

1    **Attention Mobilization as a Modulator of Listening Effort: Evidence from Pupillometry**

2    Johns, M. A.<sup>1†</sup>, Calloway, R. C. <sup>1</sup>, Karunathilake, I. M. D. <sup>2</sup>, Decruy, L. P. <sup>1</sup>, Anderson, S.<sup>3</sup>, Simon,

3    J. Z. <sup>1,2,4</sup>, & Kuchinsky, S. E. <sup>3,5</sup>

4

5    <sup>1</sup>Institute for Systems Research, University of Maryland, College Park, MD, 20742, USA

6    <sup>2</sup>Department of Electrical and Computer Engineering, University of Maryland, College Park, MD,

7    20742, USA

8    <sup>3</sup>Department of Hearing and Speech Sciences, University of Maryland, College Park, MD 20742,

9    USA

10    <sup>4</sup>Department of Biology, University of Maryland, College Park, MD, 20742, USA

11    <sup>5</sup>National Military Audiology and Speech Pathology Center, Walter Reed National Military

12    Medical Center, Bethesda, MD 20889, USA

13

14    <sup>†</sup>Corresponding Author: Michael A. Johns, Institute for Systems Research, University of

15    Maryland, College Park, MD, 20742, USA. Email: maj@umd.edu

16

1    **Abstract**

2    Listening to speech in noise can require substantial mental effort, even among younger normal-  
3    hearing adults. The task-evoked pupil response (TEPR) has been shown to track the increased  
4    effort exerted to recognize words or sentences in increasing noise. However, few studies have  
5    examined the trajectory of listening effort across longer, more natural, stretches of speech, or the  
6    extent to which expectations about upcoming listening difficulty modulate the TEPR. Seventeen  
7    younger normal-hearing adults listened to 60-s-long audiobook passages, repeated three times in  
8    a row, at two different signal-to-noise ratios (SNRs) while pupil size was recorded. There was a  
9    significant interaction between SNR, repetition, and baseline pupil size on sustained listening  
10   effort. At lower baseline pupil sizes, potentially reflecting lower attention mobilization, TEPRs  
11   were more sustained in the harder SNR condition, particularly when attention mobilization  
12   remained low by the third presentation. At intermediate baseline pupil sizes, differences between  
13   conditions were largely absent, suggesting these listeners had optimally mobilized their attention  
14   for both SNRs. Lastly, at higher baseline pupil sizes, potentially reflecting over-mobilization of  
15   attention, the effect of SNR was initially reversed for the second and third presentations:  
16   participants initially appeared to disengage in the harder SNR condition, resulting in reduced  
17   TEPRs that recovered in the second half of the story. Together, these findings suggest that the  
18   unfolding of listening effort over time depends critically on the extent to which individuals have  
19   successfully mobilized their attention in anticipation of difficult listening conditions.

20

21    **Keywords:** speech in noise, anticipatory arousal, mental effort, task-evoked pupil response,  
22    baseline pupil size

23

## 1    **Introduction**

2            Listening to and understanding speech can require substantial mental effort, even if the  
3    words are ultimately correctly perceived (McCoy et al., 2005), indicating that speech-intelligibility  
4    measures alone are insufficient to characterize the difficulty of the listening process. Listeners  
5    must use a limited set of cognitive resources to simultaneously maintain attention to the target  
6    speaker, process the linguistic content, and comprehend the intended message (Carroll et al., 2016;  
7    Kidd et al., 2014). The effort required to accomplish this can further be compounded in adverse  
8    listening conditions, such as in the presence of background noise or competing speakers (Matty's  
9    et al., 2012; Alain et al., 2018; Killion et al., 2004), even for normal-hearing younger adults  
10   (Zekveld et al., 2010). In such contexts, listeners must engage in auditory stream segregation,  
11   tuning in to the target speaker based on low-level acoustic features (e.g., pitch) and/or high-level  
12   semantic content (e.g., topic) while tuning out irrelevant acoustic signals (see Snyder & Alain,  
13   2007 for a review and discussion).

## 14    *Sustained Attention to Listening*

15           A further source of difficulty arises when listening for long periods of time—such as  
16   having a conversation in a crowded restaurant or attending a poster session in a noisy convention  
17   center. In cases of prolonged listening, sustained attention may lead to fatigue and reduced  
18   deployment of cognitive resources to meet task demands (McGarrigle et al., 2017). Sustained  
19   attention has been defined in terms of an individual's readiness to detect rare or unpredictable  
20   signals over time (Sarter et al., 2001). Depending upon one's model of cognition (for a review, see  
21   Fortenbaugh et al., 2017), sustained attention has been viewed as a separable subtype of attention  
22   (tonic and phasic alerting; Posner & Peterson, 1990), as involving multiple subtypes of attention

1 (e.g., alerting and orienting; Tang et al., 2015) or as a function of multiple sensory and cognitive  
2 functions to sustain processing to internal or external information across long periods of time  
3 (Chun et al., 2011).

4 There is increasing awareness within the hearing sciences of the need for laboratory stimuli  
5 and tasks that better reflect real-world listening situations, which includes listening to extended  
6 connected discourse (for a consensus paper, see Keidser et al., 2020). However, much of the  
7 research on sustained attention outside the domain of listening has focused on simple vigilance  
8 tasks (Kristjansson et al., 2009; Martin et al., 2022), and most research on listening effort has  
9 focused on short sentences (Winn, 2016; Winn & Moore, 2018; Zekveld et al., 2010), although  
10 some work has expanded to longer listening situations, such as strings of three connected sentences  
11 (McGarrigle et al., 2017) and 25-s long tone streams (Zhao et al., 2019). In an auditory decoding  
12 study, greater listening effort, as indicated by variation in average pupil dilation and in parietal  
13 alpha power, was observed to predict endogenous attention switches as individuals listened to 60-  
14 s-long audiobook passages (Haro et al., 2022). In two studies of hearing aid users, listeners  
15 attended to speech stimuli that were ~30-second news stories presented in 4-talker background  
16 babble while EEG and pupillometry were recorded (Fiedler et al., 2021; Seifi Ala et al., 2020).  
17 Seifi Ala et al. (2020) observed larger mean pupil sizes in the more challenging signal-to-noise  
18 ratio (SNR) condition (-5 vs. 0 dB SNR in 4-talker babble), both overall and across 5-second time  
19 bins. Fiedler et al. (2021) found an interaction between noise reduction and SNR (+3 vs. +8 dB  
20 SNR) on mean pupil size, such that a larger benefit of noise reduction was observed at the more  
21 challenging SNR. Thus, while substantial research has focused on examining listening effort in  
22 response to single words and sentences in adverse conditions (for a review, see Zekveld et al.,  
23 2018), there has been less work investigating how attention and effort are mobilized and sustained

1 throughout extended durations of connected speech, particularly within individual listening trials  
2 for younger adults with normal-hearing thresholds.

3 Examining the relationship between sustained attention and listening effort with longer  
4 stimuli may ultimately be more reflective of real-world listening situations for two reasons. First,  
5 longer passages of connected discourse may more adequately reflect listeners' day-to-day  
6 experiences with language (i.e., verisimilitude; Franzen & Wilhelm, 1996). Second, single words  
7 and disconnected sentences lack some of the higher-level semantic and pragmatic processes that  
8 are often crucial to understanding longer stretches of speech, such as keeping track of different  
9 types of information (e.g., topics, referents, and events) over long periods of time (see Sparks &  
10 Rapp, 2010 for a review and discussion). Importantly, if the listener misses crucial information  
11 due to adverse listening conditions or to the effects of fatigue, for example, then this can have  
12 downstream consequences for comprehension (Winn, 2023).

13 *Pupillometry Measures of Sustained Attention to Listening*

14 The extent to which an individual allocates their attentional resources to a listening task at  
15 a given point in time is determined by a number of factors laid out in the Framework for Effortful  
16 Listening (FUEL; Pichora-Fuller et al., 2016). FUEL defines listening effort in terms of the  
17 allocation of capacity-limited mental resources to demands of a listening task. This definition  
18 highlights that listening effort is a function of listening demands, listener capacities, and a so-  
19 called effort allocation policy. Motivation and arousal, which may be particularly expected to  
20 change over extended listening epochs, are key determinants of that policy, affecting how much  
21 and when available mental resources are applied to a task, partly determined by "the demands  
22 imposed by the activities in which the organism engages, or prepares to engage" (Kahneman, 1973,

1 p. 17). This suggests that comprehensive measures of listening effort should incorporate indices  
2 of arousal, particularly as to the extent that changes are expected over time.

3 While subjective measures of effort, intelligibility, and attention have provided useful  
4 insights into behaviors and perceived effort during listening tasks, these measures may not  
5 adequately reflect a listener's current arousal state or the amount of effort that was ultimately used  
6 to accomplish the task (Winn & Teece, 2021, 2022). Alternatively, changes in pupil dilation have  
7 been used as an online, objective measure of cognitive effort, attention, and arousal (Wagner et  
8 al., 2019; Zekveld et al., 2010; Zekveld & Kramer, 2014) and have been linked to locus coeruleus  
9 (LC) activity in the brain (Rajkowski et al., 1993; Elman et al., 2017; Murphy et al., 2014) and  
10 LC-driven patterns of behavior (Gilzenrat et al., 2010). Increased activity in the LC results in  
11 increased concentrations of norepinephrine (NE) that are present during periods of high attentional  
12 allocation and arousal (Aston-Jones & Cohen, 2005).

13 *Pupillometry Measures of Interactions Between Arousal State and Task Evoked  
14 Listening Effort*

15 Two distinct modes of LC activation—tonic and phasic—have also been linked to different  
16 aspects of the pupil response that, in turn, reflect different attentional states. Pupil size during a  
17 neutral baseline period (prior to stimulus onset) has been argued to reflect tonic LC activity and  
18 can serve as an indicator of general arousal (in an inattentive, engaged, or distractible state) as well  
19 as anticipatory arousal (Ayasse & Wingfield, 2020) or attention mobilization (Seropian et al.,  
20 2022)—the readying of cognitive resources in preparation to carry out an upcoming task.  
21 Expectations about upcoming listening challenges, as may be experienced when listening in poorer  
22 signal-to-noise ratios (SNRs) or with a hearing impairment, have been observed to alter attention  
23 mobilization as indexed by baseline pupil size (Seropian et al., 2022). For example, Ayasse and

1 Wingfield (2020) examined baseline pupil dilation over the course of a 160-trial auditory sentence  
2 comprehension task in both normal-hearing and hearing-impaired individuals. While hearing-  
3 impaired listeners began the task with larger baseline pupil sizes compared to normal-hearing  
4 listeners, baseline pupil size gradually decreased, with the two groups becoming more similar by  
5 the end of the task. Importantly, response accuracy increased across the task, suggesting that this  
6 decline was not due to fatigue or disengagement, but rather to “an increased level of arousal  
7 reflecting task anxiety or a lack of confidence in likely success” (Ayasse & Wingfield, 2020, p. 5)  
8 or to an increase in attention mobilization in anticipation of a difficult task.

9 The task-evoked pupil response (TEPR) is a measure of the relative change in pupil dilation  
10 that is time locked to the onset of an attended stimulus that is thought to reflect, in part, phasic LC  
11 activity (Joshi et al., 2016). Larger TEPRs are often associated with increased attention and task  
12 difficulty, as well as with more salient stimuli (Zekveld et al., 2018). In listening tasks, larger task-  
13 evoked pupil sizes have been shown to reflect increased listening effort, with increasing pupil size  
14 associated with greater task difficulty (McGarrigle et al., 2017; Winn, 2016; Zhao et al., 2019).  
15 Previous research has suggested that poorer SNRs result in increased TEPRs—until a tipping point  
16 when listeners begin to give up and disengage—indicative of the increased effort required to  
17 comprehend a degraded speech signal (Ohlenforst et al., 2017, Koelewijn et al., 2015). While  
18 “giving up” is generally associated with reductions in both pupil size and performance, patterns of  
19 relative disengagement (and thus reductions in effort) can also be observed with relatively good  
20 performance. Following the “principle of least effort” (Ayasse et al., 2021), individuals may exert  
21 only the minimum effort needed to perform a task when they do not feel motivated to process the  
22 speech more deeply, such as when listening to extended boring monologues (Herrmann &  
23 Johnsrude, 2020). Reductions in pupillary measures of listening effort have also been observed

1 with increasing stimulus familiarity, such as when encountering more commonly used lexical  
2 items (Papesh & Goldinger, 2012) or repeatedly encountering the same auditory (Marois et al.,  
3 2018) or visual (Ferrari et al., 2016) stimulus.

4 Tonic and phasic LC activity—and, by extension, baseline pupil size and the TEPR—are  
5 not independent of one another (e.g., Knapen et al, 2016), with their nonlinear relationship  
6 reflected on a Yerkes-Dodson curve (Yerkes & Dodson, 1908). Low tonic LC activity is related to  
7 inattentiveness and under-mobilization of attentional resources, which is associated with poorer  
8 performance, lower baseline pupil sizes, and reduced TEPRs. Intermediate levels of tonic LC  
9 activity have been linked to optimal arousal states and task performance (McGinley et al., 2015),  
10 such that intermediate baseline pupil sizes result in the largest TEPRs (Murphy et al., 2011). This  
11 state may reflect optimal mobilization of attentional resources (i.e., exploitative rather than  
12 explorative; Jepma & Nieuwenhuis, 2011). Lastly, high tonic LC activity (also known as a  
13 hyperactive tonic state) has been associated with increased distractibility, task disengagement, and  
14 decreased task performance (Kane et al., 2017; McGinley et al., 2015; Murphy et al., 2011;  
15 Unsworth & Robison, 2016). Additionally, in human models, high LC-NE tonic activity has also  
16 been associated with higher rates of self-reported mind wandering (i.e., off-task thoughts) during  
17 reading (Franklin et al., 2013). As such, this state is associated with higher baseline pupil sizes but  
18 reduced TEPRs, and may reflect over-mobilization of attentional resources (i.e., explorative rather  
19 than exploitative).

20 Recently, Relaño-Iborra et al., (2022) examined the relationship between baseline pupil  
21 size and the TEPR, using pupil recordings from a speech intelligibility task with blocked SNRs  
22 (Wendt et al., 2018). The authors found that baseline pupil size was not only modulated by time-  
23 on-task effects and SNR, but also significantly modulated the shape the shape of the TEPR derived

1 from a growth curve analysis (GCA) model. Baseline pupil size was found to increase with poorer  
2 SNRs for both four-talker babble and speech-shaped noise. The authors suggested that the increase  
3 in baseline pupil size in the more difficult conditions may have reflected preparatory control:  
4 because SNR conditions were blocked, participants could anticipate the difficulty of upcoming  
5 trials. Interestingly, however, the effects of SNR tended to diminish as the task progressed, which  
6 may indicate that “[a]fter sufficient exposure, listeners seem able to gauge whether effort  
7 deployment would result in a successful completion of the task, thus disengaging from it if success  
8 could not be achieved” (Relaño-Iborra et al., 2022, p. 12).

9 Together, these studies suggest that one’s arousal state has a critical, and strongly non-  
10 monotonic, impact on effort allocation to task demands. However, more research is needed to  
11 understand potential interactions between anticipated acoustic difficulties and stimulus repetition  
12 effects, particularly at the level of individual listening trials. Furthermore, studies that have  
13 examined the TEPR as a measure of listening effort have predominantly utilized trial-by-trial  
14 baseline pupil size to account for trial- and participant-level variability – either to be subtracted  
15 from or to normalize TEPR values (Mathôt et al., 2018). However, as noted, baseline pupil size  
16 has been observed to not only affect the height of the TEPR, but also its shape (Knapen et al.,  
17 2016; Relaño-Iborra et al., 2022). Previous research has also suggested that baseline pupil size and  
18 the TEPR may reflect different processes (Micula et al., 2021; 2022). Thus, to the extent baseline  
19 pupil size reflects anticipatory attention mobilization and effort for known upcoming listening  
20 demands, traditional baseline correction procedures may obscure or, worse, overcorrect for  
21 meaningful differences between listening conditions.

22 *Goals of the Present Study*

1 The present study examines the relationship between attention mobilization—how  
2 individuals prepare their attention in anticipation of an upcoming task—and listening effort  
3 allocation—how listeners deploy and use their attentional resources during the task—when  
4 listeners can anticipate the difficulty of the upcoming trial. Extending the results of Relaño-Iborra  
5 et al. (2022), the present study focuses on trial-level variation in attention mobilization for a  
6 sustained listening task involving exact stimulus repetitions. Participants listened to three  
7 presentations of several 60-s long audiobook passages and were instructed to attend to one of two  
8 competing speakers in an easy or difficult listening situation, determined by SNR. Participants  
9 were told that specific passages would be blocked in this fashion and thus, the first presentation  
10 effectively served as a cue regarding task difficulty for the two subsequent presentations. Longer  
11 passages were chosen both to examine longer-term changes in the TEPR and to more adequately  
12 approximate real-world listening scenarios (i.e., longer stretches of connected discourse). Our  
13 research questions (RQ) and hypotheses (H) are as follows:

14 RQ1. How is attention mobilization modulated by task difficulty to the extent that listeners  
15 can anticipate how difficult the upcoming stimulus will be?

16 H1. Attention mobilization—and thus baseline pupil size—will be larger for the  
17 harder compared to the easier SNR condition. In addition, subsequent repetitions  
18 (i.e., the second and third presentation) will increase attention mobilization, and  
19 this increase will be larger for the harder compared to the easier SNR condition.

20 RQ2. How is listening effort allocation modulated by task difficulty to the extent that listeners  
21 can anticipate how difficult the upcoming stimulus will be?

22 H2. Listening effort allocation—and thus the TEPR—will be greater for the harder  
23 compared to the easier SNR condition. Stimulus repetitions will decrease

1 listening effort, and this decrease will be larger for the harder compared to the  
2 easier SNR condition (i.e., a steeper linear decline in the TEPR).

3 RQ3. How does attention mobilization interact with listening effort allocation to the extent  
4 that listeners can anticipate how difficult the upcoming stimulus will be?

5 H3. Attention mobilization (baseline pupil size) will modulate listening effort  
6 allocation (via the TEPR) in the following ways: 1) at lower baseline pupil sizes  
7 (i.e., lower tonic LC activity), the TEPR for both SNR conditions (0 dB and -6  
8 dB) will be diminished, as will differences in the TEPR between the two  
9 conditions; 2) at intermediate baseline pupil sizes (i.e., intermediate tonic LC  
10 activity), the TEPR for both conditions will be largest, with the harder SNR  
11 condition eliciting larger TEPRs compared to the easier SNR condition; and 3) at  
12 higher baseline pupil sizes (i.e., higher tonic LC activity), while the TEPR may  
13 be elevated, differences between the two conditions will again be diminished.

## 14 **Methods**

### 15 *Participants*

16 Nineteen participants (12 women, 7 men;  $M_{age} = 21.1$  years,  $SD = 2.16$ , *range*: 18.5 to 26.1)  
17 were enrolled in the study, which was approved by the University of Maryland's Institutional  
18 Review Board. Participants received monetary compensation for their participation. Participants  
19 were administered an audiogram in each ear that included third octave band tones from 0.125 to  
20 14 kHz. All participants had audiometric thresholds within normal limits of  $\leq 25$  dB HL from 0.25  
21 to 4 kHz in their better ear. Participants self-reported having normal or corrected-to-normal vision,  
22 no psychiatric or neurological conditions, not taking psychoactive stimulants or depressants, and

1 were native English speakers with no exposure to a second language prior to the age of 12. A score  
2 in the normal range of 26 or better on the Montreal Cognitive Assessment (MoCA) was also  
3 required for participation.

4 *Measures and Stimuli*

5 The audiobook listening task was part of a larger study where magnetoencephalography  
6 (MEG) data were also collected during the audiobook listening task on the same participants. The  
7 method and discussion of the MEG data are reported in Karunathilake et al. (2023). The audiobook  
8 task consisted of 60-s long audiobook segments from a 19<sup>th</sup> century short story available in the  
9 public domain (male recording: Irving, 2006; female recording: Irving, 1977). Stimuli were  
10 presented across four blocked SNR conditions: 0 dB, -6 dB, Babble, and Clean. In the 0 dB and -  
11 6 dB conditions, participants heard two different passages in each block with each passage  
12 presented three times in a row. To avoid using a fixed order of audiobook passages (e.g., all  
13 participants hearing the same passages in the same order), four lists of stimuli were created such  
14 that, within each list, the order of the individual audiobook passages was pseudorandomized. These  
15 lists were then divided into four blocks, one for each of the SNR conditions. In the current study,  
16 only the 0- and -6-dB blocks were analyzed because they always occurred before the Babble and  
17 Clean blocks, with the order of the 0 dB and -6 dB blocks counterbalanced across lists (i.e., some  
18 participants heard the 0 dB block first while others heard the -6 dB block first). These two SNRs  
19 also showed the greatest difference in the neural reconstruction of the speech envelope in a prior  
20 MEG study using these same speech materials (Presacco et al., 2016, Fig. 6). Additionally, the  
21 Clean condition utilized repeated segments from the other conditions, while in the Babble  
22 condition the competing speech was multi-talker babble that does not convey any meaning, unlike  
23 the competing talkers in the 0 dB and -6 dB conditions. Given this difference, we opted to exclude

1 the Clean and Babble blocks from our analyses and instead focus on the effects of SNR between  
2 two competing speakers. Stimuli in the 0 dB and -6 dB conditions had participants attend to either  
3 a female or a male speaker in the presence of a competing speaker of the other gender speaking a  
4 different portion of the audiobook that was not present in any other stimuli in these conditions. In  
5 the 0 dB condition, both speakers were presented at 70 dB SPL. In the -6 dB condition, the target  
6 speaker remained at 70 dB SPL while the competing speaker was presented at 76 dB SPL. For  
7 both conditions, half of the stimuli had participants attend to the female speaker and half to the  
8 male speaker. This resulted in two audiobook segments for each SNR condition. As mentioned  
9 above, in order to allow for signal averaging in an MEG study of auditory encoding (Karunathilake  
10 et al., 2023), each stimulus was repeated three times in a row. While repetition allows for stability  
11 in MEG measures of auditory processing, shifts in attention may occur as listeners anticipate and  
12 habituate to the upcoming difficulty and content of the passage. Participants also completed a  
13 separate speech-perception-in-noise (SPIN) task at these same SNRs using sentences extracted  
14 from the audiobook that did not overlap with those used in the audiobook task. The SPIN task  
15 along with the behavioral findings from the audiobook task served as a manipulation check; for  
16 more detailed information about the SPIN task, see Karunathilake et al. (2023). The minimum time  
17 between the offset of one auditory passage and the onset of the baseline epoch for the next passage  
18 was 69 seconds. This period included time for the experimenter to ask the comprehension question  
19 and, for the first presentation, an intelligibility rating as well as to wait for the MEG signal to  
20 stabilize again following the participant's verbal responses. Specifically, after every presentation,  
21 participants answered a short comprehension question designed only to ensure participants  
22 attended to the story. There was a different question for each repetition of the audiobook passage  
23 which could be a true-or-false, open-ended, or multiple-choice question. Participants were not

1 given feedback about their response accuracy. After the first presentation of each new audiobook  
2 segment, participants were also asked to provide a subjective intelligibility rating indicating how  
3 much of the passage they understood. The rating was on a scale of 0 to 10, where 0 indicated that  
4 the participant understood none of the passage while 10 indicated that they understood all of the  
5 passage.

6 *Procedure*

7 The initial session took place in a laboratory setting. Intake assessments were administered  
8 in person as part of recruitment efforts for a larger study of neuroplasticity in auditory aging.  
9 Individuals were contacted about potential enrollment in the current study if they met the  
10 aforementioned language, audiogram threshold, vision, psychiatric and neurological history, and  
11 MoCA score requirements to be eligible for the study. In a subsequent session, participants  
12 completed the audiobook listening task. During this task, pupillometry and  
13 magnetoencephalography (MEG) data were collected; however, only the pupillometry data are  
14 presented here (refer to Karunathilake et al., 2023 for a detailed analysis of the MEG and  
15 behavioral data). Participants were situated in a magnetically shielded chamber, lying down with  
16 their eyes 790 mm from the top of a projector screen (772 mm wide x 457 mm tall) and 914 mm  
17 from its bottom. The ambient room lighting was dimmed, and visual stimuli were chosen (medium  
18 gray screen, RGB value of 128, 128, 128) to yield a luminance of 62 lux, to ensure pupil recordings  
19 were collected in the approximate middle of an average individual's expected dynamic range.  
20 Auditory stimuli were administered diotically via insert headphones that were also used by the  
21 experimenter to communicate task instructions. Finally, the SPIN task described above was  
22 administered on a separate day.

1 Pupil size data were collected using an MEG-compatible SR Research EyeLink 1000 Plus  
2 eye-tracker with a long-range mount with a sampling rate of 1000 Hz using monocular tracking.  
3 Prior to the start of the audiobook listening task, participants completed a calibration procedure in  
4 which participants were asked to fixate on a square as it moved around the screen on a nine-point  
5 grid. For the audiobook listening task, participants were instructed to focus on the center of a  
6 medium gray screen where a cartoon image of either a male or female face would be displayed to  
7 indicate the upcoming target speaker. Each of the images was an equi-luminant black line drawing  
8 centered on the screen measuring 183 mm wide by 137 mm tall. The image appeared two seconds  
9 prior to the onset of the passage (i.e., the baseline window) and remained onscreen throughout the  
10 60-s story.

11 An experimenter verbally explained that the participant's task was to listen to the target  
12 speaker and that they would be asked questions after each presentation. The experimenter provided  
13 verbal instructions about the subjective intelligibility ratings, informed participants to respond  
14 aloud, and noted that the experimenter would record responses. The experimenter began each trial  
15 (consisting of a 2-s pre-stimulus baseline and presentation of a 60-s audiobook passage) by first  
16 verbally indicating whether the participant should attend to the male or female speaker and then  
17 manually started the trial. The verbal cue was provided in addition to the visual cue (male or female  
18 face) as redundancy to ensure participants knew which speaker to attend to (because, for example,  
19 the participant might not see the screen clearly due to having removed their glasses for the MEG  
20 scan). At the conclusion of the first presentation of each audiobook segment, the experimenter  
21 asked the comprehension question followed by the subjective intelligibility rating question. For  
22 the remaining two presentations, only the comprehension question was asked. After recording the

1 responses, the experimenter again informed the participant which speaker to attend to and then  
2 manually began the next trial.

3 **Analyses**

4 *Data Preprocessing and Cleaning*

5 Pupil size data were extracted starting from the 2 s baseline period prior to stimulus onset  
6 and 60 s after stimulus onset for each presentation. Pre-processing of pupil data consisted of the  
7 following: first, samples during blinks and saccades were removed, as were any periods of  
8 excessive distortions (e.g., Winn et al., 2018, p. 20). As discussed below, gaze position was  
9 modeled as a two-dimensional univariate smooth (van Rij et al., 2019). As such, data were not  
10 excluded when samples fell away from central fixation (i.e., fixations away from the center of the  
11 screen or off of the image cue) because this multivariate smooth was able to account for the effects  
12 of gaze position on pupil size (Gagl et al., 2011). Prior to filtering, linear interpolation was  
13 performed to fill in missing data as the pupil size data could not be filtered with missing values.  
14 These data were then low pass filtered with a cutoff frequency of 5 Hz using a finite impulse  
15 response (FIR) filter (Hamming window of order 50). Interpolated data were removed after  
16 filtering. Data were then downsampled to 10 Hz.

17 For a given trial, if 30% or more of the pupil size data were excluded during the 2-s baseline  
18 period *or* 45% or more of the pupil size data were excluded during the 60-s stimulus period, that  
19 trial was excluded from analysis. Of the 228 total trials, 69 (30.26%) were excluded based on the  
20 above criteria (0 dB SNR: 33 trials excluded; -6 dB SNR: 36 trials excluded). Participants were  
21 excluded entirely if two or more trials for a given SNR were excluded, eliminating two of the 19

1 participants (total percent trials excluded: 31.58%). Analyses on the pupillometry and behavioral  
2 data included only these 17 participants.

3 *Behavioral Analyses*

4 All analyses were conducted in R (v. 4.2.2; R Core Team, 2024). The R script in its entirety,  
5 as well as the data necessary to replicate these analyses, are available on the Open Science  
6 Framework (<https://osf.io/r396t/>). Accuracy to the SPIN task, as well as accuracy to the  
7 comprehension questions following each presentation of the 60-s audiobook passages, were  
8 analyzed using logistic mixed-effects regression using the glmer function in lme4 (v. 1.1-31; Bates  
9 et al., 2015). The model for the SPIN task predicted the proportion of correctly recalled words in  
10 each sentence by SNR (0 dB, -6 dB) and included a random intercept of subject (including a  
11 random slope of SNR by subject caused the model to not converge). The model for the  
12 comprehension questions predicted accuracy by the interaction between SNR (0 dB, -6 dB) and  
13 presentation (first, second, third) and a random intercept of subject with a random slope of SNR  
14 (including random slopes of the interaction between SNR and presentation or the main effects of  
15 SNR and presentation caused the model to not converge). Self-reported intelligibility ratings after  
16 the first presentation of the 60-s audiobook passages were analyzed using a cumulative link mixed-  
17 effects model (CLMM) using the ordinal package (Christiansen, 2022). The model predicted self-  
18 reported intelligibility ratings by SNR (0 dB, -6 dB) and included a random intercept of subject  
19 (including a random slope of SNR by subject caused the model to not converge).

20 *Pupil Size Analyses*

21 Pre-trial baseline pupil size has been shown to reflect attention or arousal states (Ayasse &  
22 Wingfield, 2020; Wagner et al., 2019) and the study design includes stimulus repetition that may

1 influence such processes. As such, linear mixed-effects regression was performed using the lmer  
2 function in the lme4 package (v. 1.1-31; Bates et al., 2015), and *p*-values were calculated using  
3 lmerTest (Kuznetsova et al., 2017). This model predicted baseline pupil size (the median pupil size  
4 during the 2 s prior to stimulus onset) by the interaction between SNR (0 dB, -6 dB) and  
5 presentation (first, second, third) and included a random intercept of participant (including the  
6 interaction between SNR and presentation or the main effects of SNR and presentation caused the  
7 model to not converge). Pairwise comparisons were conducted using the emmeans function in the  
8 emmeans package (v. 1.8.4-1; Lenth, 2023).

9 The TEPR was analyzed using a generalized additive mixed model (GAMM) which allows  
10 for the modelling of non-linear trends in time series data while simultaneously accounting for  
11 autocorrelation—of particular importance for the TEPR (van Rij et al., 2019). All models were  
12 created using the bam function in the mgcv package (v. 1.8-41; Wood, 2003, 2011, 2017), while  
13 model criticism, testing, and visualization were performed using the itsadug package (v. 2.4.1; van  
14 Rij et al., 2022). The model predicted the TEPR by the ordered factor variables of presentation  
15 (first [reference level], second, third), SNR (0 dB [reference level], -6 dB), and their interaction.  
16 These ordered factors were specified in both the parametric terms—which estimate overall height  
17 differences of the TEPR across conditions—and in the smooth terms. The smooth terms also  
18 included baseline pupil size as an additional continuous predictor alongside time (see below).  
19 Importantly, since baseline pupil size was included in the model—and because baseline correction  
20 can change the shape of the TEPR (i.e., by baseline normalization) or can inadvertently obscure  
21 or even invert differences between conditions (i.e., by baseline subtraction), baseline correction  
22 was not performed on the TEPR (van Rij et al., 2019, p. 4; see also Reilly et al., 2019). As such,  
23 the TEPR is measured as raw pupil size in arbitrary units (a. u.).

1            Ordered factor smooths estimate differences between specific conditions (or combinations  
2    of conditions) similarly to linear regression but implemented within the GAMM framework. A  
3    ‘reference smooth’ estimates the TEPR for the chosen reference level (e.g., first presentation, 0 dB  
4    SNR) and has no factor specified in the ‘by’ argument (analogous to the intercept in the summary  
5    of a linear regression). Subsequent smooths are called ‘difference smooths’ and estimate the  
6    difference between the reference smooth and the condition represented by each difference smooth  
7    using an ordered factor specified in the ‘by’ argument (analogous to the estimates presented below  
8    the intercept in a linear regression). For example, the ordered factor term “SNR6.ord” is true for  
9    all data points in the –6 dB SNR condition and false for all data points in the 0 dB SNR condition.  
10   If this term were the only term in the model, the reference smooth would estimate the TEPR for  
11   the 0 dB SNR condition, while the difference smooth specified by the term “SNR6.ord” would  
12   estimate the *difference* between the 0 dB SNR condition and the –6 dB SNR condition (e.g., what  
13   must be added to the 0 dB SNR smooth in order to get the –6 dB SNR smooth). This is particularly  
14   useful given that the *p*-values provided by a GAMM indicate only if the fitted smooth is  
15   significantly different from 0.

16           The smooth terms were specified using tensor product interactions to examine both how  
17   the TEPR changes over time and also how the shape of this trajectory changes as a function of  
18   baseline pupil size. Tensor product interactions allow for modelling multiple independent variables  
19   with different scales, as a separate penalty matrix is calculated for each variable (Wood, 2017, pp.  
20   325-328). In the present study, these variables are time (e.g., on the x-axis) with units *s* and baseline  
21   pupil size (e.g., on the y-axis) with arbitrary units. We included what Sóskuthy (2021) called  
22   ‘random reference/difference smooths’. These smooths are specified to estimate by-subject factor  
23   smooths using the same ordered factors specified in the tensor product smooths mentioned above.

1 Random reference smooths can be thought of as analogous to intercept differences between  
2 subjects at the reference level of an ordered factor, whereas random difference smooths can be  
3 thought of as analogous to random slopes that represent differences between subjects as estimated  
4 for each condition comparison (Sóskuthy, 2021). In order to fully examine the interaction between  
5 baseline pupil size, SNR, and presentation on the TEPR, the model was subsequently relevelled so  
6 that each presentation (first, second, and third) in the 0 dB SNR condition served as the reference  
7 level (see Pandža et al., 2020 and Phillips et al., 2021 for examples of model releveling). An initial  
8 model was run to estimate the *rho* autocorrelation parameter, which was then used in an embedded  
9 AR1 model. The *rho* value was then adjusted manually until the autocorrelation was sufficiently  
10 accounted for (Porretta et al., 2018). The number of knots (*k*) was increased based on  
11 recommendations from the *gam.check* function in the *itsadug* package. Fitted smooths were  
12 visualized using the *plot\_smooth* function in *itsadug*, fitted heatmaps were created using the  
13 *fvisgam* function in *itsadug*, and difference heatmaps were created using the *plot\_diff2* function in  
14 *itsadug*.

## 15 **Results**

### 16 *Accuracy and Intelligibility Ratings*

17 The generalized linear mixed-model predicting accuracy on the SPIN task showed a  
18 significant main effect of SNR, such that the proportion of correctly recalled words was  
19 significantly greater in the 0 dB compared to the -6 dB SNR condition (Est. = 2.37,  $z = 11.78, p <$   
20  $.001$ ). The proportion of correctly recalled words was .81 ( $sd = .34$ ) in the 0 dB SNR condition  
21 and .42 ( $sd = .25$ ) in the -6 dB SNR condition.

1 The generalized linear mixed-model predicting accuracy to the comprehension questions

2 following each presentation of the audiobook passage suggested no effect of SNR, presentation,

3 or their interaction (all  $p$ -values  $> 0.10$ ). Estimated marginal means calculated using the emmeans

4 function in the emmeans package further suggest no effect of SNR when averaged across

5 presentations and no effect of presentation when averaged across SNRs (all  $p$ -values  $> 0.3$ ).

6 Overall accuracy across SNR and presentation was 69.2% ( $sd = 46.3\%$ ).

7 Lastly, the cumulative link mixed-model predicting self-reported intelligibility ratings

8 following the first presentation of each audiobook passages showed a significant main effect of

9 SNR, such that ratings were significantly lower in the  $-6$  dB SNR condition compared to the  $0$  dB

10 SNR condition (Est. =  $-2.10$ ,  $z = -4.14$ ,  $p < .001$ ). Average intelligibility ratings were 5.84 ( $sd =$

11 1.80) in the  $0$  dB SNR condition and 4.66 ( $sd = 1.58$ ) in the  $-6$  dB SNR condition. Combining the

12 results of the SPIN task with the behavioral results from the audiobook task suggest that the SNR

13 manipulation was successful.

#### 14 *Effects of presentation and SNR on attention mobilization via baseline pupil size*

15 The model analyzing baseline pupil size showed a significant main effect of presentation.

16 Pairwise comparisons of estimated marginal means showed that baseline pupil sizes for the first

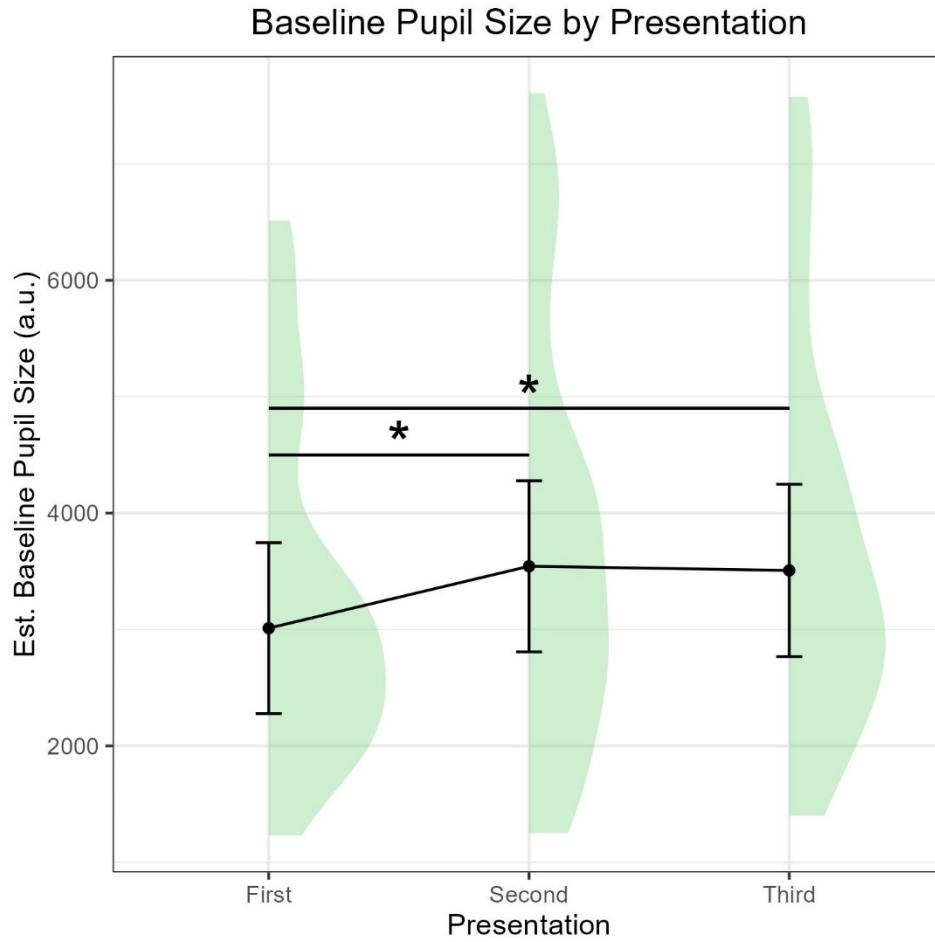
17 presentation were smaller compared to the second ( $t = 4.16$ ,  $p = .04$ ) and third ( $t = 4.07$ ,  $p < .001$ )

18 presentations. There was no difference between the second and third presentations ( $p = .76$ ) nor

19 any interactions between presentation and SNR. The model summary is provided in Table 1, and

20 model estimates of baseline pupil size are shown in Figure 1.

21 [INSERT TABLE 1 ABOUT HERE]



1

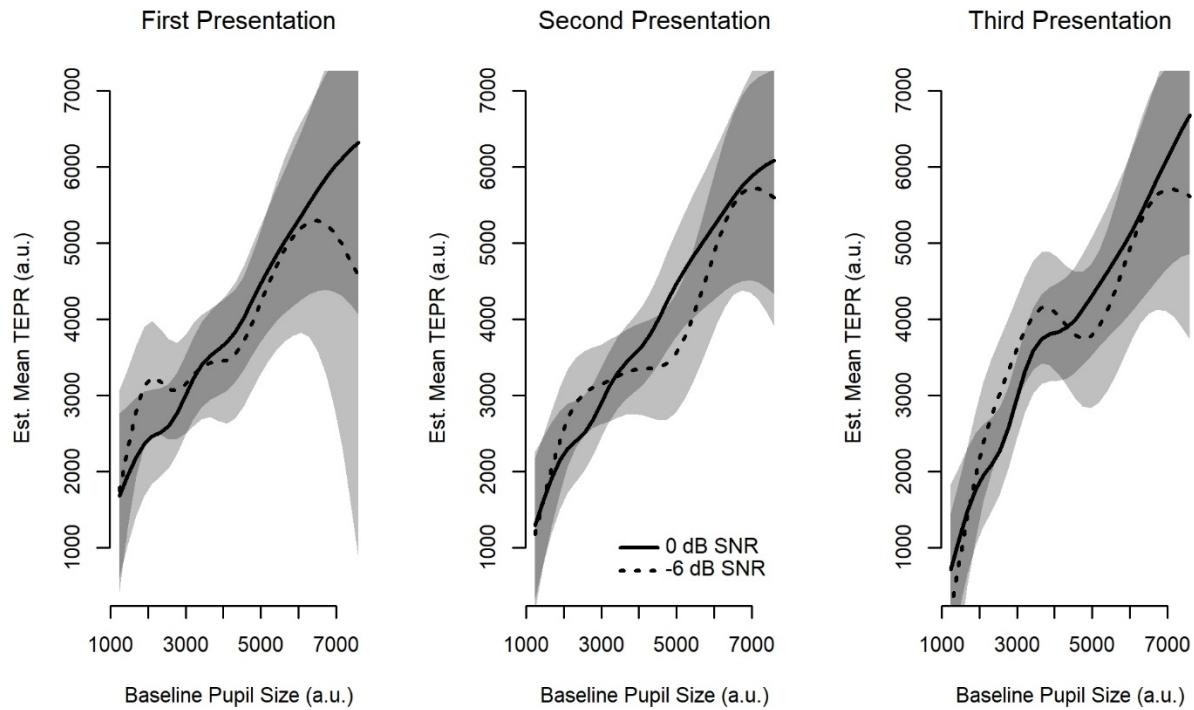
2 **Figure 1.** Model-estimated baseline pupil size values by presentation, collapsed across SNR. Baseline pupil size is  
3 based on the median pupil size during a 2-s period of silence before the start of the audio with the male or female  
4 face cue present on screen. Error bars represent the 95% confidence interval; shaded green regions represent the  
5 distribution of raw (e.g., not model-estimated) baseline pupil size values for each presentation. Horizontal lines  
6 with asterisks indicate a significant difference between the indicated presentations.

7 *Effects of presentation and SNR on sustained listening effort via dynamic pupil  
8 response*

9 The summary table for the GAMM used to analyze the TEPR, with the first presentation  
10 at 0 dB SNR as the reference level, is presented in Table 2. Summaries for when the model was  
11 relevelled to the second and third presentations are presented in Appendix A. For the parametric  
12 effects, there were no significant effects of SNR or presentation on the overall height of the TEPR.

1 A key reason for this, as detailed below, is that these effects seem to vary greatly depending on  
2 both the time within the 60-s passage as well as baseline pupil size. It is also important to note that,  
3 consistent with previous literature (Gilzenrat et al., 2010), increasing baseline pupil size was  
4 associated with overall larger TEPRs, as can be seen in Figure 2 below.

5 [INSERT TABLE 2 ABOUT HERE]



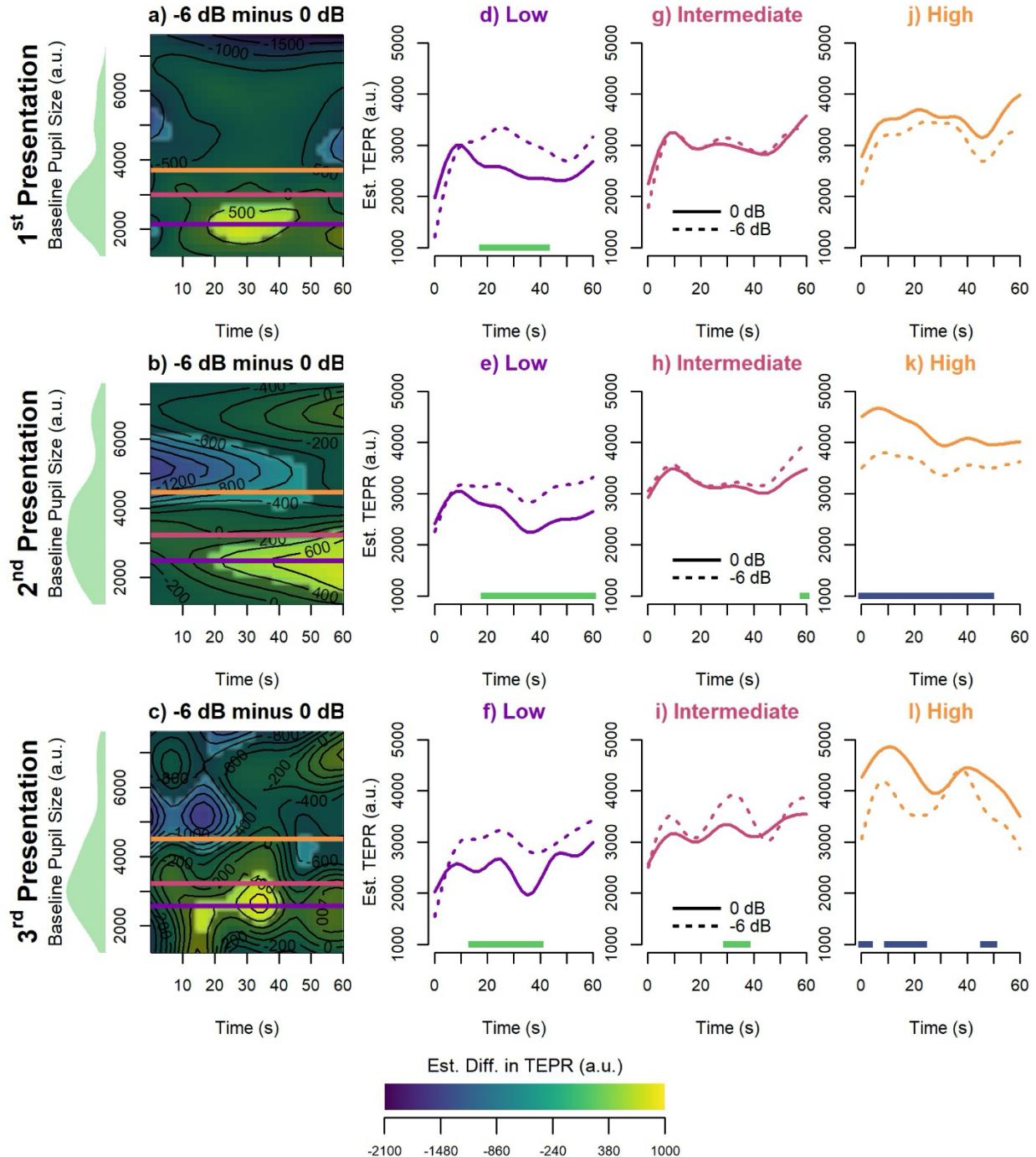
6  
7 **Figure 2. Model-estimated mean evoked pupil size as a non-linear function of baseline pupil size for each**  
8 **presentation/SNR combination.**

9 The tensor product interactions suggested significant non-linear interactions between time,  
10 baseline pupil size, presentation, and SNR (all  $p$ 's  $< .001$ ; see Table 2 and Appendix A for model  
11 summaries). Figure A in the appendix is provided to show the model estimated TEPR as a function  
12 of time (on the x-axis) and baseline pupil size (on the y-axis), with color representing the value of  
13 the TEPR (on the z-axis) at that time/baseline combination. In other words, the contour plots  
14 represent estimated wiggly two-dimensional surfaces such that taking a horizontal slice at a given  
15 baseline pupil size value would result in a one-dimensional smooth showing the estimated TEPR

1 across time at that value of baseline pupil size. Density plots to the left of each contour plot show  
2 the distribution of baseline pupil sizes (e.g., trials) values for each presentation/SNR combination.

3 Figure 3 illustrates the effect of SNR as a function of baseline pupil size for each  
4 presentation (note that the panels are ordered by column/top-to-bottom rather than by row/left-to-  
5 right for Figures 3, 4, and 5). The left-most column in Figure 3 (panels a through c) shows the  
6 model-estimated differences between the  $-6$  dB and  $0$  dB SNR conditions as a function of time  
7 (on the x-axis) and baseline pupil size (on the y-axis), with color representing the estimated  
8 difference in the values of the TEPR at that time/baseline combination—that is, as if the wiggly  
9 two-dimensional surface for the  $0$  dB SNR condition had been subtracted from that of the  $-6$  dB  
10 SNR condition. Highlighted regions indicate significant differences between the two SNR  
11 conditions. In addition, the three remaining columns (panels d through l) present horizontal slices  
12 at the low ( $1^{\text{st}}$  quartile), median, and high ( $3^{\text{rd}}$  quartile) baseline pupil sizes for the  $0$  dB and  $-6$  dB  
13 SNR conditions, represented as purple, pink, and orange lines, respectively. Given that baseline  
14 pupil size was found to significantly differ between the first and third and second and third  
15 presentations, these quartiles were calculated for each presentation separately. These slices were  
16 chosen simply to aid in the visualization of the contour plots; baseline pupil size was treated as  
17 continuous in all models and not as quartiles. Panels d through l thus show the estimated TEPRs  
18 across time at these specific baseline pupil size values. The solid lines represent the  $0$  dB SNR  
19 condition while the dashed lines represent the  $-6$  dB SNR condition. The colored horizontal bars  
20 along the x-axis show time windows of significant difference between the two conditions, with  
21 green indicating a positive difference ( $-6$  dB  $>$   $0$  dB) and blue indicating a negative difference ( $-6$   
22 dB  $<$   $0$  dB). Shaded regions around the fitted smooths indicate 95% confidence intervals. Lastly,  
23 density plots show the distribution of baseline pupil size values (e.g., trials) for each presentation

1 collapsed across SNR. Figures 4 and 5 follow this same pattern; however, instead of showing  
 2 differences between the two SNRs at each presentation, Figure 4 shows the presentation-wise  
 3 differences for the 0 dB SNR condition, and Figure 5 shows the presentation-wise differences for  
 4 the -6 dB SNR condition.

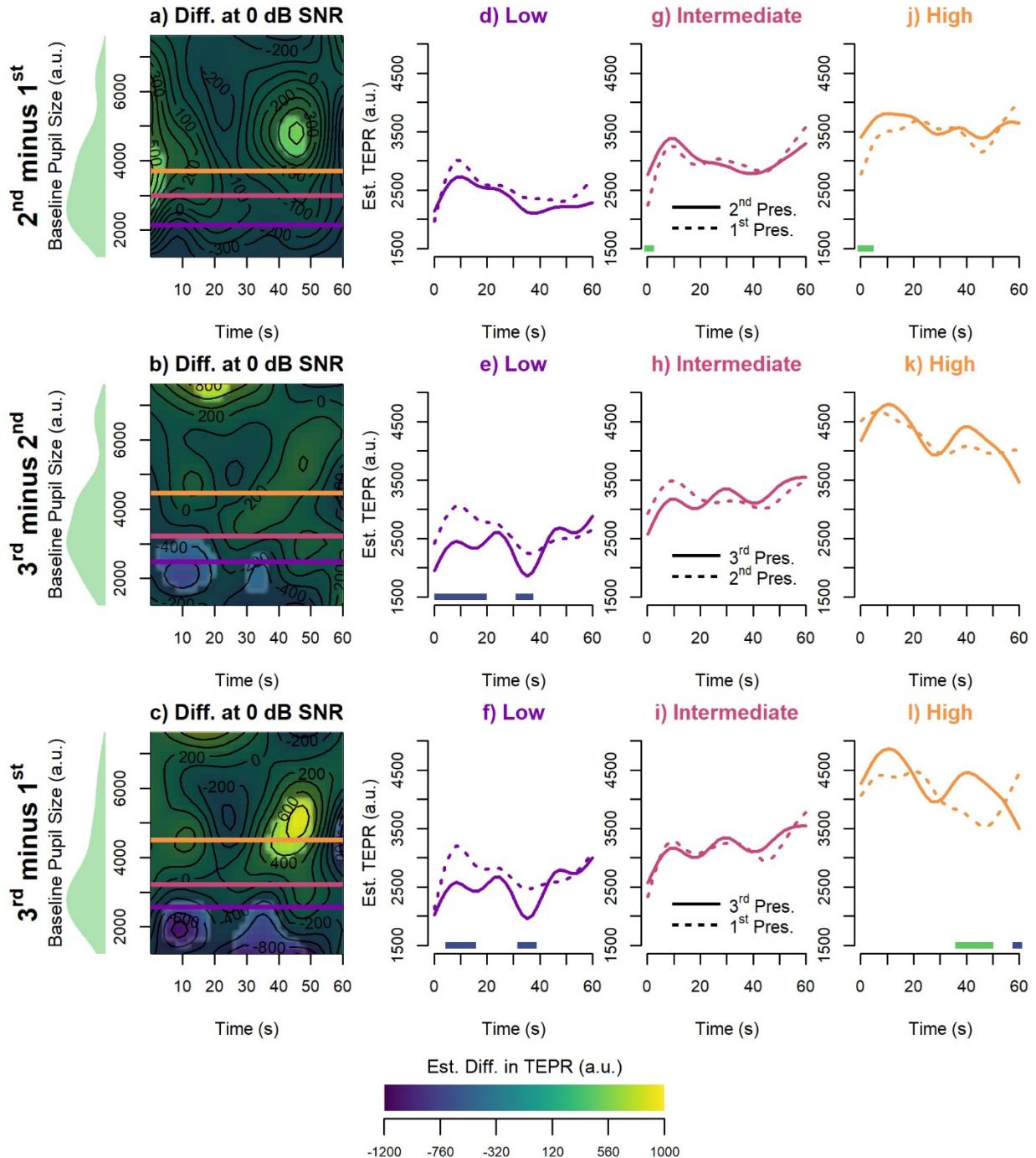


1 *Figure 3. Comparisons between the 0 dB and –6 dB SNR conditions showing the estimated difference in evoked*  
2 *pupil size (z-axis) by time (x-axis) and baseline pupil size (y-axis). Highlighted regions indicate regions of*  
3 *significant difference between the two presentations. Horizontal lines represent the low (1<sup>st</sup> quartile, purple line),*  
4 *median (pink line), and high (3<sup>rd</sup> quartile, orange line) baseline pupil size values. Fitted smooths for 0 dB (solid*  
5 *line) and –6 dB (dashed line) SNR are displayed at low, median, and high baseline pupil size values. Time periods*  
6 *of significant difference are marked by the green (positive difference) and blue (negative difference) bars at the*  
7 *bottom of the plot. An interactive version of this figure is available online at [https://michael-johns.shinyapps.io/ynh\\_pupil\\_slideshow/](https://michael-johns.shinyapps.io/ynh_pupil_slideshow/).*  
8

9 As can be seen in Figure 3 panels d through f, the –6 dB SNR condition elicited larger  
10 TEPRs than the 0 dB condition primarily for lower baseline pupil size values. This difference  
11 occurred during the approximately middle third of the passage during the first and third  
12 presentation but extends from approximately 20 s until the end of the passage during the second  
13 presentation. At intermediate baseline pupil size values, such differences between the two SNR  
14 conditions are absent during the first presentation and are relatively small and short-lived in the  
15 second and third presentations. Lastly, at higher baseline pupil size values, there is evidence that  
16 the 0 dB SNR condition elicits significantly larger TEPRs than the –6 dB SNR condition at various  
17 points throughout the passage. During the first presentation, this difference was present only in the  
18 last ~10 s of the passage. During the second presentation, however, this difference strengthened  
19 and extended for nearly the entire duration of the passage, with larger differences occurring  
20 towards the beginning of the passage and ultimately disappearing in the final ~10 s of the passage.  
21 Lastly, during the third presentation, a similar effect could be seen but was instead limited almost  
22 entirely to the first half of the passage.

23 To clarify the nature of the interactions depicted in Figure 3, Figures 4 and 5 provide an  
24 alternative visualization of these results, but instead displaying presentation-wise comparisons for  
25 the 0 dB and –6 dB SNR conditions, respectively. As in Figure 3, the left-most column presents  
26 heatmaps of the presentation-wise differences as a function of time and baseline pupil size, while  
27 the three remaining columns show fitted smooths for the two compared presentations at low (1<sup>st</sup>

1 quartile), median, and high (3<sup>rd</sup> quartile) baseline pupil size values, represented by the purple, pink,  
2 and orange lines, respectively. In the 0 dB SNR condition (Figure 4), the heatmaps show that, at  
3 low baseline pupil size values, the TEPR is lower at the third presentation compared to the second  
4 and first presentation (panels d, e, and f). In the -6 dB SNR condition (Figure 5), however, there  
5 are little-to-no differences between presentations at low baseline pupil size values (panels d, e, and  
6 f). This suggests that the effect of SNR seen for low baseline pupil sizes is a result of decreasing  
7 TEPRs for the 0 dB condition compared to relatively similar TEPRs for the -6 dB condition.



1  
2 **Figure 4. Additional visualization of the interaction presented in Figure 3 of presentation-wise estimated**  
3 **differences in evoked pupil size (z-axis) by time (x-axis) and baseline pupil size (y-axis) for the 0 dB SNR condition.**  
4 **Highlighted regions indicate regions of significant difference between the two presentations (as calculated from**  
5 **the re-referenced model presented in Table 2). Horizontal lines represent the low (1<sup>st</sup> quartile, purple line), median**  
6 **(pink line), and high (3<sup>rd</sup> quartile, orange line) baseline pupil size values. Fitted smooths for the two compared**  
7 **presentations are displayed at low, median, and high baseline pupil size values. Time periods of significant**  
8 **difference are marked by the green (positive difference) and blue (negative difference) bars at the bottom of the**  
9 **plot.**

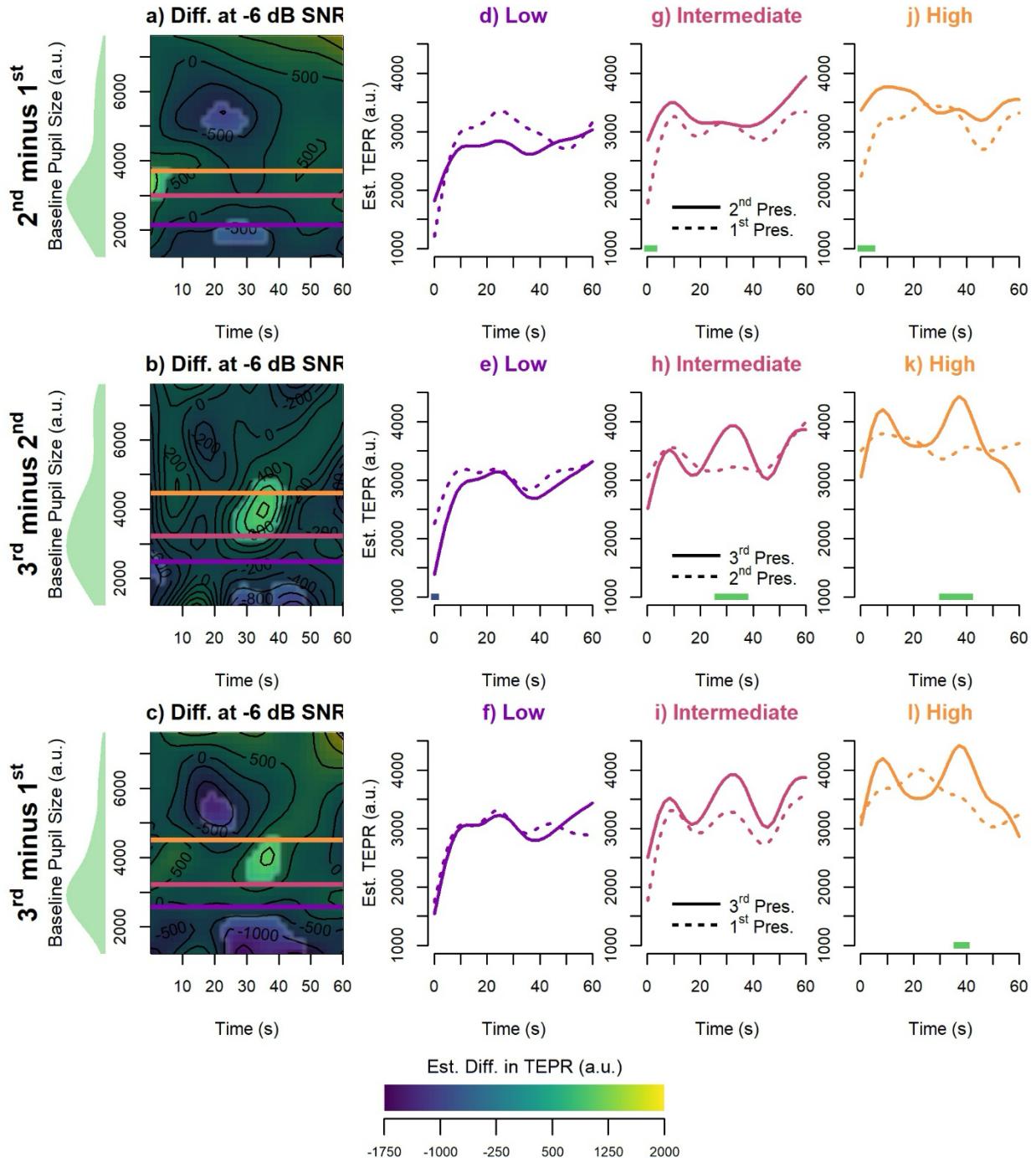


Figure 5. Additional visualization of the interaction presented in Figure 3 of presentation-wise estimated differences in pupil size (z-axis) by time (x-axis) and baseline pupil size (y-axis) for the -6 dB SNR condition. Highlighted regions indicate regions of significant difference between the two presentations (as calculated from the re-referenced model presented in Table 2). Horizontal lines represent the low (1<sup>st</sup> quartile, purple line), median (pink line), and high (3<sup>rd</sup> quartile, orange line) baseline pupil size values. Fitted smooths for the two compared presentations are displayed at low, median, and high baseline pupil size values. Time periods of significant difference are marked by the green (positive difference) and blue (negative difference) bars at the bottom of the plot.

1    **Discussion**

2    *RQ1) How is attention mobilization modulated by task difficulty?*

3            This study revealed that pre-stimulus baseline pupil size varied with stimulus repetition  
4    and impacted the TEPR measure of sustained listening effort across 60-s story listening in noise.  
5    With respect to our first research question (RQ1), we observed that pre-stimulus baseline pupil  
6    size significantly increased from the first to the second presentation and remained elevated for the  
7    third presentation but did not vary by SNR. The fact that the baseline pupil size increased in  
8    preparation for the second presentation suggests that listeners increased attention mobilization in  
9    anticipation of the subsequent repetitions, and maintained this level of mobilization until a new  
10   passage began. As such, the predictions of our first hypothesis (H1) only partially played out.

11    *RQ2) How is listening effort allocation modulated by task difficulty?*

12           With respect to our second research question (RQ2), baseline pupil size was observed to  
13   modulate not only the shape of the TEPR but also the effect of both SNR and repetition on the  
14   TEPR. However, the effects of SNR and repetition were not consistent with the predictions of our  
15   second hypothesis (H2), and instead a more complex interaction unfolded. In what follows, we  
16   discuss this interaction between baseline pupil size, SNR, and repetition on the TEPR to explore  
17   how these changes in attention mobilization affect the deployment of listening effort allocation  
18   over time (RQ3, H3).

19    *RQ3) How does attention mobilization interact with listening effort allocation?*

20           At lower baseline pupil sizes values—thought to be indicative of inattentiveness or under-  
21   mobilization of attentional resources (Hopstaken et al., 2015)—listening effort remained elevated  
22   in the harder -6 dB SNR condition compared to the 0 dB SNR conditions, even for the second and

1 third stimulus presentations. For all three presentations, the  $-6$  dB SNR condition elicited larger  
2 TEPRs than the  $0$  dB SNR condition, with the largest and most sustained difference between the  
3 two conditions occurring during the second presentation. This finding was observed despite the  
4 potential benefits of repetition, such as easier lexical access, which may have otherwise led to a  
5 gradual decrease in the SNR effect with each presentation (e.g., Calloway & Perfetti, 2020; Yang  
6 et al., 2007; Papesh & Goldinger, 2012; Marois et al., 2018). In other words, when attention  
7 mobilization remained low—even when the participant could have anticipated what the upcoming  
8 difficulty of the passage would be—the effect of SNR on listening effort allocation persisted in  
9 spite of the facilitative effects of repetition (H3).

10 At intermediate baseline pupil size values, there was evidence that listeners may have  
11 begun to mobilize their attention more optimally in both SNR conditions. Overall, differences  
12 between the two conditions were largely reduced, rather than exaggerated as originally predicted  
13 (H2, H3). While small time windows of significant difference are present for the second and third  
14 presentations (Figure 3, panels h and i), it is important to note that this occurs at these specific  
15 values of baseline pupil size. Overall, when examining the heatmaps (Figure 3, panels b and c),  
16 these differences largely disappeared for baseline pupil size values between approximately 3000  
17 and 4000 a.u..

18 At higher baseline pupil sizes, attention is thought to have been over-mobilized, resulting  
19 in a hypertonic state where listeners were more distractible and disengaged from the task  
20 (Hopstaken et al., 2015). In such a disengaged state, during the first presentation of a passage,  
21 differences between the two SNR conditions on the TEPR were largely absent. On average (i.e.,  
22 irrespective of time), the TEPR for both conditions was elevated, evidenced by the general effect  
23 that increasing baseline pupil size resulted in a higher mean TEPR (Figure 2). During the second

1 presentation (when listeners now had knowledge of upcoming listening difficulty), however, the  
2  $-6$  dB SNR condition elicited a significantly *smaller* TEPR compared to the 0 dB SNR condition  
3 for the majority of the passage—that is, the opposite of what was originally predicted (H2, H3).  
4 While this observation may suggest that, in this disengaged state, listeners had ‘given up’ (e.g.,  
5 Relaño-Iborra et al., 2022, p. 12), the behavioral responses to the comprehension questions do not  
6 fully support this interpretation – average accuracy to the comprehension questions was 69.2% ( $sd$   
7 = 46.3%) and did not significantly differ between the two SNR conditions or by presentation.

8 Rather, the observed smaller TEPR in the  $-6$  versus the 0 dB SNR condition following the  
9 first presentation may suggest that listeners engaged the least amount of effort required to perform  
10 the task (i.e., the principle of least effort; Ayasse et al., 2021) especially in the more aversive  
11 listening condition. Because each passage was repeated three times in a row, participants could  
12 have extracted enough information during the first presentation (and/or second) to be able to also  
13 answer the subsequent comprehension question (second or third presentation). Questions were  
14 designed to ensure some attention to the materials (Chapman & Hallowell, 2021), but not to be  
15 very difficult. The Model of Listening Engagement (MoLE; Herrmann & Johnsruude, 2020) notes  
16 that relative listening disengagement can occur when active participation is not required, “[e]ven  
17 when speech comprehension is easy, … for example, when listening to a long, tedious monologue”  
18 (p. 5, caption Fig. 1B) which is arguably the case in the current task. When a listener is in an over-  
19 mobilized state of attention (higher baseline pupil size), there may be little utility in exerting  
20 additional task-related effort (Eckert et al., 2016) to obtain more than a “good-enough” lexico-  
21 syntactic representation of the passage (e.g., Ferreira & Patson, 2007). Especially in the  $-6$  dB  
22 SNR condition, it may actually be aversive or, minimally, cause displeasure to sustain a deeper  
23 level of attention than necessary (Matthen, 2016).

1       Lastly, at higher baseline values during the third stimulus presentation, the results revealed  
2       that the SNR difference in the TEPR was reduced both in magnitude and in duration, localized  
3       primarily to the first half of the passage. The observation that this difference is diminished in the  
4       latter half of the passage suggests that, even in this over-mobilized state, listeners were able to re-  
5       engage and allocate more of their listening effort. One reason for this may have been that—similar  
6       to what was discussed previously for lower baseline pupil size values—a combination of the  
7       anticipation of the upcoming difficulty and the added benefit of an additional repetition led to a  
8       facilitative effect, potentially reducing the aversiveness of the  $-6$  dB SNR condition and thus  
9       reducing the differences between the two SNR conditions, even in a hypertonic state (H3). Future  
10      research to support this interpretation may benefit from manipulations of the depth of processing  
11      of the passage materials, such as with comprehension questions that require more integrative  
12      processing.

13      *Implications for theories and analyses of listening effort*

14       In line with the Framework for Understanding Effortful Listening (FUEL; Pichora-Fuller  
15      et al., 2016), the present study highlighted the importance of considering both the input-related  
16      external factors (i.e., SNR and stimulus repetition) as well as (internal) arousal state in  
17      understanding effortful listening. Particularly in cases where listeners have some knowledge about  
18      upcoming listening challenges (e.g., before entering a crowded room, listening with hearing loss),  
19      this work suggests it is critical to assess the extent to which listeners mobilize their attention to  
20      contextualize measures of listening effort.

21       From an analytical perspective, this work also highlights that the baseline epoch can  
22      contain critical information—not just a bias or noise to be subtracted or normalized out—when  
23      trying to understand the time course of effortful listening across different conditions. Although

1 exact stimulus repetition is not a frequent occurrence in real-world listening, attention mobilization  
2 comes into play in a variety of scenarios. Listeners develop expectations about upcoming listening  
3 challenges based on their knowledge of the probabilistic properties of English (Papesh &  
4 Goldinger, 2012), the ease of listening to familiar voices (Papesh, Goldinger, & Hout, 2012), cues  
5 about upcoming acoustic conditions (e.g., noise that is informative of an upcoming SNR; Seropian  
6 et al., 2022), and experience with hearing loss that leads them to expect difficulty in most  
7 conversations (Ayasse & Wingfield, 2020). Furthermore, aligned with previous results (Knapen et  
8 al., 2016; Relaño-Iborra et al., 2022), baseline pupil size was observed to affect the shape (not just  
9 the height) of the pupil response across time. Thus, performing baseline correction on the TEPR  
10 without first examining the impact of the listening condition of interest on the pre-stimulus pupil  
11 size has the potential to minimize, eliminate, or potentially artifactually reverse expected effects  
12 of listening demands on the TEPR.

13 The current study is novel in its examination of the trial-level pupil response to an extended  
14 passage of connected speech at varying SNRs. Previous studies have largely focused on examining  
15 listening effort in response to single words (e.g., Kuchinsky et al., 2013), sentences (e.g., Zekveld  
16 et al., 2010), or tone streams (Zhao et al., 2019). Some recent work on auditory decoding has  
17 examined longer stretches of speech similar to the present study, but focused on measures of effort  
18 that were predictive of attention switching between speakers (Haro et al., 2022) rather than effort  
19 associated with sustained attention to a single speaker. Studies that have examined listening to ~30  
20 second stories-in-babble in adults with hearing loss have found effects of SNR (Seifi Ala et al.,  
21 2020) and an SNR-by-noise-reduction interaction (Fiedler et al., 2021) on mean pupil dilation, but  
22 did not observe changes in these effects across time or as a function of baseline states of attention.

1 This study is also novel in its examination of the effect of baseline pupil size on the  
2 temporal dynamics of the TEPR. For example, McGarrigle et al. (2017) observed that pupil size  
3 was more sustained while listening to 12 seconds of speech at an easier (vs. harder) SNR, with the  
4 effect emerging around 9 seconds after onset, but only for the second block of the experiment.  
5 However, they concluded that baseline pupil size did not drive their TEPR effects because the  
6 baseline was not affected by SNR or block number. However, they did not investigate potential  
7 effects of the baseline on the shape of the TEPR across time, which the current study observed  
8 greatly modulate the observability and onset of SNR effects. Thus, to our knowledge, the current  
9 study represents a novel investigation of story listening of this length in younger adults with  
10 normal-hearing thresholds to better understand the relationship between attention mobilization and  
11 how effort unfolds throughout individual sustained listening trials (cf. Haro et al.'s [2022]  
12 examination of pupil dilation to predict attention switches).

13 The findings of the present study build upon prior research examining the relationship  
14 between baseline pupil size and the shape of the TEPR. We demonstrated similar findings to those  
15 of Relaño-Iborra et al. (2022) despite a few key differences. For example, Relaño-Iborra et al.  
16 (2022) found that baseline pupil size generally decreased as the task progressed. This is in contrast  
17 to the present study, where subsequent presentations of the same passage led to an increase in  
18 baseline pupil size. This discrepancy may largely be due to the design of the tasks: Relaño-Iborra  
19 et al. (2022) examined isolated, non-repeated sentences. As such, the decrease in baseline pupil  
20 size across the task may reflect aspects of fatigue or habituation (e.g., gradual overall  
21 disengagement from the task). Nonetheless, the authors also found that baseline pupil size  
22 increased with task difficulty, suggestive of increased preparatory control. This is in line with the  
23 present study: when participants can anticipate the difficulty of the upcoming stimulus (by virtue

1 of already having heard it once), they mobilize or up-regulate their attention in preparation.  
2 Similarly, Micula et al. (2021) found that baseline pupil size increased when task difficulty became  
3 more unpredictable. At first glance, this too seems to contradict the findings of the present study;  
4 however, as Micula et al. (2021, p. 1676) suggest, this increase may not be driven by predictability  
5 per se, but rather by participants' increasing alertness or engagement in response to the more  
6 difficult, unpredictable task. Ultimately, Relaño-Iborra et al. (2022), Micula et al. (2021), and the  
7 present study all demonstrate the importance of examining baseline pupil size, its relationship to  
8 performance, and its effects on the shape of the TEPR as a measure of listening effort deployment  
9 across varying listening conditions. Whether listeners can anticipate the difficulty of the upcoming  
10 stimulus and can thus determine whether they should mobilize additional resources, or if the task  
11 becomes unpredictable and requires listeners to be more alert and attentive, baseline pupil size  
12 seems to serve as an informative index of how much listeners mobilize or prepare their attentional  
13 resources during adverse listening conditions.

#### 14 *Limitations and Future Directions*

15 One limitation of this study relates to the interpretation of the TEPR: intuitively, it is  
16 expected that the more effort a task requires—and thus, the more attention that must be allocated  
17 —the larger the TEPR will be. In the present study, however, there were conditions under which  
18 the harder –6 dB SNR condition elicited *smaller* rather than larger TEPRs. We interpreted this  
19 somewhat unintuitive finding in the context of the principle of least effort (Ayasse et al., 2021).  
20 That is, participants may have had a good-enough (Ferreira & Patson, 2007) understanding of the  
21 passage by the second and/or third presentation, such that they only engaged a minimal amount of  
22 effort for the –6 dB SNR stimuli that were not enjoyable (Matthen et al., 2016) or motivating to  
23 process more deeply (Herrmann & Johnsrude, 2020). A limitation of the current study is that

1 subjective intelligibility was only assessed after the first presentation, but not the subsequent two  
2 presentations of the passage segment. In future studies, collecting presentation-level subjective  
3 intelligibility data might help to provide evidence for or against our interpretation: reduced TEPRs  
4 in the harder SNR condition correlating with lower ratings may be more indicative of giving up,  
5 while similar ratings compared to the easier SNR may be more indicative of good-enough  
6 understanding. Collecting measures of listening aversiveness or motivation, or including  
7 comprehension questions that require greater depth of story processing may provide related  
8 insights into our interpretation.

9 Another limitation of the current study is that the distribution of baseline pupil size values  
10 may not represent the full range from absolute under- to over-mobilization, and indeed this may  
11 vary on a person-by-person and day-to-day basis. For example, some individuals during the current  
12 study may have ranged only from more to less under-mobilized (i.e., they would fall on the left  
13 side of the Yerkes-Dodson curve) while others may have ranged only from more to less over-  
14 mobilized (i.e., on the right side of the Yerkes-Dodson curve). To somewhat limit potential  
15 extreme individual differences in the range of tonic arousal, inclusion criteria required that  
16 participants reported no psychiatric or neurological conditions and were not taking psychoactive  
17 stimulants or depressants. Participants were also allowed to select the time of day they preferred  
18 for testing. However, without some way of gauging an individual's attentional state (both generally  
19 in their daily lives and at that particular time of testing) or referencing their baseline pupil size  
20 values to some known range, it is difficult to ascertain what 'low' and 'high' baseline pupil sizes  
21 values actually reflect. In the present study, we opted for the 1<sup>st</sup> quartile, median, and 3<sup>rd</sup> quartile  
22 (between participants) as reference points for visually examining the effects of baseline pupil size  
23 on the TEPR, although this was modelled continuously, in order to capture where the majority of

1 the data lie. This group-level way of analyzing the data may not adequately reflect individual  
2 differences. In this vein, the limited range of SNRs may also have contributed a more limited  
3 distribution of baseline pupil sizes, as compared to prior studies that sought to capture the full  
4 psychometric function (Relaño-Iborra et al., 2022; Wendt et al., 2018).

5 A minimum of 69 seconds elapsed between one passage onset and the next passage's  
6 baseline epoch. Especially in future studies in which it is not feasible to include such a long time  
7 for the pupil to return to its physiological baseline, it may be more critical to examine the relative  
8 contribution of physiological carry-over of the pupil response (Winn et al., 2018) versus attention  
9 mobilization in anticipation of difficult listening on baseline pupil size. One way to do this could  
10 be to also include blocks in which passage difficulty is not predictable as a control (i.e., SNR  
11 and/or exact excerpts are not repeated). Future neuroimaging studies may also provide insight into  
12 our contention that any sustainment of pupil size between trials would instead be driven by the  
13 continued upregulation of performance monitoring and/or cognitive control processes to support  
14 subsequent task processing (e.g., Hsu, Kuchinsky, & Novick, 2020; Vaden et al., 2013).  
15 Regardless of the extent to which the baseline represents signal or noise, the current study  
16 highlights the importance of explicitly examining its impact on the TEPR.

17 The current study demonstrated that the anticipated difficulty of a sustained listening task  
18 modulated not only the extent to which listeners mobilized their attention in advance of listening,  
19 but also the deployment of listening effort throughout the task. Extending previous studies that  
20 have predominantly focused on single words and sentences, often presented in isolation and  
21 without context, the present experiment examined changes in effort throughout 60-s-long  
22 audiobook passages in the presence of a competing talker. Two SNRs were examined. The results  
23 suggested that when listeners had not adequately prepared for the upcoming difficulty of the trial

1 (e.g., they did not know what was next or did not sufficiently mobilize their attention), the TEPR  
2 was sensitive to differences in SNR. However, SNR effects were not observed at intermediate  
3 baseline pupil sizes, suggesting that listeners had optimally readied their attention for the  
4 upcoming task demands. At higher baseline pupil sizes, in which listeners may have over-  
5 mobilized their attention or may have been in a more distractible state, the effect of SNR was  
6 reversed. In the first half of the passage, these potentially overwhelmed listeners showed a reduced  
7 TEPR for the harder SNR condition that gradually recovered in the second half. Ultimately,  
8 however, listeners in this over-mobilized state showed reduced TEPRs to both SNR conditions by  
9 the third and final presentation, suggesting a reduction effort allocation for both SNRs. Together,  
10 these findings suggest that the time course of listening effort depends not only on how difficult the  
11 listening situation is, but also on the extent to which individuals are able to anticipate and prepare  
12 for those upcoming challenges. Future work aims to examine how these relationships change with  
13 aging and hearing loss, as these individuals in these populations may be predisposed to anticipating  
14 such difficulties with listening in their daily lives.

15

## 16 **Acknowledgements**

17 We thank Jason Dunlap and Janani Perera for assistance with data collection and Dr. Ed Smith for  
18 audio engineering support. We are grateful to Dr. Martijn Wieling for his consultation on our  
19 implementation and interpretation of GAMMs. This work was supported by the National Institute  
20 on Aging (NIA) grant P01-AG055365, the National Institute on Deafness and Other  
21 Communication Disorders (NIDCD) grant R01-DC019394 and training grant DC-00046 (to  
22 RCC), and the National Science Foundation (NSF) grant SMA-1734892 (to JZS). For Dr.  
23 Kuchinsky: The identification of specific products or scientific instrumentation is considered an

1 integral part of the scientific endeavor and does not constitute endorsement or implied endorsement  
2 on the part of the authors, DoD, or any component agency. The views expressed in this article are  
3 those of the authors and do not necessarily reflect the official policy of the Department of Defense  
4 or the U.S. Government.

5

1    **References**

2    Alain, C., Du, Y., Bernstein, L. J., Barten, T., & Banai, K. (2018). Listening under difficult  
3    conditions: An activation likelihood estimation meta-analysis. *Human Brain Mapping*,  
4    39(7), 2695–2709. Doi: 10.1002/hbm.24031

5    Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine  
6    function: Adaptive gain and optimal performance. *Annual Review of Neuroscience*, 28(1),  
7    403–450. Doi: 10.1146/annurev.neuro.28.061604.135709

8    Ayasse, N. D., & Wingfield, A. (2020). Anticipatory Baseline Pupil Diameter Is Sensitive to  
9    Differences in Hearing Thresholds. *Frontiers in Psychology*, 10. Doi:  
10    10.3389/fpsyg.2019.02947

11    Ayasse, N. D., Hodson, A. J., & Wingfield, A. (2021). The Principle of Least Effort and  
12    Comprehension of Spoken Sentences by Younger and Older Adults. *Frontiers in  
13    Psychology*, 12. Doi: 10.3389/fpsyg.2021.629464

14    Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models  
15    Using lme4. *Journal of Statistical Software*, 67(1), 1-48. Doi: 10.18637/jss.v067.i01

16    Calloway, R. C., & Perfetti, C. A. (2020). Integration and structure building across a sentence  
17    boundary: ERP indicators of definite/indefinite article, noun repetition, and  
18    comprehension skill effects. *Language, Cognition and Neuroscience*, 35(1). Doi:  
19    10.1080/23273798.2019.1637911

20    Carroll, R., Warzybok, A., Kollmeier, B., & Ruigendijk, E. (2016). Age-related differences in  
21    lexical access relate to speech recognition in noise. *Frontiers in Psychology*, 7, 1–16.  
22    Doi: 10.3389/fpsyg.2016.00990

1 Chapman, L. R., & Hallowell, B. (2021). Expecting Questions Modulates Cognitive Effort in a  
2 Syntactic Processing Task: Evidence From Pupillometry. *Journal of Speech, Language,*  
3 *and Hearing Research*, 64(1), 121–133. Doi: 10.1044/2020\_JSLHR-20-00071

4 Christiansen, R. H. B. (2022). Ordinal – Regression Models for Ordinal Data. R package version  
5 2022.11-16. <https://CRAN.R-project.org/package=ordinal>

6 Chun, M. M., Golomb, J. D., & Turk-Browne, N. B. (2011). A Taxonomy of External and  
7 Internal Attention. *Annual Review of Psychology*, 62(1), 73–101. Doi:  
8 10.1146/annurev.psych.093008.100427

9 Eckert, M. A., Teubner-Rhodes, S., & Vaden, K. I. (2016). Is listening in noise worth it? The  
10 neurobiology of speech recognition in challenging listening conditions. *Ear and Hearing*,  
11 37(Supplement 1), 101S-110S. doi: 10.1097/AUD.0000000000000300

12 Elman, J. A., Panizzon, M. S., Hagler Jr., D. J., Eyler, L. T., Ganholm, E. L., Fennema-  
13 Notestine, C., Lyons, M. J., McEvoy, L. K., Franz, C. E., Dale, A. M., & Kremen, W. S.  
14 (2017). Task-evoked pupil dilation and BOLD variance as indicators of locus coeruleus  
15 dysfunction. *Cortex*, 97, 60-69. Doi: 10.1016/j.cortex.2017.09.025

16 Ferrari, V., De Cesarei, A., Mastria, S., Lugli, L., Baroni, G., Nicoletti, R., & Codispoti, M.  
17 (2016). Novelty and emotion: Pupillary and cortical responses during viewing of natural  
18 scenes. *Biological Psychology*, 113, 75-82. Doi: 10.1016/j.biopsych.2015.11.008

19

20 Ferreira, F., & Patson, N. D. (2007). The ‘Good Enough’ Approach to Language  
21 Comprehension. *Language and Linguistics Compass*, 1(1–2), 71–83. Doi:  
22 10.1111/j.1749-818X.2007.00007.x

1 Fiedler, L., Seifi Ala, T., Graversen, C., Alickovic, E., Lunner, T., & Wendt, D. (2021). Hearing  
2 Aid Noise Reduction Lowers the Sustained Listening Effort During Continuous Speech  
3 in Noise—A Combined Pupillometry and EEG Study. *Ear and Hearing*, 42(6), 1590.  
4 Doi: 10.1097/AUD.0000000000001050

5 Fortenbaugh, F. C., DeGutis, J., & Esterman, M. (2017). Recent theoretical, neural, and clinical  
6 advances in sustained attention research. *Annals of the New York Academy of Sciences*,  
7 1396(1), 70–91. Doi:10.1111/nyas.13318Franklin, M. S., Broadway, J. M., Mrazek, M.  
8 D., Smallwood, J., & Schooler, J. W. (2013). Window to the wandering mind:  
9 Pupillometry of spontaneous thought while reading. *Quarterly Journal of Experimental  
10 Psychology*, 66(12), 2289–2294. Doi: 10.1080/17470218.2013.858170

11 Franzen, M. D., & Wilhelm, K. L. (1996). Conceptual foundations of ecological validity in  
12 neuropsychological assessment. In R. J. Sbordone & C. J. Long (Eds.), *Ecological  
13 validity of neuropsychological testing* (pp. 91–112). Gr Press/St Lucie Press, Inc.

14 Gagl, B., Hawelka, S., & Hutzler, F. (2011). Systematic influence of gaze position on pupil size  
15 measurement: analysis and correction. *Behavior Research Methods*, 43, 1171–1181. Doi:  
16 10.3758/s13428-011-0109-5

17 Gilzenrat, M. S., Nieuwenhuis, S., Jepma, M., & Cohen, J. D. (2010). Pupil diameter tracks  
18 changes in control state predicted by the adaptive gain theory of locus coeruleus function.  
19 *Cognitive, Affective and Behavioral Neuroscience*, 10(2), 252–269. Doi:  
20 10.3758/CABN.10.2.252

21 Haro, S., Rao, H. M., Quatieri, T. F., & Smalt, C. J. (2022). EEG alpha and pupil diameter reflect  
22 endogenous auditory attention switching and listening effort. *European Journal of  
23 Neuroscience*, 55(5), 1262–1277. Doi: 10.1111/ejn.15616

1 Herrmann, B., & Johnsrude, I. S. (2020). A model of listening engagement (MoLE). *Hearing*  
2 *Research*, 397, 108016. Doi: /10.1016/j.heares.2020.108016

3 Hopstaken, J. F., van der Linden, D., Bakker, A. B., & Kompier, M. A. J. (2015). A multifaceted  
4 investigation of the link between mental fatigue and task disengagement.  
5 *Psychophysiology*, 52(3), 305–315. Doi: 10.1111/psyp.12339

6 Hsu, N. S., Kuchinsky, S. E., & Novick, J. M. (2020). Direct impact of cognitive control on  
7 sentence processing and comprehension. *Language, Cognition and Neuroscience*, 36(2),  
8 211-239. Doi: 10.1080/23273798.2020.1836379

9 Irving, W. (2006) *The Legend of Sleepy Hollow*. (Chip, Narr.) [Audiobook]. LibriVox.  
10 <https://librivox.org/the-legend-of-sleepy-hollow-by-washington-irving> (Original work  
11 published 1820)

12 Irving, W. (1977) *The Legend of Sleepy Hollow*. (C. Hardin Killavey, Narr.) [Audiobook].  
13 Audible. <https://www.audible.com/pd/The-Legend-of-Sleepy-Hollow-Audiobook/>  
14 (Original work published 1820)

15 Jepma, M., & Nieuwenhuis, S. (2011). Pupil Diameter Predicts Changes in the Exploration-  
16 Exploitation Trade-off: Evidence for the Adaptive Gain Theory. *Journal of Cognitive  
17 Neuroscience*, 23(7), 1587-1596. Doi: 10.1162/jocn.2010.21548

18 Joshi, S., Li, Y., Kalwani, R. M., & Gold, J. I. (2016). Relationships between Pupil Diameter and  
19 Neuronal Activity in the Locus Coeruleus, Colliculi, and Cingulate Cortex. *Neuron*,  
20 89(1), 221-234. Doi: 10.1016/j.neuron.2015.11.028

21 Kahneman, D. (1973). *Attention and Effort*. Prentice-Hall.

22 Kane, G. A., Vazey, E. M., Wilson, R. C., Shenhav, A., Daw, N. D., Aston-Jones, G., & Cohen,  
23 J. D. (2017). Increased locus coeruleus tonic activity causes disengagement from a patch-

1 foraging task. *Cognitive, Affective and Behavioral Neuroscience*, 17(6), 1073–1083. Doi:  
2 10.3758/s13415-017-0531-y

3 Karunathilake, I. M. D., Dunlap, J. L., Perera, J., Presacco, A., Decruy, L., Anderson, S.,  
4 Kuchinsky, S. E., & Simon, J. Z. (2023). Effects of Aging on Cortical Representations of  
5 Continuous Speech. *Journal of Neurophysiology*, 129(6), 1359-1377. Doi:  
6 10.1152/jn.00356.2022

7 Keidser, G., Naylor, G., Brungart, D. S., Caduff, A., Campos, J., Carlile, S., Carpenter, M. G.,  
8 Grimm, G., Hohmann, V., Holube, I., Launer, S., Lunner, T., Mehra, R., Rapport, F.,  
9 Slaney, M., & Smeds, K. (2020). The quest for ecological validity in hearing science:  
10 what it is, why it matters, and how to advance it. *Ear and Hearing*, 41, 5S-19S. doi:  
11 10.1097/AUD.0000000000000944

12 Kidd, G., Mason, C. R., & Best, V. (2014). The role of syntax in maintaining the integrity of  
13 streams of speech. *The Journal of the Acoustical Society of America*, 135(2), 766–777.  
14 Doi: 10.1121/1.4861354

15 Killion, M. C., Niquette, P. A., Gudmundsen, G. I., Revit, L. J., & Banerjee, S. (2004).  
16 Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in  
17 normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of  
18 America*, 116(4), 2395–2405. Doi: 10.1121/1.1784440

19 Knapen, T., de Gee, J. W., Brascamp, J., Nuitjen, S., Hoppenbrouwers, S., & Theeuwes, J.  
20 (2016). Cognitive and Ocular Factors Jointly Determine Pupil Responses under  
21 Equiluminance. *PLoS ONE*, 11(5): e0155574. Doi: 10.1371/journal.pone.0155574

1 Koelewijn, T., de Kluiver, H., Shinn-Cunningham, B. G., Zekveld, A. A., & Kramer, S. E.  
2 (2015). The pupil response reveals increased listening effort when it is difficult to focus  
3 attention. *Hearing Research*, 323, 81-90. Doi: 10.1016/j.heares.2015.02.004

4 Kristjansson, S. D., Stern, J. A., Brown, T. B., & Rohrbaugh, J. W. (2009). Detecting phasic  
5 lapses in alertness using pupillometric measures. *Applied Ergonomics*, 40(6), 978–986.  
6 Doi: 10.1016/j.apergo.2009.04.007

7 Kuchinsky, S. E., Ahlstrom, J. B., Vaden, K. I., Cute, S. L., Humes, L. E., Dubno, J. R., &  
8 Eckert, M. A. (2013). Pupil size varies with word listening and response selection  
9 difficulty in older adults with hearing loss. *Psychophysiology*, 50(1), 23–34. Doi:  
10 10.1111/j.1469-8986.2012.01477.x

11 Kuznetsova, A., Brockhoff, P. B., & Christiansen, R. H. B. (2017). lmerTest Package: Tests in  
12 Linear Mixed Models. *Journal of Statistical Software*, 82(13), 1–26. Doi:  
13 10.18637/jss.v082.i13

14 Lenth, R. V. (2023). Emmeans: Estimated Marginal Means, a.k.a. Least-Squares Means. R  
15 package version 1.8.4-1. <https://cran.r-project.org/package=emmeans>

16 Marois, A., Labonté, K., Parent, M., & Vachon, F. (2018). Eyes have ears: Indexing the orienting  
17 response to sound using pupillometry. *International Journal of Psychophysiology*, 123,  
18 152-162. Doi: 10.1016/j.ijpsycho.2017.09.016

19 Martin, J. T., Whittaker, A. H., & Johnston, S. J. (2022). Pupillometry and the vigilance  
20 decrement: Task-evoked but not baseline pupil measures reflect declining performance in  
21 visual vigilance tasks. *European Journal of Neuroscience*, 55(3), 778-799. Doi:  
22 10.1111/ejn.15585

1 Mathôt, S., Fabius, J., Van Heusden, E., & Van der Stigchel, S. (2018). Safe and sensible  
2 preprocessing and baseline correction of pupil-size data. *Behavior Research Methods*,  
3 50(1), 94–106. Doi: 10.3758/s13428-017-1007-2

4 [Matthen, M. \(2016\). Effort and Displeasure in People Who Are Hard of Hearing. Ear &](#)  
5 [Hearing, 37\(1\), 28S-34S. doi: 10.1097/AUD.0000000000000292](#)

6

7 Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in  
8 adverse conditions: A review. *Language and Cognitive Processes*, 27(7-8), 953-978. Doi:  
9 10.1080/01690965.2012.705006

10 McCoy, S. L., Tun, P A., Cox. L. C., Colangelo, M., Stewart, R. A., & Wingfield, A. (2005).  
11 Hearing loss and perceptual effort: Downstream effects on older adults' memory for  
12 speech. *The Quarterly Journal of Experimental Psychology Section A.*, 58(1), 22-33. Doi:  
13 10.1080/02724980443000151

14 McGarrigle, R., Dawes, P., Stewart, A. J., Kuchinsky, S. E., & Munro, K. J. (2017).  
15 Pupillometry reveals changes in physiological arousal during a sustained listening task.  
16 *Psychophysiology*, 54(2), 193–203. Doi: 10.1111/psyp.12772

17 McGinley, M. J., David, S. v., & McCormick, D. A. (2015). Cortical membrane potential  
18 signature of optimal states for sensory signal detection. *Neuron*, 87(1), 179–192. Doi:  
19 10.1016/j.neuron.2015.05.038

20 Micula, A., Rönnberg, J., Fiedler, L., Wendt, D., Jørgensen, M. C., Larsen, D. K., & Ng, E. H. N.  
21 (2021). The Effects of Task Difficulty Predictability and Noise Reduction on Recall  
22 Performance and Pupil Dilation Responses. *Ear and Hearing*, 42(6), 1668–1679. Doi:  
23 10.1097/AUD.0000000000001053

1 Micula, A., Rönnberg, J., Książek, P., Nielsen, R. M., Wendt, D., Fiedler, L., & Ng, E. H. N.  
2 (2022). A Glimpse of Memory Through the Eyes: Pupillary Responses Measured During  
3 Encoding Reflect the Likelihood of Subsequent Memory Recall in an Auditory Free  
4 Recall Test. *Trends in Hearing*, 26, 1–12. Doi: 10.1177/23312165221130581

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

8 Murphy, P. R., Robertson, I. H., Balsters, J. H., & O'Connell, R. G. (2011). Pupillometry and P3  
9 index the locus coeruleus-noradrenergic arousal function in humans. *Psychophysiology*,  
10 48(11), 1532–1543. Doi: 10.1111/j.1469-8986.2011.01226.x

11 Ohlenforst, B., Wendt, D., Kramer, S. E., Naylor, G., Zekveld, A. A., & Lunner, T. (2018).  
12 Impact of SNR, masker type and noise reduction processing on sentence recognition  
13 performance and listening effort as indicated by the pupil dilation response. *Hearing  
14 Research*, 365, 90-99. Doi: 10.1016/j.heares.2018.05.003

15 Pandža, N. B., Phillips, I., Karuzis, V. P., O'Rourke, P., & Kuchinsky, S. E. (2020).  
16 Neurostimulation and Pupillometry: New Directions for Learning and Research in  
17 Applied Linguistics. *Annual Review of Applied Linguistics*, 40, 56-77. Doi:  
18 10.1017/S0267190520000069

19 Papesh, M. H., & Goldinger, S. D. (2012). Pupil-BLAH-metry: Cognitive effort in speech  
20 planning reflected by pupil dilation. *Attention, Perception & Psychophysics*, 74, 754-765.  
21 Doi: 10.3758/s13414-011-0263-y

1 Papesh, M. H., Goldinger, S. D., & Hout, M. C. (2012). Memory strength and specificity  
2 revealed by pupillometry. *International Journal of Psychophysiology*, 83(1), 56–64. doi:  
3 10.1016/j.ijpsycho.2011.10.002

4 Phillips, I., Calloway, R. C., Karuzis, V. P., Pandža, N. B., O'Rourke, P., & Kuchinsky, S. E.  
5 Transcutaneous Auricular Vagus Nerve Stimulation Strengthens Semantic  
6 Representations of Foreign Language Tone Words during Initial Stages of Learning.  
7 *Journal of Cognitive Neuroscience*, 34(1), 127-152. doi: 10.1162/jocn\_a\_01783

8 Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W. Y., Humes, L.  
9 E., Lemke, U., Lunner, T., Matthen, M., Mackersie, C. L., Naylor, G., Phillips, N. A.,  
10 Richter, M., Rudner, M., Sommers, M. S., Tremblay, K. L., & Wingfield, A. (2016).  
11 Hearing impairment and cognitive energy: The framework for understanding effortful  
12 listening (FUEL). *Ear and Hearing*, 37, 5S-27S. doi: 10.1097/AUD.0000000000000312

13 Porretta, V., Kyröläinen, A. J., van Rij, J., & Järvikivi, J. (2018). Visual world paradigm data:  
14 From preprocessing to nonlinear time-course analysis. In *Intelligent Decision  
15 Technologies 2017: Proceedings of the 9th KES International Conference on Intelligent  
16 Decision Technologies (KES-IDT 2017)–Part II 9* (pp. 268-277). Springer International  
17 Publishing.

18 Posner, M. I., & Peterson, S. E. (1990). The attention system of the human brain. *Annual Review  
19 of Neuroscience*, 13(1), 25–42.

20 Presacco, A., Simon, J. Z., & Anderson, S. B. (2016). Effect of informational content of noise on  
21 speech representation in the aging midbrain and cortex. *Journal of Neurophysiology*,  
22 116(5), 2356–2367. doi: 10.1152/jn.00373.2016

1 R Core Team (2024). R: A language and environment for statistical computing. R Foundation for  
2 Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

3 Rajkowski, J., Kubiak, P., & Aston-Jones, G. (1994). Locus coeruleus activity in monkey: Phasic  
4 and tonic changes are associated with altered vigilance. *Brain Research Bulletin*, 35(5-6),  
5 607-616. doi: 10.1016/0361-9230(94)90175-9

6 Reilly, J., Kelly, A., Kim, S. H., Jett, S., & Zuckerman, B. (2019). The human task-evoked  
7 pupillary response function is linear: Implications for baseline response scaling in  
8 pupillometry. *Behavior Research Methods*, 51(2), 865–878. doi: 10.3758/s13428-018-  
9 1134-4

10 Relaño-Iborra, H., Wendt, D., Neagu, M. B., Kressner, A. A., Dau, T., & Bækgaard, P. (2022).  
11 Baseline pupil size encodes task-related information and modulates the task-evoked  
12 response in a speech-in-noise task. *Trends in Hearing*, 26, 1–17. doi:  
13 10.1177/23312165221134003

14 Sarter, M., Givens, B., & Bruno, J. P. (2001). The cognitive neuroscience of sustained attention:  
15 where top-down meets bottom up. *Brain Research Reviews*, 35(2), 146–160. doi:  
16 10.1016/S0165-0173(01)00044-3

17 Seifi Ala, T., Graversen, C., Wendt, D., Alickovic, E., Whitme, W. M., Lunner, T., Seifi Ala, T.,  
18 Graversen, C., Wendt, D., Alickovic, E., Whitmer, W. M., & Lunner, T. (2020). An  
19 exploratory Study of EEG Alpha Oscillation and Pupil Dilation in Hearing-Aid Users  
20 During Effortful listening to Continuous Speech. *PLOS ONE*, 15(7), e0235782. doi:  
21 10.1371/journal.pone.0235782

22 Seropian, L., Ferschneider, M., Cholvy, F., Micheyl, C., Bidet-Caulet, A., & Moulin, A. (2022).  
23 Comparing methods of analysis in pupillometry: application to the assessment of

1 listening effort in hearing-impaired patients. *Helijon*, 8(6), e09631. doi:  
2 10.1016/j.heliyon.2022.e09631

3 Snyder, J. S., & Alain, C. (2007). Toward a neurophysiological theory of auditory stream  
4 segregation. *Psychological Bulletin*, 133(5), 780–799. doi: 10.1037/0033-2909.133.5.780

5 Sóskuthy, M. (2021). Evaluating generalised additive mixed modelling strategies for dynamic  
6 speech analysis. *Journal of Phonetics*, 84. doi: 10.1016/j.wocn.2020.101017

7 Sparks, J. R., & Rapp, D. N. (2010). Discourse processing--examining our everyday language  
8 experiences. *WIREs Cognitive Science*, 1(3), 371-381. doi: 10.1002/wcs.11

9 Tang, Y.-Y., Hölzer, B. K. & Posner, M. I. (2015). The neuroscience of mindfulness meditation.  
10 *Nature Reviews Neuroscience*, 16, 213–225. doi: 10.1038/nrn3916

11 Unsworth, N., & Robison, M. K. (2016). Pupillary correlates of lapses of sustained attention.  
12 *Cognitive, Affective and Behavioral Neuroscience*, 16(4), 601–615. doi: 10.3758/s13415-  
13 016-0417-4

14 Vaden, K. I., Kuchinsky, S. E., Cute, S. L., Ahlstrom, J. B., Dubno, J. R., & Eckert, M. A.  
15 (2013). The cingulo-opercular network provides word-recognition benefit. *The Journal of  
16 Neuroscience*, 33(48), 18979–18986. doi: 10.1523/JNEUROSCI.1417-13.2013

17 van Rij, J., Hendriks, P., van Rijn, H., Baayen, R. H., & Wood, S. N. (2019). Analyzing the Time  
18 Course of Pupillometric Data. *Trends in Hearing*, 23. doi: 10.1177/2331216519832483

19 van Rij, J., Wieling, M., Baayen, R., & van Rijn, H. (2022). Itsadug: Interpreting Time Series  
20 and Autocorrelated Data Using GAMMs. R package version 2.4.1. <https://cran.r-project.org/package=itsadug>

1 Wagner, A. E., Nagels, L., Toffanin, P., Opie, J. M., & Başkent, D. (2019). Individual Variations  
2 in Effort: Assessing Pupillometry for the Hearing Impaired. *Trends in Hearing*, 23. doi:  
3 10.1177/2331216519845596

4 Wendt, D., Koelewijn, T., Książek, P., Kramer, S. E., & Lunner, T. (2018). Toward a more  
5 comprehensive understanding of the impact of masker type and signal-to-noise ratio on  
6 the pupillary response while performing a speech-in-noise test. *Hearing Research*, 369,  
7 67–78. doi: 10.1016/j.heares.2018.05.006

8 Winn, M. B. (2016). Rapid release from listening effort resulting from semantic context, and  
9 effects of spectral degradation and cochlear implants. *Trends in Hearing*, 20. doi:  
10 10.1177/2331216516669723

11 Winn, M. B. (2023). Time Scales and Moments of Listening Effort Revealed in Pupillometry.  
12 *Seminars in Hearing*, 44(02), 106–123. doi: 10.1055/s-0043-1767741

13 Winn, M. B., & Moore, A. N. (2018). Pupillometry reveals that context benefits in speech  
14 perception can be disrupted by later-occurring sounds, especially in listeners with  
15 cochlear implants. *Trends in Hearing*, 22. doi: 10.1177/2331216518808962

16 Winn, M. B., & Teece, K. H. (2021). Listening effort is not the same as speech intelligibility  
17 score. *Trends in Hearing*, 25. doi: 10.1177/23312165211027688

18 Winn, M. B., & Teece, K. H. (2022). Effortful Listening Despite Correct Responses: The Cost of  
19 Mental Repair in Sentence Recognition by Listeners with Cochlear Implants. *Journal of  
20 Speech, Language, and Hearing Research*, 65(10), 3966–3980. doi:  
21 10.1044/2022\_JSLHR-21-00631

1 Winn, M. B., Wendt, D., Koelewijn, T., & Kuchinsky. S. E. (2018). Best Practices and Advice  
2 for Using Pupilometry to Measure Listening Effort: An Introduction for Those Who  
3 Want to Get Started. *Trends in Hearing*, 22. doi: 10.1177/2331216518800869

4 Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood  
5 estimation of semiparametric generalized linear models. *Journal of the Royal Statistical  
6 Society (B)*, 73(1), 3-36. doi: 10.1111/j.1467-9868.2010.00749.x

7 Wood, S. N. (2013). Thin-plate regression splines. *Journal of the Royal Statistical Society (B)*,  
8 65(1), 95-114. doi: 10.1111/1467-9868.00374

9 Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R* (2<sup>nd</sup> Ed.). Chapman  
10 and Hall/CRC.

11 Yang, C. L., Perfetti, C. a, & Schmalhofer, F. (2007). Event-related potential indicators of text  
12 integration across sentence boundaries. *Journal of Experimental Psychology: Learning,  
13 Memory, and Cognition*, 33(1), 55–89. doi: 10.1037/0278-7393.33.1.55

14 Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-  
15 formation. *Journal of Comparative Neurology and Psychology*, 18, 459-482. doi:  
16 10.1002/cne.920180503

17 Zekveld, A. A., & Kramer, S. E. (2014). Cognitive processing load across a wide range of  
18 listening conditions: Insights from pupilometry. *Psychophysiology*, 51(3), 277–284. doi:  
19 10.1111/psyp.12151

20 Zekveld, A. A., Koelewijn, T., & Kramer, S. E. (2018). The pupil dilation response to auditory  
21 stimuli: current state of knowledge. *Trends in Hearing*, 22. doi:  
22 10.1177/2331216518777174

1 Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2010). Pupil response as an indication of  
2 effortful listening: The influence of sentence intelligibility. *Ear and Hearing*, 31(4), 480–  
3 490. doi: 10.1097/AUD.0b013e3181d4f251

4 Zhao, S., Bury, G., Milne, A., & Chait, M. (2019). Pupillometry as an objective measure of  
5 sustained attention in young and older listeners. *Trends in Hearing*, 23. doi:  
6 10.1177/2331216519887815

7

1 **Table 1. Summary of LMER: Baseline pupil size by presentation and SNR.**

Formula: Baseline Pupil Size ~ SNR × Presentation + (1   Participant)					
<i>Fixed effects</i>	<i>Est.</i>	<i>Std. Error</i>	<i>df</i>	<i>t</i>	<i>p</i>
(Intercept)	3149.70	381.51	21.19	8.26	< .001
SNR (-6 dB)	-281.38	168.16	135.98	-1.67	.10
Presentation (2 <sup>nd</sup> )	343.23	163.28	135.37	2.10	.04
Presentation (3 <sup>rd</sup> )	348.14	168.26	135.41	2.07	.04
SNR (-6 dB) × Pres. (2 <sup>nd</sup> )	384.24	233.64	135.90	1.65	.10
SNR (-6 dB) × Pres. (3 <sup>rd</sup> )	301.47	240.24	135.54	1.26	.21

<i>Random effects</i>	<i>Variance</i>	<i>Std. Dev.</i>
1   Participant	2482076.00	1575.50

Notes. SNR = signal-to-noise ratio. Baseline pupil size is based on the median pupil size during a 2-s period of silence prior to the start of the audio with the male or female face cue present on screen. Bolded *p*-values indicate significance at  $\alpha = .05$ .

2  
3

1 **Table 2. Summary of GAMM: TEPR by time, baseline pupil size, presentation, and SNR.**

Parametric Terms	Est.	SE	t	p
(Intercept)	3315.40	163.84	20.24	< .001
-6 dB SNR (TRUE)	-55.42	75.26	-0.74	.46
2 <sup>nd</sup> Pres. (TRUE)	7.28	106.74	0.07	.95
3 <sup>rd</sup> Pres. (TRUE)	-83.34	237.99	-0.35	.73
-6 dB SNR, 2 <sup>nd</sup> Pres. (TRUE)	-7.33	111.91	-0.07	.95
-6 dB SNR, 3 <sup>rd</sup> Pres. (TRUE)	118.76	157.16	0.76	.45

Smooth Terms	EDF	Ref.df	F	p
s(Gaze X, Gaze Y)	192.99	198.59	621.03	< .001
te(Time, BPS)	50.17	57.50	3.29	< .001
te(Time, BPS): -6 dB SNR (TRUE)	31.12	37.50	1.67	< .001
te(Time, BPS): 2 <sup>nd</sup> Pres. (TRUE)	22.67	28.15	1.13	.29
te(Time, BPS): 3 <sup>rd</sup> Pres. (TRUE)	40.56	48.45	1.97	< .001
te(Time, BPS): -6 dB SNR, 2 <sup>nd</sup> Pres. (TRUE)	40.33	47.14	2.40	< .001
te(Time, BPS): -6 dB SNR, 3 <sup>rd</sup> Pres. (TRUE)	23.53	27.42	2.58	< .001

Random Smooths	EDF	Ref.df	F	p
s(BPS, Subject)	53.91	125.00	5.71	< .001
s(Time, Subject)	95.77	169.00	2.63	< .001
s(Time, Subject): -6 dB SNR (TRUE)	95.01	170.00	1.78	< .001
s(Time, Subject): 2 <sup>nd</sup> Pres. (TRUE)	73.66	170.00	1.61	< .001
s(Time, Subject): 3 <sup>rd</sup> Pres. (TRUE)	86.59	150.00	2.69	< .001
s(Time, Subject): -6 dB SNR, 2 <sup>nd</sup> Pres. (TRUE)	82.15	160.00	2.07	< .001
s(Time, Subject): -6 dB SNR, 3 <sup>rd</sup> Pres. (TRUE)	77.64	140.00	2.10	< .001

R<sup>2</sup> = 0.93; deviance explained = 78.5%; fREML = 59,481

Notes. Reference level of 0 dB SNR, 1<sup>st</sup> Presentation. SNR = signal-to-noise ratio; BPS = baseline pupil size. Baseline pupil size is based on the median pupil size during a 2-s period of silence before the start of the audio with the face cue present. Bolded p-values indicate significance at  $\alpha = .05$ .