Baseline pupil size encodes task-related information and modulates the task-evoked response in a speech-in-noise task

Helia Relaño-Iborra¹,², Dorothea Wendt³, Mihaela Beatrice Neagu², Abigail Anne Kressner²,⁴, Torsten Dau² and Per Bækgaard¹

Abstract
Pupillometry data are commonly reported relative to a baseline value recorded in a controlled pre-task condition. In this study, the influence of the experimental design and the preparatory processing related to task difficulty on the baseline pupil size was investigated during a speech intelligibility in noise paradigm. Furthermore, the relationship between the baseline pupil size and the temporal dynamics of the pupil response was assessed. The analysis revealed strong effects of block presentation order, within-block sentence order and task difficulty on the baseline values. An interaction between signal-to-noise ratio and block order was found, indicating that baseline values reflect listener expectations arising from the order in which the different blocks were presented. Furthermore, the baseline pupil size was found to affect the slope, delay and curvature of the pupillary response as well as the peak pupil dilation. This suggests that baseline correction might be sufficient when reporting pupillometry results in terms of mean pupil dilation only, but not when a more complex characterization of the temporal dynamics of the response is considered. By clarifying which factors affect baseline pupil size and how baseline values interact with the task-evoked response, the results from the present study can contribute to a better interpretation of the pupillary response as a marker of cognitive processing.

Keywords
pupillometry, listening effort, growth curve analysis, baseline pupil size

Accepted 3 October 2022

Introduction
The connection between pupil dilation & cognitive processes was documented as early as the 1800s (Schiff & Foa, 1874, for a historical review see Beatty & Lucero-Wagoner, 2000). Yet, scientific focus on the task-evoked pupillary responses (TEPRs) first truly began in the second half of the 20th century (e.g., Hess & Polt, 1964; Kahneman & Beatty, 1966; Nunnally et al., 1967; Polt & Hess, 1960). These early studies investigated the pupil responses in a myriad of experimental conditions, studying the effects of perception (Kahneman & Beatty, 1967), problem solving (Hess & Polt, 1964), decision making (Simpson & Hale, 1969) or arousal (Bradshaw, 1967), among others. These studies reported changes in the pupil size relative to a baseline value measured before the task of interest under the premise that the pupil response can be separated in two independent components: the baseline pupil size, commonly referred to as “tonic response”, and the TEPR, defined as the pupil changes (i.e., dilations or constrictions) that occur as a result of a specific cognitive task, often associated with a “phasic response”.

¹Cognitive Systems Section, Department of Applied Mathematics and Computer Science, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark
²Hearing Systems Section, Department of Health Technology, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark
³Eriksholm Research Center, Oticon, 3070 Snekkersten, Denmark
⁴Copenhagen Hearing and Balance Center, Rigshospitalet, 2100, Copenhagen, Denmark

Corresponding author:
Helia Relaño-Iborra, Cognitive Systems Section, Department of Applied Mathematics and Computer Science, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark.
Email: heliaib@dtu.dk
Some of these early studies focused on the perception of auditory stimuli e.g., Kahneman & Beatty (1967) measured pupil responses during a pitch discrimination task and Nunnally et al. (1967) explored the pupil’s reaction to pure tones presented at different sound pressure levels. However, the use of pupillometry as an indicator of cognitive activity during complex listening tasks, namely speech-in-noise processing, only erupted in the second decade of the 2000s with several studies linking the TEPR to listening effort (e.g. Ohlenforst et al., 2018; Wendt et al., 2016, 2018; Wetzel et al., 2016; Zekveld et al., 2011; see Zekveld et al., 2018 for a thorough review of the literature). Despite the time span and the methodological differences between the state-of-the-art listening effort literature and the earlier studies, listening effort, i.e., the allocation of cognitive resources to the completion of an auditory task (Pichora-Fuller et al., 2016), is still most commonly reported by the peak pupil dilation (PPD) or mean pupil dilation (MPD) extracted from a baseline corrected TEPR. In these studies, the baseline region is defined in a pre-task condition in quiet (e.g., Wetzel et al., 2016) or, more often, in the presence of a simpler acoustic signal e.g., the masker (alone) in a speech-in-noise paradigm (e.g., Ohlenforst et al., 2018; Wendt et al., 2016, 2018). Very few studies have focused on the baseline region and its relation to listening effort (Ayasse & Wingfield, 2020; Alhanbali et al., 2020).

Whereas the TEPR has been widely shown to reflect some level of cognitive processing, the sources of variability of the baseline pupil size remain manifold, such as the pupillary light reflex (e.g., Bradshaw, 1969; Lowenstein & Lowenfeld, 1964; Pysakhovich et al., 2015, 2017; Reilly et al., 2019; Steinhauer et al., 2004); tonic arousal (e.g., Gilzenrat et al., 2010; Jepma & Nieuwenhuis, 2011; Murphy et al., 2014); age (Birren et al., 1950; Kasthurirangan & Glasser, 2006; Ko et al., 2011; Piquado et al., 2010) and cognitive abilities (Aminihajibashi et al., 2020; Tsukahara et al., 2016). Tryon (1975) provided an early review of more than twenty sources of variability in pupil size. However, less is known about whether changes in the baseline values might also arise due to task-related parameters, such as mental demand. Several studies showed no evidence of these effects (e.g., Granholm et al., 1996; Trani & Verhaeghen, 2018) while others showed changes in pre-task baselines correlated with task complexity (e.g., Irons et al., 2017; Mosaly et al., 2017; Steiner & Barry, 2011).

The assumption of independence between the baseline pupil size value and the TEPR was first made by Kahneman & Beatty (1967). In their study, where test subjects performed a pitch-discrimination task, a consistent drop in the baseline values (i.e., a constriction in the pupil size) across experimental trials was found, while no decrease in the mean dilation of the TEPR was observed across trials. Bradshaw (1969) tested the hypothesis of independence of the two responses in a controlled manner by manipulating the lighting conditions during an auditory reaction time task, such that the baseline pupil size changed across measurements. He found that this manipulation did not affect the peak amplitude nor the shape of the TEPR. The findings from Bradshaw (1969) were replicated by Xu et al. (2011) using a similar paradigm for arithmetic tasks and by Reilly et al. (2019) for the perception of pure tone transitions and for a visual inspection task at different luminosity levels. In line with these findings, albeit in a less controlled manner, Beatty (1982a) observed no changes in the baseline pupil size for an auditory vigilance task that elicited significant changes in the TEPR. Similarly, Granholm et al. (1996) analyzed baseline values and TEPRs for a digit-recall task and found no correlation between baseline and pupil dilation neither at the recall onset nor during the period where listeners retained the digits in working memory.

In an influential review of the early pupillometry findings, Beatty (1982b) compared the measured peak amplitudes of the baseline-corrected TEPRs across different studies, arguing that the magnitude of the TEPRs during cognitive processing is “independent of baseline pupillary diameter over a physiologically reasonable but not extreme range of values”. Since this formalization of the TEPR as an independent measure from baseline pupil size, baseline correction has been widely accepted as the method to isolate the TEPR, allowing not only for inter-task comparisons, but also to compare across different test subject responses as well as across differences in experimental conditions.

In contrast, work by Peysakhovich and colleagues (Pysakhovich et al., 2015, 2017) showed that the TEPR, as characterized by the mean pupil dilation, was luminance-dependent (and therefore also baseline-dependent) both in short-term memory tasks and multiplication tasks. Similarly, Steinhauer et al. (2004) found a task-by-luminance interaction for a subtraction task at different luminosity levels and a modulation of the TEPR response when the baseline pupil size was altered by using pharmacological and environmental (i.e., luminosity) manipulations. Furthermore, Gilzenrat et al. (2010) found an inverse relationship between baseline pupil size and the magnitude of the TEPR. The findings of these studies, that elevated baselines lead to reduced TEPRs, might indicate adaptive adjustments of the tonic response for performance optimization (i.e., preparatory processing) by the locus coeruleus (LC) as predicted by the adaptive gain theory of tonic and phasic activation of Aston-Jones and Cohen (2005) and are consistent with models of optimal arousal (Aston-Jones et al., 1999; Teigen, 1994; Yerkes & Dodson, 1908).

Despite the contradicting evidence on the relationship between baseline pupil size and TEPR, baseline correction is still the standard in pupillometry literature. The analysis of baseline-corrected TEPRs, implies the assumption of independence between the two periods of the pupil response. Recently, the potential biases introduced in the TEPR by baseline correction were investigated by Mathôt et al. (2018). They found that reported TEPRs varied depending on the accuracy of the baseline calculation, supporting the
idea that a relationship exists between baseline values and the TEPR. Several studies have argued for a characterization of the TEPR that bypasses baseline correction (Duchowski et al., 2018, 2020; Peysakhovich et al., 2015), however baseline correction is still prevalent in pupillometry research.

In addition, most studies that investigated the relationship between baseline and TEPR used a limited characterization of the TEPR, with metrics such as the PPD and the MPD. However, despite the practicality of these static measures of the TEPR, they represent only one aspect of the pupillary response. To the knowledge of the authors, a systematical investigation of the relationship between baseline values and the entire time course of the TEPR has not yet been undertaken. With studies proposing measures of the TEPR beyond the PPD and MPD (e.g., Bianchi et al., 2019; Kuchinsky et al., 2013; Mirman et al., 2008; Wendt et al., 2018), it remains to be clarified what the implications of the baseline correction are for the TEPR and how baseline values and the TEPR are related when using a more complex characterization of the response.

The present study investigated the relationship between the baseline pupil size and the temporal dynamics of the TEPR as well as the influence of the baseline pupil size of experimental factors, such as task demand and time-on-task (defined here as the amount of time during which cognitive resources are actively invested on the task). The goal of the study was to clarify whether the assumption of independence between baseline pupil size and baseline-corrected TEPR holds for TEPR metrics that characterize the time course of the response beyond PPD and MPD during a speech in noise task. Specifically, this study focused on the analysis of pupil data estimates obtained using growth curve analysis (GCA; Mirman et al., 2008), due to their current surge in popularity in the pupillometry literature (e.g., Bianchi et al., 2019; Juul Jensen et al., 2018; Koch & Janse, 2016; Kuchinsky et al., 2013, 2014; McGarrigle et al., 2017; Mclaughlin et al., 2020; Neagu et al., 2019; Winn et al., 2015; Winn, 2016).

**Methods**

To explore experimental factors affecting baseline pupil size and the relationship between baseline pupil size and the TEPR, this study analyzed a dataset collected by Wendt et al. (2018, Experiment 2) consisting of pupil recordings obtained during a speech intelligibility in noise task. Sentences were presented in the presence of different noise-maskers. A block of 25 sentence-trials was used to test each considered signal-to-noise ratio (SNR). Wendt et al. (2018) analyzed changes to the baseline corrected TEPR to investigate the impact of SNR and noise type on listening effort. In the present study the analysis was extended to include (i) the

---

**Figure 1.** Workflow of the study. The squares represent processing and analysis stages, whereas the connecting lines indicate the type of data that is transferred between stages. Double lines represent block recordings containing multiple trials of equivalent conditions; dashed lines represent the individual time series for each trial after trial separation; solid lines represent baseline corrected traces (i.e., the task-evoked pupillary response, TEPR); dotted lines represent the extracted baseline values and dash-dotted lines represent the TEPR metrics (i.e., the peak pupil dilation and the Growth Curve Analysis estimates).
baseline values and (ii) the influence of other experimental parameters beyond task demand.

Figure 1 shows the different methodologies and analysis strategies considered in this study. Each methodological stage is detailed in the sections below.

**Experimental Data**

Wendt et al. (2018) collected pupil recordings of 29 native Danish, normal-hearing listeners during a speech-in-noise task. Sentences from the Danish HINT corpus (Nielsen & Dau, 2011) were presented in the presence of two different noise maskers: a speech-shaped-noise (SSN) and a four-talker babble noise (4TBB) at different SNRs. Blocks of 25 trials, each containing one sentence, were presented for each SNR condition, whereby the block presentation order was randomized across listeners. Each trial included 3 s of noise alone followed by the sentence in noise (average sentence duration 1.5 s; std = 0.2 s) and three seconds of noise alone following the sentence offset. After the noise offset the participants provided their response, followed by a 2-s recovery period in quiet before a new trial was initialized.

The sound pressure level (SPL) of the noise masker was fixed at 65 dB, and SNRs ranging from −20 to 8 dB in 4 dB-steps were obtained by varying the level of the speech signal. All SNR conditions for one masker were presented in a single session, and the order of which noise masker was tested first was randomized across listeners.

The recordings were collected using the iView X RED System eye tracker (SensoMotoric Instruments, Teltow, Germany) with a sampling rate of 60 Hz. Even though Wendt et al. (2018) recorded both eyes, only the left eye was used in the present study. The average correlation between the left and right eye traces was r > 0.99 (minimum 0.987), it was therefore assumed that the results presented here should be independent of the eye-choice. Listeners were tested under constant luminance conditions (~135 lux), with small adaptations for listeners that had relatively big pupil sizes at rest (Wendt et al., 2018).

**Data Preprocessing**

The raw pupil recordings were pre-processed to remove artifacts and reduce the noisiness inherent to pupil recordings. The processing workflow provided by Relañño-Iborra and Bækgaard (2020) was used which removes artifacts and noise from pupillometry recording and provides annotated data. Here, each recording (i.e., each 25-sentences SNR block) was preprocessed to remove blinks, identify saccade regions (i.e., rapid eye movements), as well as to interpolate and denoise the data. Blinks were defined as samples whose recorded value was more than three times lower than the mean of that block’s recording, as recommended by Winn et al. (2018). Subsequently, saccades were detected using the velocity-based algorithm proposed by Duchowski et al. (2002). Velocity data were not provided in the raw data of Wendt et al. (2018), thus, the angular velocity was derived from the gaze coordinates by:

$$v = \frac{\theta}{\Delta t}$$

where \(\theta\) represents the visual angle. The time increment \(\Delta t\) in this study was chosen to be 0.08 s, equivalent to five time-samples. A region of the recording was considered a saccade when the calculated velocity over that region was above 22°/s, as suggested by Holmqvist et al. (2011). Missing data and blink portions were linearly interpolated. The region of interpolation includes an additional 50 ms segment prior to and 150 ms segment after the detected missing portions to avoid measurement artifacts related to blinks (Winn et al., 2018). Additionally, if a blink occurred within a saccade, the whole saccade region was interpolated. The final denoised signal was obtained by low-pass filtering the interpolated data (\(f_{\text{cutoff}} = 10 \, \text{Hz}\)). After preprocessing, each block recording was separated into individual trial traces. The baseline was calculated as the mean pupil size during the last second of the noise-alone region prior to the sentence onset for each trial. Baseline correction was performed for each trial by subtracting the baseline value from each sample in the remaining portion of the trial.

A quality threshold was defined such that traces that contained more than 15% of interpolated data were rejected. Additionally, if missing regions were found in the baseline period, the trial was also rejected (Mathôt et al., 2018). Overall, one listener was discarded due to excessive missing data (>20% of invalid trials across all conditions). After removing this listener, 10,800 traces were analyzed (28 listeners x 8 SNRs x 25 sentences per SNR-block x 2 maskers), from which 58 were rejected as they did not meet the quality threshold (0.005%). For all analyses in this study, the first 5 trials of each block were discarded to avoid biasing the statistical analyses due to effects of initial arousal (Winn et al., 2018).

**Characterization of the TEPR**

As this study aimed to evaluate effects of baseline on commonly reported pupil metrics, and in order to obtain stable TEPRs, the trials within each SNR-block were averaged, such that one overall trace was obtained per listener and SNR-block. To characterize the resulting TEPRs, the PPD was extracted, defined as the maximum value found in the first 5 s following the sentence onset, such that the analysis window covered the listening and retention periods, i.e., the sentence duration (1.5 ± 0.2 s) and the following 3-s of noise alone, but not the response periods. In addition to the traditional PPD, growth curve analysis (GCA; Mirman, 2014) was applied to obtain estimates of the mean, slope, curvature and delay of the TEPR in the same analysis window.
GCA generates models of the TEPR that provide a reduced metric space to characterize the pupillary response. Originally inspired by its use on longitudinal studies, Mirman et al. (2008) provided a generalization for time series analysis of the pupillary response. The underlying idea behind the GCA model is that of nested mixed-effects models. First, a so-called level-1 model is built, which provides the temporal relationship and is defined as:

\[
PupilSize_{ij} = a_0 + \beta_{1i} \cdot time_{1ij} + \beta_{2i} \cdot time_{2ij} + \beta_{3i} \cdot time_{3ij} + \epsilon_{ij}
\]  

where \( i \) corresponds to each listener and \( j \) represents each time window. Here, 100-ms time windows were used to analyze the TEPR. The different time components are then defined by orthogonal polynomials to avoid co-linearity of the parameters. In other words, the vectors \( time_{1} \), \( time_{2} \) and \( time_{3} \) are generated independently and orthogonally following Mirman et al. (2008). \( a_0 \) corresponds to the intercept or mean dilation and it can be directly compared to the traditional mean pupil dilation (MPD); \( \beta_{1i} \) represents the linear or slope term and it indicates the speed of the pupil dilation; \( \beta_{2i} \) is the quadratic or curvature term indicating the width between the rising and decaying of the response; and \( \beta_{3i} \) corresponds to the cubic or delay term associated with the presence of secondary peaks. \( \epsilon_{ij} \sim N(0, \sigma^2) \) is the residual error of the model. The values of each parameter of the level-1 model are further characterized by a nested mixed effect model such that the fixed and random effects of interest can be included for each temporal characteristic independently. Listeners were considered a random effect whereas the fixed effect structure consisted of the noise type, the SNR and its interaction. Thus, the level-2 models were defined as:

\[
a_{0i} = y_{00} + y_{0,SNR} \cdot SNR + y_{0,Noise} \cdot Noise + y_{0,SNR \cdot Noise} + \xi_{0i}
\]

\[
\beta_{1i} = y_{10} + y_{1,SNR} \cdot SNR + y_{1,Noise} \cdot Noise + y_{1,SNR \cdot Noise} + \xi_{1i}
\]

\[
\beta_{2i} = y_{20} + y_{2,SNR} \cdot SNR + y_{2,Noise} \cdot Noise + y_{2,SNR \cdot Noise} + \xi_{2i}
\]

\[
\beta_{3i} = y_{30} + y_{3,SNR} \cdot SNR + y_{3,Noise} \cdot Noise + y_{3,SNR \cdot Noise} + \xi_{3i}
\]

where \( y_{00} \) is the baseline value of \( a_{0i} \), \( y_{0,SNR} \) is the fixed effect of the SNR condition on the intercept, \( y_{0,Noise} \) is the effect of the noise in the intercept, \( y_{2,Noise} \cdot SNR \) reflects the interaction term and \( \xi_{0i} \) is the random deviation from the mean for the \( i^{th} \) individual—and correspondingly for the rest of the level-2 models. In contrast to Wendt et al. (2018), where GCA was applied to the dataset separately for each noise masker, here one GCA model for the entire dataset was built in order to analyze not only the effect of SNR in the GCA components but also that of the noise type. A random effect was included in each of the level-2 models to be able to obtain unique values for each parameter for each listener, noise type and condition. A forward model selection was used as recommended by Mirman et al. (2008) in order to select the best predictive model. Fixed effects were added sequentially to each time component according to their expected effect size, starting with the SNR, then noise and finally their interaction. Model comparisons were performed using likelihood ratio tests in R (Chambers and Hastie, 1992), and the final model was chosen such as to obtain a minimal Akaike’s Information Criteria (AIC; Akaike, 1973) so that the risk of overfitting was accounted for in the model selection criteria.

### Statistical Analyses

**Factors Affecting Baseline Pupil Size.** The baseline pupil sizes, defined as the mean pupil size during the last second of the noise-alone region prior to sentence onset of each trial, were extracted as a part of the data preparation. Changes in the individual trial baseline resulting from experimental parameters, such as the block order and the trial order within a block, were investigated. The effect of task difficulty, manipulated in this data set by changes in the SNR, as well as the noise type and recording session, were also evaluated. A mixed-effects model was used with listener evaluated as a random effect. The fixed effects considered were noise type, session, SNR, block order and trial number and all interactions; quadratic effects were also included (see the appendix, for the mathematical derivation of the statistical model).

The model was reduced by backwards selection from the maximal base model, thus not assuming any a priori effect size. Likelihood ratio tests were used for model comparisons and non-significant effects were sequentially removed until all remaining effects were significant. The analysis was performed in R (R Core Team, 2018) using the package lme4 (Bates et al., 2015).

**Baseline Effects on the TEPR.** To investigate the effect of the baseline pupil size and experimental parameters on the temporal dynamics of the baseline corrected TEPR, five metrics were used to characterize the TEPR: PPD, GCA mean, GCA slope, GCA curvature and GCA delay (i.e., the intercept, linear, quadratic and cubic GCA parameters, respectively). Each of the metrics was calculated individually for each listener, noise type and SNR-block (i.e., averaged across trials), as previously described. The resulting values were used as the dependent variable in five separate mixed effects models. The base model was common for all metrics with listener as a random effect and SNR, noise type, block order and
baseline as potential fixed effects. Here, in contrast to the baseline pupil size analysis where single-trial baselines were considered, the mean baseline across all trials within a block was considered as to have the same time scale for all variables. The models were reduced by backwards selection from the maximal base model using likelihood ratio tests with the R-package lme4 (Bates et al., 2015) until all remaining effects were significant.

**Results**

**GCA Model for the Characterization of the TEPR**

The GCA model selection revealed significant improvements in the model performance (as evaluated by an increase in the log-likelihood and a decrease in the AIC) when including SNR and noise effects in all the GCA parameters (i.e., in Equations (3) to (6) for $\alpha_{0i}$, $\beta_{1i}$, $\beta_{2i}$ and $\beta_{3i}$). Additionally, the models performed better when including SNR and noise type interactions in all GCA components, except in the GCA delay (i.e., in the quadratic component). For a detailed summary of the model comparisons performed, reference is made to Table 1.

The results from the fitted model are shown in Figure 2 as a function of time from the sentence onset. The GCA model results are indicated as thick lines indicating mean values and the shadowed area representing the standard error. The left panel shows the data and model results for the 4TBB, and the right panel shows the corresponding results for the SSN. Data and model results for each SNR are shown using a color legend. Figure 3 shows the marginal means of the baseline pupil size for the fitted model, illustrating the main effects of SNR, noise type and block order, as well as their three-way interaction. The marginal means of the baseline pupil size across blocks (x-axis) are shown for each SNR (color legend) for the 4TBB (left panel) and the SSN (right panel). It can be observed that the baseline pupil size decreases as the time-on-task increases (here represented by the block order), whereby the extent of the constriction changes for each SNR, i.e., the lowest SNRs (SNR < -12 dB) have a larger constriction from Block 1 to 8 than the higher SNRs (SNR >4 dB). Larger baseline sizes were found for the hardest conditions (SNR >-16 dB) than for the easier conditions (SNR >4 dB) for the SSN, but only when they were presented earlier in the experiment, with the reverse trend emerging when these conditions are presented later in the experiment (i.e., harder conditions showed lower baselines

### Table 1. Results from the stepwise model comparisons of the growth curve analysis (GCA) for the task-evoked pupillary response (TEPR).

|                                              | AIC     | BIC     | logLik   | deviance | $\chi^2$ | p      |
|----------------------------------------------|---------|---------|----------|----------|----------|--------|
| Listener effects only                        | 65689   | 65813   | -32830   | 65659    |          |        |
| SNR effect in intercept                      | 62601   | 62783   | -31279   | 62557    | 3102.00  | <0.001***|
| SNR effect in intercept and slope            | 62418   | 62658   | -31180   | 62360    | 197.21   | <0.001***|
| SNR effect in intercept, slope and quadratic term | 61608   | 61907   | -30768   | 61536    | 823.28   | <0.001***|
| SNR effect in intercept, slope, quadratic and cubic terms | 61472   | 61829   | -30693   | 61386    | 150.48   | <0.001***|
| Noise type effect in intercept               | 61231   | 61596   | -30571   | 61143    | 243.29   | <0.001***|
| Noise type effect in intercept and slope      | 61228   | 61601   | -30569   | 61138    | 4.47     | <0.05*  |
| Noise type effect in intercept, slope and quadratic term | 61199   | 61580   | -30554   | 61107    | 31.08    | <0.001***|
| Noise type effect in intercept, slope, quadratic and cubic terms | 61181   | 61570   | -30543   | 61086    | 20.34    | <0.001***|
| SNR:Noise interaction in intercept           | 60886   | 61334   | -30389   | 60778    | 309.02   | <0.001***|
| SNR:Noise interaction in intercept and slope | 60764   | 61270   | -30321   | 60642    | 135.29   | <0.001***|
| SNR:Noise interaction in intercept, slope and quadratic term | 60760   | 61324   | -30312   | 60624    | 18.33    | <0.05*  |
| SNR:Noise interaction in intercept, slope, quadratic and cubic terms | 60765   | 61387   | -30208   | 60615    | 9.24     | 0.236  |
than the easy conditions when presented in the final blocks). For the 4TBB, a similar change in the baseline responses from the earlier to the later blocks was found; early blocks showed the highest baseline values for the most difficult SNRs (SNR < -16 dB), whereas later block showed the highest baseline values for the medium SNRs (0 < SNR < -8 dB), albeit lower overall baselines were found as time increased (i.e., the baselines for medium SNRs in later blocks are much lower than those observed for the low SNRs in early blocks). The interaction of task difficulty and presentation order in the baseline response indicates that different baseline values are measured for the most difficult conditions depending on whether they are presented early or later on in the experiment run.

Reduced baselines for the SSN masker as compared to the 4TBB masker for all corresponding SNRs and block presentation orders can also be observed in Figure 3. However, the constriction rate from the initial to the final presentation block was also smaller in the case of SSN, reflecting the interaction of these effects.

Additionally, the main effects of SNR and noise type over the baseline values are shown in Figure 4; the main effects of SNR and noise type on the PPD are also shown for comparison. The left panel in Figure 3 shows the changes in baseline pupil size as a function of SNR for the 4TBB (black circles) and the SSN (grey triangles) conditions, while the right panel in Figure 4 replots the PPDs across SNRs reported by Wendt et al. (2018) for both noises. The model of the baseline values revealed strong and significant effects of both SNR, noise type and their interaction. This is in contrast to the findings of Wendt et al. (2018) for the PPD, where very small, albeit significant (p < 0.01), effects of the noise type were found and no significant effects of the interaction between SNR and noise type (p = 0.9) were reported. Additionally, Wendt et al. (2018) reported no significant differences in performance across the two noise conditions. This indicates that baseline pupil size can encode differences in masker type characteristics, even if these differences are not reflected in the PPD.

**Baseline Effects on the TEPR**

A mixed-effects model was fitted to each of the parameters extracted in the TEPR characterization, i.e., to the PPD, GCA mean, GCA slope, GCA curvature and GCA delay. The results from the analysis are summarized in Table 4.

A highly significant (p < 0.001) negative effect of the baseline was found for the slope, curvature and delay of the TEPR. Thus, these TEPR characteristics get reduced when the baseline increases. A positive marginally

---

**Figure 2.** Pupil traces as a function of time from the sentence onset for the 4TBB masker (left) and SSN (right). The thin lines represent mean values of the raw data while the shadowed regions represent its standard errors. The thick lines represent the results from the GCA model. The SNR condition is represented by the different colors.
**Table 2.** Outputs for the mixed effect model on the TEPR. The model formula follows: \( \text{pupilSize} \sim (1 + \text{Linear + Quadratic + Cubic}) \times \text{SNR} + (1 + \text{Linear + Quadratic + Cubic}) \times \text{Noise} + (1 + \text{Linear + Quadratic + Cubic}) | \text{listener} \). The interaction of \( \text{SNR:Noise} \) on the cubic parameter was not included in the final model after model selection. Thus, estimates for the SSN at all SNRs are equal and only reported once (redundant values are shown as ‘–’).

| SNR        | Intercept Estimate | t | p  | Linear Estimate | T | p  | Quadratic Estimate | t | p  | Cubic Estimate | T | p  |
|------------|-------------------|---|----|----------------|---|----|-------------------|---|----|----------------|---|----|
| 4TBB       | −20 dB (Baseline) | −0.154 | −1.677 | 0.094 | −2.885 | −5.646 | <0.001 *** | −2.240 | −6.960 | <0.001 *** | 0.060 | 0.288 | 0.774 |
| −16 dB     | 0.102             | 4.638 | <0.001 *** | −0.008 | −0.044 | 0.965 | 0.060 | 0.320 | 0.749 | 0.336 | 2.557 | <0.05 * |
| −12 dB     | 0.270             | 12.253 | <0.001 *** | 0.542 | 2.915 | <0.01 ** | 0.327 | 1.757 | <0.05 * | 0.749 | 5.701 | <0.001 *** |
| −8 dB      | 0.482             | 21.846 | <0.001 *** | −0.038 | −0.204 | 0.839 | −0.550 | −2.958 | <0.01 ** | 0.948 | 7.213 | <0.001 *** |
| −4 dB      | 0.534             | 24.222 | <0.001 *** | 0.528 | 2.838 | <0.01 ** | −0.032 | −0.173 | 0.862 | 1.111 | 8.450 | <0.001 *** |
| 0 dB       | 0.233             | 10.556 | <0.001 *** | 0.749 | 4.030 | p < 0.001 *** | 1.326 | 7.133 | <0.001 *** | 1.308 | 9.954 | <0.001 *** |
| 4 dB       | 0.146             | 6.604 | <0.001 *** | 0.832 | 4.474 | p < 0.001 *** | 1.954 | 10.513 | <0.001 *** | 1.119 | 8.515 | <0.001 *** |
| 8 dB       | −0.045            | −2.052 | <0.05 * | 0.728 | 3.917 | p < 0.001 *** | 2.414 | 12.985 | <0.001 *** | 0.763 | 5.804 | <0.001 *** |
| SSN        | −20 dB            | −0.210 | −9.500 | <0.001 *** | −0.723 | −3.890 | <0.001 *** | 0.705 | 3.794 | <0.001 *** | 0.299 | 4.547 | <0.001 *** |
| −16 dB     | −0.050            | −1.611 | 0.107 | 0.678 | 2.577 | <0.01 ** | 0.094 | 0.357 | 0.721 | − | − | − |
| −12 dB     | −0.086            | −2.760 | <0.01 ** | −0.133 | −0.508 | 0.611 | −0.269 | −1.022 | 0.307 | − | − | − |
| −8 dB      | 0.319             | 10.219 | <0.001 *** | 2.407 | 9.157 | <0.001 *** | −0.568 | −2.161 | <0.05 * | − | − | − |
| −4 dB      | 0.161             | 5.166 | <0.001 *** | 1.132 | 4.306 | <0.001 *** | −0.566 | −2.153 | <0.05 * | − | − | − |
| 0 dB       | 0.131             | 4.196 | <0.001 *** | 1.125 | 4.278 | <0.001 *** | −0.332 | −1.264 | 0.206 | − | − | − |
| 4 dB       | −0.020            | −0.649 | 0.516 | 0.407 | 1.549 | 0.121 | −0.244 | −0.930 | 0.352 | − | − | − |
| 8 dB       | 0.238             | 7.642 | <0.001 *** | 1.291 | 4.911 | <0.001 *** | −0.801 | −3.047 | <0.01 ** | − | − | − |
significant (p = 0.049) effect was found for the PPD, suggesting that PPD increases for elevated baseline values.

**Discussion**

The results from the present study showed that for the investigated speech-in-noise listening task, baseline pupil size showed significant effects of task demand (as manipulated both by changes in SNR and differences in noise type). Furthermore, the results indicated that time-on-task affects the baseline pupil size at different time scales: across sentence trials, across presentation blocks and across measurement sessions. Additionally, it was shown that the effects of task demand and time-on-task on the measured baseline

### Table 3

| Fixed Effect                                      | Estimate | t     | p     |
|---------------------------------------------------|----------|-------|-------|
| Intercept (−20 SNR, 4TBB, 1st session)            | 5.724    |       |       |
| trial                                             | −0.027   | −6.436| <0.001*** |
| Block order                                       | 0.022    | 2.557 | <0.05*  |
| SNR                                               | −0.020   | −14.883| <0.001*** |
| 2nd Session                                       | −0.128   | −4.616| <0.001*** |
| SSN noise                                         | −0.234   | −12.382| <0.001*** |
| Trial\(^2\)                                       | 0.001    | 5.070 | <0.001*** |
| Block order\(^2\)                                 | −0.005   | −6.155| <0.001*** |
| SNR\(^2\)                                         | −0.001   | −11.110| <0.001*** |
| Block order:Session                               | −0.020   | −5.753| <0.001*** |
| Trial:Session                                     | −0.003   | −2.450| <0.05*  |
| Block order:SNR                                    | 0.001    | 5.480 | <0.001*** |
| Block order:Noise Type                            | 0.020    | 5.723 | <0.001*** |
| SNR:Noise type                                    | 0.004    | 4.086 | <0.001*** |

**Figure 3.** Marginal means of the baseline pupil size for changes in block order (x-axis), shown for each SNR (color) for the four-talker babble noise (4TBB; left panel) and the stationary speech shaped noise (SSN; right panel).
value interacted significantly. The results of this study also showed that baseline pupil size encoded differences in masker type characteristics, even if these differences are not reflected in the PPD nor in the task performance.

The observation of task complexity influencing pre-task baselines is consistent with previous studies (e.g., Ganea et al., 2020; Irons et al., 2017; Mosaly et al., 2017; Steiner & Barry, 2011) that found similar effects for a broad range of cognitive tasks. Steinhauser and colleagues linked the elevated baselines for difficult tasks with differential preparation and processing (Steinhauer et al., 2004). This preparatory control, as measured by increased baseline pupil size has also been linked to, e.g., decision-making performance (Jericic, 2019) and enhancement of stimulus detection (Steiner and Barry, 2011). In a thorough experimental analysis of baseline values, Gilzenrat et al. (2010) found that measured changes in the pupil baseline were correlated with behavior and were reliable indicators of task engagement and disengagement, arguing that this was consistent with the pupil diameter being indicative of locus coeruleus (LC) activity as predicted by the adaptive gain theory (Aston-Jones & Cohen, 2005). However, several studies did not show evidence that the baseline pupil size tracked task complexity (Granholm et al., 1996; Trani & Verhaeghen, 2018). Interestingly, Granholm et al. (1996) found that baseline pupil size did not follow task difficulty in a working memory task but did so in a visual tracking task, while baseline pupil size did not predict subsequent performance for either of them. This suggests that baseline activation might be paradigm-dependent, which may account for the differences across studies. Based on the findings of the present study, it appears that a complex task, such as understanding speech in noise, might induce preparatory control, as reflected by pre-task baseline elevation for the most difficult conditions.

Along these lines, the finding of elevated baselines for the babble noise (see Figure 4) as compared to the SSN indicated that baseline pupil size can encode differences in listening effort due to e.g., acoustic complexity of the masker, even when the PPD does not reflect any masker-type differences. These findings complement those of Wendt et al. (2018), who could not verify their hypothesis that to obtain the same performance level that they had observed across maskers, a larger allocation of effort (measured by an increase in the PPD) should be necessary for the babble masker. Therefore, the current findings support the hypothesis that elevated baselines might be a marker of performance facilitation (Steiner & Barry, 2011). Additionally, a decrease in pupil size with time-on-task was observed, which was consistent with previous findings (e.g., Ayasse & Wingfield, 2020; Hyönä et al., 1995; Steiner & Barry, 2011; for a

---

**Figure 4.** The left panel shows the extracted baselines as a function of SNR whereas the right panel shows the peak pupil dilation (PPD) as reported in Wendt et al. (2018). Data from the speech-shaped noise (SSN) masker condition are plotted in black circles and from the four-talker babble (4TBB) condition with gray triangles.
Table 4. Results from the mixed effect models for the different TEPR metrics. Each model was optimized individually, only significant effects are shown.

| PPD | GCA Mean | GCA Slope | GCA Curvature | GCA Delay |
|-----|----------|-----------|---------------|-----------|
| Variance explained by listener random effect: | 39.589% | 43.759% | 45.872% | 46.049% | 52.102% |

| | Estimate | t | p | Estimate | t | p | Estimate | t | p | Estimate | t | p | Estimate | t | p |
|---|----------|---|---|----------|---|---|----------|---|---|----------|---|---|----------|---|---|
| Baseline (4TBB, −20 dB) | −0.101 | 0.141 | 3.645 | 5.553 | 3.055 |
| SNR | 0.079 | 10.557 | <0.001*** | 0.021 | −3.731 | <0.001*** | −0.008 | −0.337 | n.s. | 0.022 | 0.416 | n.s. | 0.012 | 0.854 | <0.001*** |
| SNR² | −0.008 | −10.252 | <0.001*** | −0.002 | −5.012 | <0.001*** | −0.003 | −2.173 | <0.05* | −0.961 | −4.594 | <0.001*** | −0.629 | −4.706 | <0.001*** |
| Baseline pupil size | 0.017 | 1.993 | <0.05* | −0.042 | −0.709 | n.s. | −1.064 | −3.811 | <0.001*** | −1.624 | −8.143 | <0.01*** | −0.865 | −6.787 | n.s. |
| Noise type | 0.021 | 2.769 | <0.05* | −0.042 | −0.709 | n.s. | −1.064 | −3.811 | <0.001*** | −1.624 | −8.143 | <0.01*** | −0.865 | −6.787 | n.s. |
| SNR:baseline | 0.013 | 2.345 | <0.05* | −0.112 | −6.217 | <0.001*** | −0.043 | −3.759 | <0.001*** |
| SNR:noise type | 0.020 | 2.168 | <0.05* | −0.112 | −6.217 | <0.001*** | −0.043 | −3.759 | <0.001*** |
review see Zekveld et al., 2018). Furthermore, the results presented here showed an interaction between time-on-task and task difficulty in the pupil size (i.e., the baseline pupil sizes reflected task difficulty differently depending on the presentation order). A reduction in the pupil response for difficult conditions has previously been shown for the PPD (e.g., Ohlenforst et al., 2018; Wendt et al., 2018) and it has been argued to reflect a giving-up effect. However, its interaction with the time-on-task has not been analyzed. As shown in Figure 2, the most challenging conditions (i.e., lowest SNRs) elicited an elevated baseline only when they were presented earlier on in the experiment, whereas this elevated response was not found when presented in the later blocks. This indicates changes in the listener’s engagement (and disengagement) with the task due to familiarization with the task (engaging in the difficult tasks when they are presented earlier on, but not once they are familiar with the experiment paradigm). After sufficient exposure, listeners seem to be able to gauge whether effort deployment would result in a successful completion of the task, thus disengaging from it if success could not be achieved. This disengagement might be a result of e.g., fatigue, familiarity with the task or motivation (Pichora-Fuller et al., 2016). The interaction of task-demand and time-on-task in the measured baseline pupil size might indicate that preparatory control is less required as listeners familiarize themselves with the task at hand. Disentangling the interactions of these effects is not possible in the present dataset and requires further investigation. However, the effects of task engagement and preparatory control, as seen in the baseline in this study, further support the idea of the adaptive gain theory that pupil diameter can reflect levels of LC activity due to changes in control states (Aston-Jones & Cohen, 2005; Gilzenrat et al., 2010).

The interaction of task difficulty and presentation order in the baseline response, as well as the evidence presented here for preparatory control, is also consistent with the Framework for Understanding Effortful Listening (FUEL) of Pichora-Fuller et al. (2016) which argues for a multidimensional understanding of listening effort, where fatigue, arousal and motivation as well task demand interact in forming the listeners physiological and behavioral response to a set of stimuli and tasks. What is more, the influence of time-on-task in the relationship between baseline and task demand might explain why previous studies did not find baseline differences across task difficulties, as neither the task order nor the time-on-task was included in those analyses. Overall, the findings from this study suggest that thinking of the intra-trial baseline pupil size as a purely tonic response might not fully capture the complex cognitive processes happening before the task is presented to the listener, as also argued by Joshi and Gold (2020). Furthermore, the strong effect of presentation order across blocks, has implications for future experimental design, indicating that a randomized presentation of the different conditions might be preferable in order to factor out time-on-task influence on the pupillary response.

Indeed, a potential limiting factor of the results presented in this study is yet another side-effect of the block design which could lead to the contamination of the baseline region due to previous responses. Trials within a block were measured consecutively and, despite response and recovery times being inserted between trials, it is possible that subsequent trials were presented before the dilation from previous trials had fully converged to resting state. Thus, baseline values might reflect not only preparatory processing but also spill-over effects from the previous trials. However, as shown in Figure 4, changes in the baseline were found even when changes in the TEPR were not, supporting the idea that the baseline might encode aspects of cognitive processing that are simply not captured by the TEPR, such as e.g., task preparation, stimulus familiarity or motivation, as suggested by the FUEL (Pichora-Fuller et al., 2016).

It has been argued that elevated baselines consistent with preparatory processing and arousal (i.e., elevated LC activity) could potentially lead to reduced TEPRs, in accordance with both the adaptive gain theory (Aston-Jones & Cohen, 2005; Gilzenrat et al., 2010) and the Yerkes-Dodson model of optimal arousal (Aston-Jones et al., 1999; Teigen, 1994; Yerkes & Dodson, 1908). The present study found that elevated baseline responses corresponded to a reduced TEPR, with a significant (p < 0.001) negative effect of the baseline pupil size on the slope, curvature and delay of the TEPR. However, no effect of baseline on the GCA mean (i.e., on the average dilation, analogous to the MPD) was found, and a small and only marginally significant (p = 0.049) positive effect was found of the baseline on the PPD.

These findings contradict the earlier hypothesis of TEPR and baseline independence (e.g., Beatty, 1982a, 1982b; Bradshaw, 1969; Granholm et al., 1996; Kahneman & Beatty, 1967) but are in line with recent findings regarding the influence of baseline on TEPR metrics (e.g., Gilzenrat et al., 2010; Peysakhovich et al., 2017; Steinhauer et al., 2004). There are several findings from the current study that might explain the contradicting evidence regarding the relationship between baseline and TEPR across previous studies. First, the results from this study showed no effect of the baseline response on the GCA mean (which can be interpreted as a MPD) and found small and only marginally significant effects of baseline on the PPD (p = 0.049), suggesting a lack of stability of this effect. Given that MPD and PPD are the most common metrics of TEPRs, it is possible that previous studies that focused only in these two metrics simply overlooked the influence of the baseline in the TEPR, evident when the response is characterized in terms of estimates of its temporal dynamics. Second, task complexity might play a role on the relationship between baseline activation and TEPR. Mosaly et al. (2017) showed that different relations between baseline and TEPR exist for tasks that have variable loads, such as those measured in, e.g., Kahneman & Beatty (1967); Beatty (1982a), and tasks with sustained challenging conditions (e.g., Gilzenrat et al., 2010; Peysakhovich et al., 2017). The task considered in
the present study of speech understanding in noise is complex and, therefore, it follows that it results in a relation between baseline and TEPR. Finally, several of the studies that have discarded a relationship between baseline and TEPR were based on experimental paradigms that systematically manipulated baseline values by changing the lighting conditions only (e.g., Bradshaw, 1967; Reilly et al., 2019; Xu et al., 2011). However, Gilzenrat et al. (2010) found that when baseline changes were induced by luminosity changes, the TEPR did not show an inverse relationship with baseline, such as the one they found when manipulating the task complexity. In the present study, luminosity conditions were kept constant across task difficulty (i.e., SNRs and noise types). Thus, baseline pupil size changes (for a given individual) can be assumed to reflect task-related effects (e.g., habituation, fatigue, preparatory processing), consistent with previous findings (Gilzenrat et al., 2010; Granholm & Steinhauser, 2004).

The results from this study showed an influence of baseline pupil size on estimates of the TEPR measured using the GCA method. This modelling approach assumes a polynomial behavior of the time series and it is, furthermore, unable to account for potential autocorrelation of the response (Baayen et al., 2017; van Rij et al., 2019). Recently, generalized additive mixed modeling (GAMM; Hastie & Tibshirani, 1990) has been proposed as an alternative method to analyze pupillometry data (e.g., Algermissen et al., 2019; Aydn & Uzun, 2022; Beatty-Martínez et al., 2021; Boswijk et al., 2020; Huijser et al., 2020; Lőo et al., 2016; Pandža et al., 2020). This method overcomes some of the GCA weaknesses and could be used for similar analyses to the one presented in this study. Additionally, investigations examining task-encoding in the baseline region, and its influence on the TEPR, in auditory tasks beyond speech-in-noise paradigms are necessary in order to provide methodological recommendations regarding the role of baseline analysis and correction in pupillometry data reporting.

**Conclusion**

The findings presented here have implications for future experimental design (such as the interactions between task-demand and presentation order) and metric choices (as shown by the effect of baseline on certain aspects of the TEPR). This study might not solve the divide in the literature shown by the effect of baseline on certain aspects of the measured cognitive processes.

**Acknowledgements**

We are thankful for the comments and suggestions of two anonymous reviewers that helped improve this manuscript.

**Declaration of Conflicting Interests**

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

**Funding**

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the William Demant Foundation.

**ORCID iDs**

Helia Relaño-Iborra [i] https://orcid.org/0000-0002-2899-1673
Dorothea Wendt [i] https://orcid.org/0000-0002-3340-1180
Mihaela Beatrice Neagu [i] https://orcid.org/0000-0001-5737-4861
Abigail Anne Kressner [i] https://orcid.org/0000-0003-4274-3948
Torsten Dau [i] https://orcid.org/0000-0001-8110-4343

**Note**

1. Note that across-trial averaging was not performed in the baseline analysis as opposed to the TEPR characterization. Since the temporal changes of the baseline within the block trials was of main interest in the baseline analysis, averaging would have defeated its purpose. It was assumed that the inherent noisiness of trial-based analyses (which is often compensated for by reporting trial-averaged data), was not an issue in the baseline calculation as it is a relatively long-term measure (i.e., is obtained by averaging one second of the response).

**References**

Akaike, H. (1973). Information theory and an extension of the Maximum likelihood principle. *Proceeding of the Second International Symposium on Information Theory*, 267–281. https://doi.org/10.1007/978-1-4612-1694-0_15

Algermissen, J., Bijleveld, E., Jostmann, N. B., & Holland, R. W. (2019). Explore or reset? Pupil diameter transiently increases in self-chosen switches between cognitive labor and leisure in either direction. *Cognitive, Affective and Behavioral Neuroscience, 19*(5), 1113–1128. https://doi.org/10.3758/s13415-019-00727-x

Alhanbali, S., Munro, K. J., Dawes, P., Carolan, P. J., & Millman, R. E. (2020). Dimensions of self-reported listening effort and fatigue on a digits-in-noise task, and association with baseline pupil size and performance accuracy. *International Journal of Audiology*, 1–11. https://doi.org/10.1080/14992027.2020.1853262

Aminihajibashi, S., Hagen, T., Andreassen, O. A., Laeng, B., & Espeseth, T. (2020). The effects of cognitive abilities and task demands on tonic and phasic pupil sizes. *Biological Psychology, 156*, 107945. https://doi.org/10.1016/j.biopsycho.2020.107945

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

Aston-Jones, G., Rajkowski, J., & Cohen, J. (1999). Role of locus coeruleus in attention and behavioral flexibility. *Biological Psychiatry, 46*(9), 1309–1320. https://doi.org/10.1016/S0006-3223(99)00140-7
Ayasse, N. D., & Wingfield, A. (2020). Anticipatory baseline pupil diameter is sensitive to differences in hearing thresholds. *Frontiers in Psychology, 10*, 1–7. https://doi.org/10.3389/fpsyg.2019.02947

Aydin, Ö., & Uzun, I. P. (2022). Pupil dilation response to prosody and syntax during auditory sentence processing. *Journal of Psycholinguistic Research*. https://doi.org/10.1007/s10936-021-09830-y

Baayen, R. H., Vasishth, S., Kliegl, R., & Bates, D. (2017). The cage of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language, 94*, 206–234. https://doi.org/10.1016/j.jml.2016.11.006

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beatty, J. (1982a). Phasic tonic pupillary responses vary with auditory vigilance performance. *Psychophysiology, 19*(2), 167–172. https://doi.org/10.1111/j.1469-8986.1982.tb02540.x

Beatty, J. (1982b). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin, 91*(2), 276–292. https://doi.org/10.1037/0033-2909.91.2.276

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

Beatty-Martinez, A. L., Guzzardo Tamargo, R. E., & Dussias, P. E. (2021). Phasic pupillary responses reveal differential engagement of attentional control in bilingual spoken language processing. *Scientific Reports, 11*(1), 1–12. https://doi.org/10.1038/s41598-021-03008-1

Bianchi, F., Wendt, D., Wassard, C., Maas, P., Lunner, T., Rosenbom, T., & Holmberg, M. (2019). Benefit of higher Maximum force output on listening effort in bone-anchored hearing system users: A pupillometry study. *Ear and Hearing, 40*(5), 1220–1232. https://doi.org/10.1097/AUD.0000000000000699

Birren, J. E., Casperson, R. C., & Botwinick, J. (1950). Age changes in pupil size. *Journal of Gerontology, 5*(3), 216–221. https://doi.org/10.1097/geronj.53.3.216

Boswijk, V., Loerts, H., & Hilton, N. H. (2020). Salience is in the eye of the beholder: Increased pupil size reflects acoustically salient variables. *Ampersand, 7*, 100061. https://doi.org/10.1016/j.amper.2020.100061

Bradshaw, J. L. (1967). Pupil size as a measure of arousal during information processing. *Nature, 216*, 515–516.

Bradshaw, J. L. (1969). Background light intensity and the pupillary response in a reaction time task. *Psychonomic Science, 14*(6), 271–272. https://doi.org/10.3758/BF03329118

Chambers, J. M., & Hastie, T. J. (Eds.). (1992). *Statistical Models in S. AT&T Bell Laboratories*.

Duchowski, A. T., Biele, C., Niedzielska, A., Krejtz, K., Krejtz, I., Kiefer, P., Raubal, M., & Giannopoulos, I. (2018). The Index of Pupillary activity: Measuring cognitive load vis-à-vis task difficulty with pupil oscillation. *Conference on Human Factors in Computing Systems - Proceedings, 2018-April*, Paper 282. https://doi.org/10.1145/3173574.3173856

Duchowski, A. T., Krejtz, K., Gehrer, N. A., Bafna, T., & Bakgaard, P. (2020). The Low / High Index of Pupillary Activity. *CHI ’20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–12*. https://doi.org/10.1145/3313831.3376394

Duchowski, A. T., Medlin, E., Cournia, N., Murphy, H., Gramopadhye, A., Nair, S., Vorah, J., & Melloy, B. (2002). 3-D Eye movement analysis. *Behavior Research Methods, Instruments, and Computers, 34*(4), 573–591. https://doi.org/10.3758/BF03195486

Ganea, D. A., Bexter, A., Güntner, M., Gardères, P., Kampa, B. M., & Haiss, F. (2020). *Pupillary Dilations of Mice Performing a Vibrotactile Discrimination Task React Task Engagement and Response Confidence*. 14(September), 1–14. https://doi.org/10.3389/fnbeh.2020.00159

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

Granholm, E., Asarnow, R. F., Sarkin, A. J., & Dykes, K. L. (1996). Pupillary responses index cognitive resource limitations. *Psychophysiology, 33*(4), 457–461. https://doi.org/10.1111/j.1469-8986.1996.tb01071.x

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

Hastie, T., & Tibshirani, R. (1990). Exploring the nature of covariate effects in the proportional hazards model. *International Biometric Society, 46*(4), 1005–1016. https://doi.org/10.2307/2532444

Hess, E. H., & Polt, M. J. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science, 143*(3611), 1190–1192. https://doi.org/10.1126/science.143.3611.1190

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). Eye tracking: A comprehensive guide to methods and measures. Oxford University Press.

Huizser, S., Verkaik, M., van Vugt, M. K., & Taatgen, N. A. (2020). Captivated by thought: “sticky” thinking leaves traces of perceptual decoupling in task-evoked pupil size. *PLoS ONE, 15*(20), e0243532. https://doi.org/10.1371/journal.pone.0243532

Hyöniä, J., Tommala, J., & Alaja, A. M. (1995). Pupil dilation as a measure of higher Maximum force output on listening effort in bone-anchored hearing system users: A pupillometry study. *Ear and Hearing, 40*(5), 1220–1232. https://doi.org/10.1097/AUD.0000000000000699

Jaroddzka, H., & van de Weijer, J. (2011). Eye tracking: A comprehensive guide to methods and measures. Oxford University Press.

Joshi, S., & Gold, J. I. (2020). Pupil size as a window on neural substrates of cognition. *Trends in Cognitive Sciences, 24*(6), 466–480. https://doi.org/10.1016/j.tics.2020.03.005
Behavior Research Methods, 51(2), 865–878. https://doi.org/10.3758/s13428-018-1134-4

Relano-Iborra, H., & Baekgaard, P. (2020). PUPILS pipeline: A flexible Matlab toolbox for eyetracking and pupillometry data processing. 1–4. http://arxiv.org/abs/2011.05118.

Schiff, J. M., & Foa, P. (1874). La pupille considéré comme esthésiomètre. Mareselle Medical, 2, 736–741.

Simpson, H. M., & Hale, S. M. (1969). Pupillary changes during a decision-making task. Perceptual and Motor Skills, 29(2), 495–498. https://doi.org/10.2466/pms.1969.29.2.495

Steiner, G. Z., & Barry, R. J. (2011). Pupillary responses and event-related potentials as indices of the orienting reflex. Psychophysiology, 48(12), 1648–1655. https://doi.org/10.1111/j.1469-8986.2011.01271.x

Steinhauer, S. R., Siegle, G. J., Condray, R., & Pless, M. (2004). Sympathetic and parasympathetic innervation of pupillary dilation during sustained processing. International Journal of Psychophysiology, 52(1), 77–86. https://doi.org/10.1016/j.ijpsycho.2003.12.005

Teigen, K. H. (1994). Yerkes-Dodson: A law for all seasons. Theory & Psychology, 4(4), 525–547. https://doi.org/10.1177/0959343494004004

Trani, A., & Verhaeghen, P. (2018). Foggy windows: Pupillary responses during task preparation. Quarterly Journal of Experimental Psychology, 71(10), 2235–2248. https://doi.org/10.1177/1747021817740856

Tryon, W. W. (1975). Pupillometry: A survey of sources of variation. Psychophysiology, 12(1), 90–93. https://doi.org/10.1111/j.1469-8868.1975.tb03068.x

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

van Rij, J., Hendriks, P., van Rijn, H., Baayen, R. H., & Wood, S. N. (2019). Analyzing the time course of pupillometric data. Trends in Hearing, 23, 1–22. https://doi.org/10.1177/2331216518832483

Wendt, D., Dau, T., & Hjortkjær, J. (2016). Impact of background noise and sentence complexity on processing demands during sentence comprehension. Frontiers in Psychology, 7:345, 1–12. https://doi.org/10.3389/fpsyg.2016.00345

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

Wetzel, N., BUTTELMANN, D., SChIeler, A., & WiDDmann, A. (2016). Infant and adult pupill Dalation in response to unex Pected sounds. Developmental Psychobiology, 58(3), 382–392. https://doi.org/10.1002/dev.21377

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

Winn, M. B., Edwards, J. R., & Litovsky, R. Y. (2015). The impact of auditory spectral resolution on listening effort revealed by pupil dilation. Ear and Hearing, 36(4), e153–e165. https://doi.org/10.1097/AUD.0000000000001445

Winn, M. B., Wendt, D., Koelewijn, T., & Kuchinsky, S. E. (2018). Best practices and advice for using pupillometry to measure listening effort: An Introduction for those who want to get started. Trends in Hearing, 22, 1–32. https://doi.org/10.1177/2331216518800869

Xu, J., Wang, Y., Chen, F., & Choi, E. (2011). Pupillary Response Based Cognitive Workload Measurement under Luminance Changes. In: Campos, P., Graham, N., Jorge, J., Nunes, N., Palanque, P., Winckler, M. (eds) Human-Computer Interaction – INTERACT 2011. Lecture Notes in Computer Science (vol. 6947, pp. 178–185). Springer.

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. Journal of Comparative Neurology and Psychology, 18, 459–482.

Zekveld, A. A., Koelewijn, T., & Kramer, S. E. (2018). The pupil dilation response to auditory stimuli: Current state of knowledge. Trends in Hearing, 22, 1–25. https://doi.org/10.1177/2331216518777174

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2011). Cognitive load during speech perception in noise: The influence of age, hearing loss, and cognition on the pupil response. Ear and Hearing, 32(4), 498–510. https://doi.org/10.1097/AUD.0b013e31820512bb

Appendix

The mixed-effects model used to analyze the pre-task baseline included the block order, the trial order within a block, the task difficulty (i.e., the SNR), as well as the noise type and recording session, with listener evaluated as a random effect. A quadratic effect was assumed for all fixed effects in the maximal model before model reduction, and all interactions between main effects were also included in this base model. Such that:

\[
\text{baseline}_i = \theta_0 + \sum_{m} X_m \cdot \theta_{i,m} + u_i + \epsilon_i
\]

Where, for each \( i \)th observation, \( X \) is the model design vector, that includes all \( m \) fixed effects, \( \theta_0 \) is the intercept, \( \theta_{i,m} \) corresponds to the estimates of the \( m \)th fixed effect, \( u \) are the estimates of each individual listener (i.e., the random effect) and \( \epsilon \) is the residual error of the model. Extending the equation to describe the maximal model before model reduction, and all interactions between main effects were also included in this base model. Such that:

\[
\text{baseline}_{i,j} = \theta_{0i} + \theta_{1i} \cdot SNR + \theta_{2i} \cdot noise + \theta_{3i} \cdot block + \theta_{4i} \cdot trial + \theta_{5i} \cdot session + \theta_{6i} \cdot SNR^2 + \theta_{7i} \cdot noise^2 + \theta_{8i} \cdot block^2 + \theta_{9i} \cdot trial^2 + \theta_{10i} \cdot session^2 + \theta_{11i} \cdot SNR \cdot noise + \theta_{12i} \cdot SNR \cdot block + \theta_{13i} \cdot SNR \cdot session + \theta_{14i} \cdot SNR \cdot session + \theta_{15i} \cdot noise \cdot block + \theta_{16i} \cdot noise \cdot trial + \theta_{17i} \cdot noise \cdot session + \theta_{18i} \cdot block \cdot trial + \theta_{19i} \cdot block \cdot session + \theta_{20i} \cdot trial \cdot session + u_{i,j} + \epsilon_i
\]
With the random effects assumed to follow a normal distribution such that:

\[ u \sim N(0, \sigma_u^2) \]

And similarly for the residual error as well:

\[ \varepsilon \sim N(0, \sigma^2) \]

After the model reduction described in the main text the final model used to evaluate the effects of experimental factor on the baseline is:

\[
baseline_{ij} = \theta_{0i} + \theta_{1i} \cdot SNR + \theta_{2i} \cdot noise + \theta_{3i} \cdot block \\
+ \theta_{4i} \cdot trial + \theta_{5i} \cdot session + \theta_{6i} \cdot SNR^2 \\
+ \theta_{7i} \cdot block^2 + \theta_{8i} \cdot trial^2 + \theta_{9i} \cdot SNR \cdot noise \\
+ \theta_{10i} \cdot SNR \cdot block + \theta_{11i} \cdot noise \cdot block \\
+ \theta_{12i} \cdot block \cdot session + \theta_{13i} \cdot trial \cdot session \\
+ u_{ij} + \varepsilon_i
\]