A Primer on Design and Data Analysis for Cognitive Pupillometry
<DRAFT CHAPTER>

Jamie Reilly[1,2], Bonnie Zuckerman[1,2], and Alexandra E. Kelly[3]

Reilly J, Kelly A, & Zuckerman B (forthcoming 2021). A Primer on Design and Data Analysis for Cognitive Pupillometry. In S Goldinger & M Papesh (Eds.), Modern Pupillometry. Springer Inc.

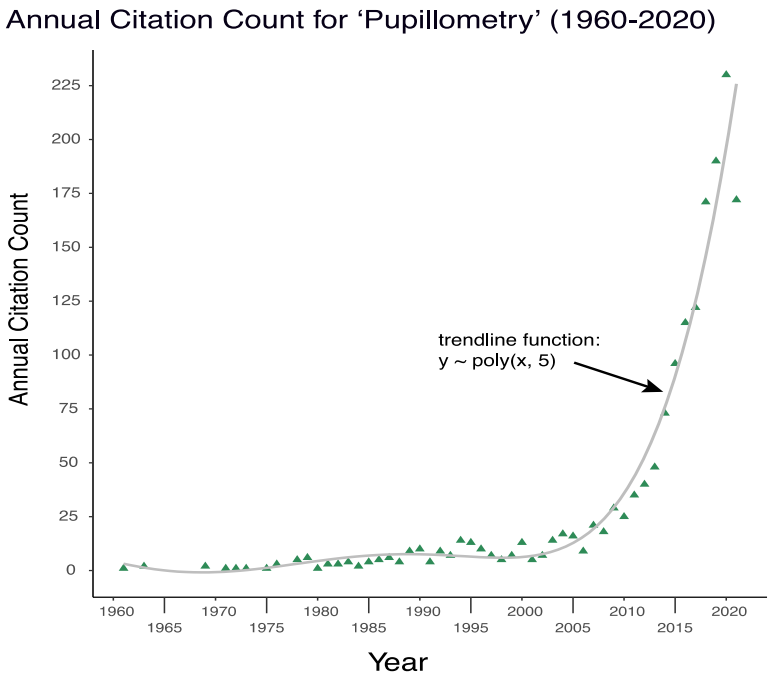
Authors' Note: This chapter was supported in part by a grant from the National Institute on Deafness and Other Communication Disorders (R01 DC0103063 to JR). Address correspondence to Jamie Reilly, PhD (reillyj@temple.edu)

[1]Eleanor M. Saffran Center for Cognitive Neuroscience, Philadelphia, Pennsylvania USA
[2]Department of Communication Sciences and Disorders, Temple University, Philadelphia, Pennsylvania USA
[3]Department of Psychology, Drexel University, Philadelphia, Pennsylvania USA

1. Introduction

This chapter is focused on cognitive pupillometry, a set of methods used for contrasting small perturbations in the surface area of the pupil evoked by cognitive demands. Cognitive pupillometry historically demanded expertise in optics, psychophysics, and engineering. Scientists were once compelled to construct their own elaborate systems for presenting stimuli and measuring pupil responses using analog film capture with sampling rates of about five images per second. Many of these early systems involved jury-rigged arrays of pulleys, mirrors, winches, and bite-bars. In contrast, infrared eyetracking systems are capable of remote recording with sampling rates exceeding 1000Hz. Much of the hardware used in modern pupillometry is commercially available and guided by intuitive point-and-click graphical user interfaces. In addition, researchers now have access to a variety of open source software packages dedicated to automated processing and statistical analyses of pupillometry data. These advances have made pupillometry a widely accessible tool whose popularity has extended beyond esoteric corners of psychophysics to a much broader range of applied cognitive science. Figure 1 illustrates this trend of exponential growth in citation counts for pupillometry in PubMed indexed articles across the interval 1960-2020.

Figure 1. PubMed Citation Counts by Year for Pupillometry-related Search Terms



Researchers unfamiliar with pupillometry may hold naïve biases about the complexity of this measurement tool. After all, pupillary data reflect the dynamics of a single channel fluctuating relatively slowly over time. How hard can that be? It may be true that collecting and analyzing raw pupillometry data is not particularly challenging. However, executing a valid and reliable cognitive pupillometry study is quite difficult. Our aims here are to introduce methodological challenges and identify solutions to common pitfalls in experimental design, execution, and analysis of this multifaceted neurobiological signal.

1.1 Physiological and behavioral indices of pupillary response functions

The morphology of the pupil and the robustness of its response to light are well-established markers of neurological and ocular pathology. For example, fixed and dilated pupils are a symptom of brainstem dysfunction incurred in severe head trauma and disorders of consciousness (Hoffmann et al., 2012; Jennett & Teasdale, 1977; Marmarou et al., 2007). Pinpoint pupils may indicate acute opiate intoxication or chemical pesticide exposure (Davies et al., 1975; Larson, 2008; Rengstorff, 1994; Rollins et al., 2014), and acute anisochoria (i.e., asymmetric pupil size) may suggest the presence of a unilateral brain tumor or glaucoma (Lam et al., 1987). These conditions highlight the utility of clinical pupillometry as a tool for inferring disease states within the eye(s) and/or the brain that guide differential diagnosis and medical management.[1] Clinical pupillometry typically involves measurement of macroscale features such as the shape of the pupil or its responsiveness to light. Many of these characteristics are observable with the naked eye or using simple handheld magnification (e.g., ophthalmoscope). In the chapter to follow, we will describe challenges involved in measuring a far more subtle pupillary response.[2]

Eckard Hess and James Polt (1960) introduced the English-speaking scientific community to a new neurophysiological response function. In this study, the authors filmed the pupils of six adult men and women as they viewed five photographs comprised of: 1) neutral landscape; 2) baby; 3) mother + baby; 4) partially nude female; 5) partially nude male. The remarkable finding was that small fluctuations in pupil size (assessed by % change) were evoked by "interest value" of the stimuli. Female participants showed the highest peak dilation when viewing the partially nude male photo, whereas male participants showed the opposite pattern. Crucially, Hess and Polt controlled illuminance during the experiment. This allowed the authors to isolate a response evoked by cognitive factors rather than light reflexes. Hess and Polt (1960) touted the far-reaching implications of this biological signal, and a star was born.

---

[1] For reviews of clinical pupillometry and applications see Barbur et al. (2004), Bremner (2009)

Throughout the latter half of the twentieth century, pupillometry research has gradually shifted focus from more nebulous constructs such as 'interest level' to better operationalized variables such as physiological arousal (Beatty, 1982; Bradley et al., 2008; Nassar et al., 2012; Peysakhovich et al., 2017). We have since learned that the human pupil dilates in response to a vast range of cognitive and perceptual challenges, including memory encoding and retrieval (Goldinger & Papesh, 2012, 2013; Papesh et al., 2012; Papesh & Goldinger, 2015), effortful listening while perceiving speech in noise (Kuchinsky et al., 2014; Van Engen & McLaughlin, 2018; Zekveld et al., 2010, 2011, 2014; Zekveld & Kramer, 2014), and difficulty manipulations during mental arithmetic (Causse et al., 2017). In fact, Tryon (1975) compiled a non-exhaustive list of 23 sources of pupil variation with the caveat that many more sources of variability lurk beneath the surface.

1.2 Cognitive pupillometry

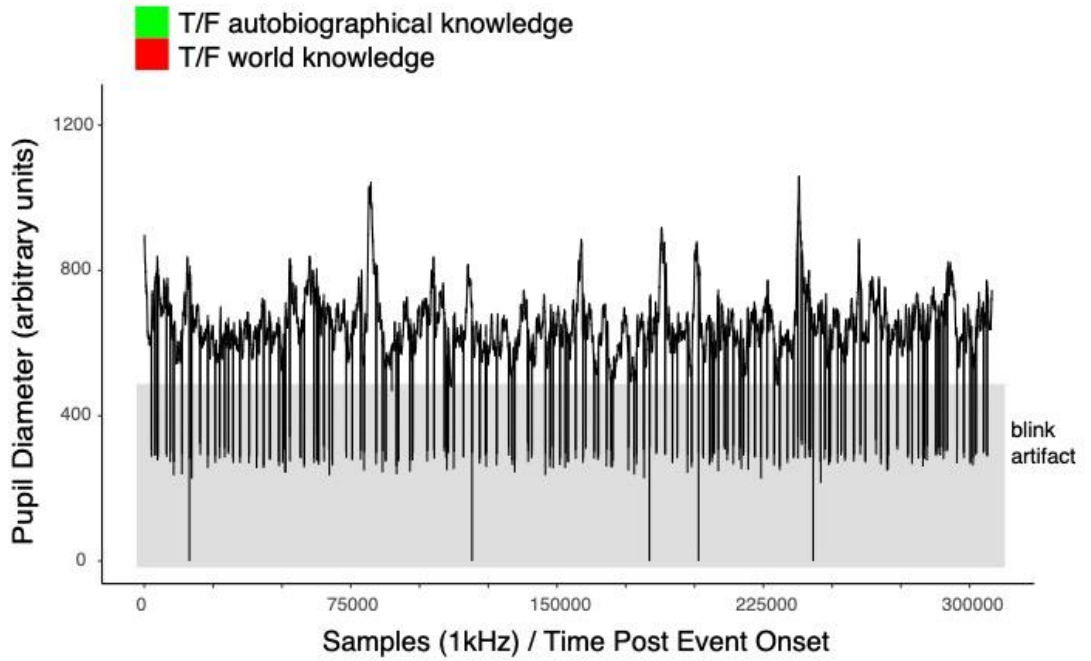
Pupillometry experienced a boom throughout the latter half of the twentieth century. Beatty (1982) coined the term task-evoked pupillary response or TEPR as a descriptive label for this particular response function. Many researchers have since adopted this nomenclature, and 'TEPR' remains in common use today. Nevertheless, the 'TEPR' distinction has not been met with universal acceptance. The 'task-evoked' component of the TEPR implies a discrete, exogeneous demand such as detecting a target word or solving a math problem. Yet, TEPR-like activation can also be observed in the absence of an external stimulus. Here we will collectively refer to such pupillary responses using the broader distinction of *cognitive pupillometry*.

Within the past twenty years, many disciplines including psychophysics, cognitive psychology, and cognitive neuroscience adopted pupillometry as a go-to tool. The promise of this technique was that non-invasive measurement of the surface of the eye could yield a proxy measure of brain activity within noradrenergic brainstem nuclei such as the locus coeruleus (LC) that modulate tonic and phasic arousal (Aston-Jones & Cohen, 2005; Gilzenrat et al., 2010; Mathôt, 2018; Wang & Munoz, 2015). Coupling between the pupil and LC has previously been demonstrated within macaque using combinations of both invasive and non-invasive neurophysiological recording techniques (Aston-Jones & Cohen, 2005; Joshi et al., 2016). Recent efforts have leveraged multimodal human neuroimaging with simultaneous pupil recording to elucidate the functional architecture of the ascending arousal network (Elman et al., 2017; Murphy et al., 2014; Wainstein et al., 2021).

Cognitive pupillometry experiments typically employ event-related designs wherein stimuli from two or more conditions are interspersed at jittered intervals. Consider, for example, a hypothetical experimental design where a psycholinguist wishes to test her hypothesis that verbs have higher processing demands than nouns in the context of background noise. She plans to test this hypothesis via a lexical decision experiment where participants hear nouns, verbs, nonwords, and filler words in the

context of interfering background noise as pupil size is continuously recorded. When she inspects the raw data, she will see a noisy, possibly non-stationary time series (i.e., rising or falling baselines) littered with missing and other complex artifacts. Figure 2 illustrates a raw pupillary time series continuously recorded from the left eye of a single subject in our own laboratory. In this particular example, the participant heard stimuli corresponding to two conditions. One condition reflected true statements about the world (e.g., Paris is the capitol of France.), whereas the other condition involved false statements about the world matched in length to the true statements (e.g., Paris is the capitol of Italy). Our aim in this ongoing study is to contrast pupil responses for true and false statements as a potential screening tool for language comprehension in severe brain injury. Figure 2 gives the reader a sneak peek at what raw pupillometry data look like sampled over a single six minute session for one participant. [3]

Figure 2. Raw single-subject data in an event-related pupillometry study



Note: These data reflect a session of a single-subject raw pupil time series sampled from the participant's left eye via an Eyelink 1000 Plus eyetracker (1000 Hz sampling rate). The color bar reflects points in the time series corresponding to event onsets within two experimental conditions. The experimental conditions involve true or false statements about the world versus the self.

[3] Raw data in pupillometry tend to approximate a hot mess (see Figure 2).

There is great heterogeneity across pupillometry studies as to what constitutes an event and how to window its subsequent pupil response. In a continuous pupillometry study, pupil dilation is modeled over an extended response interval such as during mental arithmetic (Hess & Polt, 1964; Klingner et al., 2011) or during creative problem solving tasks (Bradshaw, 1968; de Rooij et al., 2018). In contrast, discrete pupillometry involves time-locking a pupil response to an infinitesimally brief stimulus. This discrete approach is analogous to deconvolution techniques for analyzing hemodynamic response function (HRF) (Gitelman et al., 2003). In ERP and fMRI time series analyses, the stimuli are typically treated as delta functions. All subsequent activation is time-locked to that particular event onset/offset. This method has allowed researchers to develop mathematical basis functions for many biological signals. For example, the canonical HRF has a characteristic wave peaking about six seconds after event onset/offset followed by a slow decay to baseline after about 16 seconds (Handwerker et al., 2004).

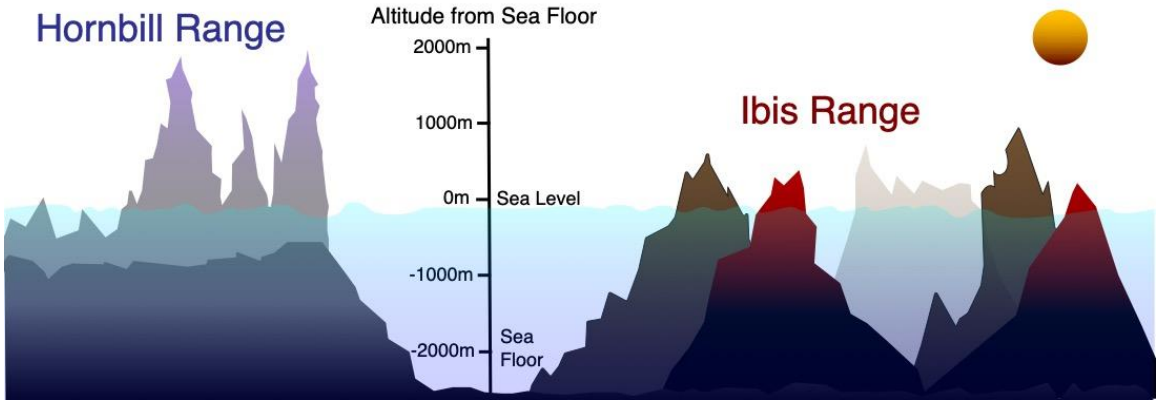
Much remains to be learned about the pupil response function including how its parameters are impacted by individual differences. However, some of its global features have gained mainstream acceptance as canonical in nature (Denison et al., 2020). In a typical pupillometry experiment, researchers contrast pupil change per unit time normalized to a 0mm baseline pupil size. When evoked change in pupil size (Y) is plotted against time (X), a canonical pupil response function bears resemblance to a time-compressed HRF or a slightly positively skewed mountain.

This chapter will focus primarily on optimizing study designs to evoke discrete pupillary response functions. Essentially, the analysis goal of cognitive pupillometry involves determining whether two or more composite mountains differ from each other. In the section to follow, we describe complexities, obstacles, and assumptions involved in the technical challenge of contrasting two composite mountains.

1.3 The two mountain problem: A measurement metaphor

Consider a scenario where you live on a 2-dimensional planet with two prominent mountain ranges. Figure 3 depicts the Ibis Range and the Hornbill Range. Mountain climbers flock to the Hornbill Range, whereas the Ibis Range is less frequented. Neither mountain range has been mapped, nor is it feasible to catalogue the entire population of mountains. Ibis Tourism Adventures, Inc. has hired you to substantiate their claim that the Ibis Range has “bigger and better” mountains than the Hornbill Range. Your budget for undertaking this investigation is enormous, and your client expects a full scientific report in six months. Where to begin?

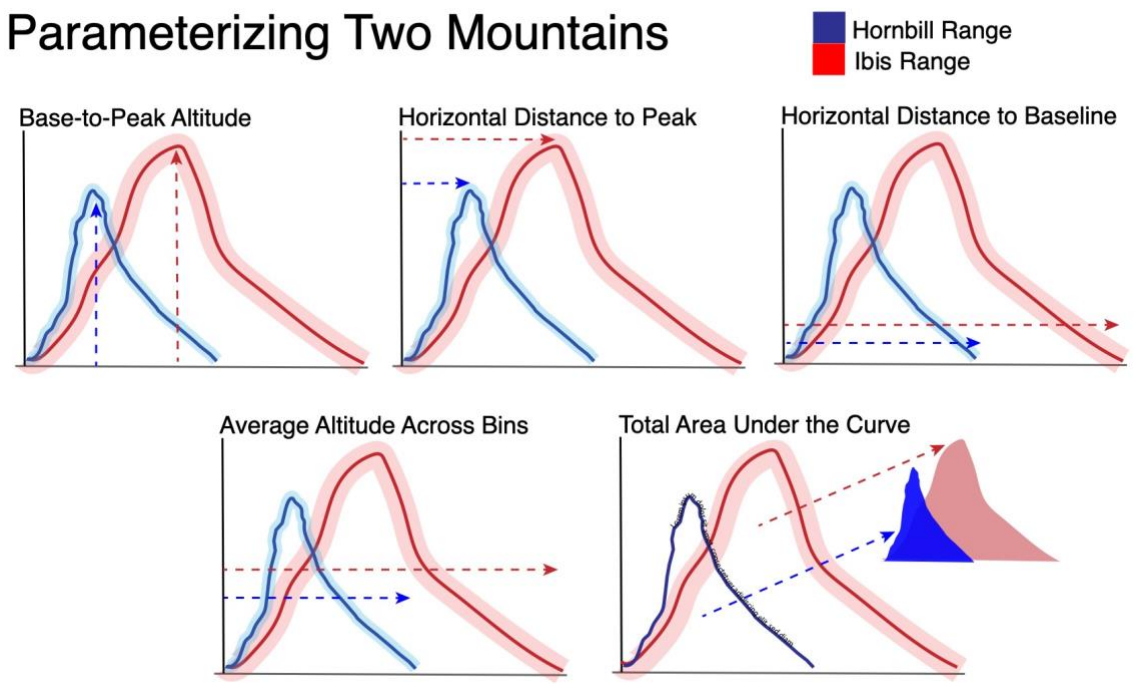
Figure 3. The Two Mountain Problem



You might start with a formal hypothesis test. Your null hypothesis (H_0) is that the Ibis and Hornbill ranges do not differ. Your alternative hypothesis (H_1) is that Ibis mountains are on average 'bigger and better' than Hornbill Mountains. Your next step is to operationally define 'bigger and better' in a principled manner that promotes both falsifiability and replication. You settle upon the following measurement parameters for 'bigger and better' illustrated in Figure 4:

1. Base-to-Peak Altitude: 'Height' of the mountain at its tallest point normalized for the initial altitude of the mountain (i.e., raw peak altitude minus raw initial altitude) (y_{max}).
2. Horizontal Distance to Peak: Horizontal distance from the origin (0,0) to the point on the x-axis corresponding to the mountain's peak altitude (x_{obs} at y_{max}).
3. Horizontal Distance to Baseline: Horizontal distance from the initial rise of the mountain (0,0) until it descends to baseline or alternatively plateaus.
4. Average Altitude: The average 'height' of the mountain across all x-values normalized for initial starting altitude (i.e., raw peak altitude minus raw starting altitude).
5. Area Under the Curve: Total approximate area of the mountain from initial rise to return to baseline.

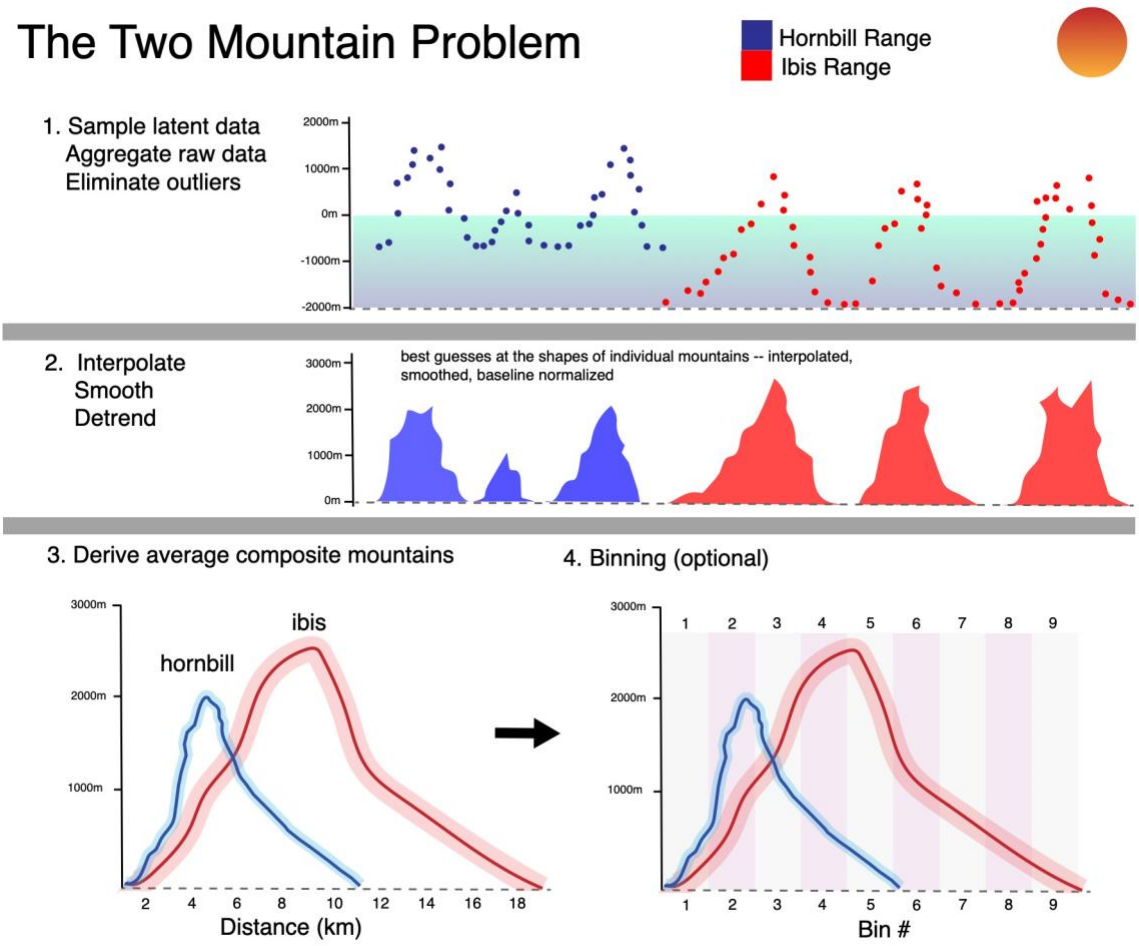
Figure 4. How to Measure a Mountain



At this point in your investigation, you established a guiding hypothesis, operationalized an abstract construct (e.g. 'bigger and better'), and identified a set of objectively measurable variables. You next devise sampling procedures that will yield accurate and unbiased estimates of the mountain ranges. GPS-equipped drones and remote submersibles are soon dispatched to the respective mountain ranges to sample altitude (y-axis) and global position (x-axis) at many points. Your hope is that extensive sampling will promote excellent source reconstruction and that your margin of measurement error will be small. However, these hopes are soon dashed when you learn of a high rate of equipment loss from winds, snow squalls, and giant squids. Your recovered data are punctuated by strange artifacts and missingness. What next?

Your next step involves data cleaning with the goal of removing outliers and imputing missing observations. Once these preprocessing steps are compete, you are now prepared to statistically evaluate your original hypothesis. You run statistical tests that are sensitive to autocorrelation (e.g., one point on a mountain is not independent of the previous point) and overfitting and interpret the results. Your data processing pipeline is illustrated in Figure 5.

Figure 5. A Data Processing Pipeline for the Two Mountain Problem



Despite being over 80% submerged, the Ibis Mountains have a higher base-to-peak altitude and greater total area than Hornbill Mountains. In contrast, Hornbill Mountains have steeper slopes. You report these findings to your client, and your job is done. Cognitive pupillometry shares many similarities with the Two Mountain Problem. When modeled as an event-related design, researchers typically examine relative change in pupil size evoked by discrete events. Each event in a pupillometry experiment corresponds to an individual mountain in the Two Mountain Problem. Many events together comprise an experimental condition.

1.4 Expertise and technical specialization in pupillometry

The engineering challenges of pupillometry are immense. A machine must detect near microscopic (~0.1mm) fluctuations of a slightly oblong disk simultaneously rotating within an eye and translating within a head. Hess & Polt (1960) evaluated changes in pupil size using an analog film camera to photograph the pupil less than five times per second. The authors then projected the images onto a screen and manually measured their dimensions. Pupillometry has since evolved an ever more sophisticated suite of hardware and software solutions for data acquisition. In addition, a number of inspiring teams of researchers have independently developed comprehensive open source pupillometry processing pipelines such as GazeR (Geller et al., 2019), CHAP (Hershman et al., 2019), and pupillometryR (Forbes, 2020). Each of these programs is free, well-supported, and transparent with respect to their processing and artifact correction algorithms.

Pupillometry has reached a level of sophistication where it is impractical if not impossible to master all of the nuances of the constituent processes (e.g., digital signal processing, optics, artifact detection). The good news is that you don't have to. Inexpensive eyetracking hardware, supportive user communities, and open source software have made pupillometry a widely accessible tool. Nevertheless, the benefits of automation are counterbalanced by a number of pitfalls. There are a great many ways to conduct a 'bad' pupillometry study (Loewenfeld & Lowenstein, 1993), even with advanced measurement and automated processing tools. Pupillometry is a methodological minefield with project-sinking danger at all stages from design through experiment execution and analysis to the global interpretation of results. An experimenter might spend months meticulously matching stimuli for illuminance and visual complexity only to overlook key individual differences (e.g., fatigue, substance abuse, motivation) or task-correlated covariates (e.g., blinks). Many of these sources of error are difficult to detect and some are impossible to retroactively correct. Biased data or faulty analyses can in turn support spurious conclusions. To follow, we focus on common threats to validity during design (before), execution (during), and analysis (after) of your pupillometry study.

2.0 Before your Pupillometry Study

2.1 Before: Theoretical considerations

Consider content and construct validity when debating the use of pupillometry as your dependent measure. Validity is typically regarded as a psychometric dimension in the development of behavioral scales and assessments. Yet, a critical evaluation of validity can also be a useful exercise in weighing the value of pupillometry as a source of evidence for your particular research question. Construct validity is the extent to which a particular tool accurately measures the latent concept it is intended to measure (Cronbach & Meehl, 1955; Sireci, 1998). For example, a wrongheaded emotion researcher might operationally define an abstract concept such as 'happiness' as 'the number of steps a person takes in a day'. This operational definition lends itself to reliable objective

measurement, but it has no construct validity. In contrast, content validity is the extent to which an instrument measures all components of a latent construct. A critical evaluation of validity forces us to answer two key questions:

a) What exactly do we hope to measure?
b) What exactly does pupillometry measure?

Construct validity has special relevance for pupillometry. When a researcher seeks to assess cognitive load, for example, she should provide a clear operational definition of this construct. A chronic lack of falsifiable and reproducible operational definitions has been a major shortcoming in pupillometry. Researchers have shifted the goalposts over time as to what the cognitive pupil response indexes (e.g., interest level, cognitive load, cognitive control, mental effort, mental workload, mental activity, resource allocation, attention changes, executive functioning, effortful mental activity, phasic arousal, adaptive gain, attentional change). One possibility is that the pupil responds selectively to each of these constructs and that these orthogonal signals all bottleneck at the pupil. A more plausible alternative hypothesis is that many of these constructs load on the same latent factor, i.e., physiological arousal. Thus, the pupillary system might be agnostic to subtle distinctions between cognitive load and cognitive control but instead fuels the metabolic demands required by each. In any case, researchers must consider the multifactorial nature of this neurobiological signal and exercise appropriate caution regarding the certainty of their inferences.

Pupillometry is typically used for confirmatory hypothesis testing. For example, in a study of listening effort and speech intelligibility, Zekveld and Kramer (2014) hypothesized that, "pupil dilation would be largest at medium intelligibility levels, and smaller in both easy conditions and in extremely difficult listening conditions resulting in cognitive overload". In this particular design, the null hypothesis (H_0) is that pupil dilation will not differ across the listening conditions. The importance of a clearly stated set of falsifiable hypotheses cannot be overstated. In the Zekveld and Kramer (2014) study, the independent variable was speech intelligibility, and the authors manipulated speech intelligibility by modulating the signal-to-noise ratio. From a philosophical standpoint, this is an outstanding manipulation because it is both content-valid and construct-valid.

Rigorous design in pupillometry should include formal hypothesis tests and validity checks. In an ideal world, experiment planning should also include principled stopping rules and well-justified sample size estimates. Power estimation remains a lacuna in pupillometry, however. Heterogeneity in measurement scales across different studies (e.g., pixels, arbitrary units, mm) and the lack of an extensive database of age-stratified norms has left researchers with few options other than to plan sample sizes by mimicking previous studies.

2.2. Before: Methodological Considerations

Numerous trait- and state-level individual differences moderate the cognitive pupil response. Trait variables known to impact pupillometry tend to remain relatively stable or evolve slowly over time (e.g., age, attention deficit disorder, memory span). In contrast, state factors involve behaviors which rapidly change in response to specific conditions (e.g., exaggerated startle reflexes to sudden sounds, agitation after consuming caffeine, migraine headache, etc.). Individual differences pose unique threats to validity especially when studying clinical populations whose levels of arousal, fatigue, motor coordination, affect, motivation, and other neurocognitive abilities fluctuate throughout the course of a day or on idiosyncratic medication dosage schedules. Thus, conducting a valid pupillometry study among special populations requires exhaustive consideration of a wide range of etiology-specific anatomical and neurobehavioral characteristics.[4] To follow, we focus on broad considerations for conducting cognitive pupillometry among neurotypical adults.

2.2.1 Before: Considering Trait-Level Individual Differences

Trait-level individual differences are sometimes the focus of pupillometry. For example, Tsukahura and colleagues reported significant correlations between baseline pupil size and fluid intelligence (a trait-level factor) (Tsukahara et al., 2016; Tsukahara & Engle, 2021). However, individual differences in trait-level variability are more often considered as 'noise' or obstacles toward the aim of making clean inferences about a particular experimental manipulation. A common method of controlling for trait-level differences involves specifying inclusion/exclusion criteria.

Age is trait-level variable known to influence the pupillary response. The levator muscles within our eyelids may stretch and weaken with age causing our eyelids to droop (Finsterer, 2003; Friedman, 2005). This phenomenon, known as ptosis, can present a conundrum for eyetracking especially for researchers who are unaware of it. In our own experience, eyetrackers tend to omit many observations and often misattribute ptosis to blink artifact. Correcting for ptosis during an experiment might involve simple steps such as recording from the less droopy eye or asking participants to open their eyes as widely as possible.

Another trait-level factor associated with aging is smaller baseline pupil size (Kim et al., 2000; Morris et al., 1997; Van Gerven et al., 2004). It remains an open question whether the amplitude of evoked pupil responses from this reduced baseline is dampened in aging. Some studies have reported higher task-evoked dilation for older adults (Piquado et al., 2010), whereas others have reported the opposite phenomenon

[4] For recent representative pupillometric studies in neuropsychiatric disorders see Burley, Snowden, & Gray (2019), Kries, Zhang, Moritz, & Pfuhl (2021), and Schneider, Leuchs, Czisch, Samann, & Spoormaker (2018).

(Gerven et al., 2004), an issue which also depends on the statistical approach taken (McLaughlin et al., 2021). In any case, researchers should be cautious when conducting between-subjects contrasts such as younger vs. older adults because their canonical pupil response functions likely differ in ways that are not yet fully understood.

Many pupillometry studies among neurotypical adults use a common set of exclusion criteria for trait-level factors. These include the presence of sensory deficits (e.g., sensorineural hearing loss), a history of neurodevelopmental disorders (e.g., dyslexia, specific language impairment, attention deficit disorder), ocular disease or trauma (e.g., cataracts), and neurological disorders (e.g., traumatic brain injury, stroke). Pupillometry researchers typically do not control for handedness as a proxy measure for language lateralization since it is assumed that the cognitive pupillary response is coupled across both eyes. Some researchers do, however, control for bilingualism. This decision tends to be motivated (however implicitly) by the need to control for the additional processing demands imposed by language proficiency and/or code switching.

Researchers interested in testing neurotypical participants have several options for confirming roughly normal global cognition. The first option involves asking participants to identify whether they have a history of learning disability or neurological disorder. It is also helpful to ask participants if they are currently experiencing difficulties in memory, language, or concentration. Although self-report can be expedient, it has limitations with respect to sensitivity and specificity. Participants might be unaware that they are experiencing declines in cognition (e.g., mild cognitive impairment), or they might fear stigma of disclosing an impairment.

A more rigorous alternative to self-report involves formally assessing global cognition using a standardized neuropsychological screening tool such as the Montreal Cognitive Assessment (MoCA) (Nasreddine et al., 2005) or the Mini Mental State Examination (MMSE) (Folstein et al., 1975). There are advantages and disadvantages to confirming normal cognition using a standardized measure. First, specificity is problematic for the MoCA because this particular measure tends to misclassify older African American adults as cognitively impaired with high false positive rates (Rossetti et al., 2017; Zahodne et al., 2017). As a consequence, your well-intended screening tool could introduce cultural and/or racial bias by screening out particular group(s) of people whose inclusion is essential for promoting representativeness.

Another ethical consideration involves follow-up for people who fail a cognitive screen. A common and significant fear in middle-aged and older-adults is memory loss and the onset of dementia (Kessler et al., 2012). Experimenters rarely consider the impact that being disqualified from a study might cause. Researchers should plan for this contingency and encourage participants to discuss subjective memory complaints with their primary care provider. In addition, we have found it helpful to provide people with

concrete resources such as contact information for helplines (e.g., Alzheimer's Association) and university clinics where they might receive free or low-cost treatment.[5]

2.2.2 Before: Considering State-Level Individual Differences

A dedicated pupillometry researcher must consider many state-level individual differences. People experience normal fluctuations in mood, arousal, and motivation throughout the day. These daily fluctuations are nested within longer cycles of maturation playing out over months and years. Some of this variability is predictable. For example, a pupillometry study conducted among college students at exam time is likely to involve elevated stress, sleeplessness, fatigue, and anxiety that could contaminate your results. Similarly, time of testing may play a significant role in the performance of people with insomnia. Although it is impossible to assess all possible trait-level factors, a useful strategy is to consider factors that impact a person's ability to intensely concentrate for a prolonged duration (e.g., alertness, fatigue, motivation, drug effects).

Wakefulness is an important consideration for planning the length and timing of a pupillometry study. Drowsiness produces an abnormal pattern of pupil oscillation known as hippus (Lüdtke et al., 1998; Wilhelm et al., 1998). Moreover, sleep-deprived participants tend to demonstrate decreased pupil diameters (Morad et al., 2000) and diminished pupillary light reflexes (Lowenstein et al., 1963) due to diminished sympathetic and parasympathetic innervation of the pupillary muscles. Possible methods for counteracting fatigue include structuring relatively brief testing sessions along with regular interaction from the experimenters. This performance boost might at least in part be attributable to the 'good-subject effect', the tendency for a person to perform better when seeking tacit approval from an observer (Nichols & Maner, 2008).

Many of us respond to unpleasant states by self-medicating. For example, people might counter headache pain with ibuprofen, fatigue with caffeine, and intermittent allergy symptoms with antihistamines. Each of these states (e.g., pain, fatigue, dry eyes) that originally motivated us to medicate can influence pupillary responses in isolation. In addition, drugs commonly used to treat these states also impact the pupil response, either through direct or indirect pathways. One example of a direct effect on pupillary behavior involves the action of anticholinergic medications such as diphenhydramine (a common ingredient in Benadryl and Nyquil) on the sphincter muscles of the iris resulting in dilation (mydriasis) (Harris et al., 1946; Jaanus, 1992). The same anticholinergic drug may produce indirect effects on the pupil response by inducing drowsiness. Thus, people who take common antihistamines to cope with allergy symptoms before a pupillometry study may experience unintended side effects such as drowsiness and over-dilated pupils. Other 'red flags' for state-induced pharmacological effects in pupillometry include opiate or ethanol intoxication (Larson, 2008).

---

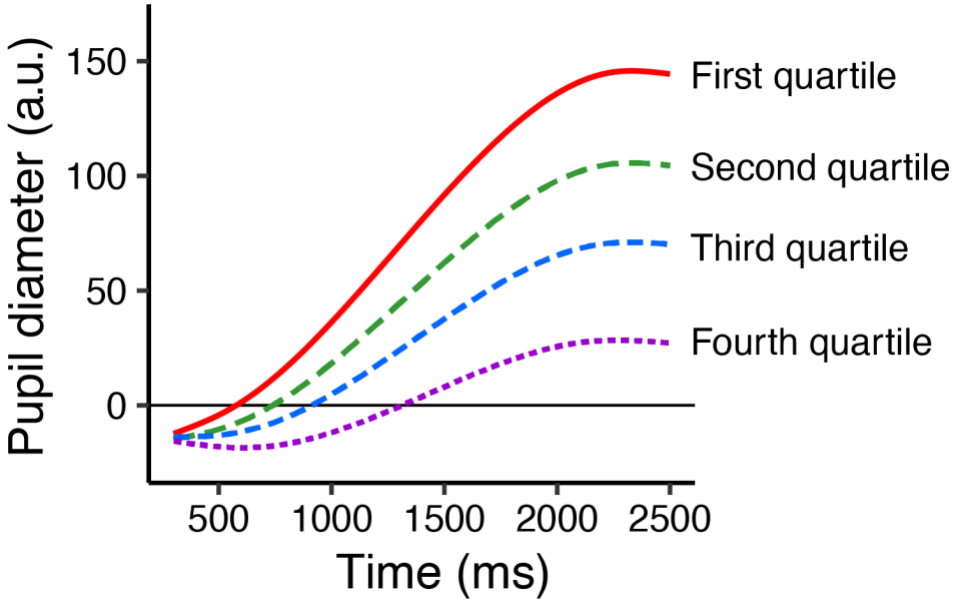
[5] This is important! Please contact the first author (reillyj@temple.edu) if he can assist with your efforts.

The range of pharmacological effects on nervous system function are substantial. Banning a drug such as caffeine solves one problem (e.g., no participants consumed caffeine) while creating an arguably worse problem (e.g., some participants experience early symptoms of withdrawal) (O'Shea & Moran, 2018). There is no clear guidance on how to approach the issue of pharmacological effects in pupillometry. Many researchers pursue a more pragmatic approach to drug exclusions. Participants are typically excluded if they are sedated and/or exhibiting altered mental status at the time of testing regardless of the drug or psychological/neurological etiology.

Other state-level factors are linked to stimulus anticipation and test anxiety. Some of these factors can be directly observed. For example, people may bite their lips, blink, or produce extraneous motor movements correlated with the onset of events. Some of these behaviors may be ameliorated by interacting with participants and adding a sufficient number of feedback-based training trials prior to conducting the true experiment. In our own experience, asking participants if they have questions or concerns, and how we can make them feel comfortable tends to put them at ease. These interactions and frequent check-ins can in turn result in better retention and higher data quality.

A final consideration that is often overlooked is how the task-evoked pupil response changes over the course of an experiment. Figure 6 shows model fits to pupil response throughout an experiment (D. McLaughlin, personal communication). The change from baseline is biggest in the beginning of the experiment, and smallest at the end. This change may reflect fatigue, habituation to the experimental paradigm, changes in tonic pupil response, or some combination of these factors. Regardless of the underlying cause(s), from a practical perspective, a response that diminishes over time is a challenge for studies in which changes in cognitive effort are of interest (for example, a perceptual learning or training study). Equally distributing experimental conditions throughout the experiment is also important, so that time is not confounded with experiment condition (if Condition A always comes first, and Condition B always last, a smaller response in Condition B might be due to changes in effort *or* simply because it occurs later, and later responses are smaller). Statistically accounting for time may provide better model fits and greater sensitivity to effects of interest.

Figure 6. Changes in task-evoked pupil response over the course of an experiment



2.2.2 Before: Considering Environmental Factors

'Let there be well-controlled light' should be the mantra of pupillometry. The pupillary system is highly photosensitive, and the dynamic range of pupillary dilation and constriction in response to light is hundreds of times larger than the magnitude of the cognitive pupil response. As such, uncontrolled luminance in the testing room and variable luminance contours within the stimuli themselves can easily eclipse the cognitive pupil response.

Recording pupillary data in extremely dark or intensely bright lighting conditions presents substantial challenges. Most pupillometry studies do not assess the full dynamic range of pupillary dilation and constriction. Instead, a more common laboratory setup involves testing participants who are seated at a computer monitor under fluorescent lighting. Winn and colleagues (2018) recommended maintaining ambient light intensities between 10 and 200 lux. Handheld light meters are inexpensive but essential tools for ensuring that light intensities fall within acceptable ranges. In addition to a moderate range of intensity, researchers should strive for relatively static lighting conditions over the course of an experiment. Testing in a windowless room can help attenuate idiosyncratic light and shadow. In our own lab, we test participants in a sound attenuated booth with blackout film covering the single window. Once the door to the sound booth is sealed, the only ambient light source is the computer monitor.

In addition to ambient lighting in the testing room, it is essential to control for systematic differences in the stimuli. For example, photographs of objects in one condition might simply be darker than another condition (e.g., snowy scenes vs. forest scenes). In this particular example, an experimenter interested in executive demands

would almost certainly expect to encounter higher evoked pupil dilation when people view photographs of dark forests. However, they could *not* attribute this response to the variable of interest (executive function). We can only be sure that the participants' pupils dilated more to forests because forest scenes tend to be darker than snowy scenes. Tools such as the SHINE toolbox for MATLAB (Willenbockel et al., 2010) are helpful for matching visual stimuli on luminance and other relevant variables (e.g., complexity).

Another method of controlling for systematic differences in luminance across two stimulus conditions is to avoid visual presentation altogether. Participants in our own pupillometry research hear stimuli (e.g., pure tones, curse words) while viewing a static gray background against which they view a central fixation cross (Reilly et al., 2018, 2020). Auditory presentation circumvents some visual confounds (luminance, accommodation), but it is not a panacea. Sudden loud, novel, and/or frightening stimuli are capable of triggering startle responses (Davis, 1984). Startle reflexes tend to be conserved across other mammals, and they are thought to represent an adaptation for promoting rapid escape (withdrawal) while protecting the eyes and neck (Geyer & Swerdlow, 1998) Moreover, control of central visual fixation remains crucial even in an auditory-only paradigm.

Researchers must also consider idiosyncratic factors unique to their own laboratories when planning their studies. Computer hardware and software constraints (e.g., sound/graphic cards, headphones, RAM) often impact the reliability of stimulus timing, buffering, and presentation. Anti-virus programs, firewalls, and automatic software updates can initiate unpredictably, resulting in heartbreaking data loss. In a busy lab with several different ongoing projects, experimenters often change software settings or switch peripherals (button boxes, joysticks). Participants themselves may be tempted to touch or adjust the monitor or infrared eyetracking sensor. Of course there is no way to plan for every possible contingency. Steps such as disabling computer updates and conducting pre-experiment diagnostics during pilot testing are integral for identifying and preventing many such issues.

2.2.3 Before: The Importance of Baselines

Pupillometry researchers are typically most interested in absolute change relative to some baseline. In a non-stationary time series, the baseline amplitude changes over time. Therefore, it is necessary to correct for the starting amplitude of each trial in order to eventually contrast the magnitude of evoked change from a uniform standard (0mm). Researchers have historically applied many different baseline correction techniques including non-linear corrections (% change) which assess relative change. Our laboratory empirically tested whether the pupil response scales linearly or non-linearly from different baseline amplitudes (Reilly et al., 2018). We manipulated baseline pupil amplitude by having participants complete tasks with comparable executive demands (e.g., counting pure tones) while situated in either dark or bright light. We discovered that

the pupil tends to scale linearly across different baselines and that subtractive scaling is warranted (e.g., Change – Baseline) (see also Mathôt et al., 2018).

Event-related pupillometry is almost always concerned with relative change in pupil amplitude from a neutral baseline. It is crucial to include a sufficient number of trials with their own baselines. A baseline period during an inter-trial interval should have no task demands. Its *raison d'etre* is to allow the pupil to settle into a steady state before the next event is initiated. In our own research, we tend to jitter inter-trial intervals so that event timing is slightly eccentric. During these intervals, participants view a static screen with a centrally positioned attention fixation cross. We later extract a 500 ms window of pupil size immediately preceding each event for use as that event's unique baseline, a point we will revisit in the analysis section later in the chapter. When designing your own study, mind your baselines.

2.2.4 Before: Be Prepared and Make Principled Decisions

Technicians must cope with numerous demands during an experiment, including monitoring the participant's wellbeing, troubleshooting technical issues, and ensuring that the experiment is proceeding as planned. Optimizing and standardizing eyetracker settings (e.g., sampling rate, which eye to sample from) beforehand will minimize the number of free parameters that will go wrong on testing day. Do not rely on the default settings of your eyetracker. Principled choices must be made before your study.

An often neglected step in planning involves ensuring that your measurement scale promotes replication. Some eyetrackers record pupil dimensions using arbitrary units or pixels, whereas other systems report pupil size using a standard metric scale (mm or mm^2). Please report pupil dimensions (diameter or area) in metric units. Since the pupil does not appear to scale non-linearly, measurements of % change from an arbitrary baseline (e.g., 8753 arbitrary units or 7653 pixels) are neither reproducible, nor particularly interpretable (Reilly et al., 2018) . In contrast, reporting evoked change in millimeters facilitates replication efforts across any eyetracking system.[6]

Another fundamental consideration is how researchers should report pupil size. Studies perhaps most often report evoked pupil change in terms of pupil diameter. The validity of pupil diameter is premised on the assumption that the human eye is roughly spherical and that on-center measurement of the pupil approximates a flat disc. This assumption justifies a simple transformation between diameter and surface area ($\pi \times d$). In reality, the human eye is slightly aspherical (Binda et al., 2013; Laeng et al., 2011).

[6] Conversion from arbitrary units or pixels to mm can be achieved by measuring a known reference such as an artificial eyeball. Our laboratory had no idea where to acquire an artificial eyeball, so we recorded diameter (in arbitrary units) of several black dots printed on cardstock and taped to eyeglasses at a fixed distance of 60 cm from the sensor. Prior to tracking the dots, we confirmed their diameters (e.g., 3mm, 4mm, 5mm, 6mm) using a measurement caliper. See your eyetracker's documentation for a recommended conversion technique. Sometimes this is a matter of simply checking a box for mm as preferred output.

Another principled decision involves determining an appropriate sampling rate. Hess and Polt (1960) sampled about four times a second, whereas many eyetracking systems today sample in excess of 1000 Hz. There is no consensus as to what constitutes either an optimal sampling rate or thresholds for reliable minima/maxima. A general principle borrowed from digital signal processing is that higher sampling rates yield better source reconstruction. Nevertheless, a sampling rate of 1000 Hz may be unnecessarily high for reconstructing the pupillary response function. Higher sampling rates are not always 'better' in that oversampling can incur costs both in terms of data proliferation and model overfitting. Researchers must balance these costs/benefits in selecting a sampling rate. Option 1 (not recommended) is to rely on the default sampling rate for your eyetracker and hope for the best. Option 2 is to reduce the default sampling rate and acquire data at a more plausible rate of biological change (e.g., 1000Hz to 250Hz). Option 3 is to downsample the raw data after acquisition through procedures such as binning. Option 4 is to retain all of the original data (e.g., 3000 observations per 3000ms event) and analyze evoked change using growth curve modeling, being cautious to avoid overfitting (Mirman, 2014).

3. During your Pupillometry Study

You have settled on a sampling rate and measurement scale that will promote replication. You have successfully screened and consented your participant who is comfortably positioned in a chinrest. You have cycled through your custom "pre-flight" checklist and ensured that all eyetracking settings are accurate as planned. It's go time! This section will cover considerations for ensuring high quality data collection during a pupillometry study.

3.1. During: Minimize Movement Artifacts

Analog film capture during the early days of pupillometry required rigid and prolonged head stabilization. Many remote infrared eyetrackers today are capable of recording data without any head stabilization whatsoever. Advances in motion compensation have yielded unprecedented flexibility in testing populations for whom head stabilization or restraint is impossible. Nevertheless, remote eyetracking is not without cost in terms of sacrificing data quality. Researchers must continue to minimize head motion as much as possible. An ophthalmological chinrest seems to represent a reasonable compromise between no head stabilization and *A Clockwork Orange* style restraint system. Many chinrests clamp to the side of a desk or table. If times are hard, you can even 3D print your own chinrest (Murphy, 2019) https://github.com/nimh-nif/SCNI_Toolbar/wiki/RestEasy:-An-open-source-chin-rest-for-human-psychophysics-experiments).

Gaze tracking studies are typically designed to analyze patterns of visual fixation and saccade dynamics as our eyes move across a screen. These large eye movements are

antithetical to pupillometry where design must minimize pupil foreshortening artifacts. That is, the most precise measurement of the pupil occurs while our eyes remain centrally fixated. Stimuli presented at larger visual angles tend to result in unreliable estimates of pupil size because the pupil is no longer spherical at oblique angles. Although several corrections have been proposed (Gagl et al., 2011; Brisson et al., 2013; Hayes & Petrov, 2016), there are also practical steps to minimizing peripheral eye movements. Visual stimuli should be presented centrally and within a narrow visual angle. Allow participants frequent breaks and ensure that they are engaged with the task (i.e., mind wandering = gaze wandering).

3.2. During: Minimize Fatigue and Maximize Engagement

Many of us find our own research fascinating. Fewer of us consider how boring our experiments might be for participants. Boringness is not a trivial consideration for data quality in pupillometry due to the pupillary system's sensitivity to an interaction between cognitive effort, arousal, and reward motivation (Aston-Jones & Cohen, 2005; Gilzenrat et al., 2010). Boring and repetitive tasks that tax sustained attention can induce fatigue and ultimately task disengagement (Granholm et al., 1996). When participants disengage from a demanding task and enter a resting state, they typically show a corresponding reduction (or absence) of task-evoked pupil dilation (Franklin et al., 2013).

Participants disengage for many reasons during an experiment. A common cause is when task demands exceed cognitive capacities. One of the earliest and most robust confirmations of this phenomenon was reported for memory encoding in immediate serial recall of digits (Beatty & Kahneman, 1966). Several studies in the subsequent decades have replicated this effect (Johnson et al., 2014), confirming that when memory load, as indexed by list length, is manipulated, participants show systematically larger pupil dilation during encoding as list lengths grow (e.g., to a length of approximately 7 digits) until maximum span of immediate recall (e.g., N=9 digits) is surpassed. That is, once participants disengage from the problem, pupil amplitude ceases to increase. A researcher who only considers the raw data might come to the erroneous conclusion that challenging math problems are no more demanding than simple math problems.

Another possible cause of disengagement is motivation. If there is no intrinsic reward to the participant, she might not care enough to invest her full effort. When participants are left alone in a testing room and exposed to long runs of boring stimuli, one can hardly blame them. Regardless of intrinsic motivation, keep sessions brief and provide ample breaks. For projects that require multiple testing sessions for the same participant dispersed over different days, consider scheduling each visit for roughly the same time if possible (see also Veneman et al., 2013). This will reduce potential confounds associated with normal fluctuations in arousal people experience throughout the day.

4.0 After your Pupillometry Study: What Next?

Meticulous planning and solid execution have gifted you with useable data. However, even under the best circumstances your data will require substantial preprocessing before statistical contrasts are possible. The section to follow covers steps in transforming a noisy raw time series into a series of smoothed, aggregated mountains. Data cleaning and analysis in pupillometry represents a dynamic area of methods development. To follow, we focus on some of the major considerations in developing a customized data processing and analysis pipeline.

4.1 After: Data inspection and Outlier Identification

Inspect your raw pupillometry data. Then inspect it again. Your output should include event codes (e.g., trial number), timestamps, and pupil size data at a bare minimum. Since you are most likely human, your ability to detect subtle trends within hundreds of thousands of numbers is probably limited. Plot your raw data as an uncorrected time series. Visual inspection of each session and every participant is essential for identifying both global and local artifacts.

One advantage that pupillometry researchers have is that the pupil's approximate dynamic range has known anatomical constraints. The human pupil diameter varies between approximately 2.0 mm and 9.0 mm in extremely bright and dark lighting conditions (Loewenfeld & Lowenstein, 1993; Wang & Munoz, 2015). Observations falling outside of this dynamic range are anatomically impossible and must, therefore, reflect artifact. These thresholds provide benchmarks for blink detection. When the eyelid briefly occludes the pupil during a blink, eyetrackers typically record pupil size as rapidly dropping to 0 mm. An abnormally high rate of constriction coupled with complete occlusion together indicate the presence of a blink (Hershman et al., 2018, 2019). Innovative methods of blink correction involve both detection of the blink event and the removal of 'pathological' observations preceding and following the blink. Blinks are not the only cause of signal dropout. People sneeze, cough, and look away from the infrared sensor when distracted. Each of these artifacts produce gaps in an otherwise continuous time series. In the next section we outline several techniques for filling these gaps.

4.2. After: Random vs, Systematic Data Missingness

Consider yet another scenario. You are an exterminator hired by a statistician to evaluate termite damage on a staircase. You must supply a damage report, but your client is also curious whether the damage is systematic or random. Your pest removal training has not prepared you well for this moment, but you reason that a random termite attack would result in a roughly uniform distribution of termite damage across the entire staircase. In contrast, a systematic attack would involve focal or concentrated attacks on specific subsections of the staircase. Staircases and pupillary time series both exemplify systems that can tolerate more random than systematic missingness. Pupillometry

researchers must accordingly consider both the source and extent of randomness within their data.

Intermittent coughing, sneezing, or unscheduled fire alarms likely constitute more random than systematic error. In a well-powered design, random missingness should not bias one condition over another because random noise is absorbed evenly across all experimental conditions. Similarly, the structural integrity of a staircase can survive more diffuse than focal (systematic) termite damage. Random signal dropout does, however, have limits. When termites decimate enough wood, the entire staircase is susceptible to collapse. Similarly, pupillary time series that require extensive imputation are subject to bias. There is no universal standard for a threshold of random data loss in pupillometry.

Non-random or systematic missingness can be especially difficult to control. Two common systematic artifacts include mental imagery and blinks. When people are engaged in spatial problem solving and/or visual imagery, their eye movements and pupil dynamics are both impacted (Grant & Spivey, 2003; Just & Carpenter, 1985; Laeng & Sulutvedt, 2014; Mathôt et al., 2017; Thomas & Lleras, 2007; Zavagno et al., 2017). When people engage in mental rotation or other complex working memory tasks, they experience changes in pupil size along with idiosyncratic eye movements. These perturbations are likely the result of perceptual simulation (i.e., visual imagery) rather than cognitive load. Researchers should be aware of these confounding effects when considering inferential validity.

Blinks represent another complex artifact in pupillometry. The frequency and duration of blinks are strongly correlated with task demands (Siegle et al., 2008) and blink rate provides a complementary index of cognitive load in its own right (Chen & Epps, 2014; Recarte et al., 2008). Although a well-designed pupillometry study should seek to minimize blink artifact, blinks are inevitable. In the section to follow, we discuss methods for correcting blinks and considerations for preventing blinks.

4.3. Artifact Detection

Even after careful planning and controlling for systematic error, your raw pupillary data will benefit from cleaning (Mathôt, 2018). As individual laboratories gain experience with cognitive pupillometry, many develop their own custom cleaning pipelines. Although these approaches differ in some respects such as the order of operations, they also share commonalities including artifact detection, imputation of missing values, and baseline correction.

One of the first steps in cleaning a raw pupillary time series involves filtering impossibly high and low values for raw pupil size and rate of change (i.e., acceleration). A simple approach to gross artifact rejection involves applying a bandpass filter to the raw pupil time series using known anatomical constraints on the dynamic range of the human pupil. For example, a bandpass filter with lower and upper bounds of 2 mm and 9 mm respectively will omit both impossibly low and high observations outside the known dynamic range of the pupil. Mind the distinction between missing values and zero when

applying filters or other artifact rejection procedures, When participants close their eyes or look away from an infrared sensor, eyetrackers often register a rapid drop in pupil size to 0 mm. Since no pupil has a 0 mm diameter, impossible observations must be omitted. Marking these data as missing (e.g., NA) is necessary for data imputation.

Some pupillometry artifacts are more subtle and will likely survive a bandpass filter. These artifacts necessitate special detection and correction techniques. Blinks compromise data acquisition both while the eyelid has completely occluded the pupil (all 0 mm observations) but also during the intervals surrounding the blink when the eyelid is closing and re-opening. A simple blink artifact rejection requires both replacing 0 mm observations with NA but also replacing leading and following observations within a temporal window surrounding the blink (Geller et al., 2019),. Blinks represent one such case where researchers have proposed artifact detection algorithms (Hershman et al., 2018; Kret & Sjak-Shie, 2019). The idea behind this correction approach is that the blink and the intervals surrounding the blink can be identified through a change in dilation speed (Kret & Sjak-Shie, 2019) or monotonic pattern (Hershman et al., 2018). Blink correction is built in to programs such as GazeR (Geller et al., 2019), CHAP (Hershman et al., 2019).

Recall that pupil measurement can deform when the eye rotates at oblique angles (Brisson et al., 2013; Gagl et al., 2011; Hayes & Petrov, 2016). These pupil foreshortening artifacts are not always detectable upon visual inspection of a raw pupil time series. Although complex mathematical corrections exist for off-center gaze, stimulus arrays should minimize visual angle as much as possible.

4.2 Making things whole again: Data imputation

Artifact rejection produces gaps in an otherwise continuous univariate time series. The purpose of imputation is to replace these missing values with reasonable estimates of what might have been. Pupillometry as a univariate time series represents a unique challenge for imputation. Since pupillary data are autocorrelated, the most accurate estimate of any missing observation can be derived by looking to its neighbors. For example, my best forecast of the high temperature in Philadelphia tomorrow is derived by the high temperature in Philadelphia today $\pm$ a small fudge factor. Single imputation techniques employed in randomly sampled datasets (e.g., replacing missing values with sample means or medians) are inappropriate for estimating missing values because of non-independence between successive observations. Pupillometry researchers typically approach the imputation problem by applying various forms of interpolation (e.g., linear, spline) over their datasets.

The general idea of interpolation (e.g. linear, cubic, spline) is that gaps in a non-continuous time series are estimated by ‘connecting’ the datapoints which endcap each side of the gap. The most simple type of interpolation (i.e., linear) involves plotting a straight line between endpoints and distributing all missing point estimates equidistantly along the line.

4.3. Smoothing Pupillary Time Series

There exist various smoothing algorithms, but all share the same overarching purpose. Smoothing reduces point-to-point variability across a time series. In our own work, we smooth the data using a simple moving average (window size=5). That is, each new observation is averaged with *n* observations surrounding it to yield a new time series. Larger window sizes produce smoother time series. Thus, what starts as a jagged but continuous mountain range should appear more rounded after smoothing. Although there are no recommended minimum/maximum thresholds for window size (n), researchers should exercise caution in specifying larger windows that obscure meaningful variability.

4.4 Baseline Correction and Event Extraction

One of the most common pupillometry designs involves comparing absolute differences in evoked pupil dilation within one or more conditions. Since pupil responses are typically initiated from different baseline amplitudes (e.g., 3mm vs. 6mm), it is necessary to normalize evoked change from a uniform standard (0 mm). Subtractive baseline correction is typically implemented with the rationale that baseline pupil size varies within and between individuals, effectively constituting non-stationary time series.

Our own research has shown that the pupil likely scales linearly independent of the baseline amplitude (Reilly et al., 2019). This property of the pupil response is best modeled using a linear scaling technique (i.e., subtractive correction). We typically conduct subtractive baseline correction by computing an average pupil diameter over the 500 ms neutral interval preceding an event. We subtract this median baseline pupil size from each subsequent observation extending outward from the event onset (0 ms) to a specified window of 3000 ms. This procedure normalizes each event to its own baseline, allowing flexibility in baseline pupil size over a session. If you apply dynamic baseline correction using this method, there is no need for detrending because all events are normalized to the same starting amplitude (0 mm).

During the final stages of preprocessing, you will extract all baseline-corrected events from the time series. You must specify the duration over which you will window the pupil response function. The event window must be of a sufficient duration to characterize the rise and fall of the pupil response. However, the event window must not be so long that it captures rest or anticipation of the next subsequent stimulus. Event duration is typically constrained both by stimulus duration and the nature of the canonical pupil response.

4.4 Data Analysis: Contrasting Two Mountains

The past decade has seen advances in the sophistication, power, and complexity of pupillometry data analysis. The vices and virtues of these particular statistical models

(e.g., generalized additive mixed models) are beyond the scope of this primer. Nevertheless, pupillometry does pose challenges for any statistical model you ultimately pursue. We address some of these issues to follow.

We have identified five parameters that characterize a two-dimensional mountain (see Figure 3). These include: peak amplitude, time-to-peak, sustained amplitude, area under the curve, and time-to-decay. Researchers face a dilemma in selecting which parameter(s) to contrast because no comprehensive neurobiological model of these parameters exists. As such, we have a limited understanding of how the canonical pupil response function might shift under different neural and behavioral challenges across the lifespan. Reciprocal relationships between brainstem activation, arousal, and pupil dilation have motivated an intense focus on peak amplitude and time-to-peak as parameters of interest in pupillometry. Nevertheless, great variability exists in how peak amplitude is derived (Tun et al., 2009) and indeed whether the pupil response might better be modeled using a two-peak solution with an early peak at approximately 600ms and a late peak at 1200ms (Steinhauer & Hakerem, 1992).

5.0 Concluding remarks, unanswered questions, future directions

A proliferation of low cost eyetracking hardware along with a supportive user community have seeded exponential growth in the popularity of pupillometry. In turn, pupillometry has recently generated remarkable insights into cognition, consciousness, and mental imagery (Laeng et al., 2012; Laeng & Sulutvedt, 2014; Mathôt et al., 2015; Zavagno et al., 2017; Mathôt et al., 2017). The advent of automated software applications has made pupillometry accessible to a wide range of disciplines not typically steeped in psychophysiology. Our goal in this primer is to highlight complexities of measuring this biological signal for non-experts.

There are many factors to consider if your aim is to conduct a rigorous pupillometry study. Newcomers to this technique might be frustrated to learn that there are no uniform best practices for design, execution, and analysis of pupillometry, although this does appear to be changing (Winn et al., 2018). Pupillometry often reflects a long-term commitment to learning the nuances of this technique over time. Our own laboratory has stumbled upon many of the pitfalls discussed here through trial-and-error over about a decade. A preventable error as simple as a technician switching eyetracker settings between studies has resulted in data loss for our lab on several occasions. We have also learned firsthand the necessity to monitor fatigue and to double check stimulus delays and response logging. Checklists are essential. Develop your own and share them with other researchers.

In conclusion, much remains to be learned about pupillometry. New methods of design and time series analysis hold promise for improving the rigor of pupillometry. Yet, the field remains limited by an anemic history of replication (but see de Winter et al., 2021) and the lack of age-stratified norms against which effect sizes might be derived. A deeper understanding of the neural and neuromuscular substrates of the

cognitive pupil response function is essential for understanding which cognitive processes are indexed by specific parameters and how these parameters might be selectively perturbed.

Acknowledgements

We thank Jason Geller, Jonathan Peelle, Drew McLaughlin, and members of the Concepts and Cognition Laboratory at Temple University. We are grateful to all of the scientists who have developed open source software applications for pupillometry. Your immeasurable support represents the best spirit of science. This work was supported in part by a grant from the US National Institute on Deafness and Other Communication Disorders (NIH/NIDCD DC013063)

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.

Barbur, J. L. (2004). Learning from the pupil-studies of basic mechanisms and clinical applications. *The Visual Neurosciences*, *1*, 641–656.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin, 91*(2), 276–292.

Beatty, J., & Kahneman, D. (1966). Pupillary changes in two memory tasks. *Psychonomic Science, 5*(10), 371–372.

Binda, P., Pereverzeva, M., & Murray, S. O. (2013). Attention to bright surfaces enhances the pupillary light reflex. *Journal of Neuroscience, 33*(5), 2199–2204. https://doi.org/10/f4j584

Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, *45*(4), 602–607.

Bradshaw, J. L. (1968). Pupil size and problem solving. *Quarterly Journal of Experimental Psychology*, *20*(2), 116–122. https://doi.org/10/cgmsd6

Bremner, F. (2009). Pupil evaluation as a test for autonomic disorders. *Clinical Autonomic Research*, *19*(2), 88–101. https://doi.org/10/dk64df

Brisson, J., Mainville, M., Mailloux, D., Beaulieu, C., Serres, J., & Sirois, S. (2013). Pupil diameter measurement errors as a function of gaze direction in corneal reflection eyetrackers. *Behavior Research Methods*, *45*(4), 1322–1331. https://doi.org/10/f5kq7b

Burley, D. T., Gray, N. S., & Snowden, R. J. (2019). Emotional modulation of the pupil response in psychopathy. *Personality Disorders: Theory, Research, and Treatment*, *10*(4), 365. https://doi.org/10/ggrjqp

Causse, M., Peysakhovich, V., & Mandrick, K. (2017). Eliciting sustained mental effort using the Toulouse N-back Task: Prefrontal cortex and pupillary responses. In *Advances in Neuroergonomics and Cognitive Engineering* (pp. 185–193). Springer.

Chen, S., & Epps, J. (2014). Using Task-Induced Pupil Diameter and Blink Rate to Infer Cognitive Load. *Human–Computer Interaction*, *29*(4), 390–413. https://doi.org/10/ghs95n

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281. https://doi.org/10/dcsjjf

Davies, J. E., Barquet, A., Freed, V. H., Haque, R., Morgade, C., Sonneborn, R. E., & Vaclavek, C. (1975). Human pesticide poisonings by a fat-soluble

organophosphate insecticide. *Archives of Environmental Health: An International Journal*, *30*(12), 608–613. https://doi.org/10/gmhnzh

Davis, M. (1984). The mammalian startle response. In *Neural mechanisms of startle behavior* (pp. 287–351). Springer.

de Rooij, A., Vromans, R. D., & Dekker, M. (2018). Noradrenergic Modulation of Creativity: Evidence from Pupillometry. *Creativity Research Journal*, *30*(4), 339–351. https://doi.org/10.1080/10400419.2018.1530533

de Winter, J. C. F., Petermeijer, S. M., Kooijman, L., & Dodou, D. (2021). Replicating five pupillometry studies of Eckhard Hess. *International Journal of Psychophysiology*, *165*, 145–205. https://doi.org/10/gmhjhb

Denison, R. N., Parker, J. A., & Carrasco, M. (2020). Modeling pupil responses to rapid sequential events. *Behavior Research Methods*, *52*(5), 1991–2007. https://doi.org/10/gg4zs2

Elman, J. A., Panizzon, M. S., Hagler, D. J., Eyler, L. T., Granholm, E. L., Fennema-Notestine, C., Lyons, M. J., McEvoy, L. K., Franz, C. E., Dale, A. M., & Kremen, W. S. (2017). Task-evoked pupil dilation and BOLD variance as indicators of locus coeruleus dysfunction. *Cortex*, *97*, 60–69. https://doi.org/10.1016/j.cortex.2017.09.025

Finsterer, J. (2003). Ptosis: Causes, presentation, and management. *Aesthetic Plastic Surgery*, *27*(3), 193–204. https://doi.org/10/bkph2x

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). Mini-Mental state: A practical method for grading the state of patients for the clinician. *Journal of Psychiatric Research*, *12*, 189–198.

Forbes, S. (2020). PupillometryR: An R package for preparing and analysing pupillometry data. *Journal of Open Source Software*, *5*(50), 2285. https://doi.org/10/gg2qmh

Franklin, M. S., Broadway, J. M., Mrazek, M. D., Smallwood, J., & Schooler, J. W. (2013). *Window to the wandering mind: Pupillometry of spontaneous thought while reading*. SAGE Publications Sage UK: London, England.

Friedman, O. (2005). Changes associated with the aging face. *Facial Plastic Surgery Clinics*, *13*(3), 371–380. https://doi.org/10/c9d2xm

Gagl, B., Hawelka, S., & Hutzler, F. (2011). Systematic influence of gaze position on pupil size measurement: Analysis and correction. *Behavior Research Methods*, *43*(4), 1171–1181. https://doi.org/10/c75xwk

Geller, J., Winn, M., Mahr, T., & Mirman, D. (2019). *GazeR: A Package for Processing Gaze Position and Pupil Size Data* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/gvcxb

Gerven, P. W. M. V., Paas, F., Merriënboer, J. J. G. V., & Schmidt, H. G. (2004). Memory load and the cognitive pupillary response in aging. *Psychophysiology*, *41*(2), 167–174. https://doi.org/10.1111/j.1469-8986.2003.00148.x

Geyer, M. A., & Swerdlow, N. R. (1998). Measurement of startle response, prepulse inhibition, and habituation. *Current Protocols in Neuroscience*, *3*(1), 8–7. https://doi.org/10/fqf37m

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*(2), 252–269.

Gitelman, D. R., Penny, W. D., Ashburner, J., & Friston, K. J. (2003). Modeling regional and psychophysiologic interactions in fMRI: The importance of hemodynamic deconvolution. *NeuroImage*, *19*(1), 200–207. https://doi.org/10/bc8ct8

Goldinger, S. D., & Papesh, M. H. (2012). Pupil dilation reflects the creation and retrieval of memories. *Current Directions in Psychological Science*, *21*(2), 90–95.

Goldinger, S. D., & Papesh, M. H. (2013). Recollection is fast and easy: Pupillometric studies of face memory. In *Psychology of Learning and Motivation* (Vol. 59, pp. 191–222). Elsevier.

Granholm, E., Asarnow, R. F., Sarkin, A. J., & Dykes, K. L. (1996). Pupillary responses index cognitive resource limitations. *Psychophysiology*, *33*(4), 457–461.

Grant, E. R., & Spivey, M. J. (2003). Eye movements and problem solving: Guiding attention guides thought. *Psychological Science*, *14*(5), 462–466. https://doi.org/10/d97qck

Handwerker, D. A., Ollinger, J. M., & D'Esposito, M. (2004). Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *NeuroImage*, *21*(4), 1639–1651. https://doi.org/10/dv6qcs

Harris, R., McGavack, T. H., & Elias, H. (1946). The nature of the action of dimethylaminoethyl benzhydryl ether hydrochloride (Benadryl): Effects upon the human eye. *The Journal of Laboratory and Clinical Medicine*, *31*(10), 1148–1152. https://doi.org/10.5555/uri:pii:0022214346901564

Hayes, T. R., & Petrov, A. A. (2016). Mapping and correcting the influence of gaze position on pupil size measurements. *Behavior Research Methods*, *48*(2), 510–527.

Hershman, R., Henik, A., & Cohen, N. (2018). A novel blink detection method based on pupillometry noise. *Behavior Research Methods*, *50*(1), 107–114.

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

Hess, E. H., & Polt, J. M. (1960). Pupil size as related to interest value of visual stimuli. *Science*, *132*(3423), 349–350.

Hess, E. H., & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science*, *143*(3611), 1190–1192. https://doi.org/10/fb3kx4

Hoffmann, M., Lefering, R., Rueger, J. M., Kolb, J. P., Izbicki, J. R., Ruecker, A. H., Rupprecht, M., Lehmann, W., & Surgery, T. R. of the G. S. for T. (2012). Pupil

evaluation in addition to Glasgow Coma Scale components in prediction of traumatic brain injury and mortality. *British Journal of Surgery*, *99*(S1), 122–130.

Jaanus, S. D. (1992). Ocular side effects of selected systemic drugs. *Optometry Clinics*, *2*(4), 73–96.

Jennett, B., & Teasdale, G. (1977). Aspects of coma after severe head injury. *The Lancet*, *309*(8017), 878–881.

Johnson, E. L., Miller Singley, A. T., Peckham, A. D., Johnson, S. L., & Bunge, S. A. (2014). Task-evoked pupillometry provides a window into the development of short-term memory capacity. *Frontiers in Psychology*, *5*, 218. https://doi.org/10/ghs96d

Joshi, S., Li, Y., Kalwani, R. M., & Gold, J. I. (2016). Relationships between pupil diameter and neuronal activity in the locus coeruleus, colliculi, and cingulate cortex. *Neuron*, *89*(1), 221–234.

Just, M. A., & Carpenter, P. A. (1985). Cognitive coordinate systems: Accounts of mental rotation and individual differences in spatial ability. *Psychological Review*, *92*(2), 137. https://doi.org/10/b8rk8j

Kessler, E.-M., Bowen, C. E., Baer, M., Froelich, L., & Wahl, H.-W. (2012). Dementia worry: A psychological examination of an unexplored phenomenon. *European Journal of Ageing*, *9*(4), 275–284. https://doi.org/10/f4cn3v

Kim, M., Beversdorf, D. Q., & Heilman, K. M. (2000). Arousal response with aging: Pupillographic study. *Journal of the International Neuropsychological Society*, *6*(3), 348–350. https://doi.org/10/c6x3gf

Klingner, J., Tversky, B., & Hanrahan, P. (2011). Effects of visual and verbal presentation on cognitive load in vigilance, memory, and arithmetic tasks. *Psychophysiology*, *48*(3), 323–332. https://doi.org/10/fqzf3m

Kreis, I., Zhang, L., Moritz, S., & Pfuhl, G. (2021). Spared performance but increased uncertainty in schizophrenia: Evidence from a probabilistic decision-making task. *Schizophrenia Research*. https://doi.org/10/gmgb5x

Kret, M. E., & Sjak-Shie, E. E. (2019). Preprocessing pupil size data: Guidelines and code. *Behavior Research Methods*, *51*(3), 1336–1342. https://doi.org/10/gf5ssx

Kuchinsky, S. E., Ahlstrom, J. B., Cute, S. L., Humes, L. E., Dubno, J. R., & Eckert, M. A. (2014). Speech-perception training for older adults with hearing loss impacts word recognition and effort. *Psychophysiology*, *51*(10), 1046–1057. https://doi.org/10/f6kpcp

Laeng, B., Orbo, M., Holmlund, T., & Miozzo, M. (2011). Pupillary Stroop effects. *Cognitive Processes*, *12*(1), 13–21. https://doi.org/10.1007/s10339-010-0370-z

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

Laeng, B., & Sulutvedt, U. (2014). The eye pupil adjusts to imaginary light. *Psychological Science*, *25*(1), 188–197. https://doi.org/10.1177/0956797613503556

Lam, B. L., Thompson, H. S., & Corbett, J. J. (1987). The prevalence of simple anisocoria. *American Journal of Ophthalmology*, *104*(1), 69–73. https://doi.org/10/gmhpbt

Larson, M. D. (2008). Mechanism of opioid-induced pupillary effects. *Clinical Neurophysiology*, *119*(6), 1358–1364. https://doi.org/10/chkw8r

Loewenfeld, I. E., & Lowenstein, O. (1993). *The pupil: Anatomy, physiology, and clinical applications* (Vol. 2). Wiley-Blackwell.

Lowenstein, O., Feinberg, R., & Loewenfeld, I. E. (1963). *Pupillary movements during acute and chronic fatigue: A new test for the objective evaluation of tiredness* (Vol. 65). Federal Aviation Agency, Office of Aviation Medicine.

Lüdtke, H., Wilhelm, B., Adler, M., Schaeffel, F., & Wilhelm, H. (1998). Mathematical procedures in data recording and processing of pupillary fatigue waves. *Vision Research*, *38*(19), 2889–2896. https://doi.org/10/d8fwmf

Marmarou, A., Lu, J., Butcher, I., McHugh, G. S., Murray, G. D., Steyerberg, E. W., Mushkudiani, N. A., Choi, S., & Maas, A. I. (2007). Prognostic value of the Glasgow Coma Scale and pupil reactivity in traumatic brain injury assessed pre-hospital and on enrollment: An IMPACT analysis. *Journal of Neurotrauma*, *24*(2), 270–280.

Mathôt, S. (2018). Pupillometry: Psychology, physiology, and function. *Journal of Cognition, 1*(1). https://doi.org/10/gfkmbc

Mathôt, S., Fabius, J., Van Heusden, E., & Van der Stigchel, S. (2018). Safe and sensible baseline correction of pupil-size data. *Behavior Research Methods, 50*(1), 94–106.

Mathôt, S., Grainger, J., & Strijkers, K. (2017). Pupillary responses to words that convey a sense of brightness or darkness. *Psychological Science*, 0956797617702699.

Mathôt, S., Melmi, J.-B., Van der Linden, L., & Van der Stigchel, S. (2015). The mind-writing pupil: Near-perfect decoding of visual attention with pupillometry. *Journal of Vision, 15*(12), 176. https://doi.org/10.1167/15.12.176

McLaughlin, D. J., Zink, M., Gaunt, L., Spehar, B., Van Engen, K., Sommers, M. S., & Peelle, J. E. (2021). *Pupillometry reveals cognitive demands of lexical competition during spoken word recognition in young and older adults*.

Mirman, D. (2014). *Growth Curve Analysis and Visualization Using R* (1st ed.). Chapman & Hall.

Morad, Y., Lemberg, H., Yofe, N., & Dagan, Y. (2000). Pupillography as an objective indicator of fatigue. *Current Eye Research*, *21*(1), 535–542. https://doi.org/10/bqc9tn

Morris, S. K., Granholm, E., Sarkin, A. J., & Jeste, D. V. (1997). Effects of schizophrenia and aging on pupillographic measures of working memory. *Schizophrenia Research*, *27*(2–3), 119–128. https://doi.org/10/bk3fzf

Murphy, P. R., O'Connell, R. G., O'Sullivan, M., Robertson, I. H., & Balsters, J. H. (2014). Pupil diameter covaries with BOLD activity in human locus coeruleus. *Human Brain Mapping*, *35*(8), 4140–4154. https://doi.org/10.1002/hbm.22466

Nasreddine, Z. S., Phillips, N. A., Badirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., & Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: A Brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, *53*(4), 695–699.

Nassar, M. R., Rumsey, K. M., Wilson, R. C., Parikh, K., Heasly, B., & Gold, J. I. (2012). Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature Neuroscience*, *15*(7), 1040–1046. https://doi.org/10.1038/nn.3130

Nichols, A. L., & Maner, J. K. (2008). The Good-Subject Effect: Investigating Participant Demand Characteristics. *The Journal of General Psychology*, *135*(2), 151–166. https://doi.org/10/c2gcm5

O'Shea, H., & Moran, A. (2018). To go or not to go? Pupillometry elucidates inhibitory mechanisms in motor imagery. *Journal of Cognitive Psychology*, *30*(4), 466–483. https://doi.org/10/gmgb5v

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). Springer, New York, NY. https://doi.org/10.1007/978-1-4939-1236-0_9

Papesh, M. H., Goldinger, S. D., & Hout, M. C. (2012). Memory strength and specificity revealed by pupillometry. *International Journal of Psychophysiology*, *83*(1), 56–64. https://doi.org/10/bw977q

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.

Piquado, T., Isaacowitz, D., & Wingfield, A. (2010). Pupillometry as a measure of cognitive effort in younger and older adults. *Psychophysiology*, *47*(3), 560–569. https://doi.org/10.1111/j.1469-8986.2009.00947.x

Recarte, M. Á., Pérez, E., Conchillo, Á., & Nunes, L. M. (2008). Mental Workload and Visual Impairment: Differences between Pupil, Blink, and Subjective Rating. *The Spanish Journal of Psychology*, *11*(2), 374–385. https://doi.org/10/gmjb35

Reilly, J., Kelly, A., Kim, S. H., Jett, S., & Zuckerman, B. (2018). The human task-evoked pupillary response function is linear: Implications for baseline response scaling in pupillometry. *Behavior Research Methods*, 1–14.

Reilly, J., Zuckerman, B., Kelly, A., Flurie, M., & Rao, S. (2020). Neuromodulation of cursing in American English: A combined tDCS and pupillometry study. *Brain and Language*, *206*, 104791. https://doi.org/10.1016/j.bandl.2020.104791

Rengstorff, R. H. (1994). Vision and ocular changes following accidental exposure to organophosphates. *Journal of Applied Toxicology*, *14*(2), 115–118. https://doi.org/10/dnv74r

Rollins, M. D., Feiner, J. R., Lee, J. M., Shah, S., & Larson, M. (2014). Pupillary effects of high-dose opioid quantified with infrared pupillometry. *Anesthesiology, 121*(5), 1037–1044. https://doi.org/10/gmhnzg

Rossetti, H. C., Lacritz, L. H., Hynan, L. S., Cullum, C. M., Van Wright, A., & Weiner, M. F. (2017). Montreal Cognitive Assessment Performance among Community-Dwelling African Americans. *Archives of Clinical Neuropsychology*, *32*(2), 238–244. https://doi.org/10.1093/arclin/acw095

Schneider, M., Leuchs, L., Czisch, M., Sämann, P. G., & Spoormaker, V. I. (2018). Disentangling reward anticipation with simultaneous pupillometry/fMRI. *Neuroimage*, *178*, 11–22. https://doi.org/10/gdxffm

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*(5), 679–687. https://doi.org/10/bc3fc3

Sireci, S. G. (1998). The Construct of Content Validity. *Social Indicators Research*, *45*(1), 83–117. https://doi.org/10/bx7dg3

Steinhauer, S. R., & Hakerem, G. (1992). The pupillary response in cognitive psychophysiology and schizophrenia. *Annals of the New York Academy of Sciences*, *658*(1), 182–204.

Thomas, L. E., & Lleras, A. (2007). Moving eyes and moving thought: On the spatial compatibility between eye movements and cognition. *Psychonomic Bulletin & Review*, *14*(4), 663–668. https://doi.org/10/bzs8j6

Tryon, W. W. (1975). Pupillometry: A survey of sources of variation. *Psychophysiology*, *12*(1), 90–93.

Tsukahara, J. S., & Engle, R. W. (2021). Is baseline pupil size related to cognitive ability? Yes (under proper lighting conditions). *Cognition*, *211*, 104643. https://doi.org/10/gjcgv3

Tsukahara, J. S., Harrison, T. L., & Engle, R. W. (2016). The relationship between baseline pupil size and intelligence. *Cognitive Psychology*, *91*, 109–123. https://doi.org/10.1016/j.cogpsych.2016.10.001

Tun, P. A., McCoy, S., & Wingfield, A. (2009). Aging, hearing acuity, and the attentional costs of effortful listening. *Psychol Aging*, *24*(3), 761–766. https://doi.org/2009-13203-027 [pii] 10.1037/a0014802 [doi]

Van Engen, K. J., & McLaughlin, D. J. (2018). Eyes and ears: Using eye tracking and pupillometry to understand challenges to speech recognition. *Hearing Research*, *369*, 56–66. https://doi.org/10/gfqg8t

Van Gerven, P. W., Paas, F., Van Merriënboer, J. J., & Schmidt, H. G. (2004). Memory load and the cognitive pupillary response in aging. *Psychophysiology*, *41*(2), 167–174. https://doi.org/10/cp66gp

Veneman, C. E., Gordon-Salant, S., Matthews, L. J., & Dubno, J. R. (2013). Age and Measurement Time-of-Day Effects on Speech Recognition in Noise. *Ear and Hearing*, *34*(3), 288–299. https://doi.org/10/gmjcs9

Wainstein, G., Rojas-Libano, D., Medel, V., Alnæs, D., Kolskår, K. K., Endestad, T., Laeng, B., Ossandon, T., Crossley, N., Matar, E., & Shine, J. M. (2021). The ascending arousal system promotes optimal performance through meso-scale network integration in a visuospatial attentional task. *Network Neuroscience*, 1–32. https://doi.org/10/gmw45x

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.

Wilhelm, B., Wilhelm, H., Lüdtke, H., Streicher, P., & Adler, M. (1998). Pupillographic assessment of sleepiness in sleep-deprived healthy subjects. *Sleep*, *21*(3), 258–265.

Willenbockel, V., Sadr, J., Fiset, D., Horne, G. O., Gosselin, F., & Tanaka, J. W. (2010). Controlling low-level image properties: The SHINE toolbox. *Behavior Research Methods*, *42*(3), 671–684. https://doi.org/10/dxddqm

Winn, M. B., Wendt, D., Koelewijn, T., & Kuchinsky, S. E. (2018). Best practices and advice for using pupillometry to measure listening effort: An introduction for those who want to get started. *Trends in Hearing*, *22*, 2331216518800869.

Zahodne, L. B., Manly, J. J., Smith, J., Seeman, T., & Lachman, M. E. (2017). Socioeconomic, health, and psychosocial mediators of racial disparities in cognition in early, middle, and late adulthood. *Psychology and Aging*, *32*(2), 118–130. https://doi.org/10.1037/pag0000154

Zavagno, D., Tommasi, L., & Laeng, B. (2017). The Eye Pupil's Response to Static and Dynamic Illusions of Luminosity and Darkness. *I-Perception*, *8*(4), 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*, *101c*, 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*(3), 277–284.

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2010). Pupil response as an indication of effortful listening: The influence of sentence intelligibility. *Ear and Hearing*, *31*(4), 480–490. https://doi.org/10/cx648q

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2011). Cognitive load during speech perception in noise: The influence of age, hearing loss, and cognition on the pupil response. *Ear and Hearing*, *32*(4), 498–510. https://doi.org/10/cfgzkx