEPR under different exogenous (e.g., environmental) and/or endogenous (e.g., tonic arousal) conditions.

In a related manipulation of the effects of illuminance on cognitive load, Peysakhovich et al. (2017) evaluated TEPRs elicited during sustained processing during *n*-back tasks (one-back, two-back) under two screen luminance levels (low, high). Participants showed higher baseline tonic pupil size in the two-back condition, but unlike for Steinhauer et al. (2004), the TEPRs were amplified in the low-luminance condition. Most relevant to the present investigation, the phasic pupillary dilation evoked by intermittent presentation of math problems scaled independent of the baseline. These results lend complexity to the interpretation of generalized state versus event-related pupil dilation, suggesting a dissociation in the response profiles of tonic (nonlinear) versus phasic (linear) pupil dilation.

Bradshaw (1969) reported perhaps the most direct manipulation of luminance's effect on phasic pupil response. The author compared pupil response functions in an auditory target detection task while participants fixated on dark versus bright backgrounds. Upon visual inspection, these respective phasic response functions were similar (see Bradshaw's Fig. 1, p. 272), supporting the conclusion that the TEPR scales linearly, independent of its baseline. Bradshaw's study employed a sample size (N = 7), sampling rate (2.7 Hz), and statistical approach (visual inspection of response functions) that prove untenable by today's standards. Thus, the question of how the TEPR scales remains open.

Our aim in the experiments to follow was to evaluate whether the pupil dilation response properties for transient events (e.g., target detection) are mediated by baseline pupil diameter. We reasoned that if cognitive task demands are held constant, then any observed differences in pupillary response functions would be attributable to differing pupil baselines. We specifically predicted that task-evoked pupillary responses elicited in lower luminance would be functionally equivalent to the responses elicited in higher luminance for stimuli with comparable cognitive demands. That is, phasic pupil dilation for a transient, discrete event would scale independent of its baseline. This prediction in favor of linear response scaling would be confirmed by *equivalence*, in that TEPRs evoked in low ambient light would be isomorphic to TEPRs elicited in bright light. In contrast, nonlinear scaling would be supported by differences such that TEPRs elicited by low illuminance would be dampened relative to responses elicited in high illuminance, secondary to scaling proportionately from a higher baseline.

Experiment 1

Method

We continuously sampled pupillary diameter while participants heard streams of pure tones punctuated by abrupt, temporally jittered frequency shifts (e.g., 150–300 Hz), contrasting TEPRs induced under two different illuminance levels (low/high). There is a rich history of the use of pure-tone auditory discrimination in pupillometry (Beatty, 1982). The obvious advantage of pure tones is that the stimuli do not vary along any visual dimension. Simple discrimination of highly contrastive tones may be vulnerable to floor effects because the task is minimally demanding. Therefore, we adapted the discrimination task to a more demanding counting paradigm in which participants were compelled to detect tone differences and to maintain a running count of the total number of distinct tones they heard.

Participants

The participants included neurotypical adults (N = 27; 14 men, 13 women) from the local Philadelphia community. Their mean age was 25.26 years (SD = 3.96). We screened for hearing impairment using pure-tone audiometry with a detection threshold of 30 dB (1, 2, 4 kHz). Additional exclusionary criteria were self-reported history of ocular trauma, nystagmus, language disability, or current use of sedating drugs. Participants were asked to remove hard contact lenses, glare-resistant glasses, and eye makeup.

Equipment and software

We sampled pupil diameter via a 120-Hz infrared table-mounted eye tracker (Sensorimotoric Instruments (SMI)

Red-M, Boston MA) positioned at the base of a 17-in. Dell LCD monitor. The tracker was controlled through SMI iView software running on a Dell Precision T7600 computer. Head stability was augmented using an optician's chinrest. Auditory stimuli were created in Audacity (wavefile format, 44.1-kHz sampling rate) and presented via over-the-ear headphones (Sennheiser Inc., Model HD 380 Pro). Illuminance levels were gauged using a light meter (UEi Testing Instruments Inc., Model DLM2). Stimuli were delivered and pupil data recorded using the Experiment Center software (SMI Inc., Boston MA).

Stimulus characteristics

Participants heard a series of four distinct prerecorded audio files, each consisting of a run of five concatenated pure tones (250, 400, 550, 700, 850 Hz). A 150-Hz difference within the primary frequency range of human speech perception far exceeds any just noticeable difference between tone pairings, making the contrasts readily discriminable (Moore, 1973). Tone pairs were presented in a fixed, pseudorandom order. The duration of each tone was varied stochastically between 9.5, 13.5, and 17.5 s, with no breaks between tones. The total duration of each run was 78 s.

In previous pupillometric studies employing pure tones, the onset of each tone was modeled as the event onset for the TEPR. In the experiment to follow, we modeled evoked responses at each switch point between tones within the audio file. For example, participants might hear a 550-Hz pure tone for 17.5 s, followed by an 850-Hz pure tone for 9.5 s. Each audio file contained four switch points, which served as event onsets for statistical modeling of the TEPR (see the Data Analyses section below). We, therefore, modeled 16 events for each participant [4 per run × 4 runs].

Experimental procedures

Participants listened to four different runs of pure tones, divided into two blocks that varied by illuminance (i.e., high/low). The bright ("lights on") condition consisted of constant fluorescent overhead lighting with an additional LED desk lamp and an LCD monitor displaying a white screen [background RGB values 255, 255, 255; HSL lightness value 100%]. Illuminance for the bright condition was metered at 753 lux. The dark condition consisted of no overhead or desk lighting but with the monitor displaying a constant midrange gray screen [background RGB 192,192, 192; HSL lightness value 75.3%]. Illuminance for this dark condition was metered at 16 lux. Participants viewed the same centrally positioned attention fixation cross (slate gray) in both conditions. The block order was counterbalanced across participants.

Participants were tested individually while seated at a monitor in a quiet, windowless room. Participants read instructions encouraging them to remain as motionless as possible and to maintain focus on a central fixation cross. We structured this experiment as an incidental counting task. That is, participants were instructed to listen to the tones and to maintain a running count of the total number of tones they heard.

Once ready, participants donned headphones and positioned their chin on a chinrest situated 60 cm from the monitor. The experimenter then altered the ambient light level to the participant's preassigned counterbalanced order and positioned herself behind a room divider at a second workstation from which she initiated the experiment.

Each experimental run began with a 5-point gaze calibration and validation sequence. The first tone within each run lasted an extended duration (30 s). The purpose of this prolonged presentation was to allow the pupil to settle to a steady state prior to initiating tone shifts. Participants then heard tone sequences as we continuously monitored pupil size. At the conclusion of each tone sequence, the visual fixation cross vanished, and participants viewed a block of text instructing them to key in the number of tones they had just heard. Figure 1 represents this trial structure.

After each tone sequence was completed, the experimenter initiated the second tone run within each ambient light condition. The experimenter then adjusted the ambient lighting in the testing room to match the next preassigned illuminance condition. The total experiment duration was approximately 30 min.

Data analyses

We developed a data-processing pipeline using the R statistical program (R Core Team, 2013). We first extracted pupil measurements from the left eye and isolated all 0-mm measurements, associated with blinks, sudden head turns, or other idiosyncratic measurement artifacts. We then linearly interpolated across all 0-mm measurements, effectively treating them as missing data.[2] Finally, we smoothed the interpolated time series by applying a simple moving average (window = 5). We modeled TEPRs corresponding to the 3,000-ms temporal window following each tone switch point (for precedence, see also Korn & Bach, 2016; Laeng, Orbo, Holmlund, & Miozzo, 2011) The 500-ms interval preceding each switch point was treated as the baseline for that respective event.

[2] Hershman, Henik, and Cohen (2018) recently demonstrated that this standard practice of blink correction is suboptimal, because linear interpolation is anchored to/from erroneous endpoints as the eyelid opens and closes. Blink correction is especially important in typical pupillometry paradigms involving manipulation(s) of cognitive load, because blink rate is modulated by the task demands. That is, the frequency of blinks is positively correlated with higher executive demands (Siegle, Ichikawa, & Steinhauer, 2008). Each of the experiments here involved fixed executive demands, thereby reducing the probability of systematic bias from blinks in one condition. As a secondary post-hoc safety measure, we report blink rates across both experiments.

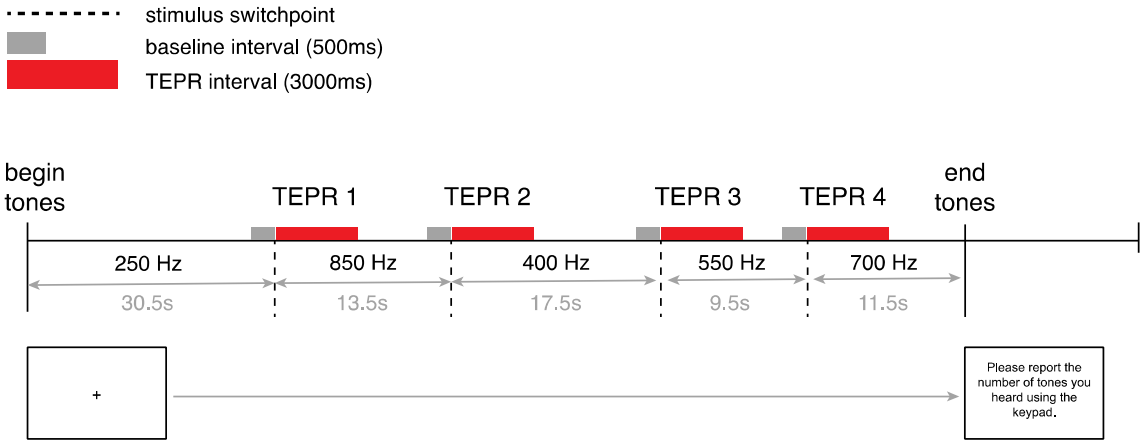

Fig. 1 Tone difference detection trial structure: Sample trial structure of one tone-counting sequence. All others adhered to the same structure, differing only by the order of pure-tone frequencies and the variable interstimulus interval. All tone sequences are freely available for inspection and use at www.reilly-coglab.com/data.

Derivation of linear subtraction scores We derived a linear time series for each illuminance condition by subtracting the mean of the baseline for each event from each successive 200-ms time bin within the 3,000-ms temporal window following that event. This procedure generated change scores (in millimeters) for the onset of tone shifts, normalizing for baseline amplitude. We then collapsed across items, such that each participant's data were composed of two time series (light and dark TEPRs). We scaled pupil responses using the following multistep process:

(a) Derive a baseline for each event (i) and participant (j) by averaging across each observation of left-eye pupil diameter during the 500-ms period preceding the onset of each stimulus.

$$BASELINE_{ij} = \frac{1}{N} \sum_{0\ ms}^{500\ ms} DIAMETER_{ij}$$

(b) Bin the event data by averaging pupil diameter observations within each 250-ms time window, extending 3,000 ms outward from each event (i) for each participant (j). These parameters yield 15 bins per event, where t represents the bin number.

$$BIN_{ijt} = \frac{1}{N} \sum_{0\ ms}^{250ms} DIAMETER_{ijt}$$

(c) Complete the baseline correction by subtracting each event's baseline mean from each successive time bin (t) for each event (i) by participant (j):

$$Corrected\ Pupil\ \Delta_{ijt} = BIN_{ijt} - BASELINE_{ij.}$$

Derivation of peak amplitudes The peak amplitude of the TEPR is parametrically modulated by task difficulty (Karatekin, Couperus, & Marcus, 2004; Peysakhovich et al., 2015; Piquado, Isaacowitz, & Wingfield, 2010; Szulewski, Roth, & Howes, 2015; Zekveld, Heslenfeld, Johnsrude, Versfeld, & Kramer, 2014; Zekveld & Kramer, 2014). Numerous laboratories have adopted strategies for measuring peak amplitude while minimizing noise confounds. Sampling one data point within noisy time series can yield unstable estimates, similar to contrasting the single hottest year within a temporally extended climate time series (Karatekin et al., 2004). In pupillometry, a common method for deriving a stable estimate of peak amplitude is to average across a specific temporal range. Steinhauer and Hakerem (1992) argued for two dissociable TEPR peaks, one attaining a maximum amplitude between 600 and 900 ms post-stimulus-onset, and another at approximately 1,200 ms. Other researchers, however, have noted significant variability in time-to-peak and maximum amplitude across age ranges, clinical populations, and task demands (Loewenfeld & Lowenstein, 1993; Tun, McCoy, & Wingfield, 2009). As such, although some commonalities exist, there is no uniform standard for isolating a peak range or delineating the width of this window. We looked to our data to isolate a temporal window that contained the range of maximum values across participants, averaging the evoked pupil dilation across a peak range from 1,500 to 2,250 ms post-stimulus-onset.

Contrasts of light–dark time series We conducted an omnibus Bayesian repeated measures analysis of variance (ANOVA) of the mean pupil diameter at each successive time bin, scaled linearly, using the JASP statistical software (JASP Team, 2017). In addition to omnibus time series contrasts, we contrasted the mean dilation and peak amplitude of the TEPR using Bayesian paired t tests. We used JASP's default

parameters and priors (*r*-scale fixed effects = .5, *r*-scale random effects = 1, *r*-scale covariates = 0.354, 10,000 samples, Cauchy prior = .707).

At the most fundamental level, Bayes factors quantify the strength of evidence for both alternative (BF_{10}: H_1) and null (BF_{01}: H_0) hypotheses (Jarosz & Wiley, 2014; Rouder, 2014). We hypothesized that the TEPR would be roughly equivalent in darkness and brightness. However, it must be noted that, similar to frequentist null hypothesis significance testing (NHST), equivalence cannot be confirmed directly via a high H_0 Bayes factor. For equivalence testing, Kruschke (2011) recommended defining a region of practical equivalence (ROPE) as part of the decision rule for a Bayesian equivalence test. If the 95% high-density interval (HDI) falls entirely within a ROPE, one might reasonably accept the null and conclude group equivalence (Kruschke, 2011; Lakens, 2016). The upper and lower bounds of a ROPE are defined by convention using informed priors (e.g., physiological constraints). Near-microscopic but nevertheless statistically significant effects of less than one pixel have been reported as evidence for differences (see Exp. II of Laeng & Sulutvedt, 2014). Thus, no definitive threshold exists upon which to establish a ROPE for Bayesian equivalence testing. We explicitly evaluated equivalence (when applicable) with frequentist tests of one-sided significance (TOST) as implemented within the Equivalence package for R (Johnson, 2016).

Results

We eliminated two participants due to equipment error and one participant who failed to correctly answer posttest probes. Among the retained participants ($N = 24$), we discarded four individual tone runs that comprised > 50% missing observations. Blinks and other artifacts such as idiosyncratic head turns constituted 5.22% of the data (9,909 of 189,745 total samples). The distributions of missing values were roughly equivalent across the bright (3.83%) and dark (4.39%) illuminance conditions.

Participants were highly accurate in detecting pure-tone shifts (99.91%), resulting in minimal data loss. As a manipulation check, we first contrasted the average uncorrected pupil diameters in low versus high illuminance. The pupil diameter was predictably smaller in the bright condition (2.91 mm, $SD = 0.34$) than in the dark (3.53 mm, $SD = 0.47$), as confirmed by a directional Bayesian paired *t* test (hypothesized bright < dark) (Bayes factor $BF_{10} = 3.103 \times 10^6$; strong evidence). Figure 2a represents a scatterplot of the range of individual differences in uncorrected baseline pupil diameter across conditions. Figure 2b illustrates the TEPRs elicited by tone switches in low versus high illuminance. Table 1 summarizes the contrasts of average dilation at each successive time bin.

We conducted an omnibus Bayesian repeated measures ANOVA on evoked pupil dilation, treating illuminance (two levels) and time (15 levels) as two within-subjects factors. The Bayes factor (BF_{01}) for the illuminance model null hypothesis was 1.01, indicating that the null hypothesis was equally likely as the illuminance-only model. Table 2 is the corresponding ANOVA table.

Peak amplitude contrasts We contrasted peak amplitudes by averaging the evoked pupil dilations within the peak range (1,500–2,250 ms) for each participant and conducting a Bayesian paired-sample *t* test at the group level. The null hypothesis was that there was no effect of illuminance on peak amplitudes. The mean peak dilation in light was 0.10 mm ($SD = 0.16$), and the average peak in darkness was 0.15 mm ($SD = 0.16$). TOST equivalence testing (variance not assumed, 95% CI, $p < .0001$) supported a conclusion of no difference between the bright and dark condition peaks (see Fig. 2a).

In addition to peak dilation, we calculated the average evoked response across each participant's bright and dark time series. The average magnitude of the TEPR over all 3-s events was 0.05 mm in high illuminance and 0.06 mm in lower illuminance. The TOST statistic [$t(23)$, mean difference = 0.01, epsilon = 1, variance not assumed, 95% CI, $p < .0001$] supported a conclusion of no difference between high- and low-illuminance TEPRs.

Interim discussion: Experiment 1

The TEPRs elicited in low versus high illuminance showed similar amplitude and dispersion properties. This response pattern is inconsistent with nonlinear scaling predictions. Rather, the TEPR was linear, peaking between 1,600 and 1,800 ms from stimulus onset. From the standpoint of cognitive load theory, two tasks with similar executive resource demands matched on other confounding variables should elicit similar pupil dilations. Participants demonstrated this pattern through equivalent TEPRs for the same task under different lighting conditions.

Several aspects of the experimental design warrant caution, including the limited number of trials ($N = 16$ per participant) and a discrete rather than continuous range of illuminance in the testing room (see Fig. 2). In addition, the audio stimuli were highly contrastive, far exceeding the just noticeable difference in audible frequencies for each tone transition. One possibility is that by making perceptual discrimination too easy, we inadvertently introduced a floor effect by essentially eliminating cognitive load. Nevertheless, the observed that the data remained consistent with linear scaling of the pupil, regardless of whether floor effects were at play. Two tasks with similar cognitive demands (however small) would, under the nonlinear model, show differences. For example, if the task were universally too easy, then responses elicited in the dark would still be dampened according to the law of initial values,

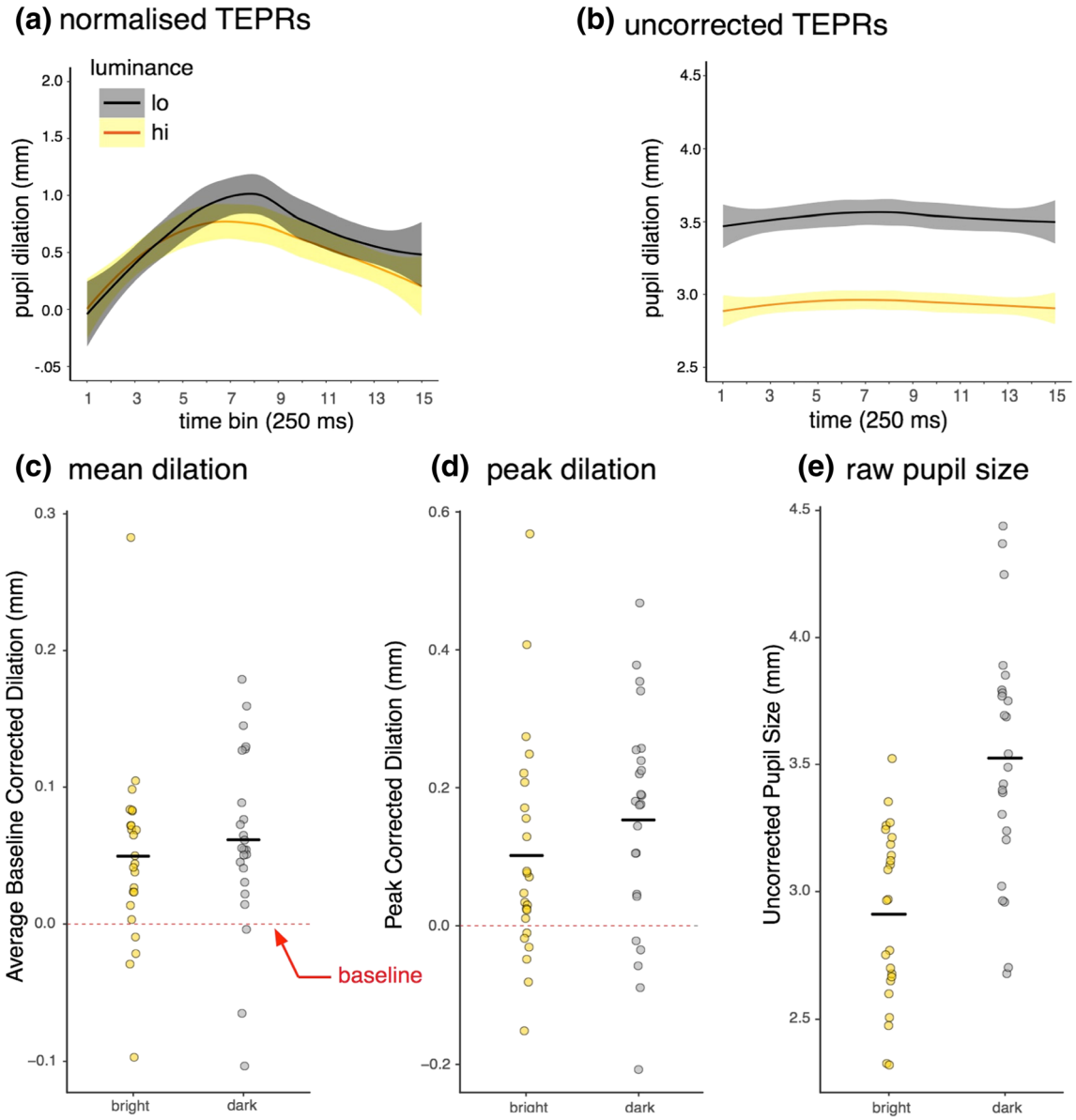

Fig. 2 TEPRs and baseline variability to tone switches (Exp. 1). Panel **a** represents the group-level baseline-normalized TEPRs elicited by tone switches in low versus high ambient light (i.e., Event Bin – Baseline Mean). The *y*-axis represents change in millimeters from baseline; the *x*-axis represents time post-target-onset. Panel **b** represents the same TEPR change scores plotted as uncorrected data (reflecting native baselines). The panel **c** scatterplot reflects the distribution of average baseline-corrected change scores across participants, where each data point represents the mean pupil dilation value across all items and time bins. The horizontal crossbars represent means of the group distributions for the different lighting conditions. Panel **d** represents the average baseline-corrected peak amplitude across participants, and panel **e** represents the average uncorrected pupil diameter across each participant's pupil time series. Several participants demonstrated negative peak and average evoked amplitudes (see **c** and **d**). Upon inspection of their raw data, these participants showed a progressive reduction in pupil size across their trials, resulting in negative change scores (i.e., the event amplitude was less than the baseline amplitude). One might speculate several possible explanations for such a trend, including a generalized reduction in tonic arousal as the experiment continued.

since dilation would initiate from a higher baseline. Participants did not show such an effect. Instead, the TEPRs elicited in low ambient lighting were roughly equivalent to those elicited in brightness. In the experiment to follow, we evaluated the stability of the TEPR in a different modality (visual word monitoring) under an additional midrange lighting constraint.

Experiment 2

Here we evaluated pupil dilation during a visual word-monitoring task in which participants read sequentially presented words and indicated detection of a specified target word via button-press. Participants completed this task under three different light intensities (low, mid, high). We contrasted

Table 1 Experiment 1 baseline-corrected task-evoked pupillary dilations across time

Bin#	Time (ms)	Bright				Dark				BF_{01}
		Amplitude		CI (95%)		Amplitude		CI (95%)		
		Mean	*SD*	Low	Up	Mean	*SD*	Low	Up	
1	0–200	.000	.025	– .010	.011	.004	.029	– .008	.016	4.181
2	200–400	.022	.057	– .002	.047	.010	.046	– .010	.029	2.957
3	400–600	.049	.056	.025	.072	.043	.044	.024	.061	4.260
4	600–800	.057	.066	.029	.085	.056	.056	.032	.080	4.650
5	800–1,000	.058	.076	.026	.091	.051	.072	.020	.081	4.197
6	1,000–1,200	.084	.081	.049	.117	.086	.079	.052	.119	4.612
7	1,200–1,400	.088	.081	.053	.122	.123	.087	.086	.160	1.304
8	1,400–1,600	.067	.090	.030	.105	.117	.111	.069	.164	0.947
9	1,600–1,800	.066	.098	.025	.108	.087	.128	.037	.137	3.633
10	1,800–2,000	.063	.110	.017	.109	.066	.107	.021	.111	4.638
11	2,000–2,200	.055	.106	.010	.099	.059	.090	.021	.097	4.601
12	2,200–2,400	.045	.098	.004	.087	.059	.096	.019	.100	3.814
13	2,400–2,600	.037	.087	2.325e –5	.073	.062	.103	.018	.106	2.602
14	2,600–2,800	.038	.094	– .012	.068	.057	.118	.007	.107	2.669
15	2,800–3,000	.021	.085	– .015	.058	.043	.121	– .008	.094	3.343

TEPRs across the three illuminance conditions, predicting equivalence across a graded range of ambient light.

Method

Participants

The participants included neurotypical adults (N = 42; 34 women, 8 men) from the Philadelphia region, six of whom had also participated in Experiment 1. The mean age was 22.05 years (SD = 3.40). We applied the same exclusionary criteria as in Experiment 1.

Stimulus characteristics

We developed a corpus for word monitoring by first querying the MRC Psycholinguistic Database (Coltheart, 1981), filtering for four-letter English nouns with concreteness values > 500 (yield N = 354). We then cross-referenced the MRC concreteness norms with arousal ratings from Warriner, Kuperman, and Brysbaert (2013), eliminating nouns with arousal ratings z > 1 and entries with a lemma frequency < 3 per million in the SUBTLEX database (N = 266) (Brysbaert & New, 2009). Finally, we removed semantically or syntactically ambiguous words (e.g., homonyms or words whose usage as verbs dominates). These procedures generated a corpus of 199

Table 2 Experiment 1 ANOVA results

Model comparison					
Models	$P(M)$	$P(M \mid \text{data})$	BF_M	BF_{10}	Error%
Null model (incl. subject)	.200	5.907e –14	2.363e –13	1.000	
Luminance	.200	5.834e –14	2.333e –13	0.988	1.000
Time Bin	.200	.409	2.768	6.924e +12	0.276
Luminance + Time Bin	.200	.586	5.664	9.923e +12	1.587
Luminance + Time Bin + Luminance × Time Bin	.200	.005	0.020	8.362e +10	1.603
Analysis of effects					
Effects	P(incl)	P(incl | data)		$BF_{Inclusion}$	
Luminance	.600	.591		0.963	
Time Bin	.600	1.000		5.681e +12	
Luminance × Time Bin	.200	.005		0.020	

All models include subject

highly frequent concrete nouns. We randomly selected three of these nouns (*boot*, *desk*, and *hawk*) to serve as the word-monitoring targets. Foils (N = 135) were randomly selected from the remaining nouns and sorted into three sets (45 foils for each condition).

Experimental procedures

The testing in Experiment 2 differed by task (word monitoring) but otherwise followed the same procedures as in Experiment 1. In addition to the bright (light meter reading = 753 lux) and dark (16 lux) conditions referenced in Experiment 1, we included a midlevel luminance condition (350 lux; desk lamp with no overhead lights). Following eye-tracker calibration, participants viewed a target word for 5,000 ms with the instruction to key-press as quickly as possible whenever that word appeared in an upcoming list. Participants then viewed individual words (N = 60), marking the targets and withholding responses for foils.

Each trial was initiated by a black central fixation cross appearing on a midrange gray screen (RGB 175, 175, 175, HSL lightness 75.3%). After 1,000 ms, the fixation cross disappeared and a word appeared centered on the screen in lowercase Arial 48-point font. Participants viewed each word for 3,000 ms, regardless of whether the stimulus item was a target or foil. Once the participant had completed an individual 60-item block, the lighting conditions were modified and a new target word was assigned for the next experimental block. Presentation order within word lists was completely randomized, and the block order was fully counterbalanced. The session duration was approximately 30 min.

Data analyses

We employed the same initial processing pipeline as described for Experiment 1. Two participants were eliminated due to error and equipment failure. Within the retained participants (N = 40), incorrect responses and individual runs with less than 50% retained observations were discarded. Response accuracy to the three target nouns was high (99.28%), and 87.46% of the original data were retained. Blinks and related missing data constituted 8.15% of all observations (185,093 of 2,272,183 total samples). The distributions of missing values were roughly equivalent across the bright (8.40%) mid (8.22%), and dark (8.27%) illuminance conditions.

We modeled TEPRs evoked by the repeated appearance of the monitored words (15 repetitions per lighting condition).

Results

Figure 3 illustrates individual differences in baseline variability (panels 3c–3e), as well as the TEPRs elicited within the three lighting conditions (panels 3a and 3b). As an initial manipulation check similar to that in Experiment 1, participants showed the predicted canonical pupillary light response with ascending average uncorrected pupil diameters across the low (3.84 mm), mid (3.53 mm), and high (3.32 mm) illuminance conditions. This omnibus linear trend was confirmed by a one-factor Bayesian repeated measures ANOVA (illuminance: three levels) revealing strong predicted support ($BF_{10} = 1.97 \times 10^{10}$) for the difference model (H_a), thus confirming variability in the event baselines from which the TEPRs were initiated. Table 3 summarizes the magnitude of evoked pupil responses at each successive time bin. Figure 3b represents group-level response functions evoked in each of the three luminance conditions.

As is illustrated in Fig. 3, the TEPRs elicited at different baseline pupil amplitudes were similar. We conducted an omnibus Bayesian repeated measures ANOVA, treating luminance (three levels) and time (15 levels) as two within-subjects factors against the null model (i.e., H_o: no effect of illuminance or time and no Time × Illuminance interaction). Table 4 summarizes the ANOVA results, supporting a robust main effect of time ($BF_{10} = 6.9 \times 10^{12}$), but showing no evidence for a main effect of illuminance ($BF_{10} = 0.96$) and no evidence of a Time × Illuminance interaction ($BF_{inclusion} = 0.02$).

Figure 3 shows scatterplots demonstrating the range of individual variability in peak amplitude of the TEPRs in the low (mean = 0.17 mm), mid (mean = 0.21 mm), and high (mean = 0.17 mm) illuminance conditions. At the group level, peak amplitudes were roughly equivalent across illuminance conditions, as gauged by an omnibus repeated measures ANOVA with a Bayes factor (BF_{01}) of 2.37 in favor of the null hypothesis. We followed up this omnibus test with tests of one-sided equivalence. Figure 3 (panels 3c–3e) demonstrate similar peak amplitudes between the low- versus high-illuminance [TOST equivalence for $t(33)$, $p < .00001$], high- versus mid-illuminance [TOST equivalence $p < .00001$], and low- versus mid-illuminance [TOST equivalence $p < .00001$] time series.

Figure 3e reflects scatterplots of the average baseline-corrected amplitudes across all of the time bins for TEPRs elicited in the low- (0.066 mm), mid- (0.057 mm), and high- (0.058 mm) illuminance conditions. There is moderate to strong evidence ($BF_{01} = 9.58$) in favor of the null hypothesis that the average baseline-corrected pupil amplitude did not differ across the three luminance conditions.

General discussion

Linear and nonlinear scaling potentially yield highly discrepant results given the same raw data. Nonlinear methods are appropriate for many autonomic functions, for which the magnitude of a response does indeed diminish as the baseline signal rises. It is not inconceivable, therefore, that the TEPR

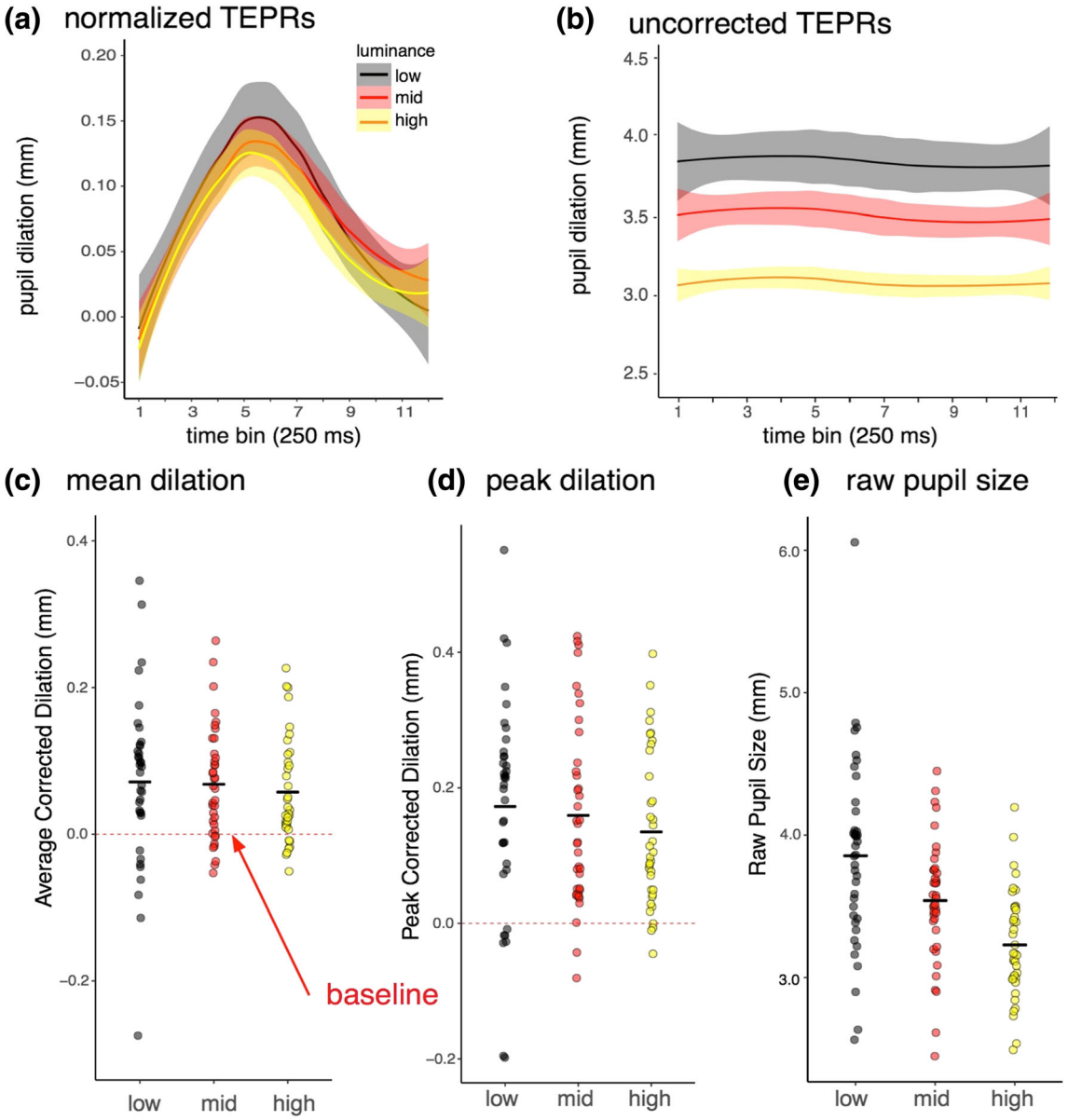

Fig. 3 TEPRs and baseline to orthographic word monitoring (Exp. 2). Panel **a** represents the group-level baseline-normalized TEPRs elicited by word matching in low versus high ambient light (i.e., Event Bin – Baseline Mean). The *y*-axis represents change in millimeters from baseline; the *x*-axis represents time post-target-onset. Panel **b** represents the same TEPR change scores plotted on uncorrected data (reflecting native baselines). The panel **c** scatterplot reflects the distribution of average baseline-corrected change scores across participants, where each data point represents the mean pupil dilation value across all items and time bins. The horizontal crossbars represent means of the group distributions for the different lighting conditions. Panel **d** represents the average baseline-corrected peak amplitude across participants, and panel **e** represents the average uncorrected pupil diameter across each participant's pupil time series.

scales in such a manner. Alternatively, the TEPR could scale linearly, independent of its baseline amplitude. Ultimately, the "correct" scaling method is an empirical question that remains paradoxically unresolved. We addressed this foundational methodological issue by characterizing response functions elicited across different target detection tasks varied by baseline amplitudes. TEPRs scaled independently of their baselines, consistent with relative stability of phasic dilation reported both by Peysakhovitch and colleagues (2015; Peysakhovitch et al., 2017) and earlier by Bradshaw (1969). Thus, detecting a tone shift or target word elicits similar absolute magnitude dilation whether the participant is sitting relatively low versus higher ambient light intensity. Empirical confirmation of a linear response profile conveys the potential for both promise and peril to cognitive pupillometry.

The peril of erroneous response scaling

The extent to which erroneous scaling procedures have impacted past work remains undetermined. Conversion

Table 3 Experiment 2 baseline-corrected task-evoked pupillary dilations across time

Bin#	Time (ms)	Bright		Mid		Dark	
		Mean	*SD*	Mean	*SD*	Mean	*SD*
1	0–250	– .011	.029	.000	.050	.005	.056
2	250–500	.007	.046	.006	.070	.021	.086
3	500–750	.069	.081	.060	.093	.077	.104
4	750–1,000	.115	.095	.119	.097	.121	.118
5	1,000–1,250	.119	.099	.130	.111	.151	.124
6	1,250–1,500	.121	.114	.131	.131	.155	.161
7	1,500–1,750	.106	.118	.120	.130	.131	.163
8	1,750–2,000	.065	.112	.088	.122	.088	.174
9	2,000–2,250	.031	.115	.059	.110	.060	.186
10	2,250–2,500	.029	.103	.046	.106	.023	.190
11	2,000	.018	.092	.030	.101	.016	.165
12	2,200	.020	.085	.032	.102	.008	.147

Amplitudes reflect dilation in millimeters from baseline (subtraction)

between linear and nonlinear metrics is not a straightforward algebraic transformation analogous to temperature (Celsius–Fahrenheit). The greatest potential for bias lies within: (1) responses proportionately scaled from a relatively constricted baseline (e.g., a proportional change from 1 to 1.5 mm is substantial) and (2) temporally extended time series with rising baselines, as would typically occur during a continuous performance task or through the buildup of interference in continuous recall over minutes. Reanalysis of past work using linear baseline correction procedures is warranted. However, such reanalysis must be undertaken on the original time series that preserves information about the original event baselines. For many studies, the lack of archival data compels replication.

Potential promise of standardizing a linear response function

The raw magnitude of a 1% pupil dilation initiated from a 6.0-mm baseline differs from a 1% dilation initiated from a 2.0-mm baseline. Yet, proportional baseline scaling cannot capture this difference. If the TEPR does indeed scale nonlinearly, then replication not only requires precise matching of luminance contours but the more intractable problem of matching baselines across different settings. Luminance is one factor among many. Individual differences in anatomy, physiological arousal, mood, caffeine intake, and even the perceived attractiveness of the examiner all conspire to influence the pupil baseline (Bradley, Keil, & Lang, 2012).

Consider, for example, the challenge of replicating the three seminal studies that launched cognitive pupillometry. Hess and Polt (1960, 1964) implemented proportional baseline correction but failed to report either the baseline pupil diameter (in millimeters) or the ambient light intensity in the testing room (Hess & Polt, 1960). Replication of Kahneman and Beatty's (1966) classic digit span experiment is impeded by a related challenge. Kahneman and Beatty reported raw change scores (in millimeters) but did not apply any baseline correction. If the pupil scales nonlinearly, then a true replication of Kahneman and Beatty would compel matching the countless number of factors required to match a grand mean baseline pupil diameter of approximately 3.70 mm (see their Fig. 2, p. 1584).

Kahneman (1973) acknowledged that replication within pupillometry is possible if the amplitude of the TEPR is independent of its initial starting point. Linearity grants an essential degree of freedom by justifying subtractive baseline scaling, a simple but powerful technique for normalizing all change scores from 0 mm, regardless of individual differences in anatomy, task demands, or discrepant luminance levels

Table 4 Experiment 2 Bayesian repeated measure ANOVA model comparison

Model comparison					
Models	*P*(*M*)	*P*(*M* | data)	BF_M	BF_{10}	Error%
Null model (incl. subject)	.200	6.058e –14	2.423e –13	1.000	
Luminance	.200	6.026e –14	2.410e –13	0.995	1.019
Time Bin	.200	.419	2.881	6.911e+12	0.260
Luminance + Time Bin	.200	.576	5.436	9.509e+12	1.014
Luminance + Time Bin + Luminance × Time Bin	.200	.005	0.021	8.575e+10	1.987
Analysis of effects					
Effects	*P*(incl)	*P*(incl | data)	$BF_{Inclusion}$		
Luminance	.600	.581	0.926		
Time Bin	.600	1.000	5.514e +12		
Luminance × Time Bin	.200	.005	0.021		

All models include subject

across laboratories.[3] Although it may be premature to abandon the tradition of cautious matching of luminance across laboratories, confirmation of a linear scaling function renders replication far more plausible.

Another potential benefit of linearity is characterization of a canonical response function. A stable pupillary response can potentially facilitate time series analyses now utilized in neuroimaging. In one such technique (i.e., convolution), a researcher or machine-learning algorithm evaluates the fit of an observed time series against a predicted time series, in which events reflect the canonical task-evoked pupillary response. This approach would be impossible if the pupil scaled nonlinearly.

Caveats

We modeled responses to target detection tasks, treating the stimuli as infinitesimally brief, on/off signals. Discrete stimuli appear to produce a canonical waveform. However, target detection is just one class of task-evoked pupillary response. Kahneman and Beatty (1966) evaluated pupillary dilation during a digit span task over a longer response interval, observing peaks at approximately 5 s post-stimulus-onset. It is unknown whether task-evoked pupillary responses to continuous tasks are unique or instead reflect temporal dispersion of a canonical phasic response function (e.g., Steinhauer et al., 2004, found an interaction between phasic pupil dilation and luminance during a sustained processing task). Thus, the present data are only generalizable to target detection. In addition, our choice to model a finite window from stimulus onset may mean we failed to capture later-occurring elements of the pupil response.

Construct validity is another source of uncertainty within our design. We operationalized the TEPR as phasic dilation time-locked to a stimulus. However, pupil dilation is only a proxy measure of phasic arousal. Since we did not measure phasic arousal directly, it is uncertain whether the observed pupillary responses index other physiological process such as the acoustic startle reflex (Exp. 1) or the pupillary light reflex (Exp. 2). This an especially important consideration in light of the complex interaction between phasic versus tonic arousal (Peysakhovich et al., 2017) and sympathetic versus parasympathetic nervous system pathways (Gilzenrat, Nieuwenhuis, Jepma, & Cohen, 2010). We perturbed baseline pupillary diameter by manipulating the intensity of ambient light, thus leveraging a parasympathetic process. It is conceivable that tonic arousal induces nonlinear phasic arousal (see also Gilzenrat et al., 2010).

[3] This does not absolve researchers from controlling illuminance within-subjects (e.g., matching luminance and stimulus complexity for Condition A vs. Condition B within session).

Another potential confound relates to a restricted range of illuminance. Figures 2 and 3 represent the baseline pupil diameters across lighting conditions. At the extremes of luminance, the human pupil can constrict to a pinprick or dilate to the size of a small marble. Figures 2 and 3 reflect the relatively narrow range of baseline pupil diameters we achieved across lighting conditions. We observed a mean uncorrected pupil diameter of 2.9 mm in the high-illuminance condition, relative to a mean of 3.5 mm in low illuminance. Although these baseline means statistically differed, it is possible that the equivalence was an artifact of limited variability in the range of illuminance. In pilot work, we attempted a more extreme range of illuminance and encountered several obstacles. In full darkness the eye tracker produced uninterpretable noise, a known limitation of dark eye tracking caused by narrow contrast between the pupil and iris (Holmqvist & Nystrom, 2011). In contrast, testing was impossible in highly intense lighting, because participants complained of significant discomfort and showed avoidance behaviors (e.g., squinting, blinking). Thus, equipment limitations and human factors constrained the range of pupil dilation. There are, however, unseen advantages to testing within this relatively narrow range of illuminance, including ecological validity and avoidance of idiosyncratic behavior of the pupil at extremes of light intensity. It is not inconceivable that pupil mechanics dictate a saturation point of luminance or pharmacologically induced dilation, at which TEPRs would be virtually unobservable. The analysis of pupillary behavior at midrange illuminance avoids such ceiling effects.

A final caveat relates to our use of pupillary diameter as the primary metric of evoked change, when in fact surface area is the true determinant (Binda, Pereverzeva, & Murray, 2013; Laeng et al., 2011; Peysakhovich et al., 2015). The validity of this measure rests on the assumption that the pupil is a perfect circle whose area is a straightforward algebraic transformation of diameter (πr^2). Pupil shape varies widely across the animal kingdom. The pupils of cats, snakes, and goats are slit-like, whereas cuttlefish pupils are shaped like the letter W (Greenfieldboyce, 2015). In contrast, the human pupil is typically regarded as round and is accordingly amenable to measurement using circle geometry. Wyatt (1995), however, cast doubt upon this assumption by reporting significant individual differences in pupil shape (e.g., elliptical), potentially compromising the validity of the circular geometry approach. Thus, surface area may yield a more accurate metric of pupil size and pupil change; we will revisit this point with the aim of resolving it.

Concluding remarks and future directions

Cognitive pupillometry is a well-worn technique, with over half a century in active use. Pupillary time series may appear simple relative to other psychophysiological signals (e.g.,

multivoxel analyses in fMRI). Yet the TEPR is a remarkably complex signal whose measurement and neural signal generators remain ambiguous. Our aim in this study was to resolve one particular measurement problem related to response scaling. During the review, many additional empirical questions were raised with respect to how pupil time series are processed and contrasted. Pupillometry lacks consensus and/or best practice guidelines for determining: (a) minimal thresholds that constitute meaningful differences between events, (b) binocular versus monocular pupil sampling, (c) temporal downsampling and smoothing procedures, (d) modeling TEPRs using diameter versus surface area of the pupil, and (e) objectively defining a peak range and/or dissociating peaks. Mathôt et al. (2018) recently contributed a great service to the field by proposing a formal set of guidelines that will help standardize baseline correction.

Author note We thank our reviewers, editor, and Timothy Shipley for showing us the light.

References

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

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

Beatty, J., & Lucero-Wagoner, B. (2000). The pupillary system. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of Psychophysiology* (2nd ed., pp. 142–162). Cambridge, UK: Cambridge University Press.

Binda, P., Pereverzeva, M., & Murray, S. O. (2013). Attention to bright surfaces enhances the pupillary light reflex. *Journal of Neuroscience*, *33*, 2199–2204.

Bradley, M. M., Keil, A., & Lang, P. J. (2012). Orienting and emotional perception: Facilitation, attenuation, and interference. *Frontiers in Psychology*, *3*, 493. https://doi.org/10.3389/fpsyg.2012.00493

Bradshaw, J. L. (1969). Background light intensity and the pupillary response in a reaction time task. *Psychonomic Science*, *14*, 271–272. https://doi.org/10.3758/BF03329118

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of present word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*, 977–990. https://doi.org/10.3758/BRM.41.4.977

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology*, *33A*, 497–505. https://doi.org/10.1080/14640748108400805

Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, *29*, 162–173.

Duñabeitia, J. A., & Costa, A. (2015). Lying in a native and foreign language. *Psychonomic Bulletin & Review*, *22*, 1124–1129. https://doi.org/10.3758/s13423-014-0781-4

Gilzenrat, M. S., Nieuwenhuis, S., Jepma, M., & Cohen, J. D. (2010). Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. *Cognitive, Affective, & Behavioral Neuroscience*, *10*, 252–269. https://doi.org/10.3758/CABN.10.2.252

Graham, R., Hoover, A., Ceballos, N. A., & Komogortsev, O. (2011). Body mass index moderates gaze orienting biases and pupil diameter to high and low calorie food images. *Appetite*, *56*, 577–586.

Greenfieldboyce N (2015) Eye shapes of the animal world hint at differences In our lifestyles. Retrieved July 8, 2018, from https://www.npr.org/sections/health-shots/2015/08/07/430149677/eye-shapes-of-the-animal-world-hint-at-differences-in-our-lifestyles

Hayashi, N., Someya, N., & Fukuba, Y. (2010). Effect of intensity of dynamic exercise on pupil diameter in humans. *Journal of Physiological Anthropology*, *29*, 119–122.

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

Hess, E. H., & Polt, J. M. (1960). Pupil size as related to interest value of visual stimuli. *Science*, *132*, 349–350.

Hess, E. H., & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science*, *143*, 1190–1192.

Holmqvist, K., & Nystrom, M. (2011). *Eyetracking: A comprehensive guide to methods and measures*. Oxford, UK: Oxford University Press.

Iqbal, S. T., Zheng, X. S., & Bailey, B. P. (2004). Task-evoked pupillary response to mental workload in human-computer interaction. In *CHI'04 extended abstracts on human factors in computing systems* (pp. 1477–1480). New York, NY: ACM Press. Retrieved from http://dl.acm.org/citation.cfm?id=986094

Janisse, M. P. (1974). Pupillometry: Some advances, problems and solutions. In *Pupillary dynamics and behavior* (pp. 1–8). Berlin, Germany: Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4757-1642-9_1

Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *Journal of Problem Solving*, *7*(1), 2. https://doi.org/10.7771/1932-6246.1167

JASP Team. (2017). JASP (Version 0.8.5). Retrieved from https://jasp-stats.org/

Johnson, A. (2016). Package "equivalence" (Version 0.7.2). Retrieved from https://cran.r-project.org/web/packages/equivalence

Kahneman, D. (1973). *Attention and effort* (Vol. 1063). Englewood Cliffs, NJ: Prentice-Hall.

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science*, *154*, 1583–1585. https://doi.org/10.1126/science.154.3756.1583

Kankipati, L., Girkin, C. A., & Gamlin, P. D. (2011). The post-illumination pupil response is reduced in glaucoma patients. *Investigative Ophthalmology and Visual Science*, *52*, 2287–2292.

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

Korn, C. W., & Bach, D. R. (2016). A solid frame for the window on cognition: Modeling event-related pupil responses. *Journal of Vision*, *16*(3), 28. https://doi.org/10.1167/16.3.28

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*, 299–312. https://doi.org/10.1177/1745691611406925

Lacey, J. I. (1956). The evaluation of autonomic responses: Toward a general solution. *Annals of the New York Academy of Sciences*, *67*, 125–163.

Laeng, B., Orbo, M., Holmlund, T., & Miozzo, M. (2011). Pupillary Stroop effects. *Cognitive Processes*, *12*, 13–21. https://doi.org/10.1007/s10339-010-0370-z

Laeng, B., & Sulutvedt, U. (2014). The eye pupil adjusts to imaginary light. *Psychological Science*, *25*, 188–197. https://doi.org/10.1177/0956797613503556

Lakens, D. (2016). *Equivalence tests: A practical primer for t-tests, correlations, and meta-analyses*. PsyArXiv preprint. https://doi.org/10.1177/1948550617697177

Loewenfeld, I. E., & Lowenstein, O. (1993). *The pupil: Anatomy, physiology, and clinical applications* (Vol. 2). Ames, IA: Iowa State University Press.

Mathôt, S., Fabius, J., Van Heusden, E., & Van der Stigchel, S. (2018). Safe and sensible preprocessing and baseline correction of pupil-size data. *Behavior Research Methods*, *50*, 94–106. https://doi.org/10.3758/s13428-017-1007-2

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

Moore, B. C. (1973). Frequency difference limens for short-duration tones. *Journal of the Acoustical Society of America*, *54*, 610–619.

Papesh, M. H., & Goldinger, S. D. (2015). Pupillometry and memory: External signals of metacognitive control. In *Handbook of biobehavioral approaches to self-regulation* (pp. 125–139). New York, NY: Springer. https://doi.org/10.1007/978-1-4939-1236-0_9

Peysakhovich, V., Causse, M., Scannella, S., & Dehais, F. (2015). Frequency analysis of a task-evoked pupillary response: Luminance-independent measure of mental effort. *International Journal of Psychophysiology*, *97*, 30–37. https://doi.org/10.1016/j.ijpsycho.2015.04.019

Peysakhovich, V., Vachon, F., & Dehais, F. (2017). The impact of luminance on tonic and phasic pupillary responses to sustained cognitive load. *International Journal of Psychophysiology*, *112*, 40–45.

Pfleging, B., Fekety, D. K., Schmidt, A., & Kun, A. L. (2016). A model relating pupil diameter to mental workload and lighting conditions. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5776–5788). New York, NY: ACM Press. https://doi.org/10.1145/2858036.2858117

Piquado, T., Isaacowitz, D., & Wingfield, A. (2010). Pupillometry as a measure of cognitive effort in younger and older adults. *Psychophysiology*, *47*, 560–569. https://doi.org/10.1111/j.1469-8986.2009.00947.x

R Core Team. (2013). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*, 301–308. https://doi.org/10.3758/s13423-014-0595-4

Siegle, G. J., Ichikawa, N., & Steinhauer, S. (2008). Blink before and after you think: blinks occur prior to and following cognitive load indexed by pupillary responses. *Psychophysiology*, *45*, 679–687.

Spitschan, M., Jain, S., Brainard, D. H., & Aguirre, G. K. (2014). Opponent melanopsin and S-cone signals in the human pupillary light response. *Proceedings of the National Academy of Sciences*, *111*, 15568–15572. https://doi.org/10.1073/pnas.1400942111

Steinhauer, S. R., & Hakerem, G. (1992). The pupillary response in cognitive psychophysiology and schizophrenia. *Annals of the New York Academy of Sciences*, *658*, 182–204.

Steinhauer, S. R., Siegle, G. J., Condray, R., & Pless, M. (2004). Sympathetic and parasympathetic innervation of pupillary dilation during sustained processing. *International Journal of Psychophysiology*, *52*, 77–86.

Szulewski, A., Roth, N., & Howes, D. (2015). The use of task-evoked pupillary response as an objective measure of cognitive load in novices and trained physicians: A new tool for the assessment of expertise. *Academic Medicine*, *90*, 981–987. https://doi.org/10.1097/ACM.0000000000000677

Tryon, W. W. (1975). Pupillometry: A survey of sources of variation. *Psychophysiology*, *12*, 90–93.

Tun, P. A., McCoy, S., & Wingfield, A. (2009). Aging, hearing acuity, and the attentional costs of effortful listening. *Psychology and Aging*, *24*, 761–766. https://doi.org/10.1037/a0014802

Wang, C.-A., & Munoz, D. P. (2015). A circuit for pupil orienting responses: Implications for cognitive modulation of pupil size. *Current Opinion in Neurobiology*, *33*, 134–140.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*, 1191–1207. https://doi.org/10.3758/s13428-012-0314-x

Weiss, M. W., Trehub, S. E., Schellenberg, E. G., & Habashi, P. (2016). Pupils dilate for vocal or familiar music. *Journal of Experimental Psychology: Human Perception and Performance*, *42*, 1061–1065. https://doi.org/10.1037/xhp0000226

Wilder, J. (1958). Modern psychophysiology and the law of initial value. *American Journal of Psychotherapy*, *12*, 199–221. https://doi.org/10.1176/appi.psychotherapy.1958.12.2.199

Wyatt, H. J. (1995). The form of the human pupil. *Vision Research*, *35*, 2021–2036.

Xu, J., Wang, Y., Chen, F., & Choi, E. (2011). Pupillary response based cognitive workload measurement under luminance changes. In *IFIP Conference on Human–Computer Interaction* (pp. 178–185). Berlin, Germany: Springer.

Zavagno, D., Tommasi, L., & Laeng, B. (2017). The eye pupil's response to static and dynamic illusions of luminosity and darkness. *i-Perception*, *8*, 2041669517717754. https://doi.org/10.1177/2041669517717754

Zekveld, A. A., Heslenfeld, D. J., Johnsrude, I. S., Versfeld, N. J., & Kramer, S. E. (2014). The eye as a window to the listening brain: Neural correlates of pupil size as a measure of cognitive listening load. *NeuroImage*, *101*, 76–86. https://doi.org/10.1016/j.neuroimage.2014.06.069

Zekveld, A. A., & Kramer, S. E. (2014). Cognitive processing load across a wide range of listening conditions: Insights from pupillometry. *Psychophysiology*, *51*, 277–284.