Trends in Hearing
Volume 22: 1–32

DOI: 10.1177/2331216518800869
journals.sagepub.com/home/tia
SAGE

Best Practices and Advice for Using Pupillometry to Measure Listening Effort: An Introduction for Those Who Want to Get Started

Matthew B. Winn[1], Dorothea Wendt[2,3], Thomas Koelewijn[4], and Stefanie E. Kuchinsky[5]

Abstract

Within the field of hearing science, pupillometry is a widely used method for quantifying listening effort. Its use in research is growing exponentially, and many labs are (considering) applying pupillometry for the first time. Hence, there is a growing need for a methods paper on pupillometry covering topics spanning from experiment logistics and timing to data cleaning and what parameters to analyze. This article contains the basic information and considerations needed to plan, set up, and interpret a pupillometry experiment, as well as commentary about how to interpret the response. Included are practicalities like minimal system requirements for recording a pupil response and specifications for peripheral, equipment, experiment logistics and constraints, and different kinds of data processing. Additional details include participant inclusion and exclusion criteria and some methodological considerations that might not be necessary in other auditory experiments. We discuss what data should be recorded and how to monitor the data quality during recording in order to minimize artifacts. Data processing and analysis are considered as well. Finally, we share insights from the collective experience of the authors and discuss some of the challenges that still lie ahead.

Keywords

pupillometry, listening effort, methods

Date received: 15 January 2018; revised: 7 August 2018; accepted: 14 August 2018

Introduction

Goal and Overview of This Article

In this introductory article, we offer advice on how to understand and incorporate pupillometry (the measurement of pupil size) as a measure of listening effort. The target audience includes researchers who have considered using pupillometry but might not be familiar with the technical or logistical challenges that are involved. For the purpose of having a standard set of recommendations in place, the authors have collected their shared experiences—both good practices as well as pitfalls—in this article. Original hypothesis-driven research can be found in numerous other publications and elsewhere in this special issue. But the *story* of how this research is done is sometimes hidden out of sight. The point of this article is to familiarize the reader with the challenges one could come up against when conducting pupillometry research for measuring listening effort.

The attraction of pupillometry is that changes in pupil dilation appear to distinguish cognitive tasks that are more or less effortful across a wide variety of domains (Beatty, 1982), including those that do not involve speech intelligibility. Pupil dilation scales with

[1]Speech-Language-Hearing Sciences, University of Minnesota, Minneapolis, MN, USA
[2]Eriksholm Research Centre, Snekkersten, Denmark
[3]Hearing Systems, Department of Electrical Engineering, Technical University of Denmark, Kongens Lyngby, Denmark
[4]Section Ear & Hearing, Department of Otolaryngology–Head and Neck Surgery, Amsterdam Public Health Research Institute, VU University Medical Center, the Netherlands
[5]National Military Audiology and Speech Pathology Center, Walter Reed National Military Medical Center, Bethesda, MD, USA

Corresponding Author:
Matthew B. Winn, Speech-Language-Hearing Sciences, University of Minnesota, Minneapolis, MN 55455, USA.
Email: mwinn@umn.edu

mathematical ability (Ahern & Beatty, 1979), short-term memory capacity (Klingner, Tversky, & Hanrahan, 2011; Zekveld, Kramer, & Festen, 2011), Stroop task interference (Laeng, Ørbo, Holmlund, & Miozzo, 2011), and resolving ambiguity in language (Vogelzang, Hendriks, & van Rijn, 2016). It therefore has the potential to add value to assessments of speech perception especially where there is reason to believe that there could be different amounts of cognitive load exerted for tasks that are not clearly distinguished by task accuracy.

One of the most important things to note about pupillometry is that pupil size is not a monotonic direct index of effort but rather a complicated mixture that reflects the combined contributions of the autonomic nervous system (ANS; Zekveld, Koelewijn, & Kramer, 2018). For cognitive-evoked dilations, the response is nonlinear (Ohlenforst et al., 2017; Wendt, Koelewijn, Książek, Kramer, & Lunner, 2018); dilations are small for easy tasks but also small for *very* hard tasks (where effort might be withdrawn because of lack of task success). Pupil dilation therefore appears to be an index of a person's willingness to exert more effort because it is *worth* the exercise of greater mental resources to achieve a goal. This important concept will return numerous times in this article, especially as it relates to fatigue and capacity. If a person is overly fatigued, there is increased likelihood that effort will be reduced because of less engagement—leading to reduced pupil size (Wang, Zekveld, Lunner, & Kramer, 2018). For the purpose of interpreting pupil size data, this means that the experimenter should be cognizant of whether changes in pupil dilation are truly indicative of changes in task-related effort or unintended participant fatigue or disengagement.

In the following sections, we begin by introducing the complicated term *effort* and the connection that pupil dilation might have with effort. We then discuss experimental design and planning, including task selection, constraints of the method, logistics for hardware and data collection, and considerations for participant inclusion. The article continues with a review of some essential components of data processing and some recent insights on physiology of the pupil response. We then contribute some advice on helpful practices for the experimenter and conclude with some recommended further reading.

What Do We Mean by Effort?

The Framework for Understanding Effortful Listening (FUEL; Pichora-Fuller et al., 2016) defines listening effort as the "deliberate allocation of mental resources to overcome obstacles in goal pursuit when carrying out a [listening] task" (p. 10 S). This definition highlights that effort arises not only as a result of the difficulty of the task itself (i.e., the intelligibility of the stimuli) but also of a result of the active application of the individual's mental capacities to overcome an obstacle. Another essential component is the participant's engagement or motivation to succeed in a task, which may vary widely across individuals. The FUEL has its roots in the classical capacity model described by Kahneman (1973), who emphasized the role of attention (and arguably used *effort* and *attention* interchangeably; Bruya & Tang, 2018). Kahneman suggested that effort is "a special case of arousal," characterizing it as effort invested in what one is doing, rather than arousal in response to what is happening to a person (e.g., from loud sounds, drugs, etc.). According to FUEL, the level of arousal related to the processing of speech in adverse conditions is reflected by activation of the ANS, which can be measured by the pupil dilation response.

Why Measure Listening Effort and Not Just Intelligibility?

Listening effort is increasingly recognized as an important aspect of hearing loss. Hearing difficulties and increased listening effort are reportedly connected with numerous medical, financial, and occupational challenges (Kramer, Kapteyn, & Houtgast, 2006; Nachtegaal et al., 2009) as well as feeling of social connectedness (Hughes, Hutchings, Rapport, McMahon, & Boisvert, 2018). Hétu, Riverin, Lalande, Getty, and St-Cyr (1988) reported interviews with individuals with hearing impairment who mentioned that fatigue related to their hearing difficulties and coping mechanisms was severe enough that they would be "... too tired for normal activities" after finishing work.

Two listeners might achieve the same intelligibility score but exert different amounts of effort to do so; pupil dilation appears consistent with subjective notions of relative difficulty even in these equally intelligible cases (Koelewijn, Zekveld, Festen, & Kramer, 2012). Because speech perception can involve a variety of cognitive linguistic skills in addition to auditory processing (Bronkhorst, 2015; Mattys, Davis, Bradlow, & Scott, 2012), the same intelligibility score can be obtained by a listener with moderate hearing loss exerting great focus, or by a person with typical hearing who is listening with less effort (Ohlenforst et al., 2017). Where audibility fails, cognitive compensation strategies (suppressing irrelevant information, relying on context, etc.) can compensate (Peelle, 2017; Rönnberg, Lunner, & Zekveld, 2013) as is often seen in the case of individuals with hearing impairment who show reliance on context in speech perception (Pichora-Fuller, Schneider, & Daneman, 1995). This greater reliance on top-down mechanisms appears to come at a cost of decline in other cognitive and physical tasks, or memory of

words heard (e.g., Koeritzer, Rogers, Van Engen, & Peelle, 2018; McCoy et al., 2005). Listeners with normal hearing also engage in potentially effortful cognitive processes when listening to acoustically challenging speech, even when the speech is highly intelligible and supported by linguistic context (Koeritzer et al., 2018). Thus, an experimenter or clinician might be interested not only in the ultimate accuracy in a task, but also the mechanisms used to accomplish the task, and how effortful it was to complete the task.

In addition to the reasons stated earlier, it is also useful to remember that not all aspects of speech perception are gauged by whether the words are correctly identified. Other aspects include analyzing and updating a talker's intention (Snedeker & Trueswell, 2004; Tanenhaus, Spivey, Eberhard, & Sedivy, 1995), predicting upcoming information (Altmann & Kamide, 1999; Tavano & Scharinger, 2015), identifying a talker (Best et al., 2018), perceiving prosodic emphasis (Dahan, Tanenhaus, & Chambers, 2001), translating speech into a different language (Hyönä, Tommola, & Alaja, 1995), and judging whether an utterance makes sense (Best, Streeter, Roverud, Mason, & Kidd, 2016). These would all be essential components of speech communication that would not be adequately quantified by a score of whether words were correctly repeated. Apart from pupillometry, other classic experimental measures like eye tracking and brain imaging show that there is value in granular responses that scale with task demands even when intelligibility is not the outcome measure of primary interest.

From the perspective of the audiologist, listening effort is arguably a worthwhile measurement *even in the absence* of intelligibility scores because effort is often the direct complaint of the patient. Gatehouse and Noble (2004) found that the disability-handicap relationship was governed by sound identification, attention, spatial aspects hearing, and "effort problems," but not "intelligibility of speech." Although it would not be controversial to say that speech intelligibility plays a role in increasing effort, we argue it is not that a word was repeated incorrectly that makes an event effortful, especially since the listener might not be aware that the perception was incorrect. Instead, effort likely arises from the related cognitive processes engaged to correct that error if the listener suspects that it might have been a mistake, or perhaps to use cognitive strategies to restore a word that was completely masked by noise.

Apart from examining the relationship between effort measures and performance accuracy measures, it is also worthwhile to consider any sign of effort as an indication of task engagement, which could be a useful outcome measure in itself. For example, Teubner-Rhodes, Vaden, Dubno, and Eckert (2017) proposed an assessment of executive function that they call "Cognitive persistence." Individuals who face listening difficulties might avoid challenging auditory environments (cf. Wu et al., 2018); tasks that evoke consistent signs of effort could indicate that the individual is at least willing to attempt the task.

Despite the showcase of pupillometry in this article, we remind the reader that it has not been conclusively established that the laboratory-based pupillometric measures of effort are directly related to symptoms such as everyday listening difficulties, susceptibility to fatigue, and poor recognition memory. Hornsby and Kipp (2016) showcase the need for systematic investigations into this connection and also highlight the concept of fatigue separately from episodic effort. However, pupillometry and other measures of effort likely play a fractional role in establishing those connections through converging sets of evidence and associations.

The Unique Value of Pupillometry

Considering the success of other measures of effort, such as dual-task paradigms (Gagné, Besser, & Lemke, 2017) and reaction times, which might not need such a complicated set of guidelines; what value is added by pupillometry? There are multiple benefits that we highlight here, which expand on the commentary on methodology by McGarrigle et al. (2014). First, pupillometry is a *time-series* measurement. Timing is an essential part of understanding listening effort because speech demands rapid auditory encoding as well as cognitive processing distributed over time, rather than being deployed all at once at the end of a stimulus. Effort might not be uniformly distributed over a perceptual event, and pupillometric measures have the benefit of showing change in dilation at different time landmarks. McCloy, Lau, Larson, Pratt, and Lee (2017) showed changes in pupil dilation in anticipation of a difficult task. Vogelzang et al. (2016) similarly showed changes in timing of pupil dilations based on pronoun ambiguity in sentences, followed by anticipatory dilations in preparation for follow-up questions. These are examples where pupillometry revealed changes in cognitive load *during* the test trial, as it related to linguistic processing during and after perception.

Measures of effort each have their own advantages and limitations. Reaction times are subject to variations in manual dexterity and speech, which might change with age or physical abilities not related to the experimental task. Pupil size and range of dilation are also affected by age (Bitsios, Prettyman, & Szabadi, 1996; Kim, Beversdorf, & Heilman, 2000; Winn, Whitaker, Elliott, & Phillips, 1994) although there are published normalization methods (discussed later) that capitalize on the reliable pupillary light reflex as a standard of dynamic range (Piquado, Isaacowitz, & Wingfield, 2010).

Pupillometry is arguably a more sensitive measure than dual-task cost, which does not provide temporal information. Compare, for example, measures of spectrally degraded speech perception by Winn, Edwards, and Litovsky (2015) using pupillometry and by Pals, Sarampalis, and Baskent (2013) using dual-task cost. We note, however, that dual-task measures can be logistically more feasible to conduct and are less affected by the methodological constraints outlined in this article. Functional magnetic resonance imaging studies have aimed to reveal the mechanisms that underlie listening effort via linking pupil dilation to the engagement of both domain-general attention and sensory-specific brain regions during speech comprehension (e.g., Kuchinsky et al., 2016; Zekveld, Heslenfeld, Johnsrude, Versfeld, & Kramer, 2014), during other cognitive tasks (e.g., Siegle, Steinhauer, Stenger, Konecky, & Carter, 2003), and during spontaneous fluctuations in alertness (e.g., Murphy, O'Connell, O'Sullivan, Robertson, & Balsters, 2014; Schneider et al., 2016).

Other neuroimaging methods with faster temporal resolutions, such as magnetoencephalography and electroencephalogram (EEG), have similarly sought to establish a neural basis for pupillary indices of listening effort. In fact, pupil dilation and EEG have been simultaneously registered in multiple studies. McMahon et al. (2016) showed that EEG alpha level was comodulated with pupil dilation for 16-channel vocoded speech, but for conditions of more-difficult six-channel vocoded speech, the relationship was much less clear. Miles et al. (2017) followed up with a related study aimed at discerning effects of intelligibility, finding that unlike EEG results, pupil dilation was related to intelligibility scores. Interestingly, the two measurements were not correlated with each other, suggesting that they tap into potentially different cognitive mechanisms. Further investigations are needed to better understand the potential connection of different measures.

There is a benefit of pupillometry in the context of testing participants who use assistive devices such as hearing aids and cochlear implants (CIs), which is that the experimenter can avoid problematic interference of the device with electrical or magnetic imaging techniques (Friesen & Picton, 2010; Gilley et al., 2006; L. Wagner, Maurits, Maat, Baskent, & Wagner, 2018). Similarly, functional magnetic resonance imaging can provide precise spatial information about the neural systems engaged during effortful speech processing (Lee, Min, Wingfield, Grossman, & Peelle, 2016; Obleser, Wise, Dresner, & Scott, 2007) but is not well suited to individuals with electronic implants, vascular disorders, or for presenting speech in relative quiet (because of machine noise). Functional near-infrared spectroscopy is unaffected by such interference and compatible with the use of implants (McKay et al., 2016), but it is slower than pupillometry (i.e., it cannot capture rapid changes), and considerably more expensive than EEG or pupillometry at the time of this writing. In all, pupillometry is not free from limitations, but is relatively easy and fast to set up, has a sufficient temporal resolution, is free from electrical artifact, and is comparatively inexpensive compared with some other imaging techniques.

Experimental Design and Planning

What Does Pupil Dilation Reflect?

Ranging between sizes of roughly 3 mm and 7 mm (Laeng, Sirous, & Gredebäck, 2012), the pupil dilates and contracts for multiple reasons (see Zekveld et al., 2018). In normal circumstances, the largest changes in pupil dilation occur in response to changes in luminance. When changing from light to dark environments, pupil diameter can increase by as much as 3 to 4 mm, or roughly 120% (Laeng et al., 2012). Conversely, the cognitive task-evoked pupil dilations that are central to this article are much smaller by comparison, on the order of 0.1 to 0.5 mm, depending on testing conditions and task. Because of these factors, one must manage the sources of dilation and constriction factors apart from the experimental task so that an evoked response can be a reliable indicator of the effort exerted during the task. In addition, the amount of pupil dilation evoked by a task can be modulated by the participant's motivation and arousal state (Stanners et al., 1979), as will be discussed in detail in this article.

It is reasonable to consider task-evoked pupil dilation to reflect not a simply unitary concept of effort but rather some amalgamation of attention, engagement, arousal, anxiety, and effort (Nunnally, Knott, Duchnowski, & Parker, 1967; Pichora-Fuller et al., 2016). While it is not within the scope of this article to clarify the distinctions between these interrelated concepts, they all have been invoked in numerous explanations of the pupillary response over the years. We use the term "Listening effort" as a useful shorthand tool that can be understood to capture a union of these concepts as they relate to hearing (difficulties), but there could be valid reasons to unpack each of these concepts individually.

In agreement with the frameworks described by Kahneman (1973) and Pichora-Fuller et al. (2016), we highlight the critical role of *intentional attentional engagement* in the study of effort. When a person has motivation to exercise more cognitive resources to a task, it can be understood in the context of goal-directed behavior, where attention not only has a target but also an intensity. Attention and effort are highly related and sometimes studied in tandem. For example, in a study conducted by Koenig, Uengoer, and Lachnit (2017), there was increased pupil dilation in early stages of

attention to consistently reinforced learning cues, while in later stages of learning when those cues did not demand as much attention, relatively larger pupil dilations were observed for ambiguous or unreinforced cues. The pupillary response was associated with a strategic shift in attention in a goal-directed task. Karatekin, Couperous, and Marcus (2004) measured significantly larger pupil dilations in conditions of divided attention in a dual-task experiment conducted to distinguish performance accuracy and efficiency (stated as ''the costs of that performance in mental effort'').

Since Kahneman's (1973) influential monograph, examination of effort has historically been tied with the concepts of attention and arousal. Bruya and Tang (2018) are critical of Kahneman's binding of attention and effort, suggesting that instead of characterizing attention as the use of cognitive or metabolic resources, we ought to instead consider it as the ''readying'' of metabolic resources in the form of adaptive gain modulation. Considering the physiological evidence in studies by Reimer et al. (2016) and McGinley, David, and McCormick (2015) and some of the speech perception work described later in this article and elsewhere in this issue, Bruya and Tang's suggestion cannot be dismissed. It should be noted however, that even in Kahneman's original book, the concept of effort *as preparation* is clearly mentioned in the introductory chapter (p. 4).

The persistent tradition is to consider larger pupil dilation to be a sign of increased listening effort, and therefore a negative outcome, compared with smaller pupil dilation. However, we should not assume that more effort (or larger pupil response) is always a negative thing. Engagement in speech communication can be a very productive and satisfying process, but only with sufficient effort or attention devoted to the input. Increased pupil dilation is a signal that a listener is at least *willing* to engage in a task and therefore could be a positive sign. Take, for example, studies of pupil dilation across a range of intelligibility. Listener will show larger pupil dilation for speech that is perceived with 70% accuracy compared with speech with 25% accuracy (Ohlenforst et al., 2017; Wendt et al., 2018). We should not conclude that the 25%-intelligible speech was *easier* to listen to, but rather that the listener was less engaged in the 25%-correct task because it was so hard that more engagement would be unlikely to return any value to the listener. This concept could help the experimenter interpret pupil dilation not as the effort demanded by a task, but rather the effort actually exerted by the participants, modulated by the perceived cost or benefit of expending more metabolic resources. We emphasize this aspect of the measurement not only to highlight nonlinearities that are less well known but also to encourage the idea that pupillometry could play a role in exploring the finding that people with hearing loss appear to select against environments with poor signal-to-noise ratio (SNR; Wu et al., 2018). Perhaps pupillometric measures of effort or engagement could reveal that a person is more capable of handling such situations with a clinical intervention, and therefore an increased dilation would be a sign of progress and increased confidence to face a wider range of communication environments.

Task Selection—What Task Properties Will Evoke Pupil Dilation?

The experimental task should ideally demand that a listener exert intentional effort beyond passive awareness of sounds in the environment. Ideally, there would be multiple experimental conditions where the participant is motivated to exert more effort in at least one condition because it will produce better results. In the following sections, we review some relevant considerations for guiding task selection.

Stimulus difficulty and listener interest. For reliable and interpretable pupillometry results, there is a balance of making the stimuli not so easy as to demand too little cognitive effort and also not so difficult as to make cognitive effort futile (see previous section, and also Wendt et al., 2018 and Eckert, Teubner-Rhodes, & Vaden, 2016 for supporting data and discussion). In addition to stimulus difficulty, the experimenter should also consider stimulus *value* to the participant. For example, Eckert et al. (2016) illustrated how a conversation with grandchildren would yield higher value than watching a documentary about lint. There will likely be more engagement (and therefore likely larger pupil dilations) when listening to the grandchildren, even if the speech is equally intelligible in both situations. Furthermore, Eckert et al. note that the more valuable conversation would likely retain its value more strongly through communication barriers, invoking extra activity from cortical regions involved in executive attentional control where boring tasks might not, since they are not worth the metabolic cost.

Basic psychoacoustics. Some basic tasks of auditory detection or discrimination might not demand cognitive resources sufficient to evoke a strong or consistent evoked pupil response although some reports do exist. For example, pitch discrimination elicits smaller pupil dilation in musicians than nonmusicians (Bianchi, Santurette, Wendt, & Dau, 2016), despite comparable peripheral sensitivity. Although no consistent pattern of pupil dilation would be expected if a participant simply hears different sounds coming from different locations, task-evoked changes do emerge in a task of explicit sound localization (Bala, Spitzer, & Takahashi, 2007). Beatty (1982) showed data from a study of selective

attention to individual pure tones, revealing a dilation pattern that was detectable (and detectably different when tones were targets or distractors), but the dilations were on the order of 0.01 mm, which is one tenth the size of those normally reported in the easiest conditions in many other articles. Without sufficiently powered experiments with a large number of trials, it is unlikely that such small effects would emerge in a consistent fashion. Beatty (1982) also notes that experimenters should take caution to distinguish between tasks of signal *detection*, in which pupil dilation increases with increased certainty (Hakerem & Sutton, 1966) and signal *discrimination*, in which pupil size increases with increased *uncertainty* (Kahneman & Beatty, 1967). Because of these complications, much of the literature on task-evoked pupil dilation concerns more tasks that are more complicated and demanding than detection of a signal, such as sentence perception, mental manipulation of input, mathematical problems, and so on.

Speech perception in quiet. For listeners with normal hearing, speech perception in quiet can be automatic or effortless if it does not come coupled demands no particular challenge (e.g., syntactic structure, auditory distortion, etc.). It therefore might not demand substantial cognitive resources to complete, producing pupil dilations that do not always reliably emerge from the noise of random pupillary oscillations. Data from Zekveld and Kramer (2014) show pupil dilations to quiet speech that hover around the baseline levels although their data were clean enough to illustrate clear interpretable morphology. In a number of published studies, speech in quiet is presented with some kind of extra cognitive demand, such as memory load (Johnson, Singley, Peckham, Johnson, & Bunge, 2014), spectral degradation (Winn et al., 2015), anomalous semantic content (Beatty, 1982), lexical competition (A. Wagner, Toffanin, & Baskent, 2016), competition from a second language (Schmidtke, 2014), conflict of prosody and syntactic structure (Engelhardt, Ferreira, & Patsenko, 2010), object-focused syntactic structure (Wendt, Dau, & Hjortkjær, 2016), and pronoun ambiguity (Vogelzang et al., 2016). In these cases, it is crucial to emphasize that the evoked dilations are likely in response to language processing and not simply auditory encoding.

Linguistic aspects of effort. In each of the aforementioned examples of speech perception studies, some aspect of language processing was the focal point of investigation. These studies establish conclusively that the cognitive activity indexed by pupil dilation does not follow merely from audition alone, but also from language processing. In another example, Hyönä et al. (1995) found increased pupil dilation in a task of sentence *translation* compared with verbatim repetition of sentences, suggesting that the pupil response reflects general processing load, not just effort in *listening* to the auditory stimulus.

Not all speech stimuli demand the same kinds of language or cognitive processing, and therefore experimenters should guard against the notion of a unitary category of "speech perception." In other words, just because stimuli are speech sounds, they might not elicit typical patterns of pupil dilation because they do not necessarily entail cognitive processes that relate to processing of natural speech. For example, it is possible that the popular style of "matrix" sentences where each word in a sentence is drawn from a closed set of choices elicits less effort, since most digits and colors can be distinguished by vowel alone (in English) and therefore might not reflect the effort needed to understand normal speech. Other sentence materials might be preferable to examine speech perception with a richer set of linguistic processes in play. Several studies have successfully used traditional speech-in-noise tests (such as the Dutch Versefeld sentences, Danish HINT test, English R-SPIN test, IEEE sentences, and others) and applied the pupillometry method.

Some linguistic stimuli might demand such limited amounts of cognitive processing that they do not elicit expected effects on pupil dilation. For example, auditory spectral degradation affects the pupillary response to sentence-length materials (Winn et al., 2015) but not recognition of individual spoken letters (McCloy et al., 2017). It is therefore worthwhile for the experimenter to consider whether the speech perception task involves some kind of linguistic computation or minimal auditory detection.

Increasing Motivation and Avoiding Boredom. Motivation will affect the pupillary response (Kahneman & Peavler, 1969). Left without a goal-directed task, a person's pupil will change size as the mind wanders (Franklin, Broadway, Mrazek, Smallwood, & Schooler, 2013), in a way that will not be aligned with stimulus presentation. If the task does not give enough reason for the participant to engage, the pupil size will likely not give useful results. Monetary incentives have been shown by Heitz, Schrock, Payne, and Engle (2008) to increase the magnitude of pupillary responses. When people are curious about the answers to trivia questions, their pupils dilate more (Kang et al. 2009)—by a small (8% vs. 4%) but detectable amount.

Although boredom is to be avoided in order to elicit pupil dilation reliably, experimenters should also consider avoiding emotional stimuli that evoke pleasure, disgust, or an otherwise strong physiological response unrelated to the planned task. For example, sentence materials can be chosen to avoid notions of violence, sexuality, or trauma. Pupillary responses to emotionally

toned or arousing stimuli were reported by Hess and Polt (1960) in an early influential paper. More recently, Partala and Surakka (2003) showed that compared with neutral stimuli, negative-valence stimuli evoked larger pupil responses, with largest dilations evoked by positive stimuli. If emotional response is not the target of investigation, then these kinds of stimuli could be avoided to reduce unwanted variability in the data.

Behavioral Considerations. In most pupillometry studies of listening effort, there is a behavioral component such as a spoken response or a button press, which can increase the measured pupil response by as much as 400% (Privitera, Renninger, Carney, Klein, & Aguilar, 2010) and this amplified response can sustain for several seconds. When the behavioral contribution is removed through deconvolution, the task-evoked pupil response is still present but is more modest and short lasting (cf. Hoeks & Levelt, 1993; McCloy, Larson, Lau, & Lee, 2016). Similar behavioral contributions to pupil size can be seen in studies of sentence recognition involving verbal responses. Winn et al. (2015) and Winn (2016) showed that pupil dilations from the verbal response were typically larger than those elicited by the listening task itself. Papesh and Goldinger (2012) carefully illustrated the effect of motor speech planning (as well as lexical frequency) on pupil dilations in a study involving cued response options that alternated between verbatim repetition or substituting "blah" in place of syllables. In numerous pupillometry studies of sentence perception, the timing of the behavioral response is so far separated from the listening response that the auditory-evoked pupil dilations recover almost completely back to baseline levels, and the behavioral-induced dilations are thus often not illustrated on published figures (Koelewijn, de Kluiver, Shinn-Cunningham, Zekveld, & Kramer, 2015; Koelewijn et al., 2012; Wendt et al., 2018; Zekveld, Kramer, & Festen, 2010).

We recommend letting pupil size return to baseline levels following a trial (though see studies that have employed deconvolution to tease apart pupillary effects arising from closely spaced visual stimuli, e.g., Wierda, Van Rijn, Taatgen, & Martens, 2012). In typical speech-in-noise testing, it is normally sufficient to wait 4 to 6 s after the completion of the participant's verbal response, but for other experiments without extensive precedent in the literature, we recommend pilot testing involving extended recording time (e.g., 10 s beyond the stimulus or response) and inspecting the data to see when the aggregated data return to baseline levels.

Experimenters should be aware of all task events that would invoke intentional attention, including physical motion. McGinley et al. (2015) found that 20% of variance in pupil size in mice was explained by locomotion; increases in pupil dilation were substantial and long lasting during motion. In addition, locomotion has been found to suppress sensory-evoked responses (Williamson, Hancock, Shinn-Cunningham, & Polley, 2015).

Experiment Logistics and Constraints

Task selection for pupillometry is somewhat constrained by the measurement technique, specifically because of the timing of the response and the challenge of avoiding changes in pupil size that are unrelated to the target task. It is therefore not advisable to simply add pupil dilation measures to an existing behavioral procedure that was not designed for pupillometry. Instead, there should be deliberate planning to design testing methods to suit the nature of the measurement technique. A compelling reason to measure pupil size should justify the cost and effort of possibly altering the experimental procedure, based on the desire to obtain information not available in behavioral methods. The pupil dilation response has complicated innervation and is affected by a wide range of experiences and stimuli, so there is an unfortunate amount of noise inherent in any pupil measurement. However, this noise can be addressed if the experimenter is careful with the experimental setup and judicious with the monitoring of factors that would affect physiological measures for any unique testing condition. Absence of these considerations will undoubtedly weaken the measurement and potentially cause distrust of the method altogether, undermining the field's confidence in the carefully produced studies that do exist.

Number of Trials

In the end, the number of trials (and participants) needed in any experiment will depend on the effect size of interest and power of the analytical approach. Generally, the experimenter will want to have at least 16 to 18 good recordings of pupil size for each condition. In any pupillometry experiment, there will be missing data because some trials will be dropped due to mistracking, contamination, or other reasons (e.g., scratch an itch or exercise a sore muscle can show up as surprisingly dramatic changes in pupil size that is unrelated to the listening task itself). Hence, it is wise to record a sufficient number of trials so that the estimation of the task-evoked response will stabilize. For sentence-perception tasks, 20 to 25 trials are normally a safe starting number. Fewer trials might be sufficient for listeners who are highly engaged in demanding tasks, where the effect is expected to be very large.

Number of trials for testing can be considered to be inversely proportional to the difficulty of the experimental task. For a very difficult task, a reliable large pupil dilation response (i.e., with a large-effect size) can be

achieved with as few as 10 trials. For distinguishing more subtle differences between similar conditions (e.g., vocoders with different number of channels, small changes in SNR, linguistic content such as semantic context), a larger number of trials is advisable. This consideration highlights the importance of authors publishing measures of effect size along with their statistical tests.

Trial Events and Timing

Trials should have consistent timing of events, for example, the onset of noise, an alerting sound, the stimulus itself, any cue to prompt a behavioral response, or any other relevant trial landmark. An illustration of an example trial timeline is given in Figure 1. Ideally, each trial should start with a drift-correction phase, in which the participant is required to look at a central fixation symbol before moving on (though this is not always possible when combining pupillometry with certain imaging modalities). Timing of each event should be planned carefully and intentionally. It is advisable to consider separating these events in time, because the pupillary responses to two events could sum together, obscuring dilations that arise from listening as opposed to those that arise from behavioral responses. Specifically, the *listening* portion of a trial could elicit a peak dilation, and a second peak in dilation could appear during the *verbal response* or *button press* portion. Sentence-repetition studies have varied in the duration of this retention interval, with times ranging from 5 s (Ohlenforst et al., 2017; Zekveld, Festen, & Kramer, 2013), 4 s (Koelewijn et al., 2012, 2015), 3.5 s (Koelewijn, Versfeld, & Kramer, 2017), 3 s (Koelewijn et al., 2012; Piquado et al., 2010), 2 s (Winn, 2016), 1.5 s (Winn et al., 2015), and some studies with variable interval durations (Zekveld et al., 2010; intervals ranged between 2.1 and 3.5 s), or no reported retention interval enforced (McMahon et al., 2016). In addition to potentially convolving the auditory and behavioral portions of pupil responses, a long retention interval might demand that a listener use short-term memory (for long intervals) or perhaps create pressure to rush to complete cognitive processing during a short interval. Shallower slopes of pupil dilation have been obtained by Zekveld et al. (2010) and Koelewijn et al. (2012, 2015) in numerous studies with longer retention intervals. A relatively shorter retention interval of 1.5 s used by Winn et al. (2015) yielded a relatively steeper slope and larger magnitude of dilation responses, perhaps because of increased pressure to respond quickly. In that study, prolonged dilations in difficult conditions of auditory degradations extended from the auditory-response peak all the way to the behavioral response peak with little recovery, while responses in easier conditions yielded quick recovery during the retention interval.

It has become more common for experimenters to introduce a cue that indicates the timing of an upcoming stimulus. For example, in tests of speech recognition in noise, there could be leading noise that lasts for 2 to 3 s before the onset of the speech (cf. Koelewijn et al., 2012, 2015, 2017; Wendt, Hietkamp, & Lunner, 2017; Wendt et al., 2018; Zekveld et al., 2010, 2013). There are at least two benefits of this practice. First, it alleviates the problem of target-masker separation, whereby simultaneous onset of speech and noise increases the difficulty of hearing the target signal. In addition, although the onset of sound could elicit a brief pupillary response, it could orient the listener so that the target signal of interest does not come as a surprise. However, the presence (or continuation) of noise after a signal, though common in published studies, could interfere with language processing, as shown by Winn and Moore (2018). As the pupillary response can be slow and long lasting, it is worthwhile to consider that the analysis window can be *after* stimulus delivery.

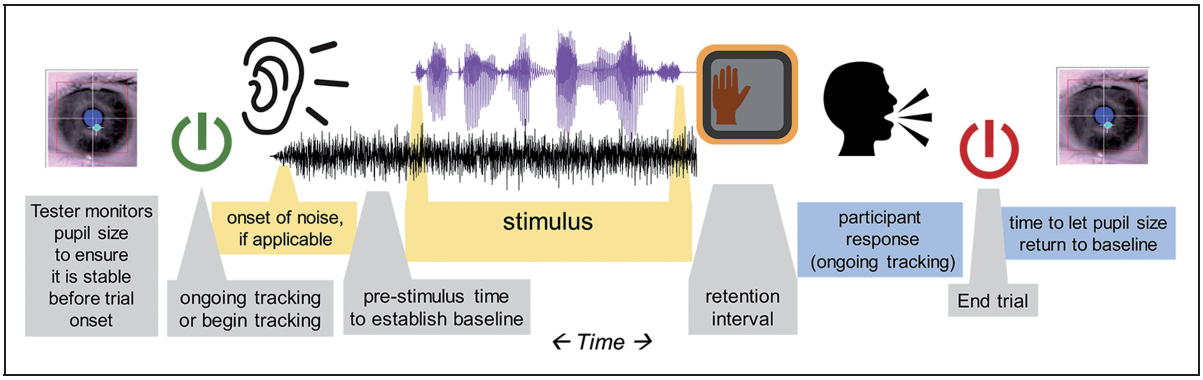

Figure 1. Events in a basic pupillometry experiment for measuring listening effort. There are other experimental paradigms that are possible, this illustrates a commonly used sequence of events.

Stimulus Duration

Most of the literature reviewed in this article features multiple trials of relatively short duration (2–6 s). For pupillometry, similar to other evoked measurements like EEG, magnetoencephalography, or auditory brainstem response, multiple stimuli of the same (or similar) type and duration are played in a testing block, and the responses are averaged in time. As long as the stimulus-driven portion of the overall evoked response is time-aligned, the part that is unrelated to the stimulus should be averaged out, leaving behind only the relevant task-evoked response. There will likely be cleaner patterns of data for these time-constrained stimuli compared with untimed stimuli, longer passages, or entire conversations, where one could not assume the same progression of cognitive processing landmarks trial-to-trial. Longer passages might not produce consistent patterns in dilation across stimuli (because of varying landmarks for parsing, resolution, or chunking) and therefore might have relevant phasic peaks neutralized by cross-trial averaging of data. The current article will focus on phasic responses to short stimuli.

Controlling the Visual Field

The amount of pupil dilation or constriction seen in response to changes in luminance far surpasses the amount of pupil dilation measured for cognitive tasks. Therefore, it is of critical importance to control the visual field when measuring task-evoked pupil dilation. Typically, the participant is stationary and visually fixated on an image that is either completely blank or with minimal stimulation. This is not to say that other protocols are impossible, but they would be subject to a higher amount of potentially confounding noise from movement, luminance effects on pupil size, and so on.

The setup in most labs includes a uniform solid color visual field that is neither too bright nor too dark. The visual field could be a plain wall, or a computer screen. Bright colors—especially white backgrounds on a computer screen—are problematic for multiple reasons. First, they could cause excessive pupil constriction; the cognitive response might not be strong enough to emerge. Second, they might cause discomfort for the participant, which we have noticed could result in a larger number of blinks and need for additional breaks during testing. Task-evoked pupil dilations have been observed reliably in dark-adapted conditions (McCloy et al., 2017; Steinhauer & Zubin, 1982). However, there are at least two cautions against testing in dark conditions. First, the pupils will dilate to accommodate low light, leaving less head room for task-evoked dilation. In addition, inspired by previous work by Steinhauer, Seigle, Condray, and Pless (2004), Wang et al. (2018) has recently shown that testing with brighter luminance elicits more reliable dilation because the parasympathetic nervous system releases its "grip" on the sympathetic nervous system's dilation-inducing projections to the pupil dilator muscles.

Combining Pupillometry With Eye Tracking

Despite risk of contamination by changes in luminance and gaze position, there are published studies where pupillometry has been used in studies of visual search or other visual recognition tasks (described in the next paragraph). These experiments offer the value of introducing the well-documented effects of lexical competition and sentence processing that have been studied with the "visual-world" paradigm, which is notable for providing precise timing information and insight on perceptual competition. Cavanaugh, Wiecki, Kochar, and Frank (2014) used a drift-diffusion model to suggest that eye tracking and pupillometry shed light on dissociable factors relating to decision-making. They found that gaze fixation time corresponds to rate of evidence accumulation, while increasing pupil size corresponds to increasing decision threshold (i.e., willingness to commit to a decision).

Visual aspects of stimuli in a gaze-tracking experiment could affect pupil size and therefore deserve extra scrutiny in the context of pupillometry. Engelhardt et al. (2010) used images in conjunction with pupillometry in a sentence comprehension task but did not publish examples of the images used. Wagner et al. (2016) used black and white line drawings in a picture-gazing task where lexical disambiguation led to changes in pupil dilation. It should be noted that in that study, concurrent gaze changes during pupillometry might have led to unknown effects on pupil size due to changes in gaze location and changes in the local luminance of the image on the retina. Using a variation of this method, Wendt et al. (2016) also used picture stimuli that were controlled to have equal luminance, and perhaps more importantly were presented *before* acoustic stimulus representation so that a pupillary response to the auditory stimuli would be measured independent of any visually driven changes in pupil size.

Although the aforementioned studies demonstrate that pupillometry could be combined with "visual-world"-style testing paradigms, there are special considerations to be made, in light of the influence of gaze position and luminance on pupil size. Kun, Palinko, and Razumenić (2012) reported that even for small targets (angular radius of 2.5°) changes in luminance can result in changes in pupil size that can obscure cognitive load-related pupil dilations. However, Palinko and Kun (2011) have also demonstrated that when the experimenter has rigorous control over the placement and luminance of objects in a visual scene, it is possible to

disentangle luminance and task-evoked changes in pupil size. In realistic everyday conditions, it might not be possible to exert such control. Kuchinsky et al. (2013) identified systematic changes in pupil size relating to gaze position, which were ultimately modeled and regressed out of the data.

Minimizing eye movement will likely lead to cleaner estimation of cognitive-evoked pupil size when using remote eye trackers, because gaze away from a remote stationary camera can cause a distorted estimation of pupil size, depending on the algorithm used. Systems that use the long axis of the ellipse fit to the pupil or that dynamically take into account the rotation of the eye away from the camera are unaffected by this issue (although one should always check their data to be sure). Methods for estimating and regressing out the degree to which a dataset is impacted by gaze position, including the proper design of a control viewing-only condition, have been described in detail by Gagl, Hawelka, and Huzler (2011) and others (e.g., Brisson et al., 2013; Hayes & Petrov, 2016; Kuchinsky et al., 2013).

Another reason to be cautious of eye movements in pupillometry tasks is that the luminance of visual field will change depending on what the participant is looking at, at any moment. If they shift gaze from a location with higher to lower luminance, pupil dilation might increase because of luminance instead of cognitive activity. Pupil size for a person shifting gaze around a room (or even around different areas of a screen) would be intractably convoluted with pupil size from luminance changes (and perhaps also with locomotion). Even if the visual scenes used are counterbalanced across the conditions of interest, one could not ensure a priori that participants would look at the displays in a consistent fashion across trials. In the best-case scenario, in which viewing patterns were relatively consistent, the added source of noise stemming from unpredictable changes in local luminance with fixations may minimize one's ability to detect differences across conditions.

Data Collection

Data Quality Monitoring

Data contamination should be detected as soon as possible—during testing. Real-time monitoring of the eye or the recorded pupil diameter shows blinks or other dropouts of data. If real-time monitoring is not an option, the estimation of pupil size could be displayed for the experimenter at the end of every trial, to see if something is amiss. Even in the absence of clear problems like head movement and shuffling posture, the pupil response can fatigue after several trials or can show a pattern of fluctuation—called *hippus* (see Figure 2). When hippus is observed, it is advisable to delay advancing to the next trial until the pupil size has stabilized. When it persists for over 10 s, this process can be aided by breaking the monotony of trials with a quick break to chat with the participant, or a brief irrelevant task (e.g., "look up to the corner of the room . . . now look back"). If the experimenter is not able to examine the time series of pupil size for each trial, it is at least recommended to monitor the eyes using a video stream of the pupil (as it is provided by most of the traditional cameras). Pupil size changes with mind wandering (Franklin et al., 2013), and the participant's mind might wander during a long and monotonous testing session. For that reason, it can be beneficial to introduce some variety or challenge to keep the participant alert.

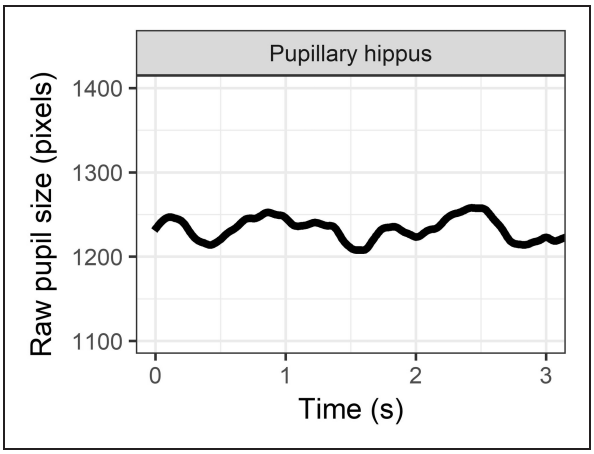

Figure 2. Pupillary *hippus*, or small ongoing fluctuations in pupil size that are unrelated to an external stimulus.

Data quality will likely change over the course of a long-testing session. It is common to observe a general decrease in pupil dilation over time, both in terms of baseline level and magnitude of dilation response. For experiments up to 1 to 1.5 h, these effects do not show up as significant. However, McGarrigle, Dawes, Stewart, Kuchinsky, and Munro (2017a) have shown an effect of task-related fatigue in pupil response during a longer sustained listening task. We have found that in typical sentence-perception experiments (with noise, or some other auditory distortion), fatigue is avoidable for most listeners if testing blocks are 2 h or shorter. Participants vary in how long they can engage and also their willingness to communicate their need for a break. Experimenters should remain vigilant for changes in participant alertness so that they can initiate breaks and avoid unwanted fatigue. Monitoring of data can reveal that the test is long enough that the participant is changing physiological state or alertness. After some reasonable number of trials (e.g., 25 trials in a sentence-recognition task), a break of a few minutes can refresh the listener.

Longer experiments could be split into different testing sessions although experimenters should be careful about splitting different compared conditions across different days, in case there are sizeable differences in pupil size or dynamic range for an individual across days. Be mindful that performance in a task can be situationally dependent and can vary by the day (Veneman, Gordon-Salant, Matthews, & Dubno, 2013). When possible, trials for different conditions could be interspersed or presented in alternating short blocks in the same testing period. The experimenter wants to ensure that the participant is in the same physiological state for each tested condition, so that any differences in pupil dilation are due to the task and not other unintended differences.

Pupillometry experiment setup and delivery improves with tester experience (just as for other methods such as EEG, where one detects when data are too noisy, develops criteria for removing an electrode, applying gel, etc.). One becomes more familiar with troubleshooting calibration and other unique situations over time, so early struggles with the method should not necessarily be taken as a sign that it will not be fruitful. Based on prior experiments and guidelines collected in this article, one could identify aspects of the testing procedure that would deviate from traditional psychoacoustics, like the increased interstimulus interval and extra attention to test difficulty and likelihood of participant disengagement.

The quality of pupil dilation measurements improves with attention to participant fatigue, comfort, readiness, and head movement. Although these factors might be noticeable in other types of behavioral psychoacoustic experiments, their effects might be even more damaging to a physiological measure like pupillometry. Examination of raw data (rather than aggregated smoothed averages) gives the experimenter a chance to identify situations that indicate that corrective action should be taken to the test protocol. For example, although blinks are normally not a problem (because they can be removed and smoothed over in postprocessing), an unusually large amount of blinks might indicate fatigue or a too-bright screen. Participants might also give long and tense blinks just at the moment of response, potentially erasing an important piece of the data. Consistently high variability in pupil baseline level before each stimulus might indicate that not enough time has passed since the last stimulus or response, as the pupil size might still be coming down from an evoked dilation.

Because attention to the aforementioned factors will likely improve with experience, we recommend that the testing procedure be at least as consistent and regimented as one would be in any other scientific procedure. It is also advisable to have testing be performed by those who have at least some practical experience with the method, perhaps by repeated practice and shadowing of more experienced lab members first.

Participant Inclusion and Exclusion Criteria and Other Considerations

Eye color. Most eye trackers are robust to differences in eye iris color, but there are occasional difficulties with very dark irises (for light-detecting systems) and light irises (for dark-detecting systems).

Makeup. Participants should be encouraged to avoid the use of mascara and eye-liner, as it can be erroneously detected as the pupil.

Age. Older listeners show generally weaker pupil dilation responses to light (Winn et al., 1994). Following Piquado et al. (2010), a control task that measures dynamic range is recommended when comparing younger and older adults.

Hearing status. Smaller amounts of pupil dilation are routinely observed in listeners with hearing loss and older listeners compared with young control groups with typical hearing (Koelewijn, Shinn-Cunningham, Zekveld, & Kramer, 2014). There is likely more than one reason for this, including listening fatigue draining a listener's cognitive resources, age-related atrophy of pupillary dilator muscles, or some other factors. It does not necessarily mean that the tasks performed by older or hearing-impaired listeners are regarded as less effortful. It could mean that they are devoting less intentional attentional engagement because they are conserving energy in a continuously exhausting task.

Pharmacological effects. Drugs can impact the ANS, which will affect the pupil dilation response. Steinhauer et al. (2004) report that blocking the sympathetically mediated alpha-adrenergic receptor of the dilator enables targeted measurement of the parasympathetic branch, while blocking of the muscarinic receptor of the sphincter muscles allows only contributions of the sympathetic branch. They showed that tropicamide (a parasympathetic ANS activity blocker) eliminated differences in the task-evoked response, while dapiprazole (a sympathetic ANS blocker) merely decreased pupil size while maintaining the phasic task-evoked response. It could therefore be especially important to guard against drugs that affect the parasympathetic nervous system. Common muscarinic antagonists that are used to treat for Parkinson's disease, peptic ulcers, incontinence, and motion sickness are all likely to inhibit the pupillary response.

Caffeine. Pupil dilations are larger after ingestion of caffeine (Abokyi, Oqusu-Mensah, & Osei, 2017).

Caffeine has been shown to affect the pupil response for up to about 6 h, particularly in people who do not routinely consume it (Wilhelm, Stuiber, Lüdtke, & Wilhelm, 2014).

Eye diseases. Some conditions might affect the biological function or appearance of the eye, such as cataracts (lowers contrast between iris and pupil), nystagmus, amblyopia (“lazy eye”), and macular degeneration.

Anything that affects visual fixation and tracking ability. Tracking can be compromised by attention deficit problems, severe fatigue. Tracking quality is sometimes affected by hard contacts and glasses (especially bifocals where refraction will change depending on eye position with respect to the lenses) although glasses do not always pose a problem and can usually be discarded in situations where there are no visual stimuli.

Head injury or any history of neurological problems. These issues can affect gaze stability, congruence of eye movements (Samadani et al., 2015), and pupil dilation (Marmarou et al., 2007).

General hearing ability (avoiding floor-level intelligibility). Participants who are unable to complete a task successfully will likely show reduced pupil dilation, because they might be more likely to abandon effort on the task.

Native language. When completing a task in a nonnative language, greater pupil dilation is observed, and some effects of language processing will deviate from those observed in native listeners (Schmidtke, 2014).

Fatigue. Although fatigue is obviously related to the study of effort, it can actually be a barrier to measurement of short-term task-evoked pupil dilation. Fatigued listeners will show a weakened pupillary response. McGinley et al. (2015) provide a clear and physiologically grounded explanation for the preference to test participants in a quiet and alert state, avoiding both fatigued and hyper-aroused states. Task-induced fatigue might be reflected in the baseline value of the pupillary response over the course of the experiment (i.e., lower baseline toward the end of the experiments). Chronic fatigue (need for recovery) effects the pupillary response as well (see Wang et al., 2018).

Measuring Pupil Dilation in Children

Relatively few published studies have used pupillometry to measure listening effort in children. Of those that have (e.g., Johnson et al., 2014; McGarrigle, Dawes, Stewart, Kuchinsky, & Munro, 2017b; Steel, Papsin, & Gordon, 2015), the age range appears to begin at 7 or 8 years. It is possible that the intentional attention mechanisms employed by adults and older school-aged children reflect cognitive activity that would simply not be invoked reliably by younger children. Furthermore, logistical constraints such as stabilized-head position, sustained attention, and patience for a very plain unstimulating visual field would certainly make pupil measurements in young children very difficult, even if their cognition and language skills were mature. It is therefore possible that pupillometry is not the ideal effort measurement to use with very young children. Later, we describe some studies of school-aged children and some related work on pupillometry in other young populations.

McGarrigle et al. (2017b) tested school-aged children (age 8–11 years old) and successfully measured differences in pupil dilation related to SNR. Notably, these SNRs did not produce differences in intelligibility, suggesting that children, like adults, can achieve the same score using different amounts of effort. Furthermore, behavioral response time did not distinguish the two listening conditions. McGarrigle et al.’s data demonstrate that it is feasible to use pupillometry for children of an age where attention and engagement are dependable, for at least 40 minutes. Incidentally, measurements in school-aged children with hearing loss might be *more* feasible, given their experience of annual (or more frequent) hearing tests that require the sustained attention and behavior that is somewhat reminiscent of pupillometry tasks.

Johnson et al. (2014) measured pupil dilation in children aged 7.5 to 14 years and obtained results that indicated reliable differences between children and adults on a short-term memory overload task. Specifically, dilation magnitude grew as memory demands increased, up to a plateau; adults’ dilations continued to grow up to a higher plateau (eight items), while children showed a reversal of dilation patterns after a smaller number of items (6) had been reached.

Steel et al. (2015) measured pupil dilation in 11 - to 15-year-old children, but the experimental design was in some ways not optimal for pupillometry as much as it was for tests of binaural fusion. They measured peak pupil diameter for a 2-s window following stimulus onset, in an experiment where average reaction times spanned a range of 2 to 3.5 s, possibly resulting in the exclusion of true peak dilation which likely occurred after the pupil data recording period. Correlations between binaural hearing and pupil dilation in that study were reported but appear to be driven by overall group differences rather than within-group binaural hearing ability and also were affected by ceiling effects and general effects of age.

Pupillometry in children younger than 8 years is rare and is typically used for purposes other than listening effort tasks. Recovery latency of pupil dilations has

been used as a biomarker for children at risk for autism spectrum disorder (ASD; Martineau et al., 2011; Lynch, James, & VanDam, 2017). Pupil size was also reported to be a biomarker for ASD by Anderson and Columbo (2009) although that study included a small number of participants, and, despite statistically detectable differences, data for the ASD group fell within the range of the control group.

Measuring pupil diameter in young children during listening tasks is a substantial challenge, for both theoretical and logistical reasons. Changes in pupil size can be measured in 8-month old infants in reaction to surprising physical events (Jackson & Sirois, 2009), and both 6- and 12-month old infants show increased pupil dilation to odd social behaviors (Gredebäck & Melinder, 2010). Thus, the pupil response can be measured; for the purpose of this article, in question is whether the assumptions that we make about the nature of language processing and goal-directed task engagement used by adults in speech recognition tasks could generalize to very young listeners.

Hardware

Trackers. It is not within the scope of this article to recommend a particular product, especially because products continue to be improved with each year. A majority of pupillometry articles in the area of listening effort have used traditional eye trackers that might more commonly be used to track eye-gaze direction. They come in many varieties, including remote cameras (that sit on a desk beneath a monitor display), tower stands (which record a reflection of the eyes akin to a teleprompter in reverse), and eyeglasses outfitted with cameras. Many of these instruments also report an estimate of pupil size, with some degree of error. Quality of the camera and quality of the software algorithms for calculating pupil size are of extreme importance, for three main reasons. First, the pupil is small, so the amount of noise in the pupil size estimation must be limited. Second, the time it takes for the system to recover from losing track of the pupils (in the case of a blink, or a look off-screen) can result in the loss of valuable data. Finally, a change in pupil size can be indistinguishable from a change in distance to the camera unless head position is stabilized, or if there are supporting measurements made, like distance. Trackers that report absolute pupil size (in millimeters) necessarily must complete such a calculation, albeit sometimes without transparency in how it is done. Some trackers instead report pupil size in arbitrary units, akin to the number of pixels that the pupil occupies on a camera image. In addition, while some eye trackers model the rotation of the eye away from center or correct pupil size for gaze position in other ways, other trackers do not, and thus extra caution (such as applying correction factors; see Brisson et al., 2013; Gagl et al., 2011; Hayes & Petrov, 2016) must be taken into account when measuring pupil size in experiments that also feature eye movements.

Clinical pupillometers. Hand-held clinical pupillometers—as used for neurology, ophthalmology, and emergency medicine—have the advantage of being user friendly (via automated routines), less expensive than some full-fledged video-based eye trackers, and designed specifically for accuracy in measuring pupil size. However, they might not have been designed for research, which could result in limitations on recording time, lack of connectivity with popular experiment delivery software, or lack of synchronized event tagging.

Chin rests or other head stabilizers. Pupil size can be estimated more reliably if the distance from the eyes to the camera remains constant (particularly for trackers that do not automatically attempt to correct for distance). It is customary to use a stabilizer such as a chin rest, akin to what could be used at an optometrist's office. However, chin rests are not always comfortable for participants, especially when they are giving verbal responses, or if it requires them to lean forward unnaturally. An alternative solution is to have the participant lean back to have her or his head position stabilized on the top of a sturdy and stationary chair.

Seating. Sturdy stationary (not rolling) chairs will make measurement easier. The participant's comfort should be taken into consideration even more than for a traditional psychoacoustic experiment, because the act of shifting posture or tensing muscles will show up as changes in pupil dilation. A height-adjustable chair akin to a hairdresser's chair (or height-adjustable table for the camera) is advisable to maintain a constant viewing angle and comfort for all participants. There are also chairs used for EEG recordings that have adjustable headrests to avoid muscle tension in the neck.

Room lighting. Light should be homogeneous in the whole room so that if a participant looks around, it won't cause a reflexive dilation in response to changing luminance on the retina. Soft lighting is best, especially if it is adjustable for individuals (Zekveld et al., 2010). A range of 10 to 200 lux, with a median for older adults around 30 lux and for younger adults around 110 lux depending on the dynamic range of their pupil. As a reference, a normal in offices is around 300 to 500 lux.

Brighter luminance produces more reliable dilations than dark settings (Steinhauer et al., 2004; Wang et al., 2018) but take caution that too-bright lighting (especially projected directly at a participant from a computer screen) might also cause discomfort and high number

of blinks. A moderate mid-range gray color background on a computer monitor or a plainly lit wall target avoids these issues of discomfort.

Handling of Raw Data

Sampling rate. The pupil changes size slowly, so a sampling frequency of 30 Hz or higher is sufficient. Very high sampling frequencies (e.g., above 120 Hz) of some trackers would be beneficial for studies of precise saccade timing but are not necessary for most pupillometry studies.

Data transfer. To ensure that stimulus timing landmarks are recorded and synchronized with corresponding timestamps in the eye tracking or pupillometry data, the experimenter should be sure that time tracking would not be compromised by the use of a single computer to handle all of the processing. There are two-computer solutions that use physically separate computers for experiment delivery and tracker data collection, with ethernet or USB links for data transfer. Timing is not as delicate an issue as it is for other methods such as EEG; there are also single-computer pupillometry solutions, which can be sufficient since pupillary responses are slow enough that a drift of 30 ms (less than the duration of one sample at 30 Hz) should not affect the quality of data.

Monocular and binocular tracking. The pupils should show congruent dilation patterns (Purves et al., 2004), so binocular tracking might not offer any substantial advantage over monocular tracking, apart from the opportunity to pick the eye that produces the fewest missing data samples.

Stimulus Timing

Of critical importance is waiting for the pupil to return to baseline size before the next trial. The duration of this interval will depend on the experimental task. Heitz et al. (2008) found that larger dilations on difficult test trials affected baseline levels for subsequent trials, even with interstimulus intervals of 3.5 s. Sentence repetition tasks might require nearly 4 to 6 s following the end of the participant's verbal response (discussed in further detail later).

Response Timing

The pupillary response takes up to 1 s to emerge, with estimates ranging from roughly 500 ms to 1.5 s (Hoeks & Levelt, 1993; Verney, Granholm, & Marshall, 2004). McGinley et al. (2015) found that the derivative of the pupil was correlated to the pupil diameter 1.3 ± 0.7 s after corresponding cortical oscillations. Peak timing in sentence-recognition experiments appears to follow the same time course, emerging typically 0.7 to 1 .2 s following stimulus offset (Winn, 2016; Winn et al., 2015). Systematically longer stimuli elicit longer latency to peak in situations where duration differences are known by the participant before the trials begin (Borghini, 2017; Winn & Moore, 2018).

What Data to Record

In addition to pupil dilation, the experimenter should record accurate timestamps of the onset and offset of a stimulus, the timing of behavioral response (if any), and the horizontal and vertical gaze positions of the eye. Timestamps will be used to aggregate and align data, and the gaze coordinates can be used to ensure fixation at a target, as well as to covary gaze position with pupil size estimation.

Data Processing

Raw pupil data must be processed in several steps before analysis and visualization. Figure 3 illustrates common steps in treating pupil data, described later.

De-blinking

Blinks are generally not a problem if they are quick (<125 ms) and uncorrelated with stimulus timing landmarks. They are typically brief enough that they can be identified, removed, and interpolated in the data without substantial change to the overall pattern. It is therefore

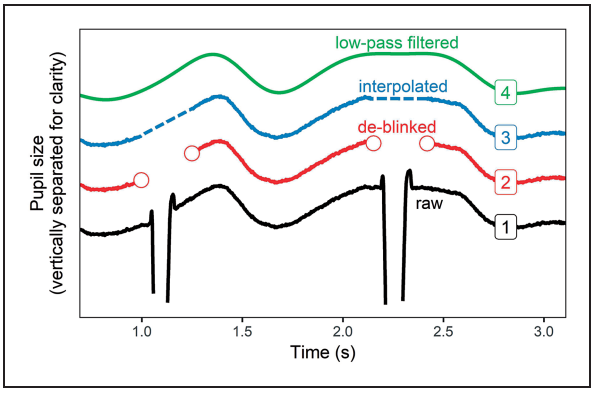

Figure 3. Sequential steps of data processing. Raw data (black, marked no. 1) contain blinks that appear as transient changes in pupil dilation separated by a blank stretch of missing data. De-blinked data (no. 2, in red) expands the gap of missing data to remove the transient excursions. The gaps are interpolated (no. 3 in blue, interpolations in dashed lines). Finally, the data are low-pass filtered (no. 4, green).

typical to not give any explicit acknowledgment of blinks during testing, to avoid conscious awareness or patterned blinks. In the case that participants exhibit blinks that are time-locked to trial events (e.g., always blinking at stimulus offset, at onset of verbal response, etc.), then there is risk of data distortion. A. Wagner et al. (2016) addressed this issue by incorporating a *Blink now* event at the end of each trial.

Klingner et al. (2011) performed a thorough analysis of 20,000 blinks to estimate the expected perturbation of pupil size. They report that the pupil size before the blink undergoes a very brief dilation of about 0.04 mm, followed by a contraction of about 0.1 mm and then a gradual recovery to preblink diameter over the next 2 s. They also reported that the difference in statistical results did not change with the incorporation of a blink correction algorithm. It is likely the case that interpolating across blinks is a safe practice that will not affect results in any meaningful way. However, we recommend that interpolation begin roughly 50 ms before the blink and end at least 150 ms after the blink in order to avoid task-uncorrelated high-frequency changes in pupil size. Note that when a pupil trace consists of a larger percentage of blinks (>15%–25% of the relevant recorded time), interpolation might result in a flat trace that no longer shows a pupil dilation response. These traces should not be used for further analysis.

Low-Pass Filtering

Klingner et al. (2011) analyzed binocular pupil measurements and found that changes faster than 10 Hz are uncorrelated across the eyes, thus justifying low-pass filtering at 10 Hz. High sampling rates are therefore not essential for pupillometry. However, it would be advisable to maintain a sampling rate high enough to distinguish between fast and slow pupil responses (in terms of derivative or rate), as they have been shown to be driven by different neural systems (Reimer et al., 2016). Numerous studies report using an n-point smoothing average filter in place of an explicit low-pass filter.

Baseline Correction

The most common method of quantifying pupil dilation is not to report absolute pupil size but instead to report *change* in pupil size relative to the time immediately before the stimulus (Beatty & Lucero-Wagner, 2000). Collecting literature from a variety of studies (Bradshaw, 1969, 1970; Kahneman & Beatty, 1967), Beatty (1982) argued that task-evoked changes in pupil size are independent of baseline pupil size, at least for intermediate tonic sizes. Reporting baseline-subtracted absolute pupil size is common (Beatty, 1982; Kramer et al., 1997; Zekveld et al., 2010) but we are unaware of any empirical verification of Beatty's claim. Since baseline pupil size typically will vary across people, vary within people across time, and will gradually diminish over the course of a testing session, this is a topic deserving of exploration. One approach is to drop the first few trials because baseline levels are substantially higher during the onset of a testing session but quickly stabilize after roughly five trials (Wendt et al., 2016; 2018). Apart from discarding trials at the onset of a session, it appears that the common practice is to handle these unknown sources of variability as one would handle other population-level and trial-level sources of noise, by recruiting a sufficiently large number of participants and presenting a large number of trials.

Duration of baseline intervals ranges from 100 ms (Karatekin et al., 2004) to 2 s (Ayasse, Lash, & Wingfield, 2017). Figure 4 illustrates how variation in the absolute baseline duration should play no substantial role in reporting pupil dilation. For baseline durations extending from 100 to 3000 ms, the baseline-corrected data are virtually identical. However, there are some factors that probably contribute to the common practice of using 1 s. First, it is a duration long enough that a single blink would not eliminate all data from the baseline window. However, it is short enough that to hopefully minimize influence of pupil dilations from a previous trial, as long as there is a sufficient intertrial interval.

Baseline is typically computed for each trial, rather than for a whole test session, since the baseline level typically will drift downward over the course of a session (which, when using a single baseline average, would result in underestimation of dilations for early trials, and overestimation of dilation for late trials). To avoid single-trial erratic deviations in baseline that are unrelated to the stimulus (e.g., random large excursions in baseline size, or a long blink), one could also compute each trial's baseline as a rolling average over a number of adjacent trials.

To elicit a more consistent baseline level across trials within a participant, one could use consistent timing landmarks and alerting stimuli to signal for trial onsets. For experiments of speech in noise, this could mean having a consistent duration of noise before stimulus onset so that the participant knows when to listen. For speech in quiet, Winn and Moore (2018) have used an alerting beep denoting trial onset, with the intention of avoiding any *surprise* responses to the target speech. Unexpected patterns on baseline levels should be explored by viewing raw trial-level data.

Baseline pupil size, while not always reported in many published studies of listening effort, has been a subject of investigation. Some authors have suggested that tonic pupil size is a measure of global arousal levels (Granholm & Steinhauer, 2004), with alternative viewpoints framing tonic pupil size as an indicator of

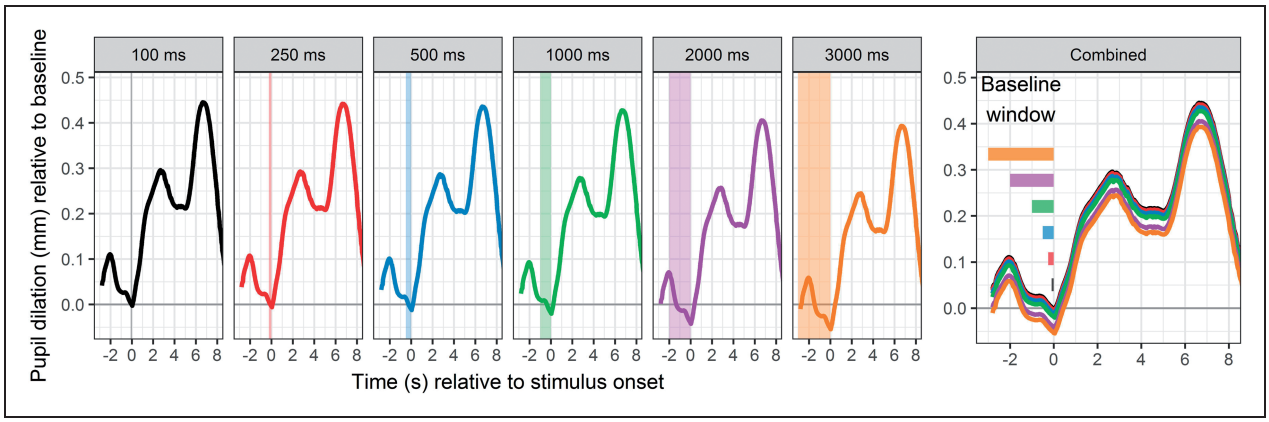

Figure 4. Different baseline intervals end at the onset of the stimulus and extend backwards by variable durations (highlighted in each panel by a shaded vertical area). Comparison of the resulting baseline-corrected data is shown on the far-right panel, revealing negligible differences across baseline durations.

attentional capacity (Kahneman, 1973). Based on a review of various physiological studies done with animals, Laeng et al. (2012) suggest that tonic activity (i.e., not stimulus-time-locked) of the locus coeruleus (LC; indexed by pupil dilation, to be discussed further later) indicates the likelihood of abandoning a current task for another, while phasic activity signals the processing of attended task-relevant events. Expanding and updating this idea, physiological work with mice (McGinley et al., 2015) suggests that tonic pupil size is related to moment-to-moment readiness for sensory detection, with intermediate sizes measured during trials with better task performance and reduced response variability. Conversely, very low tonic size was interpreted as a sign of indicating drowsiness, and very high tonic size was also found to be suboptimal, consistent with hyper-activity that would cause asynchronous cortical activity.

Figure 5 illustrates two approaches to baseline correction methods that have been observed in the literature: absolute subtraction and proportional transformation (which can be considered an additional follow-up step following baseline subtraction). In the top panel, changes in raw pupil size are shown from two hypothetical participants. Participant 1 shows greater pupil size and greater apparent difference between dilation sizes in response to two different stimulus conditions (indicated by line color). However, the apparent lack of condition effect for Participant 2 is simply hidden by baseline differences. When subtracting baseline size to yield an absolute change in pupil size, Participant 1 retains a higher overall change in dilation, but the differences between conditions are now apparent for Participant 2 as well. When analyzing proportional differences relative to baseline, the two participants' responses are transformed to look virtually identical. The consequences of these and other methods should be considered in situations where participants in different comparison groups have different overall dynamic range of pupil reactivity or substantial changes in baseline (as a result of, e.g., age differences, or testing at different times). It is worthwhile to consider Beatty's (1982) aforementioned assertion that absolute dilation is independent of baseline size (which would undermine the proportional method) but to also consider data from Piquado et al. (2010) who normalized dynamic range to overcome substantial differences in pupil reactivity across younger and older participants. Systematic study and replication will hopefully discern the optimal way to treat data with variable baseline size.

Normalization

An equivalent amount of pupil dilation across two participants could be more meaningful for the one whose pupil has a smaller dynamic range. One such known contributor to overall dynamic range is aging. Older individuals tend to have pupils that are smaller in size, with a more restricted range of dilation, and which take longer to reach maximum dilation or constriction (Bitsios et al., 1996). In light of interindividual differences in pupil dynamic range, normalization methods are sometimes applied. Following baseline correction (e.g., subtraction of baseline level), local deviation from baseline can then be expressed in numerous ways, including percent or proportional change from baseline (Hess & Polt, 1964; Johnson et al., 2014), percent change of average experimental trial value versus average control trial values (Payne, Parry, & Harasymiw, 1968), within-trial mean scaling (Kuchinsky et al., 2013), grand-mean scaling (Engelhardt et al., 2010), *z*-score transformation (McCloy et al., 2016), expressing dilation as a proportion of the individual's dynamic range of pupil size

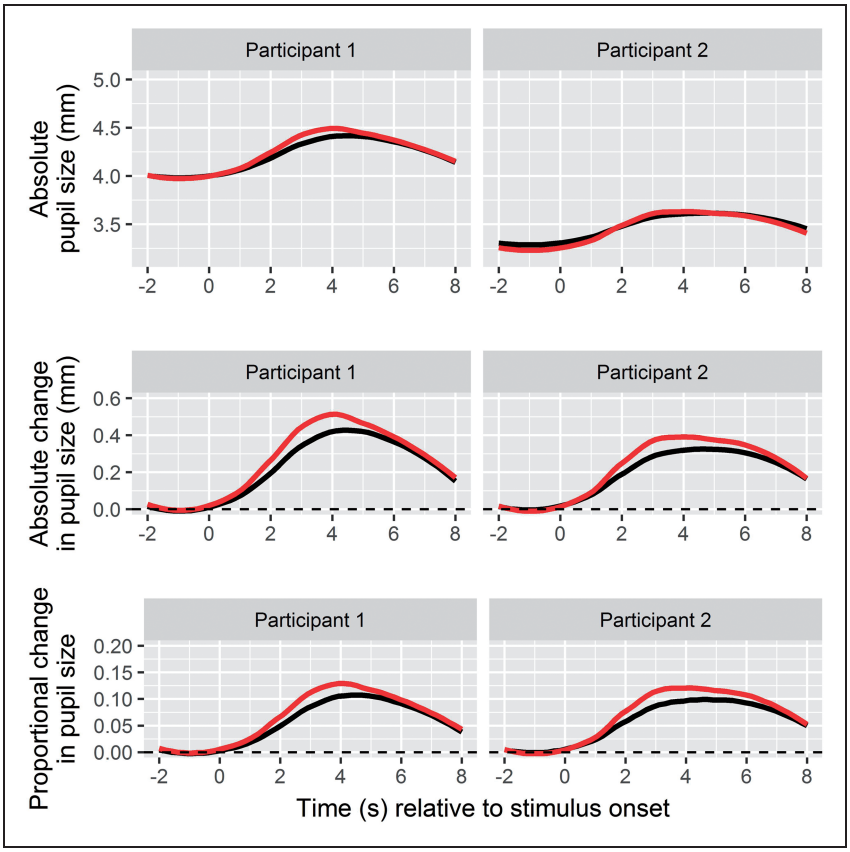

Figure 5. Illustration of baseline correction and proportionalization of data for two hypothetical individuals each participating in two testing conditions indicated by line color. Raw pupil size is shown on the upper panels, absolute change (mm) in pupil size is shown in the middle panels, and proportional change in shown in the lower panels. Baseline intervals consisted of the 1-s of data prior to stimulus onset indicated at time 0.

(Piquado et al., 2010), or as a proportion of a reference peak dilation within the individual (Winn, 2016).

Each of these methods has advantages that might suit some experimenters' needs. For example, the percent-of-range and *z*-score calculations address interindividual differences in *variability* in dilation (which might be useful in situations where engagement is different, such as for responses by normal-hearing and hearing-impaired listeners), whereas the others can correct for *average* differences and the method used by Piquado addresses *reactivity* or dynamic range, which could be important for experiments pertaining to aging. It remains unknown whether the percent or proportional methods are immune to overall differences in baseline magnitude, but these methods—as well as completely nonnormalized methods used for instance by Zekveld et al. (2010) have still been found to be effective in identifying task-related differences in pupil size within participants even when averaged across a number of participants who likely differ in pupil reactivity.

Figure 6 illustrates a hypothetical situation that illustrates an application of Piquado et al.'s (2010) method, in which change in pupil size is expressed as a proportion of the full dynamic range elicited by the pupillary light reflex. The apparently larger change in amount of evoked pupil dilation in "Participant 3" is rendered equivalent to that obtained in "Participant 4"; twice the change was observed for a pupil with twice the dynamic range. An alternative method has been used by Winn (2016) and Winn and Moore (2018), who compared the pupil dilation in one condition to the peak dilation obtained in a reference condition. Proportional change between the two conditions was considered to be normalized within individuals and therefore free from individual differences in pupillary reactivity. An intriguing area of future research could be to create a control task that determines the cognitive dynamic range of the pupil (e.g., a working memory task ranging from one item until cognitive overload). The pupil response could then be reported as a percentage of this *cognitive-evoked* range rather than the *light-evoked* range.

Although there is no consensus gold standard method of normalization, this problem has been commonly addressed (or rather *avoided*) by (a) analysis of differences within the same participants across conditions,

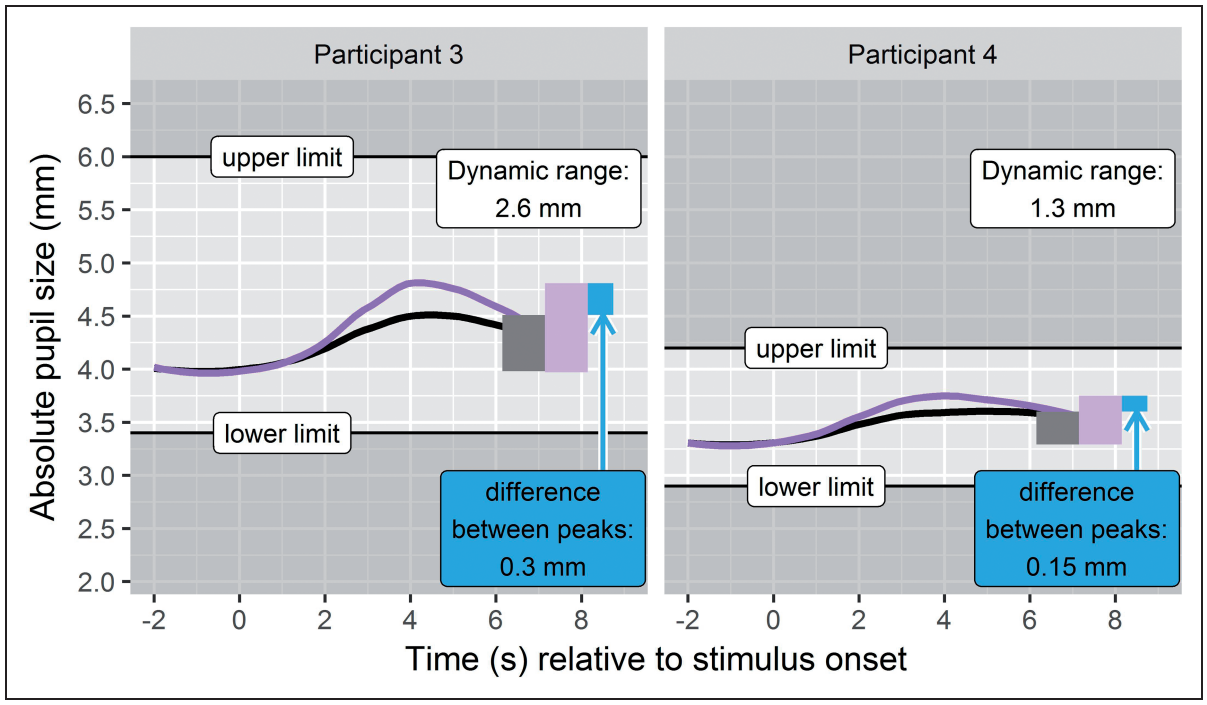

Figure 6. Different amounts of change in pupil dilation for two hypothetical individuals who have different dynamic range of pupil size. In each panel, the difference in peak pupil dilation occupies the same proportion of the overall dynamic range.

with the assumption that individual effects would be common to all tested conditions or (b) the use of sample sizes large enough to overcome the variability, with the assumption that individual differences in pupil reactivity would be approximately equally represented in comparison groups.

Contrary to the aforementioned attempts to normalize differences in pupil reactivity, these differences might be a meaningful outcome measure. For example, Koelewijn, van Haastrecht, and Kramer (2018) found that participants with history of traumatic brain injury showed reduced phasic pupil dilations, despite reporting generally higher subjective effort ratings. In addition, peak pupil dilation correlated negatively with participants' speech reception threshold, suggesting a decrease in phasic activity with a decrease in hearing acuity. Another example is given in by Jensen et al. (2018) who examined the impact of tinnitus and a noise-reduction scheme on the pupil response. They found that participants with tinnitus generally reported a greater need for recovery and showed smaller task-evoked pupil dilations (in contrast to the hypothesis of greater effort—greater dilations). The results of these studies suggest that there is a possibility of interpreting reduced pupil dilation not merely as reduced effort, but potentially as lower *capacity* (consistent with Kahneman's [1973] framework), or perhaps as reduced ability to maintain engagement. The multitude of possible interpretations highlights the need to design experiments that differentiate related concepts such as effort, attention, engagement, and fatigue.

Baseline Analysis

Numerous studies propose that the baseline (or *tonic*) pupil size is not simply noise to be discarded or normalized but rather inspected for its own sake. Gilzenrat, Nieuwenhuis, Jepma, and Cohen (2010) suggest that increases in baseline pupil diameter reflects disengagement from a task (specifically, to explore more useful tasks), whereas smaller baseline diameter would indicate increased engagement, and also reveal relatively larger task-evoked dilations. McGinley et al. (2015) have observed that tonic pupil size is related to the accuracy of performance in psychoacoustic tasks by mice. They further provide a framework for using tonic pupil size (i.e., pupil size not evoked by a specific task or stimulus) as an index of general arousal (or "brain state"), with medium-level arousal yielding optimal performance for sensory-cognitive tasks. Heitz et al. (2008) corroborated this finding in humans, finding larger prestimulus tonic pupil size for participants with higher working memory span compared with those with shorter memory spans. Even researchers primarily interested in the task-evoked pupil response should examine baseline epochs to ensure differences in peaks are not due to systematic, condition-specific biases in the baseline window. If such patterns emerge, then the baseline-corrected and normalized

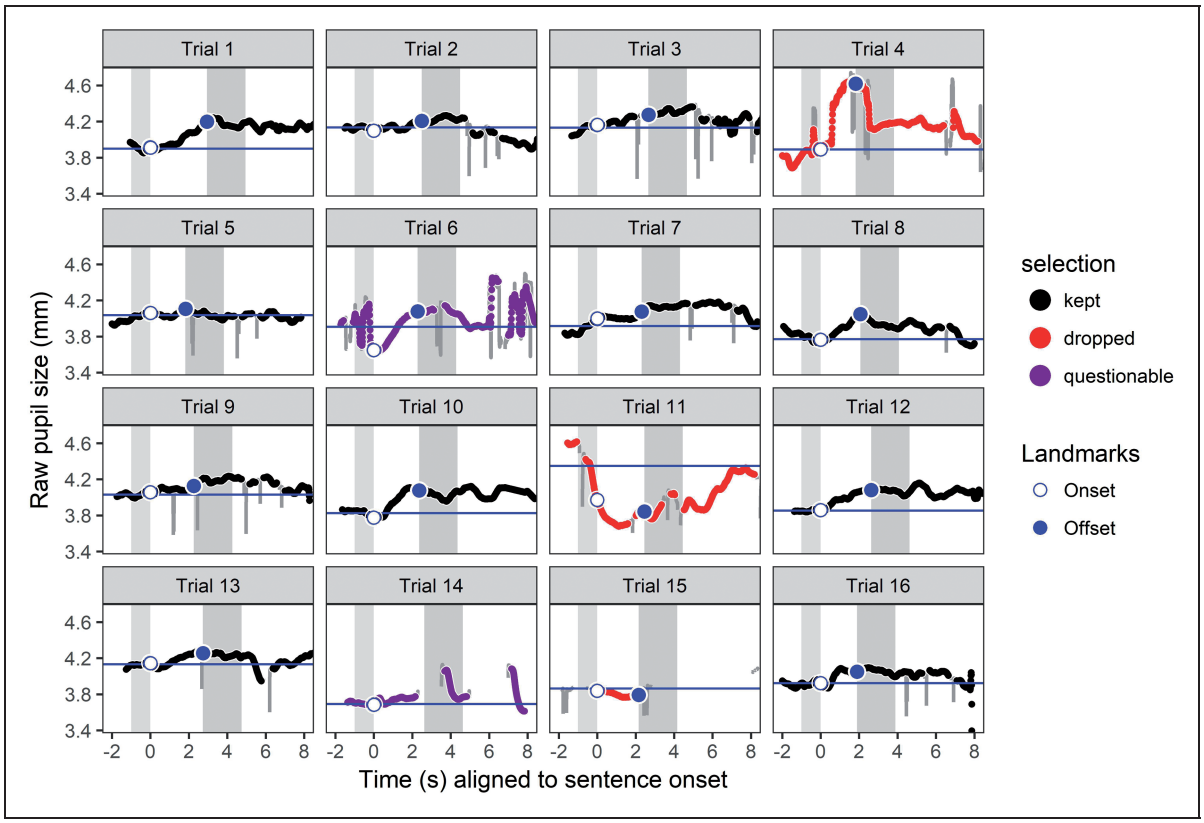

Figure 7. Sixteen individual trials of pupil data, with baseline period marked as thin gray vertical bar, and retention interval marked as thick gray vertical bar. The stimulus is played between these two bars. Baseline level for each trial is marked as the horizontal blue line. Lines plotted in color are low-pass filtered data overlaid on gray raw data that include transient vertical displacements that indicate blinks. Data in red are marked to be dropped from the data set due to excessive data loss or contamination (excessive nontask-evoked dilation) during baseline interval.

dilations should be (a) interpreted with extra caution and (b) motivate inspection of the experimental protocol to discover any unintended bias. We do not necessarily recommend discarding these data, since the baseline pattern might be indicative of an effect that is worth interpreting or using to motivate a follow-up experiment. For example, if experiments are to be conducted at very different times of day, differences in arousal or alertness might affect results (Veneman et al., 2013).

Data Alignment

Time alignment is standard in pupillometric analysis, because task-irrelevant changes in pupil size are likely to neutralize when averaged across multiple trials, leaving only the evoked changes relevant to the experiment. In cases where stimuli are of variable duration (like sentences in a standard corpus), alignment can be done by onset or offset. Alignment by stimulus onset has been used to examine the effect of intelligibility (Zekveld et al., 2010) and masker type (Koelewijn et al., 2012) on pupil size. Klingner et al. (2011) aligned their data to stimulus offset for purposes of looking at peak pupil dilation resulting from stimuli of systematically longer durations. Winn et al. (2015) and Winn (2016) aligned data to stimulus offset for the purpose of separating listening responses from speaking responses and serendipitously found temporally specific effects of prolonged effort following challenging stimuli, which would have been otherwise partially obscured by onset alignment. Figure 7 shows a series of trials with stimulus onset and offset marked, and Figure 8 shows the aggregate of these trials, aligned by both onset and offset. There is arguably a more distinct onset slope for the onset-aligned data, and more distinct shape of dilation at offset for the offset-aligned data, but the data take the same general shape regardless of alignment position.

On a technical note, it is important to consider that there could be some timing drift or numerical rounding that occur with timestamps. For example, for 60 Hz data collection, some timestamps might be multiples of 16 while others are multiples of 17, even if they refer to the same ordered index (i.e., 1st, 2nd, 3rd, etc.) of sampled time across trials. To ensure that such data points are numerically aligned when aggregating across

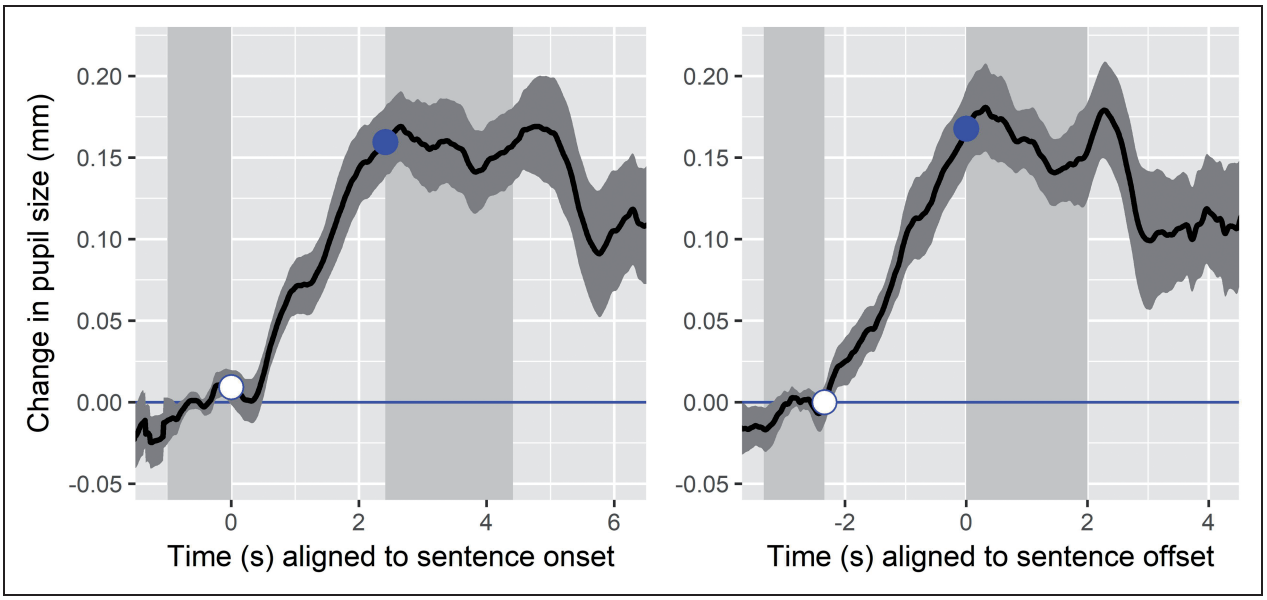

Figure 8. Aggregation of trials displayed in Figure 7, excluding "dropped" trials. The left and right panels display data aligned to stimulus onset and offset, respectively. The thin gray bar (to the left in each panel) corresponds to the baseline interval, and the thicker gray bar (to the right) corresponds to the retention interval. Because stimuli were of variable duration, average values were used for the offset time for onset-aligned data, and for onset in offset-aligned data.

trials, the experimenter might need to transform the timing data to maintain consistent timestamps.

Identifying and Dropping Contaminated Trials

Trial-level data that are contaminated (e.g., by gross excursions during baseline) should be removed from analysis using automatic rules that are consistent and motivated by realistic constraints of task-evoked pupil responses. There is no substitute for visualizing and becoming familiar with the *trial-level* data so that unexpected deviations can be detected and used to design algorithms to filter data. The nature of these deviations can vary among individuals and among eye-tracking systems. The strategy taken to drop trials should be consistent and blind to test condition, to avoid experimenter bias. There are some useful heuristics that can drive an automated trial exclusion process. For example, when a considerable amount of data samples are lost (e.g., because of long blinks), a trial should be dropped. Exact missing data criteria vary across studies, but generally range from 10% to 30%, and might be especially important for parts of the trial where the peak dilation would occur; the criteria for trial dropping might therefore be weighted by trial event time.

Often, the first few trials are excluded from analysis because the pupillary responses look considerably different from those in the rest of the block (Wendt et al., 2018). When pupil dilation slopes steeply downward during stimulus playback, it is a good time to consider dropping the trial, because the "peak" dilation will likely be more than three standard deviations from the mean. Figure 7 illustrates an example of this pattern, as well as some other examples of contaminated or mis-tracked trials that could be dropped from a data set. We recommend being very conservative with dropping trials, especially if there is no obvious artifact like an eye movement or something explaining a deviating response. All remaining unexplained noise should be addressed by event-related averaging akin to other evoked responses.

Just as for other evoked-response methods, outliers should be detected and removed from the data set. In Figure 7, data for Trial 11 are marked to be dropped because an event-unrelated brief and excessive dilation that occurs during baseline driving all subsequent data downward following baseline correction. This contamination can be detected by identifying the average or peak dilation as an outlier, or by detecting deviation of the baseline value relative to adjacent baselines. Trial 15 is dropped because it contains a large amount of missing data due to blinks or tracker error. Trial 4 is marked to be dropped because of excessive distortion precisely during the time where there would be valuable dilation information. The range and rate of dilation observed in this trial is uncharacteristic of task-evoked changes but likely to detrimentally affect data aggregation. Trial 6 is marked as "questionable" because it contains excessive distortion during baseline (and later), which propagate forward to affect the calculation of subsequent data samples in the trial (just as for Trial 11). Standard criteria

(such as absolute value of peak dilation or baseline levels being more than 3 standard deviations from the mean) could help to identify such situations. Measures of dissimilarity of individual trials might be applied in the future to characterize such deviant trials and evaluate them for contamination.

Although most of the group-averaged data from publications cited in this article take the general form illustrated in the figure earlier, it is important to note that morphology of responses will vary substantially across individuals. Lõo, van Rij, Järvikivi, and Baayen (2016) describe distinctive group patterns including differences in the number of peaks, the timing of peaks relative to trial landmarks, as well as the tendency for pupil size to either rise or *fall* after stimulus onset.

Analysis Techniques and Time Windows

As mentioned earlier, the pupil will start to dilate between roughly 0.5 to 1.3 s following the stimulus onset, and the peak dilation occurs typically roughly 700 ms to 1 s following the end of the stimulus (at least for sentence-recognition experiments where stimulus duration was relatively constant and therefore easy to predict by the listener). It is customary to measure peak pupil dilation, peak pupil latency, and mean pupil dilation in a fixed window of time around stimulus presentation. The popularity of these measurements (and peak dilation in particular) should not limit the creativity of experimenters to develop and use measurements that relate to specific hypotheses regarding the timing of the response. Some investigators have also measured the shape of the pupillary response over time (Kuchinsky et al., 2013; Wendt et al., 2018; Winn, 2016; Winn et al., 2015) using growth-curve analysis (Mirman, 2014) or generalized additive models (van Rij, 2012; van Rij, Hendriks, van Rijn, Baayen, & Wood, 2018). Alternative approaches include principal components analysis used in pupillometry experiments by Schluroff et al. (1986) with seven factors and Verney et al. (2004) with three factors.

In the case of stimuli that do not have any specific internal landmarks (such as conventional speech-in-noise tasks), mean dilation, peak dilation, and latency should suffice. However, sometimes stimuli should elicit cognitive load at specific times, and this should be approached with time-series methods or carefully chosen time windows. For example, in an experiment by Bradshaw (1968), latency to peak dilation depending on task instructions; in an open-ended problem-solving condition, latency was locked to the time of solution and reflected stimulus difficulty, but in a condition where the solution was elicited with a predictably timed probe, latency to peak was rather consistent despite differences in stimulus difficulty.

Analysis windows can vary according to the design of the test procedure. In many cases, a single large analysis window including stimulus and response preparation can be used to detect main effects such as speech intelligibility (Zekveld et al., 2010) and masker type (Koelewijn et al., 2012). In some cases, a briefer analysis window can reveal how the growth from baseline to peak can reveal differences that can be subsequently neutralized as the participant prepares a response (Winn et al., 2015). Two explicit analysis windows have been used to separate listening and rehearsal phases of sentence-repetition tests (Winn, 2016). Wendt et al. (2016) used three separate analysis windows designed to track listening, linguistic processing, and decisions. Alternatively, analysis of pupil dilation could be locked to stimulus events such as disambiguating information during a stimulus (Wagner et al., 2016).

Various language-related and aptitude-related factors can affect pupil dilation in ways that might be best described by factors other than mean, peak, and latency. For example, in a study of mathematical problem solving by Ahern and Beatty (1979), there were two groups of listeners that were found to have equivalent peak dilation and latency, but different slopes of recovery after peak; there was quicker recovery from peak for higher aptitude students. Similar patterns were described in a pitch perception task by Bianchi et al. (2016), where musicians had quicker recovery than nonmusicians. Bradshaw (1968) showed roughly equivalent peak dilations more difficult math problems regardless of whether participants obtained a solution, but *prolonged* dilation was observed when problems remained unsolved, suggesting a useful role for late-stage dilation analysis. In cases of using semantic context, the timing of changes in pupil dilation *after* stimulus delivery has been proposed as a potentially meaningful distinction across listener groups (Winn, 2016; Winn & Moore, 2018). Previous work by Verney et al. (2004) proposed using principal components analysis to identify early, middle, and late contributions to dilation, specifically mentioning a late factor responsible for dilation magnitude following the peak.

If the experimenter is interested in a listener's anticipation and online prediction, analysis could target the pupil response *during* or *before* the stimulus. For example, anticipation of difficult trials elicits larger pupil dilation even before stimuli are presented (McCloy et al., 2017). If a listener already knows the spatial location of the target signal, pupil dilation will grow at a slower rate than if the location is unknown (Koelewijn et al., 2017). Anticipation of longer stimuli brings shallower growth of dilation, presumably to distribute cognitive resources over the entire stimulus, both for digit spans (Kahneman & Beatty, 1966; Klingner et al., 2011) and speech signals (Winn & Moore, 2018). Studies by Kahneman and Beatty (1966) and Piquado et al. (2010)

reported that in conditions where listeners expected longer stimulus length (based on a blocked-condition design), larger *overall* pupil dilation was recorded during the baseline period, as if to indicate a greater tonic state of arousal in anticipation of a more challenging task. Sentences that afford the opportunity to predict upcoming words elicit shallower pupil dilation than sentences that are unpredictable, in terms of semantic content and syntactic structure (Schluroff et al., 1986).

In addition to examining stimulus onset or offset, more precise methods can be used, which track stimulus-related information or participant behavior. Ayasse et al. (2017) used pupillometry along with a concurrent eye-tracking paradigm to verify the moment of sentence comprehension (e.g., a participant would gaze at an image that related to the sentence). They examined pupil size at moment of comprehension and found significant combined effects of age and hearing loss. Notably, they found that speed of comprehension was not statistically different across the groups, but the amount of pupil dilation was reliably different. Wagner et al. (2016) similarly aligned their pupil dilation analysis to the onset of target words under lexical competition in a study that examined how spectral degradation affected lexical access. In some of these aforementioned examples, it is probable that conventional measures of overall mean, peak, and latency would identify consistent differences across conditions or participant groups, but leave out other interesting layers that are deserving of interpretation.

Stimulus difficulty will also affect the precise timing of dilations and the clarity of the overall morphology of the response. In an unpublished pilot study, Winn and colleagues presented listeners with a slowly spoken five-word sentence consisting of a name, verb, number, adjective, and plural noun (e.g., "Bill sold four red hats"). Either all words were unpredictable, or the second word (the verb) was always the same word "found." The premise was to examine whether the predictable second word would reduce pupil dilation in the moments just after that word. Overall, the pupil dilations of listeners with CIs were far more precisely time locked to the stimuli compared with dilations measured in listeners with normal hearing, for whom the task was trivially easy. The CI group showed distinct dilation responses for individual words (see Figure 9), consistent with greater demand on auditory processing in these listeners. While there was an effect of word predictability in both groups, the effect was more time constrained for the CI group, where reduction in dilation began just following the second local peak in the time series response. For the listeners with normal hearing, the reduction was spread across a larger portion of time rather than being constrained to a particular moment. The results of this small pilot study are consistent with the aforementioned

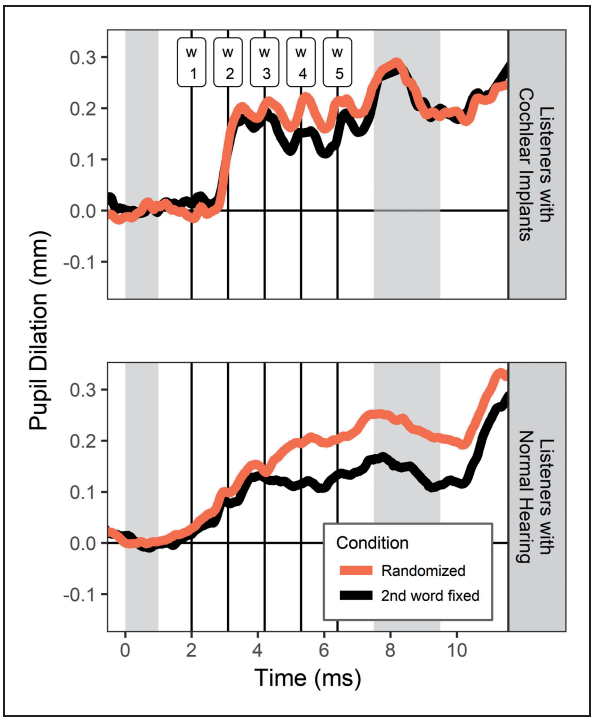

Figure 9. Temporal precision of pupil dilation morphology is related to difficulty of a task. Listeners with cochlear implants (for whom auditory perception can be quite challenging) show individuated dilations to slowly spoken words that are time locked across trials. Affecting the predictability of the second word in the sequence causes a reduced response shortly after the corresponding word. Such patterns do not emerge clearly for listeners with normal hearing, for whom the task is very easy.

capacity model and FUEL; in situations where the task-evoked boredom or low cognitive demand (i.e., the normal-hearing listeners in this slow-paced easy speech task), the pupil response should be less reliable or more difficult to interpret because the participant's attention is not locked to the stimulus (or perhaps is divided among the task and anything else on their mind). Conversely, in the case of the CI listeners for whom even recognition of words in quiet can be challenging, a steep and reliable time-locked dilation response was strong, likely because the task demanded more of their attentional capacity, and pupil size was therefore dictated primarily by the stimuli.

Effort Does Not Stop When the Stimulus Is Over

There are numerous interesting effects of cognitive load that are found in the pupillary response as it continues past the end of a stimulus. The retention interval—the time between the end of a stimulus and the behavioral response prompt—has been examined in detail in only a few studies. Piquado et al. (2010) argued that the retention interval is a good reflection of the cumulative

memory load. Indeed, when participants in a digit-span memory task report their responses, pupil size appears to decrease in a stepwise fashion with each reported digit, as if the memory is ''unloaded'' (Beatty, 1982).

The retention interval appears to show differences relating to intelligibility or confidence in speech perception. Winn et al. (2015) found that when participants heard severely degraded sentences, pupils remained dilated, but quickly constricted in cases of reduced degradation, and also in the case of the severely degraded condition when intelligibility was perfect. That study suggested that the *auditory* processing demands were reflected by the slope of pupil dilation to its peak value, whereas the continued linguistic processing needed to repair misperceptions in the difficult condition was reflected in the dilation during the retention window. Further exploration into the retention window—and its susceptibility to interference—is presented by Winn and Moore (2018).

Insight From Behavioral Economics

Humans are not machines, and their exact motivations moment-to-moment will affect their willingness to put forth extra effort in listening tasks. As that willingness—that intentional *engagement*—appears to be the central driver of the pupillary response, it should be given special consideration in experimental design. Pupil dilation should be observed during tasks that reward the listener for putting in more effort. It is no coincidence that some of the pioneering work in pupillometry was done by a psychologist-turned-behavioral economist—Daniel Kahneman—who contributed the *Capacity Model* of attention and effort that eventually gave rise to the more recent FUEL (Pichora-Fuller et al., 2016). Appreciation of the basic concepts of these models will likely aid in the planning and execution of listening effort experiments.

Eckert et al. (2016) adopt the framework of behavioral economics—the study of choice relative to the value of options—to suggest that the level of effort exerted during speech communication reflects expected value of return on effort. In other words, effort might not be exerted *if the effort is not worth it*. This is an important principle for experimenters to bear in mind as they choose their stimuli and testing conditions, as nonmonotoniticies in the difficulty-effort function could result in very complicated data that is not easy to interpret. Using this framework for a series of brain imaging studies, Eckert and colleagues have placed listening effort in the domain of *neuroeconomics*.

At some level of difficulty, listeners do not continue to exert more effort; they begin to disengage (Granholm, Asarnow, Sarkin, & Dykes, 1996; Peavler, 1974), because there is no value obtained by expending more effort. The pupils will therefore dilate less when the task is too hard to complete successfully. At intelligibility levels below 40%, pupil dilation tends to decrease as listeners disengage. Similar patterns of nonlinearity also appear in other measures of listening effort, such as dual-task cost (Wu, Stangl, Zhang, Perkins, & Eilers, 2016), notably with the same cutoff of performance level. In digit-span memory tasks, sequences longer than 7 to 9 items are generally not attainable by typical listeners. Johnson et al. (2014) found that for adults in such a task, the pupil typically does not dilate any more for 11 items than for 7 items. Furthermore, pupillary responses for longer digit sequences were actually *smaller* (especially in children), possibly reflecting abandonment of the task.

In particularly difficult laboratory experiments, it is reasonable to suspect that people might either (a) actively strategize to change their performance in ways that would not be necessary in easy conditions or (b) decide to withdraw their effort either because it does not seem worth it (Teubner-Rhodes et al., 2017), or because the task has been rendered unrealistic. Consider also that testing in very difficult conditions (e.g., at 30% to 40% accuracy) is likely not a useful situation to test because people might rarely find themselves in that situation. Individuals with hearing loss, who would certainly struggle in noisy complex listening environments, might instead change their own behavior and social activities rather than participate in those environments (Demorest & Erdman, 1986; Weinstein & Ventry, 1982; Wu et al., 2018).

Understanding the Physiology of Pupil Dilation

The physiology of the pupillary response implicates an interesting psychological framework for understanding human behavior in pupillometry experiments. Changes in pupil size are correlated to changes in activity in neurons of the LC (Rajkowski, Kubiak, & Aston-Jones, 1993; Rajkowski, Majczynski, Clayton, & Aston-Jones, 2004). This physiological connection corroborates the connection between pupil diameter and phasic attention or effort. Following from theories about the role of the (LC) in the modulation of attention, individuals will continue to perform a behavior so long it is rewarding or provides utility. Importantly, the valuation of performing a particular task may vary across groups or individuals. During such focused attention, the LC exhibits a phasic burst of activity (tied to the task-evoked pupil response) to support making a behavioral response. As task utility declines, scanning attention takes over as the individual searches for new rewards in the environment and the LC enters a high-tonic mode (tied to the baseline pupil size). Reflecting this pattern in cortex, the anterior

cingulate cortex (ACC), which is a primary input to and receives feedback from the LC (Aston-Jones & Cohen, 2005; Gilzenrat et al., 2010), has been associated with effortful attention. In particular, the ACC is part of the cingulo-opercular performance-monitoring network, which includes bilateral frontal opercula or anterior insulae. This network is engaged across a wide range of tasks (e.g., Dosenbach et al., 2006; Eckert et al., 2009) to actively monitor for response uncertainty and errors, so that other systems can be brought online to adapt and improve performance (i.e., fronto-parietal, cognitive-control network; Kerns et al., 2004). Eckert et al. (2016) thus argue for the increasing engagement of the cingulo-opercular network as an indicator of greater value or utility in engaging in speech recognition (cf. the earlier section on Behavioral economics). Interestingly, *inactivation* of cingulate neurons is associated with impairment of adjusting to errors in a task (Vaden, Kuchinsky, Ahlstrom, Dubno, & Eckert, 2015). Pupil size has been linked with activity of the cingulo-opercular system (Schneider et al., 2016; Zekveld et al., 2014). In light of this converging evidence, it seems reasonable to interpret the phasic pupillary response as a sign that a listener feels the need to ''take action'' mentally. Circling back to studies of pupil dilation and intelligibility, this framework helps to explain why there is increased dilation in moderately challenging conditions (where ''taking action'' could overcome acoustic challenges), but little dilation when the stimulus is so easy as to not demand action and also little dilation where no action is taken because the task is too difficult for performance to be successful.

Recent work suggests that the notion of understanding pupil size as an index of LC activity should be expanded and updated. For years, many pupillometry publications have cited Aston-Jones and Cohen (2005) to link pupil dilation with the activity of the LC-norepinephrine system. However, as McGinley et al. (2015) point out, the compelling figure from that publication reflected the recording of a single neuron and was from the abstract of an unpublished study. McGinley et al. suggest that while LC activity is related to pupil size, other mechanisms and neural substrates linked with pupil dilation have not been ruled out. Furthermore, they characterize pupil size as an indicator of ''brain state,'' possibly intentionally avoiding linkage with a specific local structure. Consistent with this, Reimer et al. (2016) found that pupil dilation was comodulated with cortical activity *in general*, complicating the process of using pupillometry to assess any specific function. A study in nonhuman primates also found that although LC stimulation reliably proceeds changes in pupil dilation, that similar, though weaker, patterns can be observed in other regions connected with the LC (i.e., ACC and colliculi; Joshi, Li, Kalwani, & Gold, 2016). Thus, given the extensive connectivity within the LC-norepinephrine system, these results suggest that, rather than a single brain area controlling pupil dilation, the LC acts as a hub that coordinates attention-related neural activity.

Reimer et al. (2016) showed evidence that *rapid* pupil dilations (roughly 0.25 Hz) are associated with phasic norandrenergic activity, while slower long-lasting dilations are linked more closely with cholinergic activity. Furthermore, the *rate* of change in pupil size was linked with norandrenergic activity, while overall diameter was linked with both, but more strongly with cholinergic activity. Pupil dilation was observed to reliably lag behind neural activity by roughly 1 s, consistent with behavioral studies where onset of pupil dilation and peak dilation follow stimulus onset and offset by about 1 s, respectively (Hoeks & Levelt, 1993). Steinhauer et al. (2004) suggest that the rapid early component of the dilation response is driven by parasympathetic activity while the later-occurring component is driven by the sympathetic system.

What Is Still Beyond Our Reach at This Time?

Individualized Measures or Cross-Person Comparisons

The range of variability observed in most physiological recordings will carry over to pupillometry. Data from a single person are rarely as clean as the grouped data displayed in popular pupillometry articles. Given a sufficient number of trials, more confidence can be obtained for single-participant test sessions, but data should be interpreted with caution. The absolute size of the pupil response varies across people, and there are people for whom measurements will not be reliable, sometimes for unknown reasons.

Unsurprisingly, individualized measures of listening effort are the goal of many audiologists and experimenters alike, for the purposes of tracking progress with clinical intervention or to compare treatment approaches. This goal might become attainable as techniques continue to be refined and appropriate baseline tasks can be developed. However, though there will still be some difficulties. In particular, the use of certain medications that affect that sympathetic or parasympathetic nervous system will render pupillometry unreliable (cf. Steinhauer et al., 2004). In addition, some participants exhibit markedly different morphologies in their pupillary responses over time, leaving experimenters unsure how to make a fair comparison. It is possible that comparison of within-subjects conditions (e.g., proportional differences between responses in two

conditions within an individual) will be the key to drive reliable individualized analysis.

Single-Trial Analysis and Adaptive Tracking

As for other evoked physiological measures, single trials are not sufficient to draw reliable conclusions. The reason for the recommendation of ~25 trial is that there is a wide range of variability at the individual trial level. Each individual trial will produce a time series (''trace'') of pupil dilation that is the result of many factors, some of which are beyond the experimenter's control, and perhaps unrelated to the task itself. Substantial deviations from the expected pattern of the task-evoked pupillary response are common in any series of trials and are not necessarily indicators of a specific level of cognitive load. Only when averaging together multiple trials can the experimenter get a reliable estimate of the pupillary response separate from any idiosyncratic effects on an individual trial.

The pupillary response alone should not be used as a criterion for adaptive tracking in speech perception experiments, for the reasons outlined earlier. Although it would be ideal to dynamically adjust SNR (or speech rate, spectral resolution, bandwidth, or any other signal property) to arrive at a target level of pupil dilation, it is not feasible using the standard analysis techniques used in behavioral experiments. The problem is that each unique condition (SNR, etc.) requires a sufficiently large number of trials to estimate the pupillary response, rendering the tracking anything but adaptive. Contrary to this suggestion, Marshall (2002) described a framework for virtually real-time assessment of cognitive activity using pupil dilation. However, in that article, cognitive activity was discretized as *low, medium,* or *high*; we suggest that these categories might lead to oversimplifications of important factors that experimenters might want to explore in finer detail.

At least one report exists of using pupillometry as an adaptive tracking mechanism for real-time feedback. Choi et al. (2017) used the pupil response to govern feedback in a visual scanning task in individuals at risk for psychosis. However, instead of measuring listening effort, they sought to elicit a broad indication of how much a person was actively engaged in the task, to monitor for lapses of attention or cognitive overload.

Consensus Technique for Automatic Trial Dropping

As discussed earlier, at this time, there is no universally used algorithm for detecting and removing aberrant trials. Hands-on experience with your own raw data and consultation with previous literature will inform the most appropriate filtering used to identify contaminated trials. Perhaps machine learning or other modern approaches can be used to devise algorithms to identify questionable trials (Książek, Wendt, Alickovic, & Lunner, 2018), but at the time of this writing, there is no consensus approach.

Testing in Unconstrained or Conversational Situations

As the task-evoked pupillary response is a time-locked, aggregated response, it is not conducive to conversational situations or other situations that are *unplanned* or unconstrained by trial timing. The problem is that during a conversation, it could be hard to find regular timing landmarks upon which trial data could be aligned and aggregated.

Reducing Listening Effort

The measurement of pupil size is not the same as the reduction of listening effort. Readers should be cautioned that studies designed to *measure* effort are not in themselves a tool to alleviate effort. Furthermore, changes in pupil size might not yield clear actionable conclusions about how to alleviate effort.

Linking Subjective and Objective Measures of Effort

Both subjective report and a variety of objective measures have been used to track changes in listening effort (for reviews see McGarrigle et al., 2014; Ohlenforst et al., 2017). However, to the extent that pupillometry and subjective report have been collected in the same experiment, researchers have generally observed weak to no correlations between these measures of effort (e.g., Wendt et al., 2016; Zekveld & Kramer, 2014; Zekveld et al., 2011; though cf. Koelewijn et al., 2015). Currently, it is unclear whether this pattern is due to methodological limitations (e.g., comparing offline vs. online measures, biases that are inherent in self-report), or whether conscious awareness of effort engages reflects underlying mechanisms (Mulert, Menzinger, Leicht, Pogarell, & Hegerl, 2005) that may not be tracked by objective measures. Francis, MacPherson, Chandrasekeran, and Alvar (2016) suggest that physiological measures in general likely reflect constructs other than subjective or perceived effort and can represent the listener overcoming signal distortion, the separation of target and noise, or the affective response to the difficulty of the situation. Hornsby and Kipp (2016) report that reports of fatigue in people with hearing impairment are related not to degree of hearing loss but to perceived difficulty or handicap. These relationships will continue to demand further exploration as the topic of

hearing, effort, and fatigue continue to flourish in the literature.

A Caution About Equating Pupil Size With "Effort"

Pupil size reflects multiple things, and there is no consensus on what percentage of change in pupil size corresponds to a particular change in proportion of effort capacity. In addition, note that smaller pupil dilation is routinely observed in listeners who are older, and listeners with hearing impairment (Koelewijn et al., 2017), and listeners with traumatic brain injury (Koelewijn et al., 2018), despite common reports of elevated effort in these populations.

Summary of Helpful Practices and Advice

As a start, it is a smart thing to allow your design to replicate an effect shown in a previous study. Klingner et al. (2011) provide an excellent case study of replicating known results in the midst of exploring a new problem. By doing this, the experimenter can avoid the mystery of unclear or null results by verifying that her or his data collection system is working.

Keep the luminance of the testing area as steady as possible, with few changes in visual input. When recording, either avoid eye moments or control for their effects on the estimation of pupil size. Limit physical locomotion. A normal amount of random blinking is okay but beware of stimulus-timed blinks; consider having a "blink" instruction at the end of trials. Make sure the participants have enough breaks and avoid long fatiguing sessions. Make sure data are well annotated with the appropriate time stamps. Check if you can read and processes the data after you have recorded the first few participants, to ensure that tracking and annotation are sound. Use a consistent amount of time for each trial and leave a sufficient amount of time between trials to allow the pupil to return to baseline. Make sure participants can anticipate when a trial will start; consider using an alerting signal or a consistent amount of leading noise time for each trial.

Avoid conditions of very low intelligibility or task performance, as participants are likely to disengage from the task. In line with this, consider a participant's anxiety about failure, and whether anxiety is central to the research question or just a byproduct of the testing environment. Be mindful of various sources of individual variability (cf. Tryon, 1975), and also that sometimes pupil dilation is influenced more strongly by behavioral response rather than by listening. In line with this, deliberately plan the amount of time between auditory stimulus and behavioral response.

When possible, it is advisable to test within-group effects to control for individual differences in pupillary reactivity. Use a sufficient number of trials (20–25) for each condition and leave out the first few trials when participants are still easing in to the task. Expect some amount of missing or contaminated data. Engaging tasks keep participants motivated and have shown prominent pupil responses, but be aware that emotional stimuli evoke a big response that might lead to unplanned distortion of the results. Finally, be aware of age effects when designing your study and analysis techniques.

Review Articles or Recommended Reading

General reviews of pupillometry to assess cognitive load have been written by Beatty (1982) and Laeng et al. (2012). For a more general review of the pupillary system as a whole, the chapter by Beatty and Lucero-Wagoner (2000) is especially helpful. Detailed information about the physiology of pupil dilation can be found in articles by McGinley et al. (2015), Reimer et al. (2016), and Eckstein, Guerra-Carrillo, Miller Singley, and Bunge (2017). A review of pupillometry to assess listening effort in particular is found elsewhere in this issue, by Zekveld et al. (2018). In addition to these articles, there are also other introductory reviews for new experimenters (like the current article), for the fields of second-language acquisition (Schmidtke, 2017) and for assessment of brainstem function by anesthesiologists (Larson & Behrends, 2015).

Summary and Concluding Remarks

In this article, we have aimed to provide a series of recommendations on best practices (or at least common practices) for research on conducting pupillometry studies of listening effort. We have highlighted the importance of conducting theory-driven research and employing careful research design and analytical approaches informed by well-established ideas of attention, motivation, and economy of effort, as well as the constraints of the measurement technique itself. Pupillometry has particular strengths as a noninvasive physiological measurement that can track changes in effort with some temporal precision. It is also compatible with electronic hearing devices used by many listeners who are the focus of hearing research. Although implementing any new methodology can seem daunting, we hope that by providing this (nonexhaustive) guide, researchers will be encouraged to get started and will attain success quickly. Addressing the outstanding questions in the field will require a multidisciplinary approach, involving both clinicians and basic researchers as well as perceptual, linguistic, cognitive, and neuroeconomic perspectives. It will require studies examining variability in a variety of task conditions, task goals, populations, and within and across individuals.

Given the ultimate goal of improving individuals' quality of life through better communication, we hope we have convinced readers that pupillometry is worth the effort.

Acknowledgments

This article was originally planned during the "Pupillometry in Hearing Science" workshop in Amsterdam, 2017. We would like to thank our many colleagues who have contributed advice, ideas, and opinions to the authors as this manuscript was planned and prepared. In particular, the ideas in this article were improved by discussions with Adriana Zekveld, Sophia Kramer, Graham Naylor, Matthew McGinley, Daniel McCloy, Giulia Borghini, Ashley Moore, Mark Eckert, Yang Wang, Nicole Ayasse, and Ronan McGarrigle.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded by NIH-NIDCD NIH-NIDCD R03DC014309 (M. B. W.), Oticon Fonden (Foundation) Grant 16-0463 (T. K.), and NIH-NIDCD R03 DC015059 (S. E. K.).

References

Abokyi, S., Oqusu-Mensah, J., & Osei, K. (2017). Caffeine intake is associated with pupil dilation and enhanced accommodation. *Eye*, *31*, 615–619. doi:10.1038/eye.2016.288

Ahern, S., & Beatty, J. (1979). Pupillary responses during information processing vary with scholastic aptitude test scores. *Science*, *205*, 1289–1292. doi:10.1126/science.472746

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264. doi:10.1016/S0010-0277(99)00059-1

Anderson, C., & Columbo, J. (2009). Larger tonic pupil size in young children with autism spectrum disorder. *Developmental Psychobiology*, *51*, 207–211. doi:10.1002/dev.20352

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

Ayasse, N., Lash, A., & Wingfield, A. (2017). Effort not speed characterizes comprehension of spoken sentences by older adults with mild hearing Impairment. *Frontiers in Aging Neuroscience*, *8*, 329. doi:10.3389/fnagi.2016.00329

Bala, A., Spitzer, M., & Takahashi, T. (2007). Auditory spatial acuity approximates the resolving power of space-specific neurons. *PLoS One*, *2*, e675. doi:10.1371/journal.pone.0000675

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

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

Best, V., Ahlstrom, J., Mason, C., Roverud, E., Perrachione, T., Kidd, G., & Dubno, J. (2018). Talker identification: Effects of masking, hearing loss and age. *The Journal of the Acoustical Society of America*, *143*, 1085–1092. doi:10.1121/1.5024333

Best, V., Streeter, T., Roverud, E., Mason, C., & Kidd, G. (2016). A flexible question-and-answer task for measuring speech understanding. *Trends in Hearing*, *20*, 1–8. doi:10.1177/2331216516678706

Bianchi, F., Santurette, S., Wendt, D., & Dau, T. (2016). Pitch discrimination in musicians and non-musicians: Effects of harmonic resolvability and processing effort. *Journal of the Association for Research in Otolaryngology*, *17*(1), 69–79. doi:10.1007/s10162-015-0548-2

Bitsios, P., Prettyman, R., & Szabadi, E. (1996). Changes in autonomic function with age: A study of pupillary kinetics in healthy young and old people. *Age and Ageing*, *25*, 432–438. doi:10.1093/ageing/25.6.432

Borghini, G. (2017). Listening effort during speech understanding in a second language. *Presented at the Pupillometry in Hearing Science Workshop,* Amsterdam, The Netherlands.

Bradshaw, J. (1968). Pupil size and problem solving. *The Quarterly Journal of Experimental Psychology*, *20*, 116–122. doi:10.1080/14640746808400139

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

Bradshaw, J. (1970). Pupil size and drug state in a reaction time task. *Psychonomic Science*, *18*, 112–113. doi:10.3758/BF03335723

Brisson, J., Mainville, M., Mailloux, D., Beaulieau, C., Serres, J., & Sirois, S. (2013). Pupil diameter measurement errors as a function of gaze direction in corneal reflection eyetrackers. *Behavioral Research*, *45*, 1322–1331. doi:10.3758/s13428-013-0327-0

Bronkhorst, A. (2015). The cocktail-party problem revisited: Early processing and selection of multi-talker speech. *Attention Perception and Psychophysics*, *77*, 1465–1487. doi:10.3758/s13414-015-0882-9

Bruya, B., & Tang, Y.-Y. (2018). Is attention really effort? Revisiting Daniel Kahneman's influential 1973 book attention and effort. *Frontiers in Psychology*. doi:10.3389/fpsyg.2018.01133

Cavanaugh, J., Wiecki, T., Kochar, A., & Frank, M. (2014). Eye tracking and pupillometry are indicators of dissociable latent decision processes. *Journal of Experimental Psychology General*, *143*, 1476–1488. doi:10.1037/a0035813

Choi, J., Corcoran, C., Fiszdon, J., Stevens, M., Javitt, D., Deasy, M., ... Pearlson, G. (2017). Pupillometer-based neurofeedback cognitive training to improve processing speed and social functioning in individuals at clinical high risk for psychosis. *Psychiatric Rehabilitation Journal*, *40*, 33–42. doi:10.1037/prj0000217

Dahan, D., Tanenhaus, M., & Chambers, C. (2001). Accent and reference resolution in spoken-language comprehension. *Journal of Memory and Language*, *47*, 292–314. doi:10.1016/S0749-596X(02)00001-3

Demorest, M., & Erdman, S. (1986). Scale composition and item analysis of the communication profile for the hearing impaired. *Journal of Speech and Hearing Research*, *29*, 515–535. doi:10.1044/jshr.2904.535

Dosenbach, N. U. F., Visscher, K. M., Palmer, E. D., Miezin, F. M., Wenger, K. K., Kang, H. C.,... Petersen, S. E. (2006). A core system for the implementation of task sets. *Neuron*, *50*, 799–812. doi:10.1016/j.neuron.2006.04.031

Eckert, M., Menon, V., Walczak, A., Ahlstrom, J., Denslow, S., Horwitz, A., & Dubno, J. R. (2009). At the heart of the ventral attention system: The right anterior insula. *Human Brain Mapping*, *30*, 2530–2541. doi:10.1002/hbm.20688

Eckert, M., Teubner-Rhodes, S., & Vaden, K. (2016). Is listening in noise worth it? The neurobiology of speech recognition in challenging listening conditions. *Ear and Hearing*, *37*(Suppl 1): 101S–110S. doi:10.1097/AUD.0000000000000300

Eckstein, M., Guerra-Carrillo, B., Miller Singley, A., & Bunge, S. (2017). Beyond eye gaze: What else can eye tracking reveal about cognition and cognitive development? *Developmental Cognitive Neuroscience*, *25*, 69–91. doi:10.1016/j.dcn.2016.11.001

Engelhardt, P., Ferreira, F., & Patsenko, E. (2010). Pupillometry reveals processing load during spoken language comprehension. *Quarterly Journal of Experimental Psychology*, *63*, 639–645. doi:10.1080/17470210903469864

Francis, A., MacPherson, M., Chandrasekeran, B., & Alvar, A. (2016). Autonomic nervous system responses during perception of masked speech may reflect constructs other than subjective listening effort. *Frontiers in Psychology*, *7*, A263. doi:10.3389/fpsyg.2016.00263

Franklin, M., Broadway, J., Mrazek, M., Smallwood, J., & Schooler, J. (2013). Window to the wandering mind: Pupillometry of spontaneous thought while reading. *The Quarterly Journal of Experimental Psychology*, *66*, 2289–2294. doi:10.1080/17470218.2013.858170

Friesen, L., & Picton, T. (2010). A method for removing cochlear implant artifact. *Hearing Research*, *259*, 95–106. doi:10.1016/j.heares.2009.10.012

Gagl, B., Hawelka, S., & Huzler, F. (2011). Systematic influence of gaze position on pupil size measurement: Analysis and correction. *Behavioral Research*, *43*, 1171–1181. doi:10.3758/s13428-011-0109-5

Gagné, J.-P., Besser, J., & Lemke, U. (2017). Behavioral assessment of listening effort using a dual-task paradigm: A review. *Trends in Hearing*, *21*, 1–25. doi:10.1177/2331216516687287

Gatehouse, S., & Noble, W. (2004). The speech, spatial and qualities of hearing scale (SSQ). *International Journal of Audiology*, *43*, 85–99. doi:10.1080/14992020400050014

Gilley, P., Sharma, A., Dorman, M., Finley, C., Panch, A., & Martin, K. (2006). Minimization of cochlear implant stimulus artifact in cortical auditory evoked potentials. *Clinical Neurophysiology*, *117*, 1772–1782. doi:10.1016/j.clinph.2006.04.018

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

Granholm, E., Asarnow, R., Sarkin, A., & Dykes, K. (1996). Pupillary responses index cognitive resource limitations. *Psychophysiology*, *33*, 457–461. doi:10.1111/j.1469-8986.1996.tb01071.x

Granholm, E., & Steinhauer, S. (2004). Pupillometric measures of cognitive and emotional processes. *International Journal of Psychophysiology*, *51*, 1–6. doi:10.1016/j.ijpsycho.2003.12.001

Gredebäck, G., & Melinder, A. (2010). Infants' understanding of everyday social interactions: A dual process account. *Cognition*, *114*, 197–206. doi:10.1016/j.cognition.2009.09.004

Hakerem, G., & Sutton, S. (1966). Pupillary response at visual threshold. *Nature*, *212*, 485–486. doi:10.1038/212485a0

Hayes, T., & Petrov, A. (2016). Mapping and correcting the influence of gaze position on pupil size measurements. *Behavioral Research Methods*, *48*, 510–527. doi:10.3758/s13428-015-0588-x

Heitz, R., Schrock, J., Payne, T., & Engle, R. (2008). Effects of incentive on working memory capacity: Behavioral and pupillometric data. *Psychophysiology*, *45*, 119–129. doi:10.1111/j.1469-8986.2007.00605.x

Hess, E., & Polt, J. (1960). Pupil size as related to interest value of visual stimuli. *Science*, *132*, 349–350. doi:10.1126/science.132.3423.349

Hess, E., & Polt, J. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science*, *143*, 1190–1192. doi:2.10.1126/science.143.3611.1190

Hétu, R., Riverin, L., Lalande, N., Getty, L., & St-Cyr, C. (1988). Qualitative analysis of the handicap associated with occupational hearing loss. *British Journal of Audiology*, *22*, 251–264. doi:10.3109/03005368809076462

Hoeks, B., & Levelt, W. (1993). Pupillary dilation as a measure of attention: A quantitative system analysis. *Behavioral Research Methods, Instruments & Computers*, *25*, 16–26. doi:10.3758/BF03204445

Hornsby, B., & Kipp, A. (2016). Subjective ratings of fatigue and vigor in adults with hearing loss are driven by perceived hearing difficulties not degree of hearing loss. *Ear and Hearing*, *37*, e1–e10. doi:10.1097/AUD.0000000000000203

Hughes, S., Hutchings, H., Rapport, F., McMahon, C., & Boisvert, I. (2018). Social connectedness and perceived listening effort in adult cochlear implant users: A grounded theory to establish content validity for a new patient-reported outcome measure. *Ear and Hearing*, *39*, 922–934. doi:10.1097/AUD.0000000000000553

Hyönä, J., Tommola, J., & Alaja, A. (1995). Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. *Quarterly Journal of Experimental Psychology*, *48*, 598–612. doi:10.1080/14640749508401407

Jackson, I., & Sirois, S. (2009). Infant cognition: Going full factorial with pupil dilation. *Developmental Science*, *12*, 670–679. doi:10.1111/j.1467-7687.2008.00805.x

Jensen, J. J., Callaway, S. L., Lunner, T., Wendt, D. (2018). Investigating the impact of tinnitus: A pupillometry study. *Trends in Hearing*.

Johnson, E., Singley, A., Peckham, A., Johnson, S., & Bunge, S. (2014). Task-evoked pupillometry provides a window into the development of short-term memory capacity. *Frontiers in Psychology*, *5*, A218. doi:10.3389/fpsyg.2014.00218

Joshi, S., Li, Y., Kalwani, R., & Gold, J. (2016). Relationships between pupil diameter and neuronal activity in the locus coeruleus, colliculi, and cingulate cortex. *Neuron*, *89*, 221–234. doi:10.1016/j.neuron.2015.11.028

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

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

Kahneman, D., & Beatty, J. (1967). Pupillary responses in a pitch-discrimination task. *Perception & Psychophysics*, *2*, 101–105. doi:10.3758/BF03210302

Kahneman, D., & Peavler, W. S. (1969). Incentive effects and pupillary changes in association learning. *Journal of Experimental Psychology*, *79*, 312–318. doi:10.1037/h0026912

Kang, M., Hsu, M., Krajbich, I. M., Loewenstein, G., McClure, S. M., Wang, J. T., & Camerer, C. F. (2009). The wick in the candle of learning: Epistemic curiosity activates reward circuitry and enhances memory. *Psychological Science*, *20*, 963–973. doi:10.1111/j.1467-9280.2009.02402.x.

Karatekin, C., Couperous, J. W., & Marcus, D. J. (2004). Attention allocation in the dual-task paradigm as measured through behavioural and psychophysiological responses. *Psychophysiology*, *41*, 1–11. doi:10.1111/j.1469-8986.2004.00147.x

Kerns, J., Cohen, J., MacDonald, A., Cho, R. Y., Stenger, V. A., & Carter, C. S. (2004). Anterior cingulate conflict monitoring and adjustments in control. *Science*, *303*, 1023–1026. doi:10.1126/science.1089910

Kim, M., Beversdorf, D., & Heilman, K. (2000). Arousal response with aging: Pupillographic study. *Journal of the International Neuropsychological Society*, *6*, 348–350. doi:10.1017/S135561770000309X

Klingner, J., Tversky, B., & Hanrahan, P. (2011). Effects of visual and verbal presentation on cognitive load in vigilance, memory, and arithmetic tasks. *Psychophysiology*, *48*, 323–332. doi:10.1111/j.1469-8986.2010.01069.x

Koelewijn, T., de Kluiver, H., Shinn-Cunningham, B., Zekveld, A., & Kramer, S. (2015). The pupil response reveals increased listening effort when it is difficult to focus attention. *Hearing Research*, *323*, 81–90. doi:10.1016/j.heares.2015.02.004

Koelewijn, T., Shinn-Cunningham, B. G., Zekveld, A. A., & Kramer, S. E. (2014). The pupil response is sensitive to divided attention during speech processing. *Hearing Research*, *312*, 114–120. doi:10.1016/j.heares.2014.03.010

Koelewijn, T., van Haastrecht, J., & Kramer, S. (2018). Pupil responses of adults with traumatic brain injury during processing of speech in noise. *Trends in Hearing*.

Koelewijn, T., Versfeld, N., & Kramer, S. (2017). Effects of attention on the speech reception threshold and pupil response of people with impaired and normal hearing. *Hearing Research*, *354*, 56–63. doi:10.1016/j.heares.2017.08.006

Koelewijn, T., Zekveld, A., Festen, J., & Kramer, S. (2012). Pupil dilation uncovers extra listening effort in the presence of a single-talker masker. *Ear and Hearing*, *33*, 291–300. doi:10.1097/AUD.0b013e3182310019

Koenig, S., Uengoer, M., & Lachnit, H. (2017). Pupil dilation indicates the coding of past prediction errors: Evidence for attentional learning theory. *Psychophysiology*, *55*(4), e13020. doi:10.1111/psyp.13020

Koeritzer, M., Rogers, C., Van Engen, K., & Peelle, J. (2018). The impact of age, background noise, semantic ambiguity and hearing loss on recognition memory for spoken sentences. *Journal of Speech Language and Hearing Research*, *61*, 740–751. doi:10.1044/2017_JSLHR-H-17-0077

Kramer, S., Kapteyn, T., & Houtgast, T. (2006). Occupational performance: Comparing normally-hearing and hearing-impaired employees using the Amsterdam Checklist for Hearing and Work. *International Journal of Audiology*, *45*, 503–512. doi:10.1080/14992020600754583

Kramer, S., Kapteyn, T., Festen, J., & Kuik, D. (1997). Assessing aspects of hearing handicap by means of pupil dilation. *Audiology*, *36*, 155–164. doi:10.3109/00206099709071969

Książek, P., Wendt, D., Alickovic, E., & Lunner, T. (2018). Analysis of the individual listening effort reflected by the pupillary responses during speech perception in noise. *Presented at the 10th Speech in Noise Workshop*, Glasgow, UK

Kuchinsky, S., Ahlstrom, J., Vaden, K., Cute, S., Humes, L., Dubno, J., & Eckert, M. A. (2013). Pupil size varies with word listening and response selection difficulty in older adults with hearing loss. *Psychophysiology*, *50*, 23–34. doi:10.1111/j.1469-8986.2012.01477.x

Kuchinsky, S., Vaden, K., Ahlstrom, J., Cute, S., Humes, L., Dubno, J., Eckert, M. (2016). Task-related vigilance during word recognition in noise for older adults with hearing loss. *Experimental Aging Research*, *42*, 50–66. doi: 10.1080/0361073X.2016.1108712

Kun, A., Palinko, O., & Razumenić, I. (2012). Exploring the effects of size and luminance of visual targets on the pupillary light reflex. *Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 183–186). New York, NY: ACM. doi:10.1145/2390256.2390287.

Laeng, B., Ørbo, M., Holmlund, T., & Miozzo, M. (2011). Pupillary stroop effects. *Cognitive Processing*, *12*, 13–21. doi:10.1007/s10339-010-0370-z

Laeng, B., Sirous, S., & Gredebäck, G. (2012). Pupillometry: A window into the preconscious? *Perspectives on Psychological Science*, *7*, 18–27. doi:10.1177/1745691611427305

Larson, M., & Behrends, M. (2015). Portable infrared pupillometry: A review. *Anesthesia & Analgesia*, *120*, 1242–1253. doi:10.1213/ANE.0000000000000314

Lee, Y.-S., Min, N., Wingfield, A., Grossman, M., & Peelle, J. (2016). Acoustic richness modulates the neural networks supporting intelligible speech processing. *Hearing Research*, *333*, 108–117. doi:10.1016/j.heares.2015.12.008

Lõo, K., van Rij, J., Järvikivi, J., & Baayen, H. (2016). Individual differences in pupil dilation during naming task. *Proceedings of the Annual Meeting of the Cognitive Science Society*. Retrieved from https://mindmodeling.org/cogsci2016/papers/0106/index.html

Lynch, G., James, S., & VanDam, M. (2017). Pupillary response and phenotype in ASD: Latency to constriction discriminates ASD from typically developing adolescents. *Autism Research*, *31*, 1–12. doi:10.1002/aur.1888.

Marmarou, A., Lu, J., Butcher, I., McHugh, G., Murray, G., Steyerberg, E., . . . Maas, A. (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*, 270–280. doi:10.1089/neu.2006.0029

Marshall, S. (2002). The index of cognitive activity: Measuring cognitive workload. *Proceedings of the 7th IEEE Human Factors Meeting*, Scottsdale, AZ. doi:10.1109/HFPP.2002.1042860

Martineau, J., Hernandez, N., Hiebel, L., Roché, L., Metzger, A., & Bonnet-Brilhault, F. (2011). Can pupil size and pupil responses during visual scanning contribute to the diagnosis of autism spectrum disorder in children? *Journal of Psychiatric Research*, *45*, 1077–1082. doi:10.1016/j.jpsychires.2011.01.008

Mattys, S., Davis, M., Bradlow, A., & Scott, S. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, *27*, 953–978. doi:10.1080/01690965.2012.705006

McCloy, D., Larson, E., Lau, B., & Lee, A. K. C. (2016). Temporal alignment of pupillary response with stimulus events via deconvolution. *Journal of the Acoustical Society of America*, *139*, EL57–EL62. doi:10.1121/1.4943787

McCloy, D., Lau, B., Larson, E., Pratt, K., & Lee, A. K. C. (2017). Pupillometry shows the effort of auditory attention switching. *Journal of the Acoustical Society of America*, *141*, 2440–2451. doi:10.1121/1.4979340

McCoy, S., Tun, P., Cox, L., Colangelo, M., Stewart, R., & Wingfield, A. (2005). Hearing loss and perceptual effort: Downstream effects on older adults' memory for speech. *Quarterly Journal of Experimental Psychology*, *58*, 22–33. doi:10.1080/02724980443000151

McGarrigle, R., Dawes, P., Stewart, A., Kuchinsky, S., & Munro, K. (2017a). Pupillometry reveals changes in physiological arousal during a sustained listening task. *Psychophysiology*, *54*, 193–203. doi:10.1111/psyp.12772

McGarrigle, R., Dawes, P., Stewart, A., Kuchinsky, S., & Munro, K. (2017b). Measuring listening-related effort and fatigue in school-aged children using pupillometry. *Journal of Experimental Child Psychology*, *161*, 95–112. doi:10.1016/j.jecp.2017.04.006

McGarrigle, R., Munro, K., Dawes, P., Stewart, A. J., Moore, D. R., Barry, J. G., & Amitay, S. (2014). Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology Cognition in hearing special interest group 'white paper'. *International Journal of Audiology*, *53*, 433–445. doi:10.3109/14992027.2014.890296

McGinley, M., David, S., & McCormick, S. (2015). Cortical membrane potential signature of optimal states for sensory signal detection. *Neuron*, *87*, 179–192. doi:10.1016/j.neuron.2015.05.038

McKay, C. M., Shah, A., Seghouane, A. K., Zhou, X., Cross, W., & Litovsky, R. (2016). Connectivity in language areas of the brain in cochlear implant users as revealed by fNIRS. *Advances in Experimental Medicine and Biology*, *894*, 327–335. doi:10.1007/978-3-319-25474-6_34

McMahon, C., Boisvert, I., de Lissa, P., Granger, L., Ibrahim, R., Lo, C., . . . Graham, P. (2016). Monitoring alpha oscillations and pupil dilation across the performance-intensity function. *Frontiers in Psychology*, *7*, 745. doi:10.3389/fpsyg.2016.00745

Miles, K., McMahon, C., Boisvert, I., Ibrahim, R., de Lissa, P., Graham, P., & Lyxell, B. (2017). Objective assessment of listening effort: Coregistration of pupillometry and EEG. *Trends in Hearing*, *21*, 1–13. doi:10.1177/2331216517706396

Mirman, D. (2014). *Growth curve analysis and visualization using R*. New York, NY: CRC Press.

Mulert, C., Menzinger, E., Leicht, G., Pogarell, O., & Hegerl, U. (2005). Evidence for a close relationship between conscious effort and anterior cingulate cortex activity. *International Journal of Psychophysiology*, *56*(1), 65–80. doi:10.1016/j.ijpsycho.2004.10.002

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*, 4140–4154. doi:10.1002/hbm.22466

Nachtegaal, J., Kuik, D., Anema, J., Goverts, T., Festen, J., & Kramer, S. (2009). Hearing status, need for recovery after work, and psychosocial work characteristics: Results from an internet-based national survey on hearing. *International Journal of Audiology*, *48*, 684–691. doi:10.1080/14992020902962421

Nunnally, J., Knott, P., Duchnowski, A., & Parker, R. (1967). Pupillary response as a general measure of activation. *Perception & Psychophysics*, *2*, 149–155. doi:10.3758/BF03210310

Obleser, J., Wise, R., Dresner, M., & Scott, S. (2007). Functional integration across brain regions improves speech perception under adverse listening conditions. *Journal of Neuroscience*, *27*, 2283–2289. doi:10.1523/JNEUROSCI.4663-06.2007

Ohlenforst, B., Wendt, D., Kramer, S., Naylor, G., Zekveld, A. A., & Lunner, T. (2018). Impact of SNR, masker type and noise reduction processing on listening effort as indicated by the pupil dilation. *Hearing Research*, *365*, 90–99. doi:10.1016/j.heares.2018.05.003

Ohlenforst, B., Zekveld, A., Lunner, T., Wendt, D., Naylor, G., Wang, Y., . . . Kramer, S. E. (2017). Impact of stimulus-related factors and hearing impairment on listening effort as indicated by pupil dilation. *Hearing Research*, *351*, 68–79. doi:10.1016/j.heares.2017.05.012

Palinko, O., & Kun, A. (2011). Exploring the influence of light and cognitive load on pupil diameter in driving simulation studies. *Proceedings of the Sixth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, Lake Tahoe, CA. doi:10.17077/drivingassessment.1416

Pals, C., Sarampalis, A., & Baskent, D. (2013). Listening effort with cochlear implant simulations. *Journal of Speech Language and Hearing Research*, *56*, 1075–1084. doi:10.1044/1092-4388(2012/12-0074)

Papesh, N., & Goldinger, S. (2012). Pupil-BLAH-metry: Cognitive effort in speech planning reflected by pupil

dilation. *Attention Perception and Psychophysics*, *74*, 754–765. doi:10.3758/s13414-011-0263-y

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

Payne, D., Parry, M., & Harasymiw, S. (1968). Percentage of pupillary dilation as a measure of item difficulty. *Perception & Psychophysics*, *4*, 139–143. doi:10.3758/BF03210453

Peavler, W. (1974). Pupil size, information overload and performance differences. *Psychophysiology*, *11*, 559–566. doi:10.1111/j.1469-8986.1974.tb01114.x

Peelle, J. E. (2017). Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior. *Ear and Hearing*, *39*, 204–214. doi:10.1097/AUD.0000000000000494

Pichora-Fuller, M. K., Schneider, B., & Daneman, M. (1995). How young and old adults listen to and remember speech in noise. *Journal of the Acoustical Society of America*, *97*, 593–608. doi:10.1121/1.412282

Pichora-Fuller, M. K., Kramer, S., Eckert, M., Edwards, B., Hornsby, B., Humes, L., . . . Wingfield, A. (2016). Hearing impairment and cognitive energy: The framework for understanding effortful listening (FUEL). *Ear and Hearing*, *37*(Suppl. 1): 5S–27S. doi:10.1097/AUD.0000000000000312

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

Privitera, C., Renninger, L., Carney, T., Klein, S., & Aguilar, M. (2010). Pupil dilation during visual target detection. *Journal of Vision*, *10*, 1–14. doi:10.1167/10.10.3

Purves, D., Augustine, G., Fitzpatrick, D., Hall, W., LaMantia, A., McNamara, J., & Williams, S. (Eds.). (2004). *Neuroscience* (3rd ed.). Sunderland, England: Sinauer Associates, Inc.

Rajkowski, J., Kubiak, P., & Aston-Jones, G. (1993). Correlations between locus coeruleus (LC) neural activity, pupil diameter and behavior in monkey support a role of LC in attention. *Society of Neuroscience Abstracts*, *19*, 974.

Rajkowski, J., Majczynski, H., Clayton, E., & Aston-Jones, G. (2004). Activation of monkey locus coeruleus neurons varies with difficulty and performance in a target detection task. *Journal of Neurophysiology*, *92*, 361–371. doi:10.1152/jn.00673.2003

Reimer, J., McGinley, M., Liu, Y., Rodenkirch, C., Wang, Q., McCormick, D., & Tolias, A. (2016). Pupil fluctuations track rapid changes in adrenergic and cholinergic activity in cortex. *Nature Communications*, *7*, 13289. doi:10.1038/ncomms13289

Rönnberg, J., Lunner, T., & Zekveld, A. (2013). The ease of language understanding (ELU) model: Theoretical, empirical and clinical advances. *Frontiers in Systems Neuroscience*, *13*, 7–31. doi:10.3389/fnsys.2013.00031

Samadani, U., Ritlop, R., Reyes, M., Nehrbass, E., Li, M., Lamm, E., . . . Huang, P. (2015). Eye tracking detects disconjugate eye movements associated with structural traumatic brain injury and concussion. *Journal of Neurotrauma*, *32*, 548–556. doi:10.1089/neu.2014.3687

Schluroff, M., Zimmerman, T., Freeman, R., Hofmeister, K., Lorscheid, T., & Weber, A. (1986). Pupillary responses to syntactic ambiguity of sentences. *Brain and Language*, *27*, 322–344. doi:10.1016/0093-934X(86)90023-4

Schmidtke, J. (2014). Second language experience modulates word retrieval effort in bilinguals: Evidence from pupillometry. *Frontiers in Psychology*, *5*, 1–16. doi:10.3389/fpsyg.2014.00137

Schmidtke, J. (2017). Pupillometry in linguistic research: An introduction and review for second language researchers. *Studies in Second Language Acquisition*. Advance online publication. doi:10.1017/S0272263117000195

Schneider, M., Hathway, P., Leuchs, L., Sämann, P., Czisch, M., & Spoormaker, V. (2016). Spontaneous pupil dilations during the resting state are associated with activation of the salience network. *NeuroImage*, *139*, 189–201. doi:10.1016/j.neuroimage.2016.06.011

Siegle, G., Steinhauer, S., Stenger, V., Konecky, R., & Carter, C. (2003). Use of concurrent pupil dilation assessment to inform interpretation and analysis of fMRI data. *NeuroImage*, *20*, 114–124. doi:10.1016/S1053-8119(03)00298-2

Snedeker, J., & Trueswell, J. (2004). The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive Psychology*, *49*, 238–299. doi:10.1016/j.cogpsych.2004.03.001

Stanners, R. F., Coulter, M., Sweet, A. W., & Murphy, P. (1979). The pupillary response as an indicator of arousal and cognition. *Motivation and Emotion*, *3*, 319–339. doi:10.1007/BF00994048

Steel, M., Papsin, B., & Gordon, K. (2015). Binaural fusion and listening effort in children who use bilateral cochlear implants: A psychoacoustic and pupillometric study. *PLoS One*, *10*, e011761. doi:10.1371/journal.pone.0117611

Steinhauer, S. R., Seigle, G., Condray, R., & Pless, M. (2004). Sympathetic and parasympathetic innervation of pupillary dilation during sustained processing. *International Journal of Psychophysiology*, *52*, 77–86. doi:10.1016/j.ijpsycho.2003.12.005

Steinhauer, S. R., & Zubin, J. (1982). Vulnerability to schizophrenia: Information processing in the pupil and event-related potential. In E. Usdin, & I. Hanin (Eds), *Biological markers in psychiatry and neurology* (pp. 371–385). Oxford, England: Pergamon Press.

Tanenhaus, M., Spivey, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 1632–1634.

Tavano, A., & Scharinger, M. (2015). Prediction in speech and language processing. *Cortex*, *68*, 1–7. doi:10.1016/j.cortex.2015.05.001

Teubner-Rhodes, S., Vaden, K., Dubno, J., & Eckert, M. (2017). Cognitive persistence: Development and validation of a novel measure from the Wisconsin Card Sorting Test. *Neuropsychologia*, *102*, 95–108. doi:10.1016/j.neuropsychologia.2017.05.027

Tryon, W. (1975). Pupillometry: A survey of sources of variation. *Psychophysiology*, *12*, 90–93. doi:10.1111/j.1469-8986.1975.tb03068.x

Vaden, K., Kuchinsky, S., Ahlstrom, J., Dubno, J., & Eckert, M. (2015). Cortical activity predicts which older adults recognize speech in noise and when. *Journal of Neuroscience*, *35*(9), 3929–3937. doi:10.1523/JNEUROSCI.2908-14.2015

van Rij, J. (2012). *Pronoun processing: Computational, behavioral, and psychophysiological studies in children and adults* (doctoral thesis). University of Groningen, The Netherlands.

van Rij, J., Hendriks, P., van Rijn, H., Baayen, R. H., & Wood, S. (2018). Analyzing the time course of pupillometric data. *Trends in Hearing*.

Veneman, C., Gordon-Salant, S., Matthews, L., & Dubno, J. (2013). Age and measurement time-of-day effects on speech recognition in noise. *Ear and Hearing*, *34*, 288–299. doi:10.1097/AUD.0b013e31826d0b81

Verney, S. P., Granholm, E., & Marshall, S. P. (2004). Pupillary responses on the visual backward masking task reflect general cognitive ability. *International Journal of Psychophysiology*, *52*, 23–26. doi:10.1016/j.ijpsycho.2003.12.003

Vogelzang, M., Hendriks, P., & van Rijn, H. (2016). Pupillary responses reflect ambiguity resolution in pronoun processing. *Language, Cognition and Neuroscience*, *31*, 876–885. doi:10.1080/23273798.2016.1155718

Wagner, L., Maurits, N., Maat, B., Baskent, D., & Wagner, A. (2018). The cochlear implant EEG artifact recorded from an artificial bran for complex acoustic stimuli. *IEEE Transactions on Neural Systems and Rehabilitative Engineering*, *26*, 392–399. doi:10.1109/TNSRE.2018.2789780

Wagner, A., Toffanin, P., & Baskent, D. (2016). The timing and effort of lexical access in natural and degraded speech. *Frontiers in Psychology*, *7*, 398. doi:10.3389/fpsyg.2016.00398

Wang, Y., Zekveld, A., Lunner, T., & Kramer, S. (2018). Pupil light reflex evoked by light-emitting diode and computer screen: Methodology and association with need for recovery in daily life. *PLoS One*, *13*, e0197739. doi:10.1371/journal.pone.0197739

Weinstein, B., & Ventry, I. (1982). Hearing impairment and social isolation in the elderly. *Journal of Speech and Hearing Research*, *25*, 593–599. doi:10.1044/jshr.2504.593

Wendt, D., Dau, T., & Hjortkjær, J. (2016). Impact of background noise and sentence complexity on processing demands during sentence comprehension. *Frontiers in Psychology*, *7*, 345. doi:10.3389/fpsyg.2016.00345

Wendt, D., Hietkamp, R., & Lunner, T. (2017). Impact of noise and noise reduction on processing effort: A pupillometry study. *Ear and Hearing*, *38*, 690–700.

Wendt, D., Koelewijn, T., Książek P., Kramer S., & Lunner, T. (2018). Toward a more comprehensive understanding of the impact of masker type and signal-to-noise ratio on the pupillary response while performing a speech-in-noise test. *Hearing Research*.

Wierda, S., Van Rijn, H., Taatgen, N., & Martens, S. (2012). Pupil dilation deconvolution reveals the dynamics of attention at high temporal resolution. *Proceedings of the National Academy of Sciences*, *109*, 8456–8460. doi:10.1073/pnas.1201858109

Wilhelm, B., Stuiber, G., Lüdtke, H., & Wilhelm, H. (2014). The effect of caffeine on spontaneous pupillary oscillations. *Ophthalmic & Physiological Optics*, *34*, 73–81. doi:10.1111/opo.12094

Williamson, R. S., Hancock, K. E., Shinn-Cunningham, B. G., & Polley, D. B. (2015). Locomotion and task demands differentially modulate thalamic audiovisual processing during active search. *Current Biology*, *26*, 1885–1891. doi:10.1016/j.cub.2015.05.045

Winn, B., Whitaker, D., Elliott, D., & Phillips, J. (1994). Factors affecting light-adapted pupil size in normal human subjects. *Investigative Ophthalmology & Visual Science*, *35*, 1132–1137.

Winn, M. (2016). Rapid release from listening effort resulting from semantic context, and effects of spectral degradation and cochlear implants. *Trends in Hearing*, *20*, 1–17. doi:10.1177/2331216516669723

Winn, M., Edwards, J., & Litovsky, R. (2015). The impact of auditory spectral resolution on listening effort revealed by pupil dilation. *Ear and Hearing*, *36*, e153–e165. doi:10.1097/AUD.0000000000000145

Winn, M., & Moore, A. (2018). Pupil dilation reveals ongoing effort in speech comprehension that is vulnerable to interference from later-occurring sounds: A comparison of listeners with normal hearing and listeners with cochlear implants. *Trends in Hearing*.

Wu, Y.-H., Stangl, E., Chipara, O., Hasan, S., Welhaven, A., & Oleson, J. (2018). Characteristics of real-world signal to noise ratios and speech listening situations of older adults with mild to moderate hearing loss. *Ear and Hearing*, *39*, 293–304. doi:10.1097/AUD.0000000000000486

Wu, Y.-H., Stangl, E., Zhang, X., Perkins, J., & Eilers, E. (2016). Psychometric functions of dual-task paradigms for measuring listening effort. *Ear and Hearing*, *37*, 660–670. doi:10.1097/AUD.0000000000000335

Zekveld, A., Festen, J., & Kramer, S. (2013). Task difficulty differentially affects two measures of processing load: The pupil response during sentence processing and delayed cued recall of the sentences. *Journal of Speech Language and Hearing Research*, *56*, 1156–1165. doi:10.1044/1092-4388(2012/12-0058)

Zekveld, A., Heslenfeld, D., Johnsrude, I., Versfeld, N., & Kramer, S. (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. doi:10.1016/j.neuroimage.2014.06.069

Zekveld, A., & Kramer, S. (2014). Cognitive processing load across a wide range of listening conditions: Insights from pupillometry. *Psychophysiology*, *51*, 277–284. doi:10.1111/psyp.12151

Zekveld, A., Kramer, S., & Festen, J. (2010). Pupil response as an indication of effortful listening: The influence of sentence intelligibility. *Ear and Hearing*, *31*, 480–490. doi:10.1097/AUD.0b013e3181d4f251

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

Zekveld, A., Koelewijn, T., & Kramer, S. (2018). Pupil dilation response to auditory stimuli: Current state of knowledge. *Trends in Hearing*.