Review Article

Trends in Hearing
Volume 27: 1–20
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/23312165231153288
journals.sagepub.com/home/tia
SAGE

Investigating the Reliability of Pupillometry as a Measure of Individualized Listening Effort

Mihaela-Beatrice Neagu[1], Abigail A. Kressner[1,2], Helia Relaño-Iborra[1,3], Per Bækgaard[3], Torsten Dau[1,2] and Dorothea Wendt[1,4]

Abstract

Recordings of the pupillary response have been used in numerous studies to assess listening effort during a speech-in-noise task. Most studies focused on averaged responses across listeners, whereas less is known about pupil dilation as an indicator of the individuals' listening effort. The present study investigated the reliability of several pupil features as potential indicators of individual listening effort and the impact of different normalization procedures on the reliability. The pupil diameters of 31 normal-hearing listeners were recorded during multiple visits while performing a speech-in-noise task. The signal-to-noise ratios (SNRs) of the stimuli ranged from −12 dB to +4 dB. All listeners were measured twice at separate visits, and 11 were re-tested at a third visit. To examine the reliability of the pupil responses across visits, the intraclass correlation coefficient was applied to the peak and mean pupil dilation and to the temporal features of the pupil response, extracted using growth curve analysis. The reliability of the pupillary response was assessed in relation to SNR and different normalization procedures over multiple visits. The most reliable pupil features were the traditional mean and peak pupil dilation. The highest reliability results were obtained when the data were baseline-corrected and normalized to the individual pupil response range across all visits. Moreover, the present study results showed only a minor impact of the SNR and the number of visits on the reliability of the pupil response. Overall, the results may provide an important basis for developing a standardized test for pupillometry in the clinic.

Keywords

pupillometry, listening effort, speech-in-noise, reliability

Received 16 February 2022; Revised 25 December 2022; accepted 9 January 2023

Introduction

Listening effort has been a growing topic in the auditory field over the last couple of decades. It is often defined as "the deliberate allocation of mental resources to overcome obstacles in goal pursuit when carrying out a task, with listening effort applying more specifically when tasks involve listening" (Pichora-Fuller et al., 2016). Among different measures of listening effort, pupillometry, that is, tracking of the pupil's size, has been recognized to be the "most useful autonomic indication" of effort (Kahneman, 1973). It has been shown that the pupil response is regulated by the autonomic nervous system (Wang et al., 2016; Bremner, 2009; May et al., 2019) which plays an important role in maintaining stability and balance in the body. Its activity consists of both sympathetic and parasympathetic responses (Loewenfeld, 1993; Wang et al., 2018). The relative contribution of sympathetic versus parasympathetic activity to the pupil response, however, can vary as a function of cognitive activity. Reimer et al. (2016) suggested that nonluminance-related

[1]Department of Health Technology, DTU Hearing Systems, Denmark
[2]Copenhagen Hearing and Balance Centre, Rigshospitalet, Copenhagen University Hospital, Denmark
[3]Department of Applied Mathematics and Computer Science, DTU Cognitive systems, Denmark
[4]Eriksholm Research Centre, Denmark

Corresponding author:
Mihaela-Beatrice Neagu, Department of Health Technology, DTU Hearing Systems, Denmark.
Email: mnea@dtu.dk

changes in pupil size might be determined by the locus coeruleus (LC), which has been shown to be a noradrenergic source for the cortex (Aston-Jones & Cohen, 2005; Carter et al., 2010; Jones, 2004; Lee & Dan, 2012). Several studies showed that LC activity determines parasympathetic inhibition, which thereby causes inhibition of the constrictor muscle of the pupil and ultimately leads to a dilation of the pupil (Eckstein et al., 2017; Wang et al., 2016). Furthermore, pupillometry has been demonstrated to provide a measure of listening effort during speech-in-noise tests both in normal-hearing (NH) and hearing-impaired (HI) listeners (Kramer et al., 1997; Zekveld et al., 2010, 2011; Koelewijn et al., 2012). For example, Ohlenforst et al. (2017) demonstrated that HI listeners showed an increased pupil diameter, indicating an increased allocation of resources to reach similar speech intelligibility performance compared to NH listeners. Several studies examined the impact of the level of speech intelligibility, signal-to-noise-ratio (SNR), linguistic complexity and hearing-aid signal processing on listening effort (Zekveld et al., 2011; Kuchinsky et al., 2013, 2014; McGarrigle et al., 2014; Winn, 2016; Wendt et al., 2018). For instance, pupillometry has been shown to be sensitive to changes in the acoustic signal caused by hearing-aid signal processing. Specifically, a reduction in listening effort has been reported with noise-reduction schemes for HI listeners at SNRs reflecting ecologically valid listening situations at a high level of speech intelligibility (Ohlenforst et al., 2017; Wendt et al., 2017). These studies support the hypothesis that a complete characterization of the difficulties in speech understanding arising as a consequence of hearing impairment, and the potential benefit of hearing aid interventions, can be gained when measuring listening effort in addition to speech intelligibility.

So far, pupillometry as a measure of listening effort during a speech-in-noise task has only been evaluated on a listener group level (as averaged responses across listeners), and little is known about the sensitivity and reliability of this method for individual listeners. However, such sensitivity and reliability of the method on an individual listener's level would be crucial for pupillometry to be used as a basis for individualized rehabilitation strategies. The transition from pupillometry assessed on a group level to an individual listener level is challenging because the pupil response has numerous sources of variation (Zekveld et al., 2011, 2018; Koelewijn et al., 2012; Wang et al., 2018; Partala & Surakka, 2003). For example, pupil response is affected by environmental factors, such as luminance, masking noise, or communication technologies (e.g., hearing aids). Furthermore, listener-specific factors, such as cognitive abilities, hearing impairment, or the level of fatigue, can affect the pupil response (Zekveld et al., 2018; Kuchinsky et al., 2016; Wang et al., 2018; Pichora-Fuller et al., 2016).

A few studies investigated the reliability of the pupil response assessed during speech recognition. Alhanbali et al. (2019) explored the reliability of several physiological measures during a digit-in-noise recognition task performed under individualized listening conditions, whereby the level of speech intelligibility performance was fixed at 71%. The authors reported that among the assessed physiological measures, pupillometry (specifically, the mean pupil dilation (MPD) and the peak pupil dilation (PPD) of the response) showed the highest reliability with an intraclass correlation coefficient (ICC > 0.85) as compared to electroencephalogram and skin conductance measurements. Similarly, Giuliani et al. (2020) investigated the sensitivity and reliability of different measures of listening effort (including skin conductance, pupillometry, and self-reported listening effort using a dual-task paradigm). The authors assessed listening effort during sentence recognition at SNR levels of 0, −3, and −5 dB. Consistent with Alhanbali et al. (2019), Giuliani et al. (2020) reported the highest reliability for pupillometry among all tested measures, even though the corresponding level of ICC was only fair (< 0.5). ICC is a reliability index that reflects the degree of agreement between similar measurements. Both studies showed that investigated pupil features were equally reliable to the perceived listening effort measures (NASA Task Load Index —NASA-TLX, Hart & Staveland (1988)) and a self-rated effort question.

These studies focused on the analysis of the MPD and PPD only, following the traditional characterization of the pupil response (Zekveld et al., 2010, 2011; Koelewijn et al., 2012). However, more recently, Kuchinsky et al. (2013) showed that growth curve analysis (GCA) could be used to detect changes in the shape of the pupil response over time, allowing for an independent evaluation of different temporal characteristics of the pupil response (Mirman et al., 2008; Winn et al., 2015). GCA fits orthogonal polynomial terms to time series data to show different variations in the function among individuals (Mirman et al., 2008). Not much is known, though, about the reliability of the traditional or GCA pupil features across multiple visits.

Only a few studies evaluated the reliability of various measures other than pupil features over more than two visits (e.g., psychophysiological measures: intrinsic attentive selection of one of two lateralized visual cues, Aday & Carlson (2019); daytime sleepiness, Zwyghuizen-Doorenbos et al. (1988)). Aday & Carlson (2019) showed that attention biases were not reliable until participants had fairly extensive experience with the task. They suggested that more visits could reduce the noise in the data related to task familiarity and increase reliability. These studies showed, in fact, an increase in the reliability of the tests with an increasing number of visits. However, the reliability of pupillometry assessed within a speech-in-noise task paradigm over multiple visits has not yet been studied. Furthermore, Alamia et al. (2019) and Widmann et al. (2018) showed that the pupil dilates following increased surprise or, more generally, following global arousal and that emotional arousal to novel sounds enhances the sympathetic contribution to the pupil dilation response. Thus, it follows that an arousal effect observed in the pupil response when performing a novel task (i.e., at the first visit) could result in lower reliability of pupillometry between the first and second visit than a comparison between the responses in subsequent visits. A common approach to avoid arousal effects has been to remove the first trials (within a condition) from the analysis and, thus, to reduce the impact of any initial effects (Winn et al., 2018). However, a more general arousal effect (i.e., novel task, novel environment, unknown experimenter) is difficult to control. Thus, the present study investigated the reliability of pupil response over multiple visits.

Furthermore, regarding the changes in the reliability of the pupil response with changing SNR, results differed remarkably across studies. Giuliani et al. (2020) found fair reliability irrespective of their considered SNR changes from 0 to −3 dB and from −3 to −5 dB, respectively. In contrast, other studies suggested that task demands impact reliability such that increasing task demands lead to a higher index of pupillary activity (Duchowski et al., 2018), higher inter-trial change in pupil diameter (Krejtzid et al., 2018) and prospective memory (Einstein et al., 1997).

Finally, different methods of pupil diameter normalization have been proposed in the literature (e.g., Winn et al. (2018)). A common approach when assessing listening effort in a speech-in-noise task paradigm is baseline correction. Baseline-corrected responses represent a change in pupil size relative to a particular temporal window before the stimulus, known as baseline (Winn et al., 2018). However, while some studies argued that the normalization of task-evoked changes in pupil size should be done independently of the baseline pupil size (Beatty, 1982; Bradshaw, 1969), others stated that different ways of baseline scaling could produce disparities in the reported pupil size results (Reilly et al., 2019; Mathôt et al., 2018). Moreover, relatively large interindividual differences in the dynamic range of pupil diameter have been observed, and several other approaches have been proposed to target these differences. For example, Piquado et al. (2010) obtained a dynamic range of pupil response based on changes in the luminance (dark versus light), which was then used for range normalization. Furthermore, McCloy et al. (2016) applied *z*-score transformation and Winn (2016) considered a proportional change within the individual between a reference condition and the task condition. However, the impact of the normalization procedure on the reliability of different pupil features has not been studied.

The present study aimed to obtain a better understanding of the reliability of pupillometry as an objective indicator of an individual's listening effort. Different features of the pupil response, assessed in a speech-in-noise paradigm, were extracted and the impact of task demands (i.e., changing SNR) and data normalization procedures on the reliability of those features were systematically investigated. The test-retest reliability of pupillometry was investigated by assessing pupil response over three visits. It was hypothesized that the reliability of different pupil features would increase with decreasing SNR (i.e., higher task demands). Furthermore, it was hypothesized that the reliability of different pupil features would be affected by applying distinctive normalization procedures.

Methods

Participants

Thirty-five participants (aged from 18 to 65 years, mean 38) took part in this study. All participants were native Danish speakers. They had pure-tone hearing thresholds of 20 dB hearing level (HL) or better at low frequencies (below 6 kHz) in both ears and thresholds of 30 dB HL or better at frequencies above 6 kHz. The participants had no history of eye diseases or eye operations. Exclusion criteria also included caffeine intake less than 3 hours prior to the test time. The data of four participants out of the thirty-five were excluded from the analysis because of their withdrawal from the study after the first visit. The research procedures were approved by the Science-Ethics Committee for the Capital Region of Denmark (reference H-16036391), and all participants provided written informed consent for the study procedures and received monetary compensation for their participation.

Procedure and Stimuli

Participants were asked to perform a speech-in-noise test with sentences from the Danish Hearing in Noise Test (HINT, Nielsen & Dau (2011)). HINT sentences were presented in a 4-talker babble masker, which was created by overlapping two male and two female talkers (all of them reading different excerpts from a newspaper) with the same long-term average frequency spectrum as the HINT sentences. For each measurement trial, the masker onset started 3 s prior to sentence onset and stopped 3 seconds after sentence offset, as the vertical lines in Figure 1 indicate. The length of each trial varied depending on the length of the presented HINT sentence, which has a mean duration of about 1.5 s. After the masker offset, the participants were asked to repeat back the HINT sentence. Two seconds of silence were established before noise onset to allow for the pupil to return to pre-task levels (i.e., recovery). Sentences were presented at five different SNRs: 4, 0, −4, −8 and −12 dB. Different conditions were presented in a block design with 25 trials containing 25 sentences for each SNR. Trials were randomized within each block, and the presentation order of each condition was randomized across participants. The stimuli were presented through Sennheiser HD650 headphones using an sound pressure level (SPL) Audio Phonitor Mini amplifier. The noise level was fixed to a SPL of 70 dB for both ears while the level of the target speech varied depending on the SNR.

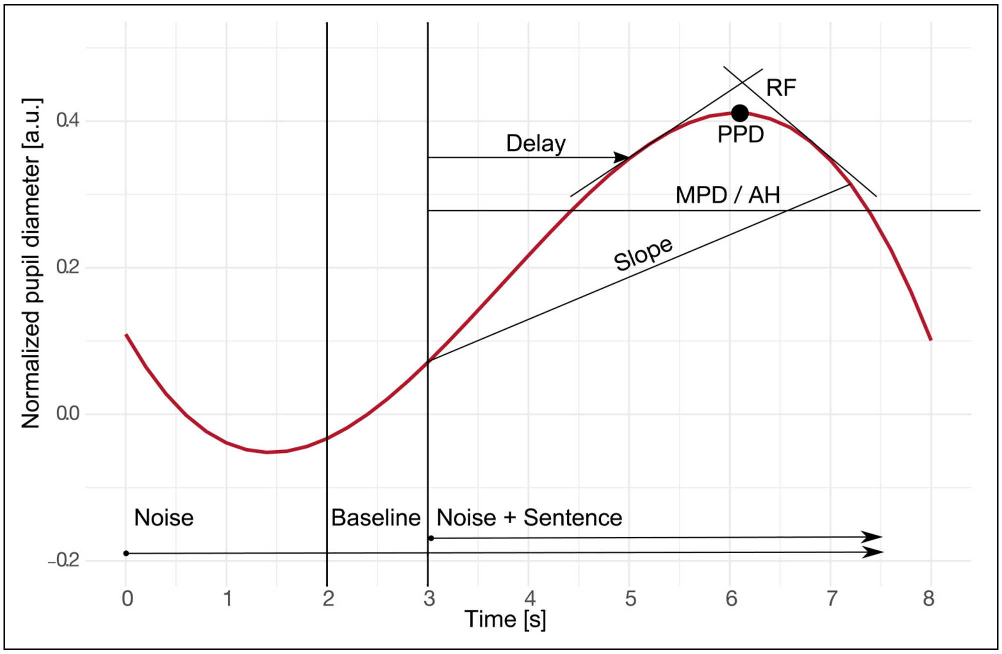

Figure 1. Schematic illustration of the pupil response within the speech-in-noise test with sentence onset at second 3. All analyzed pupil features (traditional and growth curve analysis [GCA] features) are schematically represented.

The participants were instructed to fix their gaze at a grey cross in the middle of a black screen during the speech-in-noise task and to repeat the HINT sentence or as many words as they could recall after the noise offset. The responses were scored on a word-level basis (all recognized words from the sentence were marked as correct). Speech-reception-thresholds (SRTs) were calculated by extracting the corresponding SNR value recorded at 50% correct performance by fitting a psychometric function to the data of each participant.

The participants were tested at two different visits (Visits 1 and 2) using a repeated measures design. Eleven out of the thirty-one participants were re-tested additionally at a third visit (Visit 3). The visits were spaced three to six weeks apart to avoid any learning effects of the sentence material (Bramsløw et al., 2016). The subsequent visits were scheduled at the same time of the day and at the same period of the week (i.e., beginning, middle, or end) as for Visit 1 to minimize the potential effect of fatigue at different times during a day or at different days of the week and to control for circadian rhythm effects (Daguet et al., 2019). The procedure was the same at the second and third visits with the same presentation order of the conditions and the same sentences but in a different order per condition for each of the listeners.

Apparatus and Pupillometry Data Processing

Eye-tracking data were continuously recorded during the speech-in-noise test using a desktop-mounted eye-tracker (EyeLink 1000; SR-Research Ltd., Mississauga, Ontario, Canada). Pupil sizes were recorded from the left eye with a sampling frequency of 500 Hz. The measurements were performed in the same booth with the same luminosity levels across visits (screen and ambient light). The screen's luminance and ambient light were controlled to prevent any changes in pupil response that could be attributed to changes in ambient or screen light intensity. The ambient light was measured at 75 lx for the tasks performed in light. The screen had an approximate brightness of $9\,\mathrm{cd}/m^2$ during the speech-in-noise task, where the screen displayed a black background with a grey cross in the middle. The distance from the middle of the participant's eyes to the center of the screen varied between 50 and 70 cm, as a maximum acceptable distance variation criterion within the remote mode usage of the Eye-link equipment, with 16 mm lens aperture size. Participants were asked to keep their position fixed during the eye-tracking recordings.

The pupil data were processed using (MATLAB, 2018) and R (R Core Team, 2019). In order to remove any initial arousal effects, the pupil traces of the first three trials within a block were excluded from the analysis. Since a decreasing trend of the pupil within each block was observed, the entire block recording was linearly detrended. For the eye-blink removal, the MPD with standard deviation was calculated across the whole trial. Pupil diameter values more than three standard deviations smaller than the mean were coded as eye-blinks. Eye-blinks were removed by a linear interpolation that started about 80 ms before and ended 150 ms after the blinks. Data were then smoothed using a moving average filter with a symmetric rectangular window of 117 ms. Trials with more than 20% missing data, eye blinks or artifacts were removed from the analysis. All remaining traces were scaled using each of the four normalization procedures presented in Section Data Normalization below.

Perceived Effort

After each block, participants were asked to answer the NASA-TLX (Hart & Staveland, 1988) questionnaire to assess a measure of the perceived listening effort. The NASA-TLX uses a 0-20 scale (low/high). NASA-TLX has six subitems: mental demand, physical demand, temporal demand, perceived performance, effort, and frustration. The score was calculated as a mean score of each of the subitems. Additionally, another measure of a self-rated listening effort was provided by each participant after each SNR block. On a 0-to-10 scale (with 0 indicating low effort and 10 indicating high effort), participants were asked to answer the following question: "Hvor meget anstrengte du dig for at høre sætningerne?" which translates to English as, "How much effort did you put into hearing the sentences?". Both measurements were rescaled to a 0-1 scale for further analysis.

Data Normalization

Four different normalization procedures were applied. First, baseline correction (equation 1) was applied at a trial level by subtracting the mean pupil size measured in the 1 s period preceding the sentence onset from each data point of the trial.

$$x_{baseline\ corrected} = x - \mu_{baseline} \tag{1}$$

where x is the pupil diameter at a given sample, and μ is referring to the mean pupil size within the baseline time window (i.e., between the second and third second). The baseline was established 1 s prior to the sentence onset, as recommended by Winn et al. (2018).

Alternatively, a range normalization procedure was applied for each participant for each trial. The pupil range was calculated by extracting the maximum and the minimum pupil diameter across all trials of all conditions and visits for each participant. All trials were then range normalized at a trial level (equation 2).

$$x_{range} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{2}$$

where x is the pupil diameter at a given sample and x_{max} and x_{min} refer to the overall maximum and minimum pupil diameter over all trials and visits.

As another option, a z-score normalization was applied at a trial level after the aggregates, mean and standard deviation were computed at a participant level. The z-score subtracts the MPD for an individual from each pupil sample and divides the result by the standard deviation of the MPD (equation 3).

$$xz - score = \frac{x - \mu}{\sigma} \tag{3}$$

where x is the pupil diameter at a given sample, μ refers to the mean of the pupil diameter per individual and σ to the standard deviation of the pupil diameter per individual.

Finally, a range normalization procedure (equation 4) was applied on the baseline corrected data, for each trial, using formulas (equation 1) and (equation 2) and was referred to here as "baseline range" normalization. The individual traces were, therefore, firstly baseline corrected (i.e., subtraction of the MPD in the baseline time window) and afterward, the range normalization procedure was applied, at the individual trial level. The maximum and minimum pupil size values of the range were extracted per participant across all conditions and visits.

$$x_{baseline\ range} = \frac{x_{baseline\ corrected} - x_{min}}{x_{max} - x_{min}} \tag{4}$$

Feature Extraction

The MPD was calculated as the average pupil diameter in the interval between sentence onset and masker offset (see Figure 1 arrow Noise+Sentence), after the traces were averaged across all trials within a block. The PPD was calculated as the maximum dilation in the same interval, similarly, after pupil traces were averaged within a block.

In order to account for effects reflected in the time-course of the pupillary response, GCA was applied (Mirman et al., 2008). GCA is a multilevel regression technique that fits orthogonal polynomials to time course data. A third-order (cubic) orthogonal polynomial was applied to the overall time course of the pupil diameter within a time window starting at 2 s (i.e., at the baseline onset) until 8 s of stimulus presentation (see Figure 1). A third-order polynomial function, including the intercept through cubic terms, was considered to provide a good fit to the shape of the pupil response across time (Kuchinsky et al., 2014, 2016). The model was applied on a trial-by-trial basis, with the trial representing the random effect, and the estimates were calculated on a block-level. Examples of modelled pupil responses using GCA are provided in Figure 6 in the Supplemental material. The feature extraction is described in (equation 5). Pupil size was considered as a dependent variable in the model, predicted by a series of fixed and random effects (individual and trial number, respectively).

$$\begin{aligned} \text{pupilfeature} \sim (1 + \mathrm{p}_1 + \mathrm{p}_2 + \mathrm{p}_3) * \text{participant} \\ + \left(1 + p_1 + p_2 + p_3 | \text{trial}\right) \end{aligned} \tag{5}$$

A schematic representation of the GCA features can be seen in Figure 1. The intercept term represents the average height (AH) of the pupil response, the linear term (p_1) reflects the slope, the quadratic term (p_2) reflects the rise and fall (RF) around the central inflexion point of the response function, and the cubic term (p_3) reflects the inflexions at the extremities of the curve, referred to as delay in the current study.

Reliability Analysis

The reliability of the pupil features was assessed using Spearman's correlation coefficient, which reveals how consistent the results are across the different visits, as well as the ICC, which evaluates the test-retest reliability (Cicchetti, 1994; Koo & Li, 2016). Sperman's correlation sorts the observations by rank and evaluates how similar the ranks are. Their values lie between −1 and 1 with 1 indicating strong relationship. Sperman's correlation coefficient is calculated as in equation 6.

$$\mathrm{Spearman}_{\mathrm{coef}} = \frac{\mathrm{Cov}(\mathrm{rank}_{V_1}, \mathrm{rank}_{V_2})}{\sigma_{V_1}\sigma_{V_2}} \quad (6)$$

where $Cov(\mathrm{rank}_{V_1}, \mathrm{rank}_{V_2})$ are the covariances between the ranks of the pupil measures at Visit 1, respectively Visit 2, while σ refers to the standard deviation of the same ranks.

The ICC assesses the group reliability by comparing the variability within different visits of the same participant's pupil diameter to the total variation across all visits and all participants. Here, the ICC was calculated to evaluate the reliability of different features of the pupil response (see subsection *ICC - Pupil Features* from the *Results* section) between Visits 1 and 2 for 31 participants, and between Visits 2 and 3 for the subgroup of 11 participants who came for a third visit. The latter was compared to the ICC values measured for the same 11 participants between Visits 1 and 2.

The ICC was calculated as a two-way mixed-effects model with two measurements, as reflected in (equation 7), where MS_B is the mean square between subjects, MS_C is the mean square between trials, MS_E is the mean square error, n is the number of subjects and k is the number of measurements.

$$\mathrm{ICC}_{\mathrm{agreem}} = \frac{\mathrm{MS_B} - \mathrm{MS_E}}{\mathrm{MS_B} + (\mathrm{k}-1)\mathrm{MS_E} + \frac{\mathrm{k}}{\mathrm{n}}(\mathrm{MS_T} - \mathrm{MS_E})} \quad (7)$$

To assess the test-retest reliability between two visits, ICC was calculated for each combination of normalization technique (i.e., baseline correction, range normalization, z-score, baseline range normalization) and feature (i.e., PPD, MPD, and GCA features), and between Visits 1 and 2, and Visits 2 and 3, for all combinations of normalization type and pupil feature.

Results

Group Average Data

Although this study focused on the reliability of individual's pupil features, a group-level analysis was conducted first to provide an anchor to previous literature and to gauge group-level reproducibility. The pupil traces shown in Figure 2 for the different normalization procedures were first averaged across trials at the participant level. Thereafter, the single-subject average traces were averaged across listeners for each condition and visit to form the group average. Overall, it can be seen that the general trend of increasing pupil response with decreasing SNR remains, regardless of the normalization procedure. Figure 2 shows that the pupil response changes with varying SNR on a group level. These results were in line with results from previous

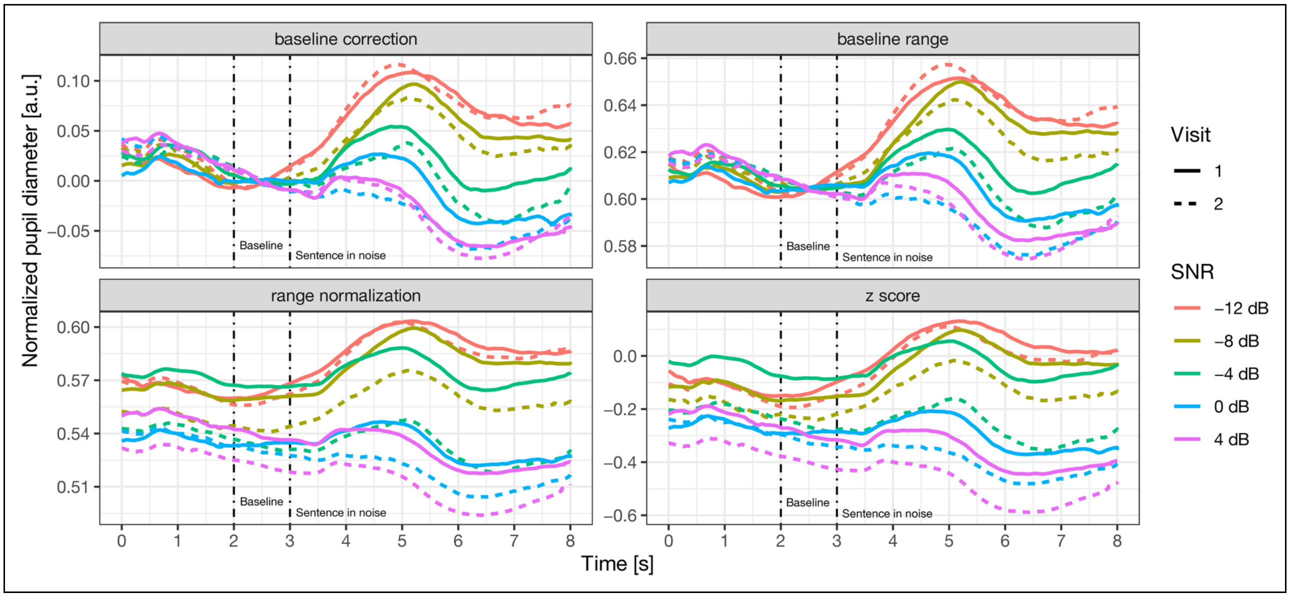

Figure 2. Pupil traces averaged across participants, normalized with different procedures. The SNRs tested are presented in different colors and the Visits are presented in different line types. SNR = signal-to-noise ratio.

studies (Wendt et al., 2018), showing that SNR manipulation impacts pupil dilation in an auditory task (Zekveld et al., 2010). Furthermore, the difficulty level manipulated through the SNR was compared to the individual listeners' task performance and the individuals' perceived effort measures using a Pearson correlation analysis. By visually inspecting the traces, it appears that larger differences between the two visits occur in the z-score (low right panel) and range normalization (low left panel) procedure compared to the other two, especially for −8, −4 and 4 dB SNR. A quantitative analysis of these differences will be provided below on an individual level (Results subsections *Consistency across visits and normalization procedures* and *ICC - Pupil Features*).

Group Level Pupil Features Across Visits and SNRs

The six different pupil features extracted from the group-averaged, baseline-corrected pupil response are displayed in Figure 3 for all three visits and all five SNRs (−12, −8, −4, 0, and 4 dB). The visits are presented in different colors, such that the figure depicts how the distribution over each feature varies as a function of SNR and visit. All features except delay showed a slightly decreasing trend with increasing SNR. An increasing trend of delay with increasing SNR indicates that the peak dilation is reached later with increasing SNR.

A two-way ANOVA was performed to investigate the impact of SNR and visit on each pupil feature for each normalization procedure. The two factors considered within this method were SNR and Visits. The SNR factor had five levels (−12, −8, −4, 0, 4 dB), while the Visit factor consisted of three levels (Visits 1, 2, 3). The results are displayed in Table 1. Significant effects are highlighted in bold. After correcting for family-wise Type 1 error by conducting Bonferroni correction, no significant impact of the visit number on the group-level analysis for any of the pupil features except delay was found, suggesting that average features were reliable across multiple visits. There was an effect of SNR for some of the features when certain normalization procedures were applied, that is, slope, RF, and delay for all normalization procedures and MPD only for baseline correction and range normalization ($p - \text{value} < 0.05$). Interestingly, a significant effect of SNR on PPD occured only for some normalization procedures for high SNRs (i.e., baseline range −4 to 4 dB SNR and range normalization 0 to 4 dB SNR).

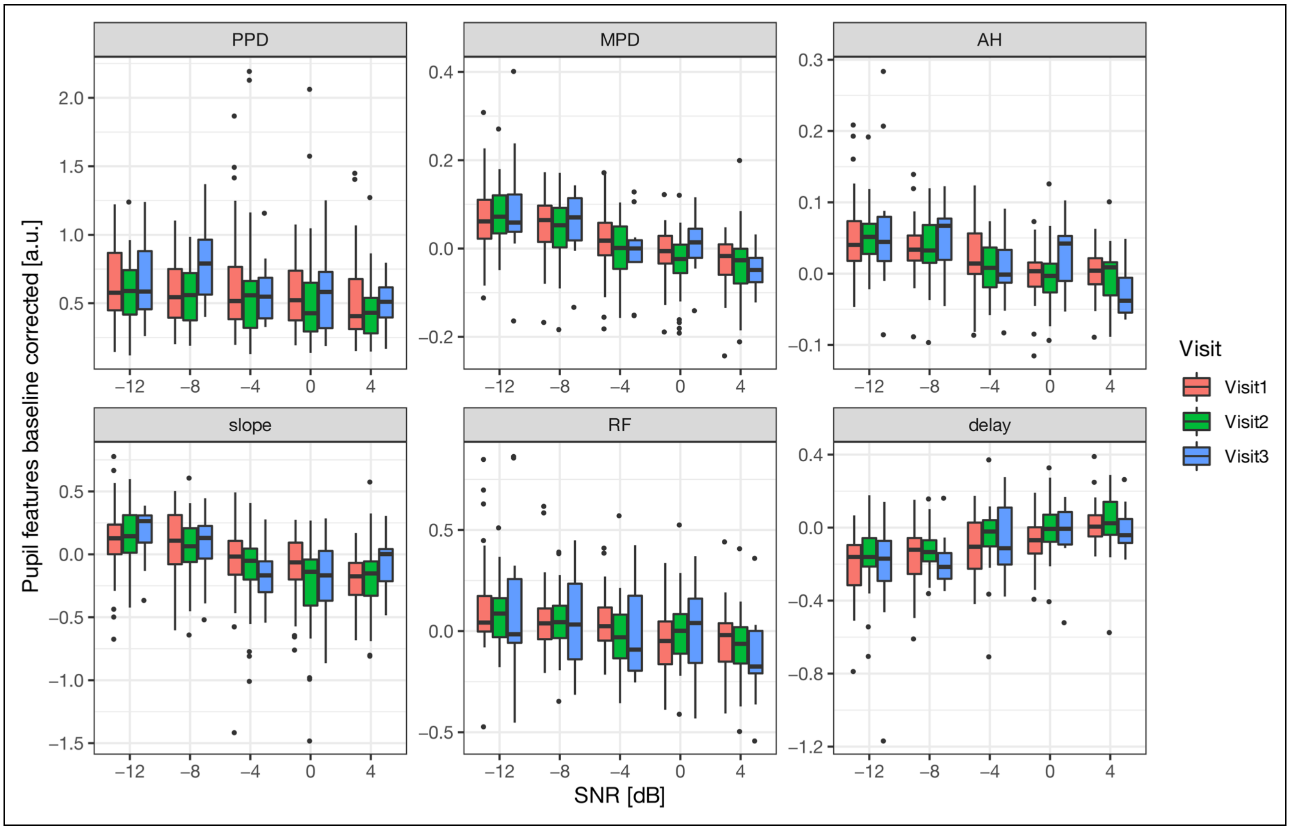

Figure 3. Boxplots of the pupil features PPD, MPD, AH, slope, RF, and delay indicated in the different panels are shown as a function of SNR for three different visits (Visits 1, 2, and 3) indicated by different colors. The mid-line of the boxes represents the median values while the vertical line is the standard deviation. PPD = peak pupil dilation; MPD = mean pupil dilation; AH = average height; RF = rise and fall; SNR = signal-to-noise ratio.

Table 1. Estimated Coefficients of the Model Obtained when Applying a Two-way ANOVA to Investigate the Effect of SNR and Visit on Different Pupil Features for Different Normalization Procedures.

Normalization	Parameters	PPD	MPD	AH	Slope	RiseFall	Delay
Baseline correction	Intercept	**0.622** ***	**0.080** ***	**0.057** ***	**0.145** ***	**0.128** ***	**−0.211** ***
	Visit2	−0.034	−0.011	−0.005	−0.033	−0.021	0.047 *
	Visit3	0.038	0.002	0.001	−0.003	−0.018	−0.002
	−8 dB	−0.006	**−0.028** *	**−0.016** *	−0.061	−0.058	0.041
	−4 dB	0.013	**−0.072** ***	**−0.04** ***	**−0.242** ***	**−0.107** **	**0.114** ***
	0 dB	−0.054	**−0.095** ***	**−0.052** ***	**−0.312** ***	**−0.128** ***	**0.156** ***
	4 dB	**−0.117** *	**−0.112** ***	**−0.061** ***	**−0.311** ***	**−0.185** ***	**0.21** ***
Range normalization	Intercept	**0.653** ***	**0.621** ***	**0.65** ***	**0.099** ***	**0.06** ***	**−0.116** ***
	Visit2	−0.022	−0.023	−0.03	−0.027	−0.007	**0.026** **
	Visit3	0.007	0.004	−0.004	−0.015	−0.009	0.009
	−8 dB	−0.012	−0.012	−0.015	−0.033	−0.027	0.019
	−4 dB	−0.034	−0.034	−0.028	**−0.125** ***	**−0.059** ***	**0.063** ***
	0 dB	**−0.065** **	**−0.065** **	−0.014	**−0.165** ***	**−0.067** ***	**0.084** ***
	4 dB	**−0.077** **	**−0.077** **	−0.027	**−0.17** ***	**−0.089** ***	**0.11** ***
z-score	Intercept	**2.382** ***	**0.23** ***	0.077	**0.641** ***	**0.415** ***	**−0.754** ***
	Visit2	0.019	**−0.051** *	**−0.143** *	−0.153	−0.054	**0.144** *
	Visit3	−0.061	−0.012	−0.08	−0.0512	−0.058	0.049
	−8 dB	0.023	−0.066	0.001	−0.251	**−0.21** *	**0.163** *
	−4 dB	0.015	**−0.239** ***	−0.016	**−0.796** ***	**−0.382** ***	**0.408** ***
	0 dB	−0.051	**−0.294** ***	−0.004	**−1.063** ***	**−0.465** ***	**0.542** ***
	4 dB	−0.059	**−0.326** ***	−0.001	**−1.086** ***	**−0.625** ***	**0.742** ***
Baseline range	Intercept	**0.626** ***	**0.588** ***	**0.581** ***	**0.105** ***	**0.066** ***	**−0.127** ***
	Visit2	−0.006	−0.008	−0.004	−0.029	−0.01	**0.027** **
	Visit3	0.003	0.004	−0.009	−0.016	−0.013	0.0129
	−8 dB	−0.011	−0.011	0.004	−0.031	−0.028	0.02
	−4 dB	**−0.039** ***	**−0.039** ***	−0.026	**−0.129** ***	**−0.062** ***	**0.068** ***
	0 dB	**−0.051** ***	**−0.052** ***	0.007	**−0.175** ***	**−0.073** ***	**0.092** ***
	4 dB	**−0.06** ***	**−0.059** ***	0.006	**−0.178** ***	**−0.096** ***	**0.122** ***

The intercept is represented by Visit 1, −12 dB SNR. Significant effects are highlighted in bold ($p< .05$ *, $p< .01$ **, $p< .001$. PPD = peak pupil dilation; MPD = mean pupil dilation; AH = average height; ANOVA = analysis of variance; SNR = signal-to-noise ratio).

Validity of the Pupil Measures

In order to assess the validity of all pupil features as markers of listening effort, Pearson correlations between the pupil features and the SNR, the perceived effort measures, and the performance in the task were analyzed (Table 2). All significant values were corrected for family-wise Type 1 error by conducting Bonferroni correction. SNR and task performance were negatively correlated with most of the pupil features. At the same time, a positive correlation was obtained between the perceived effort measures and most of the pupil measures. Note that the correlation coefficients found for the delay (with SNR, perceived effort and task performance) are against the hypothesized direction since a more delayed peak of the response is expected for more unfavorable SNRs (Wendt et al., 2018; Kuchinsky et al., 2013) and with higher perceived effort ratings. To compare the strength of the correlations among the different measurements and different pupil features, a conventional interpretation (i.e., absolute values) of the correlations was used, as the directions of the correlations were as expected. The perceived effort measures, both the self-rated effort scale and the NASA-TLX, showed low correlations provided their absolute values with the pupil features (between 0.1 and 0.35), as compared to SNR (between 0.1 and 0.51) and task performance (between 0.15 and 0.5). SNR and performance provided the highest correlations with the pupil features for the baseline correction (Corr = 0.51) and the baseline range normalization procedures (Corr = 0.49). Overall, the correlation coefficient varied depending on the pupil features. Of all pupil features, MPD, delay, and slope provided the highest absolute values of the correlations for the SNR manipulation technique (between 0.3 and 0.51). The average correlation computed across features was found for the SNR in the case of baseline correction (Corr = 0.24), *z*-score (Corr = 0.14), and baseline range (Corr = 0.15).

Speech Recognition Performance

Figure 4 shows the psychometric functions of the performance data (the averaged recognition scores) over the HINT words, averaged across all participants as a function of SNR, for each Visit represented in different colors.

Table 2. Pearson Correlation Coefficients for Different Pupil Features, Calculated through Different Normalization Procedures with SNR, Self-rated Effort Scale, NASA-TLX and Task Performance.

Normalization	Feature	SNR	Self-rated effort scale	NASA-TLX	Task performance
Baseline correction	**Average**	−**0.24**	0.18	0.11	−0.22
	PPD	−**0.18****	0.19**	0.2***	−0.17**
	MPD	−**0.51*****	0.36***	0.18***	−0.47***
	AH	−**0.45*****	0.34***	0.17*	−0.41***
	slope	−**0.45*****	0.32***	0.18***	−0.44***
	RF	−**0.3*****	0.2***	0.11*	−0.27***
	delay	**0.46*****	−0.35***	−0.16**	0.44***
Range normalization	**Average**	−0.12	0.10	0.09	−**0.14**
	PPD	−0.12*	0.12*	0.097	−**0.16****
	MPD	−**0.32****	0.24***	0.15**	−0.31***
	AH	−0.033	0.069	**0.11***	−0.07
	slope	−0.45***	0.34***	0.19***	−**0.46*****
	RF	−**0.3*****	0.2***	0.13*	−0.26***
	delay	**0.48*****	−0.37***	−0.14**	0.45***
z-score	**Average**	−**0.14**	0.11	0.08	−**0.14**
	PPD	−0.037	**0.052**	0.015	−0.025
	MPD	−0.49***	0.41***	0.2***	−**0.5*****
	AH	−0.0042	0.044	**0.085**	−0.02
	slope	−0.45***	0.33***	0.17*	−**0.46*****
	RF	−**0.31*****	0.2**	0.12*	−0.27***
	delay	**0.47*****	−0.35***	−0.12*	0.44***
Baseline range	**Average**	−**0.15**	0.07	0.01	−0.14
	PPD	−**0.32*****	0.098	−0.094	−0.28***
	MPD	−**0.33*****	0.11*	−0.085	−0.3***
	AH	0.034	0.03	**0.088**	0.019
	slope	−0.45***	0.34***	0.19**	−**0.46*****
	RF	−**0.3*****	0.2***	0.12*	−0.27***
	delay	**0.49*****	−0.37***	−0.14*	0.46***

The highest values per row are highlighted in bold. Significant effects are represented as follows: *p*< .05*, *p*< .01**, *p*< .001 ***. Values were corrected for repeated measurements. NASA-TLX = NASA Task Load Index; PPD = peak pupil dilation; MPD = mean pupil dilation; AH = average height; SNR = signal-to-noise ratio; RF = rise and fall.

Psychometric functions are displayed together with the performance data, to extract SRT for each Visit. Participants achieved high recognition performance (100% correct) at the SNRs between 0 and 4 dB. With decreasing SNR (0 to −4 dB), speech recognition dropped to approximately 92%. At lower SNRs, in particular, at −8 dB, the participants were able to perform 62%–80% correctly, while at −12 dB the performance dropped to 25%–40%. Overall, an improvement of 1.6 dB was observed in the SRT from Visits 1 to 3. A paired *t*-test was applied to the SRT values extracted at 50% correct responses to evaluate the learning effects. The results show a significant difference in the performance from Visits 1 to 2 (t = 3.57, p-value = .0012), while no significant difference was found from Visits 2 to 3 (t = 0.0313, p-value = .9756).

Consistency Across Visits and Normalization Procedures

To investigate the impact of the normalization procedure on the consistency of each pupil feature across visits, a Spearman's correlation analysis was performed with each of the pupil features. Spearman's correlation coefficients for Visit 1 versus 2 (31 participants), and Visit 1 versus 2, and 2 versus 3 (11 participants) are shown in Table 3 and the individual correlations are shown in Figures 5. For 31 participants, the highest correlation coefficients between Visit 1 and 2 were observed for two pupil features, MPD and PPD, for three out of four normalization procedures (i.e., for baseline correction, baseline range, and range normalization but not for the *z*-score). From the GCA features, the delay and slope were the most consistent features across the normalization procedures, with correlations above 0.5.

After correcting for family-wise Type 1 error by conducting Bonferroni correction, significant correlations with $p-$values < 0.0001 (***) and $p-$values < 0.001 (**) were found, as indicated in Table 3. The lowest correlation, and even some negative correlations, were observed for the *z*-score normalization procedures (ranging between −0.78 and 0.5). Among all the normalization procedures applied in this study, the baseline-corrected data combined with a

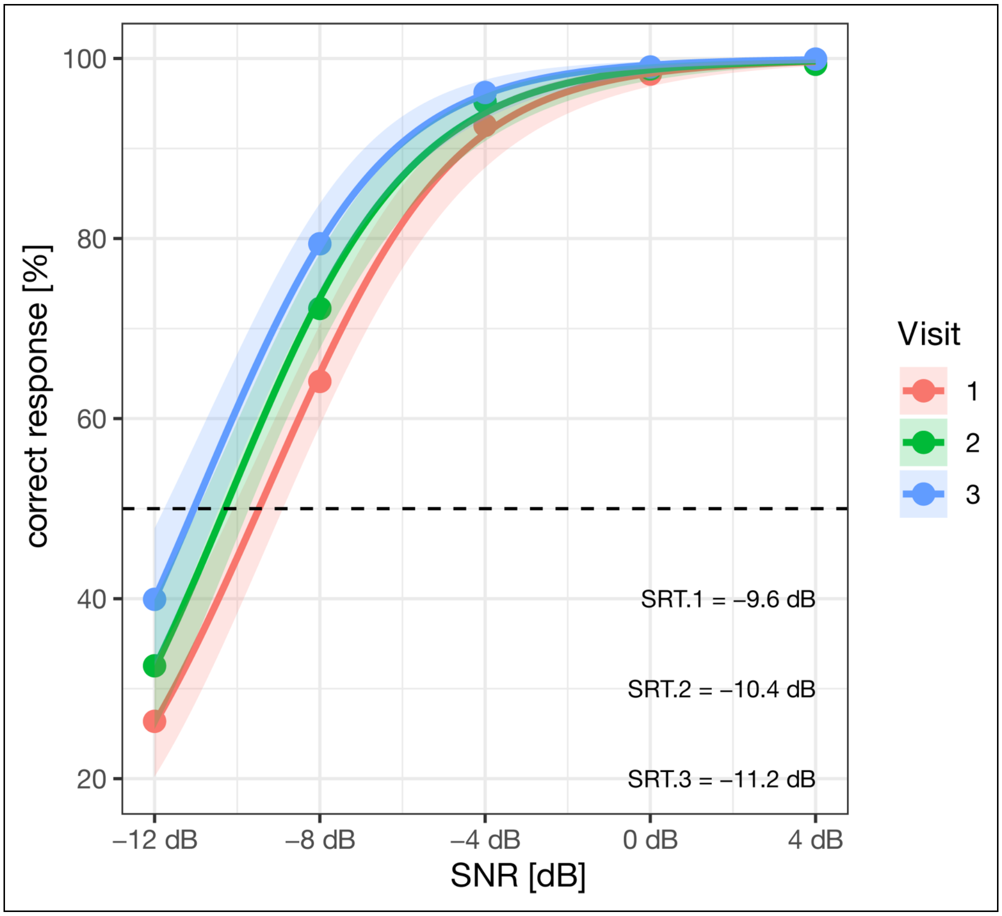

Figure 4. The proportion of correct responses as a function of SNR. Results are averaged across listeners. Measured values are shown as filled circles, while fitted psychometric functions are shown by the corresponding-colored solid functions representing the Visits. Red corresponds to measurements obtained during Visit 1, data from Visit 2 is shown in green and blue is used for data measured during Visit 3. SRT was estimated as the first tested SNR where the confidence interval of the psychometric function exceeded the 50% correct threshold (dotted line). Performance data of 31 listeners were measured in Visit 1 and 2, and of 11 listeners in Visit 3. SNR = signal-to-noise ratio; SRT = speech-reception-threshold

range normalization procedure showed the highest correlations across visits (between $R = 0.43$ and $R = 0.94$).

Due to differences in the sample size (i.e., 31 participants for Visit 1 versus Visit 2 and 11 participants for Visit 2 versus Visit 3), a comparison was also made with the pupil features of the same 11 participants at Visit 1 versus 2. Overall, for MPD and PPD, higher consistency was obtained between Visit 1 and 2 than between Visits 2 and 3 for all normalization procedures except for the baseline range normalization. The GCA features showed no clear trend in consistency across visits. Among GCA features a high correlation was only observed in the delay values for the subsample of 11 participants between both, Visit 1-Visit 2 and Visit 2-Visit 3.

ICC - Pupil Features

To examine the reliability of the pupil features on an individual level, ICC values were calculated with 95% confidence intervals and are summarized in Table 4 for Visits 1 and 2 and in Table 7 from the Supplemental material for the subsample of 11 participants for the three sessions. The results were categorized according to Cicchetti (1994), who defined excellent reliability for ICCs above the value of 0.75 and good reliability for ICCs above 0.6. Good reliability is indicated in bold, while excellent results are highlighted in bold italic in the table. Negative ICC values were truncated to zero.

For all features using baseline correction, the ICC analysis showed good to excellent reliability, with ICC values equal to or greater than 0.6. The ICCs for all SNRs have comparable values to Spearman correlations. However, the ICC values varied across SNR without following a general trend. For both, the PPD and MPD, high ICC values were observed for most of the SNRs (see Table 4) when comparing Visits 1 and 2. When applied to the GCA features of the pupil traces, good-to-excellent reliability (ICC above 0.6) was only found for 2 out of 5 of the SNRs for the slope and for 1 out of 5 of the SNRs for the other features (AH, RF, delay). Thus,

Table 3. Spearman Correlations Between Two Consecutive Visits for all Pupil Features Calculated through Different Normalization Procedures.

Spearman correlation	PPD	MPD	AH	Slope	RF	Delay	PPD		MPD		AH		Slope		RF		Delay	
	Visits 1-2 (31)	Visits 1-2 (31)	Visits 1-2 (31)	Visits 1-2 (31)	Visits 1-2 (31)	Visits 1-2 (31)	Visits 1-2 (11)	Visits 2-3 (11)	Visits 1-2 (11)	Visits 2-3 (11)	Visits 1-2 (11)	Visits 2-3 (11)	Visits 1-2 (11)	Visits 2-3 (11)	Visits 1-2 (11)	Visits 2-3 (11)	Visits 1-2 (11)	Visits 2-3 (11)
Baseline correction	**0.6** ***	0.59 ***	0.45 ***	0.5 ***	0.41 ***	0.53 ***	0.52 ***	0.5 ***	**0.64** ***	0.53 ***	0.57 ***	0.5 ***	0.48 ***	0.39 **	0.53 ***	0.45 **	**0.72** ***	**0.72** ***
Range normalization	**0.63** ***	**0.63** ***	0.12	0.51 ***	0.44 ***	0.54 ***	**0.66** ***	0.46 ***	**0.66** ***	0.46 ***	0.32 *	0.28 *	0.49 ***	0.39 **	0.58 ***	0.47 ***	**0.64** ***	**0.67** ***
z-score	0.21 **	0.54 ***	−0.78 ***	0.5 ***	0.45 ***	0.5 ***	−0.03	0.036	**0.6** ***	0.52 ***	−0.47 ***	−0.4 **	0.45 ***	0.37 **	0.59 ***	0.45 **	**0.65** ***	**0.67** ***
Baseline Range	**0.87** ***	**0.87** ***	**0.94** ***	0.54 ***	0.43 ***	0.58 ***	**0.73** ***	**0.81** ***	**0.73** ***	**0.81** ***	**0.91** ***	**0.93** ***	0.49 ***	0.38 **	0.59 ***	0.53 ***	**0.65** ***	**0.68** ***

The values above 0.6 are highlighted in bold, representing good correlation. PPD = peak pupil dilation; MPD = mean pupil dilation; AH = average height; RF = rise and fall; SNR = signal-to-noise ratio.

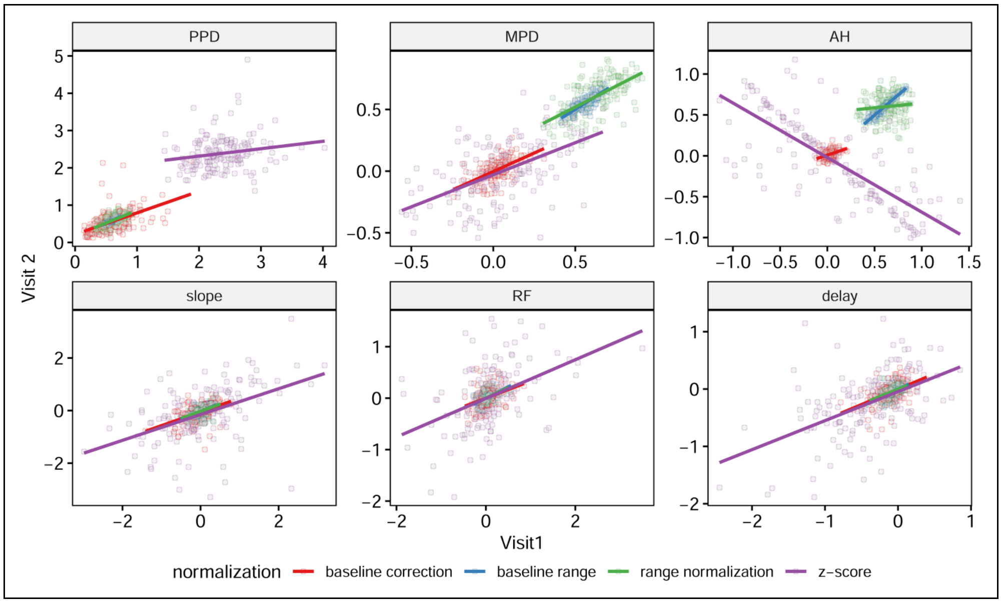

Figure 5. Scatter plot depicting the correlation between Visits 1 and 2 per individual across all SNRs and for each pupil feature (PPD, MPD, AH, slope, RF, delay) indicated in the different panels and for each normalization procedure (baseline correction, range normalization, z-score, and baseline range) as indicated by different colors. PPD = peak pupil dilation; MPD = mean pupil dilation; AH = average height; RF = rise and fall; SNR = signal-to-noise ratio.

across all SNRs, the PPD and the MPD showed overall higher ICC values compared to the GCA features.

The range normalization provided good-to-excellent reliability for the traditional PPD with 4 out of the 5 SNRs when comparing Visits 1 and 2. Interestingly, none of the ICC values was above 0.6 for the MPD. The GCA features showed, overall, poor-to-fair reliability between Visits 1 and 2 with a few exceptions (delay at −12 and −4 dB, slope at −12 dB and RF at 4 dB). When *z*-score was applied as a normalization procedure, poor-to-fair reliability was obtained for PPD and MPD for all SNRs between Visits 1 and 2. Good-to-excellent reliability was obtained for only some of the GCA features (i.e., for RF, slope, and delay), at only 2 out of the 5 SNRs.

When the data were baseline corrected and then range normalized within individuals, very high ICC values were observed for PPD, MPD, and AH, indicating that these were the most reliable features across all SNRs between Visits 1 and 2.

ICC - Perceived Effort

The NASA-TLX was analyzed to assess the perceived effort for each condition (Hart & Staveland, 1988). Participants were also asked to evaluate their effort on a scale from 0 to 10 after each condition. Both perceived effort measures were rescaled to a 0-1 scale. Reliability values (ICC) for the perceived effort measures are summarized in Table 5 for Visits 1 and 2 and in Table 6 from the Supplemental material for the subsample of 11 subjects for the three visits. For both NASA-TXL and the self-rated effort scale, good-to-excellent ICC values (above 0.6) were observed for −12 dB, −8 dB and 4 dB SNR between Visits 1 and 2 but not for −4 dB and 0 dBs.Together with the ICC coefficients, the mean, standard deviation, and range of each of the two perceived effort measures are presented in Table 6. NASA-TLX shows a mean between 0.4 and 0.52, depending on the SNR, with a higher NASA-TLX coefficient recorded for low SNRs and a standard deviation of 0.13, determining a coefficient of variation (standard deviation divided by the mean) of up to 30% for each SNR. Overall, there is no specific trend between the aggregated values (i.e., mean, standard deviation, variation) and the ICC coefficients. The mean self-rated effort scale varied depending on the SNR, with a higher perceived effort at the low SNRs and a relatively constant standard deviation of 0.2 across conditions, leading to a coefficient of variation from 19% to 80% across SNRs. Overall, low variation in the self-rated effort scale was obtained for a higher mean perceived effort. No specific trend is observed between the aggregated values and the ICC coefficients.

Discussion

The present study examined the reliability of the evoked pupil response in a speech-in-noise test paradigm to identify test conditions and analysis techniques that provide the highest test re-test reliability. Specifically, it was analyzed how task demands (manipulated through SNR changes) and data normalization impact the reliability of the evoked pupil response. Overall, the results showed that data

Table 4. ICC Values for all Normalization Procedures and SNRs between Visits 1 and 2.

ICC	Feature	PPD	MPD	AH	Slope	RF	Delay
Baseline correction	All SNRs	**0.65**	**0.73**	0.51	**0.70**	0.52	**0.66**
	−12 dB	**0.67**	**0.72**	0.56	**0.66**	0.5	0.56
	−8 dB	**0.72**	0.56	0.5	0.46	0.48	0.5
	−4 dB	**0.71**	**0.7**	***0.77***	***0.78***	0.17	***0.77***
	0 dB	0.58	0.16	0.44	0.31	0.29	0.44
	4 dB	**0.71**	***0.81***	0.49	0.53	***0.79***	0.49
Range normalization	All SNRs	0.59	0.58	***0.97***	**0.64**	**0.74**	**0.67**
	−12 dB	***0.77***	0.00	0.00	***0.77***	0.47	**0.7**
	−8 dB	**0.72**	0.19	0.19	0.33	0.58	0.46
	−4 dB	0.43	0.00	0.00	**0.62**	0.44	**0.68**
	0 dB	***0.8***	0.58	0.58	0.35	0.43	0.00
	4 dB	***0.8***	0.3	0.3	0.56	***0.82***	0.45
z-score	All SNRs	0.39	0.33	0.00	0.55	**0.71**	**0.66**
	−12 dB	0.36	0.49	0.00	***0.76***	0.46	**0.72**
	−8 dB	0.55	0.42	0.00	0.23	**0.61**	0.32
	−4 dB	0.00	0.34	0.00	**0.62**	0.49	**0.67**
	0 dB	0.26	0.00	0.00	0.37	0.28	0.00
	4 dB	0.24	0.51	0.00	0.45	***0.81***	0.52
Baseline Range	All SNRs	***0.88***	***0.90***	***0.98***	**0.71**	**0.64**	**0.69**
	−12 dB	***0.98***	***0.98***	***0.96***	***0.79***	0.51	**0.68**
	−8 dB	***0.98***	***0.98***	***0.99***	0.49	**0.6**	**0.61**
	−4 dB	***0.98***	***0.98***	***0.97***	**0.6**	0.54	**0.74**
	0 dB	***0.98***	***0.98***	***0.98***	0.5	0.5	0.00
	4 dB	***0.99***	***0.99***	***0.99***	**0.61**	***0.76***	0.54

Values between 0.6 and 0.75, representing good reliability, are highlighted in bold and values above 0.75, representing excellent reliability, are highlighted in italic bold. The negative ICC values were truncated to zero. ICC = intraclass correlation coefficient; PPD = peak pupil dilation; MPD = mean pupil dilation; AH = average height; RF = rise and fall; SNR = signal-to-noise ratio.

normalization procedures have a strong impact and that certain procedures lead to high reliability in the pupil response.

It was hypothesized that reliability would be affected by the normalization procedure of the extracted pupil response. Thus, various normalization procedures that were recommended in previous literature were considered (Winn et al., 2018; Piquado et al., 2010; McCloy et al., 2016). These procedures included baseline correction, two different range normalization procedures, and a *z*-score normalization. The results indicate that the baseline correction procedure combined with range normalization provides the highest reliability results. High agreement (ICC results) was observed for the stationary features (i.e., PPD and MPD), but also for the AH feature extracted from the GCA. Similar values of AH and MPD were obtained using this normalization procedure, as expected. However, the *z*-scores produced totally different results that might be explained by the different time period considered for the GCA features extraction than for MPD. A normalization procedure that takes into account the dynamic range of the pupil response has been suggested when comparing groups of different ages, or even when testing on different days (Winn et al., 2018; Piquado et al., 2010). The combination of a baseline correction and range normalization addresses the reactivity of the pupil response (i.e., high versus small dynamic range) and removes variance in the individual pupil response, which provides high within-subject reliability across different visits as shown by the results presented here.

The lowest agreement across all conditions was obtained with the *z*-score calculations, which use the two statistical values (i.e., mean and standard deviation) to address inter-individual differences in variability of the pupil responses. However, the *z*-score assumes a normal distribution of the data points considered when applying this procedure, and not all pupil traces (i.e., all samples within a trace) fulfilled this assumption. The normality of the samples of each pupil trace (between the sentence onset and sentence offset) was verified using Shapiro-Wilk tests. In addition, not having a baseline on a trial level established when calculating the *z*-score prior to the normalization process produces higher disparities across SNRs and visits.

It was hypothesized that changes in task demands (manipulated through the SNR) would affect the reliability of the pupil features such that higher reliability would be obtained for higher task demands. This was based on previous literature indicating increased reliability with increasing task difficulty (Aday & Carlson, 2019; Zwyghuizen-Doorenbos et al., 1988). Overall, the ICC values varied widely across SNRs, ranging from poor agreement to excellent agreement, and

Table 5. ICC Values for the Perceived Effort Measures, Comparisons Between Visits 1 and 2, Together with the Mean, Standard Deviation and the Range of these Measures.

	NASA-TXL			Self-rated effort scale				
SNR	ICC	Mean	Std	Range	ICC	Mean	Std	Range
All SNRs	***0.77***	0.46	0.13	0.22–0.78	***0.84***	0.50	0.20	0.0–1.0
−12 dB	***0.87***	0.52	0.14	0.23–0.76	**0.67**	0.84	0.16	0.5–1.0
−8 dB	***0.76***	0.51	0.12	0.26–0.75	**0.68**	0.70	0.20	0.2–1.0
−4 dB	0.55	0.45	0.12	0.25–0.78	0.14	0.42	0.20	0.1–0.9
0 dB	0.42	0.41	0.12	0.22–0.67	0.57	0.30	0.21	0.0–0.8
4 dB	***0.84***	0.40	0.12	0.23–0.68	***0.75***	0.26	0.21	0.0–0.8

ICC values between 0.6 and 0.75, representing good reliability, are highlighted in black bold, and values above 0.75, representing excellent reliability, are highlighted in italic bold. ICC = intraclass correlation coefficient; SNR = signal-to-noise ratio; NASA-TLX = NASA Task Load Index.

there was no clear trend between the SNR and the agreement. This is in line with other previous literature suggesting that reliability is independent of SNR. For example, Giuliani et al. (2020) reported fair reliability for all test conditions, independent of SNR, with no clear trends across SNRs. While Giuliani et al. (2020) studied only relatively high SNR conditions ranging between 0 and −5 dB, the present study addressed a broader range of SNRs, including more challenging SNRs up to −12 dB (corresponding to an average of 25%–40% intelligibility). The results obtained in the present study were, thus, unexpected rejecting the hypothesis of increasing reliability with SNR.

Note that the task demands were manipulated by varying the SNR. However, participants differed in their performance for a given SNR, meaning that the task demands could differ across individuals at similar SNRs. The correlation analysis indicated that the pupil response in such a speech-in-noise task is related to both SNR manipulations as well as task performance. Correlations with the listener's perceived effort were, although significant, comparably low. Those results might suggest a better agreement of the pupil response with performance and/or acoustic manipulation as compared to the perceptual effort of the listener. However, the relationship between different measures of effort is unclear. For instance, inconsistencies have been reported between objective and perceived listening effort measures. While several studies on perceived effort have reported decreasing effort investment with the addition of context (Johnson et al., 2015; Holmes et al., 2018), literature assessing objective measures of listening effort (such as reaction times) reported similar levels or increased effort with the addition of context (Tun et al., 2009; Desjardins & Doherty, 2014; Borghini & Hazan, 2020). Furthermore, previous literature suggested that perceived effort measures are only weakly correlated with objective measures. For example, Alhanbali et al. (2019) assessed several measures of listening effort and found only weak average correlations between different measures (see furthermore Strand et al. (2018) for similar findings). Hence, it has been speculated that those different measures of effort may tap into different dimensions of listening effort rather than assessing the same construct. Other literature suggested that listening effort is multi-dimensional and made the most common distinction between objective and subjective effort (Hornsby, 2013; Johnson et al., 2015; Francis & Love, 2020; Herrmann & Johnsrude, 2020). The findings of the current study, showing comparable weak correlations between pupil response and subjective effort, are in line with the literature reporting a rather inconsistent or weak agreement of objective and perceived effort measures.

The scope of this study was to investigate the reliability of pupillometry toward a diagnostic tool. In order to identify the potential benefits of HA interventions on individual's listening effort, and due to the fact that current hearing aid processing schemes, such as noise reduction, can effectively reduce the SNR, the current study chose a SNR manipulation for changing task demands. However, reliability only constitutes one aspect of evaluating pupillometry as a listening effort measure. A limitation of the method arises since a mental process, such as effort, is difficult to perfectly reproduce in consecutive sessions.

In contrast to Wendt et al. (2018), there was no evidence of disengagement in the group-level analysis, which would have been illustrated by a reduced pupil response at the lowest SNRs (e.g., −12 dB) where speech recognition performance tends to be low. Despite this, some individuals did show some level of disengagement, as larger pupil responses were observed at higher (e.g. −8 or −4 dB SNR as compared to e.g., −12 dB SNR) indicating a reduction in effort investment when processing and studying individual's pupil response. The fact that disengagement was observed in only some individuals and that task demands seemed to differ across individuals for a given SNR could, taken together, explain why reliability was not increasing with SNR and, as it was originally hypothesized.

A higher reliability for each of the pupil features was expected to be obtained between Visits 2 and 3 compared to Visits 1 and 2. This expectation was attributed to potential global arousal or to a learning effect due to the novelty of the task that could occur in the first visit compared to the subsequent visits (Alamia et al., 2019; Widmann et al., 2018). In general, the results (ICC and correlation analysis) suggested no significant impact of the number of visits on most of the pupil features. At the same time, the *t*-test analysis of the SRT measures indicated a significant improvement from Visits 1 to 2, which indicated a learning effect. However, this learning effect was not reflected in the pupil response. In other words, there was no clear trend between the reliability and the visit number in this study which further suggests that no overall arousal effect occurred across the visits. Moreover, the results of the current study suggest that within a minimum of 3 weeks between the visit, no systematic change in the pupil response is seen with respect to its reliability, and high reliability can already be obtained within two visits depending on the normalization procedure.

Note that only 11 participants out of the 31 were tested on the third visit, and, consequently, a comparison between the reliability at different visits was performed for only a subsample of 11 participants. Since ICC analysis requires a minimum of 30 participants in order to provide sufficient power Koo & Li (2016), a Spearman correlation on this subsample of participants was performed to verify the conclusions. The ICC and Spearman's correlation results for Visits 1-2 were similar, such that no trend of correlation coefficients was found with an increasing number of visits. Further testing with a larger sample of subjects participating in three visits would be needed to better clarify how the reliability changes with more than two visits.

Overall, it seems that the traditional pupil features (i.e., PPD and MPD) are more reliable than the temporal features. This finding is in line with other studies that only considered PPD and MPD as relevant features (Kramer et al., 1997; Zekveld et al., 2010; Wendt et al., 2018). Nonetheless, all the pupil features in the current study were, in one way or another, aggregated values of a time series of the pupil response. The aggregation of the pupil response over all trials and within the final trial can limit the understanding of the entire time series and its associated reliability. This aspect was partly addressed by including the GCA temporal features. However, assessment of the reliability of the pupil response using nonaggregating methods could lead to a different conclusion.

The reliability of the perceived listening effort measures (i.e., NASA-TLX and the self-rated effort scale) was assessed, and the perceived listening effort showed, in most of the cases, a reliability that was on par with the pupil features, in line with previous literature (Alhanbali et al., 2019; Giuliani et al., 2020). This study reported slightly higher, or similar reliability between measures of perceived effort and the PPD or MPD, irrespective of the normalization procedure applied. Similarly with the pupil features results, no clear patterns in the reliability of perceived effort across all of the SNRs and normalization procedures were found.

The current study assumes that effort investment will be approximately constant over repeated visits, as long as task demands (SNR manipulation) and other external factors (i.e., luminosity, caffeine intake, day and time of the test) are well-controlled. However, the assumption that a listener will invest the exact same amount of listening effort with a given difficulty level in both sessions, and that the pupil will perfectly reflect this effort investment, might not reflect the complexity of the mental processes involved in effort deployment. It has been argued that other aspects related to the personal or mental state of the listener, including arousal or cognitive capacity limits due to being fatigued, can affect effort investment (Pichora-Fuller et al., 2016). In fact, it is argued that those factors most likely vary over time. For example, the arousal level might change due to the listener getting familiar with the task which, in turn, influences effort allocation. However, it was outside the scope of the current study to strictly control for these factors. Instead, our study aimed to explore the reliability of pupillometry, which is a prerequisite to the potential use of pupillometry as a diagnostic tool to, for example, identify potential benefits of HA interventions on an individual's listening effort. However, it is suggested that future studies should examine the role of the mental state of the listener (including aspects of motivation and fatigue) with regards to effort investment in a speech-in-noise task.

This investigation provided a systematic approach to assess the reliability of various pupil features, involving external manipulations and data processing. A strong effect on reliability is shown by data normalization. However, since many methodological aspects are involved in pupil data preprocessing (i.e., detrending, blink detection and removal by interpolation, smoothing, and trial rejection), further research is needed to clarify the impact of each data-preprocessing step on the reliability of the extracted pupil features. Eventually, those findings could contribute to establishing a standardized pupil preprocessing methodology.

Overall, several pupil features as potential indicators of listening effort, revealed high reliability only in some particular cases (i.e., baseline range normalization procedure). Therefore, careful consideration of the data normalization procedure used when processing and studying individual's pupil response is recommended.

Conclusion

The current study examined the reliability of pupillometry with several normalization procedures and feature extraction methods, while also assessing the impact of SNR and the number of visits on the resulting reliability. Overall, the results suggest that SNR and the number of visits only have a minor impact on the reliability of the pupil response, at least within a speech-in-noise test paradigm. Moreover, to obtain the highest reliability across SNRs, baseline correction combined with range normalization is recommended when analyzing the pupil response of individual listeners. Moreover, the stationary features (i.e., PPD and MPD) are the most reliable features. Overall, these reliability results may provide valuable insights for determining the future of pupillometry as a potential diagnostic tool in the clinic. The data will be made available upon reasonable request.

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

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

Supplemental Material

Supplemental material for this article is available online.

References

Aday, J. S., & Carlson, J. M. (2019). Extended testing with the dot-probe task increases test–retest reliability and validity. *Cognitive Processing*, *20*(1), 65–72. https://doi.org/10.1007/s10339-018-0886-1. http://link.springer.com/10.1007/s10339-018-0886-1

Alamia, A., VanRullen, R., Pasqualotto, E., Mouraux, A., & Zenon, A. (2019). Pupil-linked arousal responds to unconscious surprisal. *The Journal of Neuroscience*, *39*(27), 5369–5376. https://doi.org/10.1523/JNEUROSCI.3010-18.2019. http://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.3010-18.2019

Alhanbali, S., Dawes, P., Millman, R. E., & Munro, K. J. (2019). Measures of listening effort are multidimensional. *Ear and Hearing*, *40*(5), 1084–1097. https://doi.org/10.1097/AUD.0000000000000697. http://journals.lww.com/00003446-201909000-00004

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. https://pubmed.ncbi.nlm.nih.gov/16022602/

Beatty, J. (1982). Task-Evoked Pupillary Responses, Processing Load, and the Structure of Processing Resources. Technical Report 1.

Borghini, G., & Hazan, V. (2020). Effects of acoustic and semantic cues on listening effort during native and non-native speech perception. *The Journal of the Acoustical Society of America*, *147*(6), 3783. https://doi.org/10.1121/10.0001126. https://asa.scitation.org/doi/abs/10.1121/10.0001126

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. https://link.springer.com/article/10.3758/BF03329118

Bramsløw, L., Simonsen, L. B., El Hichou, M., Hashem, R., & Hietkamp, R. K. (2016). Learning effects as result of multiple exposures to Danish HINT. In: *Poster presented at the International Hearing Aid Conference, Lake Tahoe, CA, USA.*

Bremner, F. (2009). Pupil evaluation as a test for autonomic disorders. *Clinical Autonomic Research*, *19*(2), 88–101. https://doi.org/10.1007/S10286-009-0515-2/FIGURES/13. https://link.springer.com/article/10.1007/s10286-009-0515-2

Carter, M. E., Yizhar, O., Chikahisa, S., Nguyen, H., Adamantidis, A., Nishino, S., Deisseroth, K., & De Lecea, L. (2010). Tuning arousal with optogenetic modulation of locus coeruleus neurons. *Nature Neuroscience*, *13*(12), 1526–1533. https://doi.org/10.1038/nn.2682. https://www.nature.com/articles/nn.2682

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284. http://doi.apa.org/getdoi.cfm?doi=10.1037/1040-3590.6.4.284

Daguet, I., Bouhassira, D., & Gronfier, C. (2019). Baseline pupil diameter is not a reliable biomarker of subjective sleepiness. *Frontiers in Neurology*, *10*(FEB). https://doi.org/10.3389/FNEUR.2019.00108. https://pubmed.ncbi.nlm.nih.gov/30858817/

Desjardins, J. L., & Doherty, K. A. (2014). The effect of hearing aid noise reduction on listening effort in hearing-impaired adults. *Ear and Hearing*, *35*(6), 600–610. https://doi.org/10.1097/AUD.0000000000000028. https://journals.lww.com/ear-hearing/Fulltext/2014/11000/The_Effect_of_Hearing_Aid_Noise_Reduction_on.3.aspx

Duchowski, A. T., Krejtz, K., Krejtz, I., Biele, C., Niedzielska, A., Kiefer, P., & Raubal, M. (2018). *CHI Conference on Human Factors in Computing Systems (CHI '18)* ACM, New York, NY, Article 282, 13 pages. https://doi.org/10.1145/3173574.3173856.

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

Einstein, G. O., Smith, R. E., McDaniel, M. A., & Shaw, P. (1997). Aging and prospective memory: The influence of increased task demands at encoding and retrieval. *Psychology and Aging*, *12*(3), 479–488. https://doi.org/10.1037/0882-7974.12.3.479. https://pubmed.ncbi.nlm.nih.gov/9308095/

Francis, A. L., & Love, J. (2020). Listening effort: Are we measuring cognition or affect, or both?. *Wiley Interdisciplinary Reviews: Cognitive Science*, *11*(1), e1514. https://doi.org/10.1002/WCS.1514. https://onlinelibrary.wiley.com/doi/full/10.1002/wcs.1514

Giuliani, N. P., Brown, C. J., & Wu, Y. H. (2020). Comparisons of the Sensitivity and Reliability of Multiple Measures of Listening Effort. *Ear & Hearing* Publish Ah: 1–10. https://doi.org/10.1097/AUD.0000000000000950. https://journals.lww.com/10.1097/AUD.0000000000000950.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. *Advances in Psychology*, *52*(C), 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

Herrmann, B., & Johnsrude, I. S. (2020). Absorption and enjoyment during listening to acoustically masked stories. *Trends in Hearing*, *24*. https://doi.org/10.1177/2331216520967850. https://journals.sagepub.com/doi/full/10.1177/2331216520967850

Holmes, E., Folkeard, P., Johnsrude, I. S., & Scollie, S. (2018). Semantic context improves speech intelligibility and reduces listening effort for listeners with hearing impairment. *International Journal of Audiology*, *57*(7), 483–492. https://doi.org/10.1080/14992027.2018.1432901

Hornsby, B. W. (2013). The effects of hearing aid use on listening effort and mental fatigue associated with sustained speech processing demands. *Ear and Hearing*, *34*(5), 523–534. https://doi.org/10.1097/AUD.0B013E31828003D8. https://journals.lww.com/ear-hearing/Fulltext/2013/09000/The˙Effects˙of˙Hearing˙Aid˙Use˙on˙Listening˙Effort.1.aspx

Johnson, J., Xu, J., Cox, R., & Pendergraf, P. (2015). A comparison of two methods for measuring listening effort as part of an audiologic test battery. *American Journal of Audiology*, *24*(3), 419–431. https://doi.org/10.1044/2015_AJA-14-0058. https://pubs.asha.org/doi/abs/10.1044/2015_AJA-14-0058

Jones, B. E. (2004). Activity, modulation and role of basal forebrain cholinergic neurons innervating the cerebral cortex. *Progress in Brain Research*, *145*, 157–169. https://doi.org/10.1016/S0079-6123(03)45011-5. https://pubmed.ncbi.nlm.nih.gov/14650914/

Kahneman, D. (1973). *Attention and effort*. Citesser. ISBN 0-13-050518-8. https://doi.org/10.1.1.398.5285. https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.398.5285.

Koelewijn, T., Zekveld, A. A., Festen, J. M., Rönnberg, J, & Kramer, S. E. (2012). Processing load induced by informational masking is related to linguistic abilities. *International Journal of Otolaryngology*, *2012*, 1–11. https://doi.org/10.1155/2012/865731

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012. https://linkinghub.elsevier.com/retrieve/pii/S1556370716000158

Kramer, S. E., Kapteyn, T. S., Festen, J. M., & Kuik, D. J. (1997). Assessing aspects of auditory handicap by means of pupil dilatation. *International Journal of Audiology*, *36*(3), 155–164. https://doi.org/10.3109/00206099709071969. http://www.tandfonline.com/doi/full/10.3109/00206099709071969

Krejtzid, K., Duchowski, A. T., Niedzielska, A., Cezary, B., & Krejtz, I. (2018). Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. https://doi.org/10.1371/journal.pone.0203629.

Kuchinsky, S. E., Ahlstrom, J. B., Cute, S. L., Humes, L. E., Dubno, J. R., & Eckert, M. A. (2014). Speech-perception training for older adults with hearing loss impacts word recognition and effort. *Psychophysiology*, *51*(10), 1046–1057. https://doi.org/10.1111/psyp.12242. http://doi.wiley.com/10.1111/psyp.12242

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

Kuchinsky, S. E., Vaden, K. I., Ahlstrom, J. B., Cute, S. L., Humes, L. E., Dubno, J. R., & Eckert, M. A. (2016). Task-related vigilance during word recognition in noise for older adults with hearing loss. *Experimental Aging Research*, *42*(1), 64–85. https://doi.org/10.1080/0361073X.2016.1108712

Lee, S. H., & Dan, Y. (2012). Neuromodulation of brain states. *Neuron*, *76*(1), 209–222. https://doi.org/10.1016/J.NEURON.2012.09.012

Loewenfeld, I. E. 1993). *The pupil: Anatomy, physiology, and clinical applications*, 2. Iowa State University Press.

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

MATLAB. (2018). *9.7.0.1190202 (R2019b)*. The MathWorks Inc,

May, P. J., Reiner, A., & Gamlin, P. D. (2019). Autonomic regulation of the eye. *Oxford Research Encyclopedia of Neuroscience*. https://doi.org/10.1093/ACREFORE/9780190264086.013.276. https://oxfordre.com/neuroscience/view/10.1093/acrefore/9780190264086.001.0001/acrefore-9780190264086-e-276

McCloy, D. R., Larson, E. D., Lau, B, & Lee, A. K. C (2016). Temporal alignment of pupillary response with stimulus events via deconvolution. *The Journal of the Acoustical Society of America*, *139*(3), EL57–EL62. https://doi.org/10.1121/1.4943787. http://asa.scitation.org/doi/10.1121/1.4943787

McGarrigle, R., Munro, K. J., Dawes, P, Stewart, A. J., Moore, D. R., Barry, J. G., & Amitay, S. (2014). Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group 'white paper'. https://doi.org/10.3109/14992027.2014.890296. https://www.tandfonline.com/doi/abs/10.3109/14992027.2014.890296.

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, *59*(4), 475–494. https://doi.org/10.1016/j.jml.2007.11.006. https://linkinghub.elsevier.com/retrieve/pii/S0749596X07001313

Nielsen, J. B., & Dau, T. (2011). The danish hearing in noise test. *International Journal of Audiology*, *50*(3), 202–208. https://doi.org/10.3109/14992027.2010.524254. http://www.tandfonline.com/doi/full/10.3109/14992027.2010.524254

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

Partala, T, & Surakka, V (2003). Pupil size variation as an indication of affective processing. *International Journal of Human Computer Studies*, *59*(1-2), 185–198. https://doi.org/10.1016/S1071-5819(03)00017-X

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W., Humes, L. E., Lemke, U., Lunner, T., Matthen, M., Mackersie, C. L., Naylor, G., Phillips, N. A., Richter, M., Rudner, M., Sommers, M. S., Tremblay, K. L., & Wingfield, A. (2016). Hearing impairment and cognitive energy: The framework for understanding effortful listening (FUEL). *Ear and Hearing*, *37*, 5S–27S. https://doi.org/10.1097/AUD.0000000000000312. https://pubmed.ncbi.nlm.nih.gov/27355771/

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

R Core Team. (2019). R: A Language and Environment for Statistical Computing. https://www.r-project.org/.

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

Reimer, J., McGinley, M. J., Liu, Y., Rodenkirch, C., Wang, Q., McCormick, D. A., & Tolias, A. S. (2016). Pupil fluctuations track rapid changes in adrenergic and cholinergic activity in cortex.

Nature Communications, *7*(1), 1–7. https://doi.org/10.1038/ncomms13289. https://www.nature.com/articles/ncomms13289

Strand, J. F., Brown, V. A., Merchant, M. B., Brown, H. E., & Smith, J. (2018). Measuring listening effort: Convergent validity, sensitivity, and links with cognitive and personality measures. *Journal of Speech, Language, and Hearing Research*, *61*(6), 1463–1486. https://doi.org/10.1044/2018_JSLHR-H-17-0257. https://pubs.asha.org/doi/abs/10.1044/2018_JSLHR-H-17-0257

Tun, P. A., McCoy, S., & Wingfield, A. (2009). Aging, hearing acuity, and the attentional costs of effortful listening. *Psychology and Aging*, *24*(3), 761–766. https://doi.org/10.1037/A0014802. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2773464/

Wang, Y., Kramer, S. E., Wendt, D., Naylor, G., Lunner, T., & Zekveld, A. A. (2018). the pupil dilation response during speech perception in dark and light: The involvement of the parasympathetic nervous system in listening effort. *Trends in Hearing*, *22*, 1–11. https://doi.org/10.1177/2331216518816603. http://journals.sagepub.com/doi/10.1177/2331216518816603

Wang, Y., Zekveld, A. A., Naylor, G., Ohlenforst, B., Jansma, E. P., Lorens, A., Lunner, T., & Kramer, S. E. (2016). Parasympathetic nervous system dysfunction, as identified by pupil light reflex, and its possible connection to hearing impairment. *PLoS ONE*, *11*(4), 1–26. https://doi.org/10.1371/journal.pone.0153566

Wendt, D., Hietkamp, R. K., & Lunner, T. (2017). Impact of noise and noise reduction on processing effort: A pupillometry study. *Ear & Hearing*, *38*(6), 690–700. https://doi.org/10.1097/AUD.0000000000000454. https://journals.lww.com/00003446-201711000-00007

Wendt, D., Koelewijn, T., Ksia¸żek, P., & Lunner, T. (2018). Toward a more comprehensive understanding of the impact of masker type and signal-to-noise ratio on the pupillary response while performing a speech-in-noise test. *Hearing Research*, *369*, 67–78. https://doi.org/10.1016/j.heares.2018.05.006. https://linkinghub.elsevier.com/retrieve/pii/S0378595517305294

Widmann, A., Schröger, E., & Wetzel, N. (2018). Emotion lies in the eye of the listener: Emotional arousal to novel sounds is reflected in the sympathetic contribution to the pupil dilation response and the P3. *Biological Psychology*, *133*, 10–17. https://doi.org/10.1016/j.biopsycho.2018.01.010. https://linkinghub.elsevier.com/retrieve/pii/S0301051118300498

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*, 233121651666972. https://doi.org/10.1177/2331216516669723. http://journals.sagepub.com/doi/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.0000000000000145

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*, 233121651880086. https://doi.org/10.1177/2331216518800869. http://journals.sagepub.com/doi/10.1177/2331216518800869

Zekveld, A. A., Koelewijn, T., & Kramer, S. E. (2018). The pupil dilation response to auditory stimuli: Current state of knowledge. *Trends in Hearing*, *22*, 233121651877717. https://doi.org/10.1177/2331216518777174. http://journals.sagepub.com/doi/10.1177/2331216518777174

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2010). Pupil response as an indication of effortful listening: The influence of sentence intelligibility. *Ear & Hearing*, *31*(4), 480–490. https://doi.org/10.1097/AUD.0b013e3181d4f251. https://journals.lww.com/00003446-201008000-00004

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 & Hearing*, *32*(4), 498–510. https://doi.org/10.1097/AUD.0b013e31820512bb. https://journals.lww.com/00003446-201107000-00008

Zwyghuizen-Doorenbos, A., Roehrs, T., Schaefer, M., & Roth, T. (1988). Test-retest reliability of the MSLT. *Sleep*, *11*(6), 562–565. https://doi.org/10.1093/sleep/11.6.562. https://academic.oup.com/sleep/article-lookup/doi/10.1093/sleep/11.6.562

Appendices

A Perceived Effort Measures

Table 6. ICC Values for the Perceived Effort Measures, Comparisons Between Visits 1, 2 and 3 for a Subsample of 11 Participants.

		NASA-TLX		Self-rated effort scale	
ICC	Feature	Visits 1-2 (11)	Visits 2-3 (11)	Visits 1-2 (11)	Visits 2-3 (11)
	All SNRs	***0.84***	***0.88***	***0.86***	***0.86***
	−12 dB	**0.64**	***0.93***	0.52	**0.67**
	−8 dB	***0.83***	***0.89***	***0.9***	0.45
	−4 dB	***0.85***	***0.91***	0.4	***0.76***
	0 dB	***0.93***	***0.91***	**0.72**	***0.75***
	4 dB	***0.88***	***0.86***	***0.94***	***0.78***

Values between 0.6 and 0.75, representing good reliability, are highlighted in black bold and values above 0.75, representing excellent reliability, are highlighted in italic bold. NASA-TLX = NASA Task Load Index; ICC = intraclass correlation coefficient; SNR = signal-to-noise ratio.

B Pupil Features

Table 7. ICC Values for all Normalization Procedures, SNRs and Comparisons Between Visits 1, 2 and 3 for a Subsample of 11 Participants.

		PPD		MPD		AH		Slope		RF		Delay	
ICC	Feature	Visits 1-2 (11)	Visits 2-3 (11)	Visits 1-2 (11)	Visits 2-3 (11)	Visits 1-2 (11)	Visits 2-3 (11)	Visits 1-2 (11)	Visits 2-3 (11)	Visits 1-2 (11)	Visits 2-3 (11)	Visits 1-2 (11)	Visits 2-3 (11)
Baseline correction	All SNRs	**0.67**	***0.80***	***0.75***	**0.74**	0.59	***0.75***	0.57	0.56	**0.64**	**0.64**	**0.68**	***0.84***
	−12 dB	0.59	**0.64**	**0.6**	***0.83***	0.28	***0.83***	**0.68**	**0.74**	**0.68**	***0.83***	**0.62**	***0.89***
	−8 dB	***0.79***	**0.65**	0.52	0.34	***0.8***	***0.82***	0.27	0.00	**0.73**	**0.67**	**0.6**	***0.8***
	−4 dB	0.11	0.45	**0.64**	**0.65**	0.00	0.06	**0.67**	0.56	0.00	0.00	***0.78***	***0.91***
	0 dB	**0.62**	0.56	0.41	**0.6**	0.47	0.57	0.3	0.26	0.2	**0.71**	0.42	***0.78***
	4 dB	***0.87***	0.3	**0.72**	0.27	**0.74**	0.14	0.23	0.59	***0.92***	**0.68**	0.51	0.36
Range normalization	All SNRs	0.59	0.59	0.58	0.59	***0.97***	***0.98***	**0.64**	0.57	**0.74**	**0.69**	**0.67**	***0.77***
	−12 dB	**0.67**	***0.82***	**0.67**	***0.79***	0.00	0.27	**0.62**	**0.71**	***0.79***	***0.79***	***0.79***	**0.66**
	−8 dB	**0.74**	**0.64**	***0.81***	**0.71**	**0.66**	0.35	**0.65**	0.00	0.00	***0.77***	0.42	**0.66**
	−4 dB	***0.77***	***0.85***	***0.8***	**0.74**	0.33	0.41	0.39	0.41	0.5	0.00	**0.7**	***0.85***
	0 dB	***0.83***	**0.81**	***0.8***	**0.73**	**0.64**	0.42	0.49	0.45	0.42	**0.62**	0.00	0.49
	4 dB	***0.78***	***0.83***	***0.81***	***0.85***	0.53	**0.68**	***0.91***	**0.64**	0.31	**0.63**	0.48	0.41
z-score	All SNRs	0.39	0.36	0.33	0.34	0	0.00	0.55	0.46	**0.71**	0.53	**0.66**	**0.72**
	−12 dB	**0.64**	0.34	0.17	**0.74**	0.00	0.00	***0.79***	**0.66**	**0.62**	***0.8***	***0.77***	**0.66**
	−8 dB	***0.75***	0.47	0.23	0.12	0.00	0.00	0.00	0.00	**0.68**	***0.78***	0.35	**0.64**
	−4 dB	0.00	0.00	0.00	0.11	0.00	0.00	0.5	0.39	0.48	0.00	**0.72**	***0.85***
	0 dB	0.25	0.00	0.29	0.51	0.00	0.00	0.19	0.26	0.17	0.5	0.00	0.41
	4 dB	0.59	0.00	0.47	0.34	0.00	0.00	0.00	0.55	***0.92***	0.51	0.5	0.5
Baseline range	All SNRs	***0.82***	***0.87***	***0.86***	***0.87***	***0.97***	***0.98***	**0.64**	0.57	**0.74**	**0.69**	**0.67**	***0.77***
	−12 dB	**0.72**	0.59	***0.9***	***0.95***	***0.86***	***0.95***	***0.78***	**0.72**	**0.64**	***0.81***	**0.72**	***0.79***
	−8 dB	***0.86***	**0.68**	***0.9***	***0.88***	***0.99***	***0.99***	0.1	0	0.58	***0.79***	0.52	**0.69**
	−4 dB	0.00	0.00	***0.92***	***0.94***	***0.95***	***0.95***	0.17	0.28	0.5	0.00	**0.71**	***0.85***
	0 dB	0.23	0.18	***0.92***	***0.95***	***0.99***	***0.98***	**0.63**	0.54	**0.62**	***0.76***	0.00	0.47
	4 dB	***0.83***	0.44	***0.95***	***0.93***	***0.98***	***0.98***	0.49	**0.62**	***0.91***	**0.64**	0.55	0.47

Values between 0.6 and 0.75, representing good reliability, are highlighted in black bold and values above 0.75, representing excellent reliability, are highlighted in italic bold. The negative ICC values were truncated to zero." ICC = intraclass correlation coefficient; SNR = signal-to-noise ratio; PPD = peak pupil dilation; MPD = mean pupil dilation; AH = average height; RF = rise and fall.

C GCA Modelled Pupil Traces

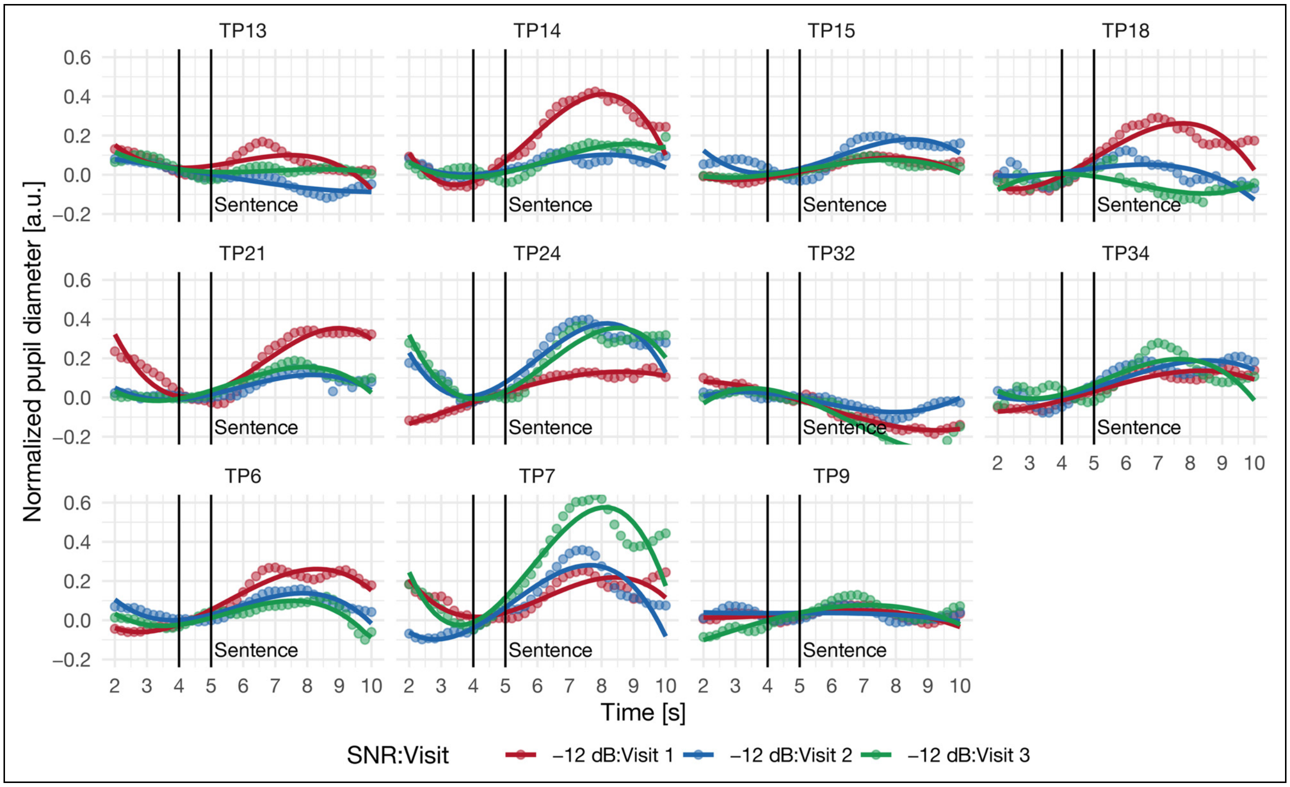

Figure 6. GCA for individual NH listeners. Examples of the pupil responses a function of time, on the three different visits in different colors for the −12 dB SNR condition. The open circles represent the actual data, while the solid lines indicate the fitted GCA model. The numbers in the figure represent the TP. GCA = growth curve analysis;NH = normal-hearing; SNR = signal-to-noise ratio; TP = test participant.