OPEN

Let's See If You Can Hear: The Effect of Stimulus Type and Intensity to Pupil Diameter Response in Infants and Adults

Amanda Saksida,[1,2] Sašo Živanović,[3] Saba Battelino,[4] and Eva Orzan[1]

Objectives: Pupil dilation can serve as a measure of auditory attention. It has been proposed as an objective measure for adjusting hearing aid configurations, and as a measure of hearing threshold in the pediatric population. Here we explore (1) whether the pupillary dilation response (PDR) to audible sounds can be reliably measured in normally hearing infants within their average attention span, and in normally hearing adults, (2) how accurate within-participant models are in classifying PDR based on the stimulus type at various intensity levels, (3) whether the amount of analyzed data affects the model reliability, and (4) whether we can observe systematic differences in the PDR between speech and nonspeech sounds, and between the discrimination and detection paradigms.

Design: In experiment 1, we measured the PDR to target warble tones at 500 to 4000 Hz compared with a standard tone (250 Hz) using an oddball discrimination test. A group of normally hearing infants was tested in experiment 1a (n = 36, mean [ME] = 21 months), and a group of young adults in experiment 1b (n = 12, ME = 29 years). The test was divided into five intensity blocks (30 to 70 dB SPL). In experiment 2a (n = 11, ME = 24 years), the task from experiment 1 was transformed into a detection task by removing the standard warble tone, and in experiment 2b (n = 12, ME = 29 years), participants listened to linguistic (Ling-6) sounds instead of tones.

Results: In all experiments, the increased PDR was significantly associated with target sound stimuli on a group level. Although we found no overall effect of intensity on the response amplitude, the results were most clearly visible at the highest tested intensity level (70 dB SPL). The nonlinear classification models, run for each participant separately, yielded above-chance classification accuracy (sensitivity, specificity, and positive predictive value above 0.5) in 76% of infants and in 75% of adults. Accuracy further improved when only the first six trials at each intensity level were analyzed. However, accuracy was similar when pupil data were randomly attributed to the target or standard categories, indicating over-sensitivity of the proposed algorithms to the regularities in the PDR at the individual level. No differences in the classification accuracy were found between infants and adults at the group level, nor between the discrimination and detection paradigms (experiment 2a versus 1b), whereas the results in experiment 2b (speech stimuli) outperformed those in experiment 1b (tone stimuli).

Conclusions: The study confirms that PDR is elicited in both infants and adults across different stimulus types and task paradigms and may thus serve as an indicator of auditory attention. However, for the estimation of the hearing (or comfortable listening) threshold at the individual level, the most efficient and time-effective protocol with the most appropriate type and number of stimuli and a reliable signal to noise ratio is yet to be defined. Future research should explore the application of pupillometry in diverse populations to validate its effectiveness as a supplementary or confirmatory measure within the standard audiological evaluation procedures.

Key words: Adults, Audiometry, Auditory attention, Infants, Pupillometry.

(Ear & Hearing 2025;XX;00–00)

INTRODUCTION

Hearing, or auditory perception, is an attentional response to sounds. In clinical practice, hearing threshold assessment usually consists of presenting acoustic stimuli (i.e., pure or warble tones) and measuring the subject's behavioral response. The procedure is called pure-tone audiometry (PTA) and is considered the standard form of audiometry. However, this procedure can generally be performed on patients older than 5 years. In infants and children, subjective responses are difficult to obtain; therefore, the description of the auditory status of an infant is obtained through a combined analysis of various physiological methods such as otoacoustic emissions, auditory brain stem response, auditory steady-state response, and electrocochleography. Each of the objective measures has limitations regarding either difficulty in administration or poorer correlation levels with behavioral methods (Gordon et al. 2004; Ahn et al. 2007; Visram et al. 2015). Therefore, a behavioral measure of hearing is considered to be the gold standard against which these methods are compared. For infants, visual reinforcement audiometry, a version of PTA, has been developed, and can be used for the assessment of behavioral responses to sounds in infants older than 6 months (Widen et al. 2000). Nonetheless, the dependence on overt behavioral responses, the habituation of the procedure, and the requirement for highly skilled professionals to interpret the results represent limitations of the method. An additional time- and cost-efficient approach that could directly capture attention to audible stimuli would therefore represent a necessary bridge between the existing objective and behavioral measures.

Among the methods that have been recently proposed to objectively capture the attentional response to auditory stimulation is pupillometry (Joshi et al. 2016; Strauch et al. 2022). Research has recently shown that the pupillary dilation response (PDR) is a valid index of auditory orienting response elicited by a rare and unexpected occurrence of a sound deviating from the auditory background (Marois et al. 2018; Bala et al. 2019), regardless of whether the tracking of unexpected stimuli was conscious or unconscious (Liao et al. 2016b; Quirins et al. 2018). The PDR can be used to assess attentional capture by a deviant sound in contexts where the pupil diameter can be

[1]Pediatric Audiology and Otolaryngology Unit, Institute for Maternal and Child Health - Istituto di Ricovero e Cura a Carattere Scientifico "Burlo Garofolo" - Trieste, Trieste, Italy; [2]Centre for discourse studies, Educational Research Institute Ljubljana, Slovenia; [3]Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia; and [4]Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and text of this article on the journal's Web site (www.ear-hearing.com).

<zdoi; 10.1097/AUD.0000000000001651>

modulated by the visual environment (Marois & Vachon 2018). Moreover, the PDR is strongly correlated with the subjective feeling of the salience/loudness of sounds (Liao et al. 2016a).

The PDR as an index of auditory attention in the pediatric population has been explored in a limited amount of recent studies. Research reports infants' and toddlers' (age range 6 to 30 months) pupil responses to auditory stimuli in the contexts of the sensitivity to phonological detail (Tamási et al. 2016, 2017), categorical sound perception (Hochmann & Papeo 2014; Calignano et al. 2021), and unexpected sounds (Wetzel et al. 2016). In these studies, infants exhibited attention to sounds during a passive listening test. Two of these studies examined infants' involuntary attention to unexpected (target or deviant) sounds in a discrimination task called oddball procedure, while, to our knowledge, no study has examined infants' PDR when detecting sounds in silence. Indirectly, results of these studies nonetheless indicate that PDR could be taken as an index of auditory orienting response to unexpected audible stimuli also in the pediatric population.

In line with this research, the PDR to unexpected acoustic stimuli has been recently proposed as an additional physiological measure of hearing, for both infants (US 2020/0253526 A1 2020) and hearing-impaired adults as a method for adjusting hearing aids (EP 3 481 086 B1 2019; US 10,609,493 B2 2020). The proposed methods predict the PDR to be an index of hearing at threshold levels for individual subjects. These methods imply that the amplitude of the PDR to target stimuli at different intensity levels (response amplitude), or the model success in correctly classifying the PDR to audible stimuli can be modeled with a psychometric function and that the auditory threshold can be defined somewhere between the quietest sound that elicits a statistically significant PDR and the amplitude of the quiet sound that does not yield such a dilation (Bala et al. 2019). However, no data are reported on the actual reliability of the method at the individual level, neither for adults nor for the pediatric population. It is also unclear how much data is necessary to obtain a reliable response from an individual, and how reliable the measure is at various intensity levels. Namely, in pupillometry, a single exposure at a certain intensity level may not be enough to reliably estimate the pupil response to target sounds. Rather, several exposures (a short block of trials) may be necessary at each intensity level (Burg et al. 2021; Mathôt & Vilotijević 2023). Furthermore, to obtain the auditory threshold (i.e., the volume) at which the pupil reliably responds to an auditory stimulus, an adaptive testing procedure may be needed (Leek 2001). Such a psychoacoustic procedure involves repeatedly varying sound intensity to find the level at which the participant stops responding, and is a common procedure in both research and clinical practice (Sinnott et al. 1983; Sinnott & Aslin 1985; Olsho et al. 1987; Leek 2001). All this contributes to the increased duration of a potentially reliable pupillometric test of auditory threshold. Especially for infants and young children, the short attention span during which pupil size can be effectively measured can pose a serious limitation to the method (Hochmann 2013; Hepach & Westermann 2016). Lastly, whereas the method has been proposed as an index of hearing for both speech and nonspeech stimuli, it remains unclear whether the PDR to speech and nonspeech sounds is comparable (cf., Wetzel et al. 2016), and whether the detection task, which is more common in clinical use (PTA), elicits similar responses as the discrimination task, which is common in infant research literature (Olsho et al. 1987; Hochmann & Papeo 2014; Wetzel et al. 2016). As reported in the recent literature, the PDR in the detection task may not only elicit similar responses but may possibly more consistently reflect hearing abilities (Bala et al. 2019) and correlate well with the perceived loudness of the stimulus (Liao et al. 2016), possibly in a similar way as other physiological measures of hearing such as ABR (Schilling et al. 2019).

To address these questions, we prepared a series of tests in which participants—a group of normally hearing infants and three groups of normally hearing young adults—passively listened to warble tones or speech sounds at various intensity levels while watching a video with unrelated content. Experiment 1 consisted of a discrimination task with an oddball procedure in which one of the warble tones was presented frequently (nontargets), while others were presented rarely (targets). To address the question of the amount of data necessary to obtain a reliable response from an individual, both full and reduced datasets were analyzed. In experiment 2, the procedure was modified to verify whether the detection task, compared with the discrimination task, elicits more systematic responses that correlate more consistently with the intensity of the stimulus (experiment 2a) and whether speech sounds elicit similar PDR responses as tones (experiment 2b). Stimuli were presented in five blocks at different (audible) intensity levels. We hypothesized that the increased PDR would be observed both at the group and at the individual level for all target stimuli, potentially with an increased response amplitude at higher intensity levels. We further hypothesized that the reduced amount of input data will negatively affect the results, and that the PDR to speech stimuli would systematically differ from the PDR to tones.

EXPERIMENT 1

In experiment 1a, we tested the reliability of the method at the individual level in a group of normally hearing infants (experiment 1a) and in a group of young normally hearing adults (experiment 1b). In both groups, we analyzed (a) the average PDR to target and nontarget sounds at different levels of intensity, (b) the reliability of the measure in individuals at various intensity levels, and (c) how reliable the PDR is in individuals when only the first six trials are analyzed at each intensity level.

Methods

Participants • For experiment 1a, we tested 36 normally hearing infants (19 girls, mean age: 21 months, range: 5-37 months) in one regional center in Italy. Their mother tongue was Italian or Slovenian, and they had normal or corrected visual acuity, and no relevant neurological or psychiatric disease. Their socioeconomic and cultural background was not assessed. Two additional infants were excluded from data analysis because no data were obtained due to the infants' fussiness and the calibration failures. Parents were recruited for the study through social media. They were informed about the study and signed the informed consent before the beginning of the study. While sample size calculation was not performed, the sample size was based on previously published articles with a comparable number of participants using pupillometry in auditory or audiovisual tasks (Wetzel et al. 2016; Bala et al. 2019; Zhao et al. 2019; Calignano et al. 2023).

Young adults (n = 12, 4 females, mean age: 29 years, SD: 4.3) from various regions of Italy were tested in experiment 1b. In all groups, their education level was university degree or higher, obtained in fields related to audiology or otorhinolaryngology, speech therapy, or audiometry. Written consent to participate in the study was obtained before the beginning of the testing. They were unaware of the specific scope of the study beforehand and were subsequently informed about it. All participants had normal hearing, normal or corrected visual acuity, and no history of psychiatric diseases or neurological problems. The sample size was smaller than in experiment 1a because of the limited availability of adult participants at the pediatric institute.

The study was approved by the regional ethical committee in 2021. All methods were performed in accordance with the relevant guidelines and regulations (e.g., the 1964 WMA Declaration of Helsinki and its later amendments). The testing took place in 2021 and 2022.

Stimuli • Auditory stimuli were warble tones. Warble tones were created using the Function generator (https://www.wavtones.com/functiongenerator.php) with a ±5% frequency modulation and the center frequencies at 250, 500, 1000, 2000, and 4000 Hz. The duration of each sound was 400 msec. The decision to use warble tones was based on the Guidelines on the Acoustics of Sound Field Audiometry in Clinical Audiological Applications (British Society of Audiology, 2008). The warble tone of the lowest frequency (250 Hz) was selected as the standard nontarget stimulus, whereas other tones were used as target stimuli.

The visual stimulus was a video recording of a girl playing with puppets, presented in black and white, at a constant brightness level (i.e., the luminosity, measured in lumens, remained constant throughout the video presentation when measured from the position of the participant). The size of the video matched the size of the screen (1600 × 900 pixels), and the maximum viewing angle was 21º to the left or to the right of the center of the screen.

Apparatus • Auditory stimuli were stored offline as uncompressed wav files and converted to analog signals (44,100 samples/s; 16-bit resolution) using a MacBook Air 2017. They were presented in free field from two active wide-range speakers (3-inch, 6.5 Ohm, 70 Hz to 20 kHz) positioned to the left and to the right of the screens at ca. 90 cm distance from each other and from the participant, forming an equilateral triangle. Participants were tested in a sound-isolated audiological testing chamber at the center, with a 60 dB reverberation time (RT60) of 200 ms and a brightness level of 7650 to 7700 lumens at the participants' position. The stimuli were calibrated with a sound-level meter (NTI Audio XL2 using an NTI M2230 measurement microphone class 1, positioned at the sweet spot (position of participants' heads) during the test. The ambient sound (background noise) level near the subject's head ranged between 28 and 38 dB SPLZ, measured using the same equipment as for stimulus calibration, depending on the period of the day when the testing took place. The stimuli were presented at various intensity levels (30, 40, 50, 60, and 70 dB HL), transformed from the sound pressure levels using the normal equal-loudness-level contours for pure tones under free-field listening conditions (SS-ISO 226:2003). The lowest intensity level was decided based on the average background noise level in the testing chambers.

The tests were performed with a portable Tobii Pro Nano corneal reflection eye tracker placed under a (1600 × 900) TFT 60 Hz 23-inch screen. Participants were seated at ca. 60 cm from the screen. Infants were seated in their parents' laps, while their parents had their eyes closed and were instructed to remain still throughout the test; however, their ears were not shielded from the sounds. As such, the procedure was comparable to the visual reinforcement audiometry procedure at our institute. The gaze calibration procedure was performed before the test. The test consisted of an oddball procedure that captures the attention system's responses to unexpected sounds by frequently (80%) presenting stimuli of one type (standard stimulus: 250 Hz tone), and rarely (20%) stimuli of interest (target stimuli: 500 to 4000 Hz) (Wetzel et al. 2016). Stimuli were presented in five blocks of trials, each block representing an intensity level. The order of the intensity blocks was pseudo-randomized across participants, but no experiment started with a 30- or 70-dB block. In total, participants listened to 10 target and 40 standard stimuli at each intensity level, of which the responses to 10 target and 10 preselected standard stimuli were analyzed. Each target and each relevant standard sound were followed by a filler standard sound, and 10 more filler standard sounds were inserted pseudo-randomly to reach the 40 standard sounds. The order of the stimuli was pseudo-randomized such that the block never started with a target sound and that no more than two target sounds, accompanied by the filler standard sound, were heard in succession. For each block, the response to target sounds was measured in comparison to the standard sound. The stimulus onset interval was jittered around 1000 msec (700 to 1300 msec). The segmented epoch for each trial was 1500 msec to account for the stimulus onset interval between two relevant stimuli (≥1900 msec), thus spanning across a relevant (target or preselected standard) and a subsequent filler standard stimulus. The length of the epoch was selected based on the existing literature that shows pupil response to sounds at around 1 to 1.5 sec after stimulus onset (Wetzel et al. 2016). The total testing duration was around 10 min and was divided into two consecutive testing sessions such that each intensity level was repeated twice for participants who managed to successfully conclude both sessions. The experimental procedure is presented in Figure 1.

Preprocessing and Statistical Analysis • Only the time samples in which both eyes were successfully captured by the eye-tracker were considered for the analysis. The time samples in which only one eye-was recorded (because the blink has already started in the second eye, or for some other reason) were excluded. The mean of left and right pupil size was taken as the pupil measurement. The correction of the pupil size based on the viewing angle was already done by the internal algorithm of the eye-tracker. The 1500 msec after the stimulus onset were analyzed for each trial. We retained trials in which at least 80% of the data were recorded. Outliers exceeding the 10th to 90th percentile, and values higher than 8 and lower than 2 mm were removed. The response in this time window was checked for eye-blinks through a custom script that identified rapid decrease or increase of pupil size that surpassed ±0.3 mm in 40 msec. Such dilation speed outliers preceding and following blinks were replaced by missing values (Kret & Sjak-Shie 2019). Taken together, these steps were considered sufficient to avoid possible artifacts in calculating the mean of the two eyes, especially given that the pupil diameters of both eyes are highly correlated (Jackson & Sirois 2009), especially locally, and

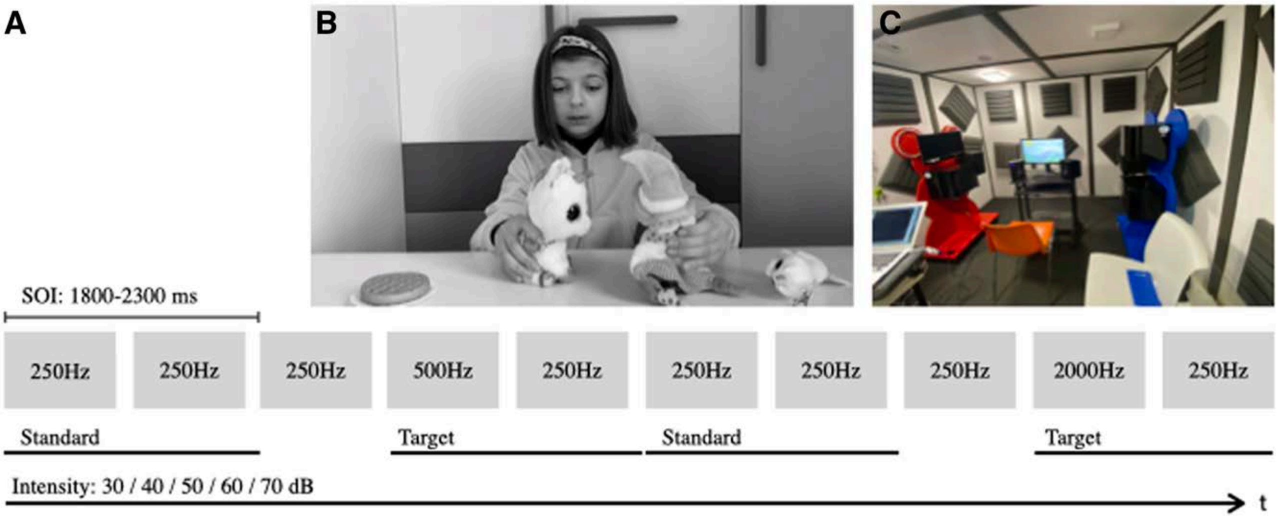

Fig. 1. A, The pictogram of the experimental procedure and stimuli in experiment 1. B, The video shot presented on the screen. C, The photo of one of the four experimental setups.

given the low sampling rate of our eye-tracker (60 Hz). Missing samples were linearly interpolated and then extrapolated with the median pupil value. The data were smoothed using a cubic-spline function model and baseline-corrected with a 300 msec baseline period. For each intensity level, only participants with at least 60% valid responses to target and 60% valid responses to standard stimuli were retained (data from 25 participants in the infant group and all 12 participants from the adult group were further analyzed).

Statistical analysis was performed in R. The analysis script and the anonymized raw dataset are available on the OSF (https://osf.io/cywq7/?view_only=c33ead0d6c364e55b9f392316e43b4f2). To assess the time window in which the presence of a target sound elicited PDR across different intensity levels, we performed permutation tests based on the restricted likelihood ratio test statistic (Lee et al. 2012). Time windows were estimated by computing the cluster-based statistic using the permuted likelihood ratio tests (Maris & Oostenveld 2007; Voeten 2018). The averaged values of the selected time window in each trial were standardized (transformed into *z* scores based on the mean value for each participant at each intensity level) and served as an input for the analysis of variance using Wilcoxon test and for the initial group-based logistic regression analysis.

The non-averaged pupil values in the selected time-window were an input for the within-participant logistic regression (GLM), the linear discriminant analysis (LDA), and the logistic general additive mixed models (GAMM). GLM and LDA are robust classification models with interpretable results and can account for the relatively small datasets. GAMM is an extension of typical regression methods—however, instead of forcing the relation between the dependent variable and the predictor to be linear, as is the case in typical linear regression, this relation is modeled as a smooth function—and has recently been used in pupillometry studies (Porretta & Tucker 2019; van Rij et al. 2019; Fink et al. 2023). The models were run for each individual in each intensity block.

The quality of the model predictions was measured through the non-parametrical receiver operator characteristic (ROC) curve analysis for the overall representation of true (sensitivity) and false positive (specificity) rates at a different cutoff values. On the basis of the area under the ROC curve (AUC), the optimum cutoff value of a model for the highest sensitivity (recall) and specificity is selected using the Youden's *J* statistic. In addition, positive predictive value (PPV or precision; percentage of correctly predicted observations) and *F*1 score (the harmonic mean of sensitivity and PPV) were computed for each model's predictions.

Results

The PDR to Target Stimuli at Different Levels of Intensity • The permutation tests based on the restricted likelihood ratio test statistic revealed the 500 to 1500 msec time window in which target sounds elicited increased PDR compared with standard sounds in experiment 1a at 70 dB, and the 500 to 1500 msec time window in experiment 1b. The average trial time course of the responses is presented in Figure 2A. The averaged values of selected time windows for each trial were aggregated by intensity block and participant.

To evaluate whether less frequent target stimuli elicit an increased PDR both in infants (experiment 1a) and adults (experiment 1b) and whether the amplitude of the response differs at different intensity levels, we first ran the mixed-effect regression models over the averaged data by Trial type (target versus standard stimuli) and intensity level. The R pseudocode was: glm (trial type ~ PDR (standardized *z* scores) × Intensity (30, 40, 50, 60, 70 dB), family = "binomial," data = averaged values). In experiment 1a, there was no significant change in PDR between target and standard trials (analysis of variance: deviance [1, 391] = 1.49, resid = 544.69, $p = 0.22$), but the interaction between intensity and the PDR was significant (deviance [1, 390] = 7.26, resid = 537.42, $p = 0.007$), whereas in experiment 1b, the trial type elicited a significant change in PDR (deviance [1, 237] = 8.16, resid = 324.56, $p = 0.004$), but there was no interaction with the intensity.

To further verify whether the amplitude of the PDR to target stimuli at different intensity levels (response amplitude) can be modeled with a psychometric function, we visualized average *z* scores for target trials at each intensity level (Fig. 2B, mid

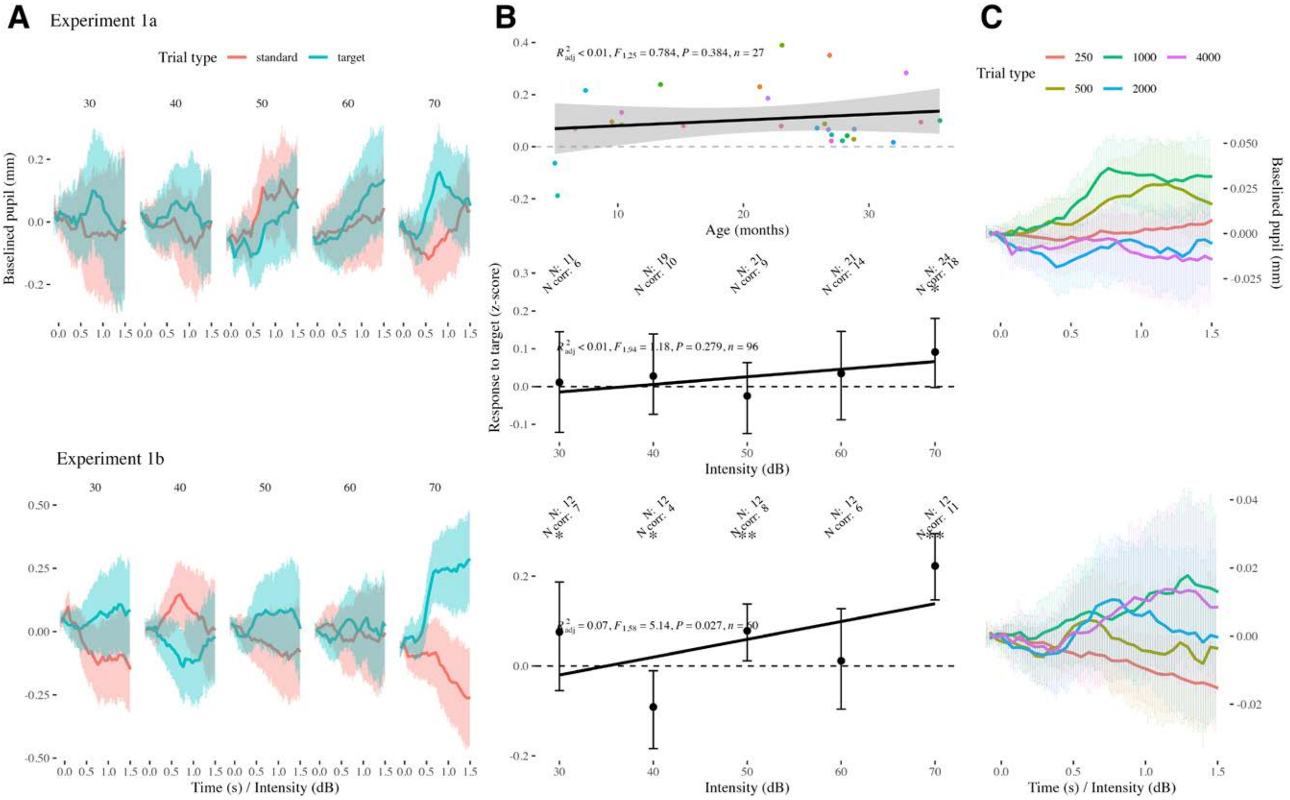

Fig. 2. A, The average trial timeline of the PDR during the 1500 msec post stimulus onset at each tested intensity level for both experiments. Red lines represent average responses to standard, and green to target trials. Shades around the lines of respective colors represent 95% CIs. B, Upper and mid panels: the average response amplitude (*z* scores of responses to target subtracted by the responses to standard trials), plotted as a function of participants' age (black line represents the linear regression line [adj $R^2 < 0.01$, $p = 0.93$]), and as a function of intensity in experiment 1a. Each participant is represented with a different color dot; black dots with error bars represent means and *t*-corrected 95% CI. Lower panel, The average response amplitude for each subject in experiment 1b, plotted as a function of intensity. Error bars represent 95% CIs. C, The average (averaged across participants and stimulus intensity levels) trial timeline of the PDR during the 1500 msec post stimulus onset to target frequencies. CIs indicate confidence intervals; PDR, pupillary dilation response.

and lower panels) and performed a simple test of the average difference between target and standard trials. Eighteen of 25 participants in experiment 1a showed overall higher PDR to target compared with standard stimuli at 70 dB (Welch two-sample test: $t[49.72] = -2.29$, $p = 0.03$, Cohen's $d = 0.63$), while there was no significant difference at other intensity levels. In experiment 1b, 12 of 12 participants showed higher PDR to target compared with standard stimuli for stimuli at 70 dB (Welch two-sample test: $t[21.99] = -6.69$, $p < 0.001$, Cohen's $d = 2.73$), and 8 of 12 at 50 and 30 dB SPL (50 dB: Welch two-sample test: $t[21.88] = -2.88$, $p = 0.008$, Cohen's $d = 1.76$; 30 dB: $t[21.95] = -2.92$, $p = 0.001$, Cohen's $d = 1.19$) (Fig. 2B, see also Table 1 in Supplemental Digital Content 1, http://links.lww.com/EANDH/B617). Given the relatively large age span in the infant group, we also verified whether the age of participants influenced their PDR to unexpected sounds. The standardized pupil responses (*z* scores) to target trials, plotted as a function of age, are presented in Figure 2B, upper panel, and we find no significant effect of age on the average pupil response to target sounds (linear regression model: coeff = 0.002, SE = 0.002, $t = 1.06$, $p = 0.29$) although the average response to target trials was significantly above the mean of the two groups ($t[30] = 2.97$, $p = 0.006$).

In experiment 1, target stimuli were warble tones with the central frequencies of 500, 1000, 2000, and 4000 Hz. To verify whether the PDR was equally elicited by all target tones, two additional generalized mixed-effect models were created with the PDR as the dependent variable and stimulus as the fixed factor. For experiment 1a, the model showed no significant effect of the target stimuli, while in experiment 1b, the 1000 Hz target sound elicited somewhat higher pupil response, although the overall variance explained by the model was not significant (estimate = 0.13, SE = 0.05, $t = 2.5$, $p = 0.02$; $F[4,55] = 1.64$, p value = 0.18). The average trial time course of the responses per stimulus is presented in Figure 2C. Model summaries are available in Supplemental Digital Content 2, http://links.lww.com/EANDH/B618.

Within-Participant Classification at Different Intensity Levels • Another possible method to assess the reliability of the PDR to detect target sounds is to predict whether the PDR can be successfully classified based on the stimulus at the individual level. The results of a classification model can be compared with the behavioral responses because model predictions can be quantified as hits, misses, correct rejections, and false alarms, just like the behavioral results. Here, the response amplitude might not play a central role in the model's success because smaller but consistent responses can be equally successfully classified as stronger ones.

To assess whether the PDR can be correctly classified based on the stimulus type (standard or target tones) in each individual participant, we first applied a simple linear classification of the standardized z scores of the averaged values per trial at each intensity (values above the mean in target trials and values below the mean in the standard trials were considered as true positives and true negatives, and the remaining values as false alarms and misses). The average sensitivity, specificity, and PPV of such classification per experiment and per intensity level, as well as the above-chance classification accuracy (proportion of participants with sensitivity, specificity, and PPV above 0.5), are represented in Figure 3A (green lines), and in Table 1 in Supplemental Digital Content 1, http://links.lww.com/EANDH/B617.

Subsequently, we also applied three classification algorithms on the non-averaged values (GLM, LDA, GAMM). The random effect of time, which would indicate that a different regression line is fitted over time for target and standard trials, was only assessed in the GAMM model (the GLM models, when the random effect of time was added, reported a singular fit or a non-convergence). The R pseudocode for the GAMM models was: trial type ~ PDR + s (time, by = PDR, k = 10), family = "binomial." The GAMM models significantly outperformed the other two models (sensitivity measure: GAMM – GLM: z = 5.43, p < 0.001, GAMM – LDA: z = 6.98, p < 0.001, Kruskal–Wallis multiple comparisons of the Dunn test, p values adjusted with the Bonferroni method, and similarly for specificity and PPV measures), with sensitivity and specificity above-chance level in 25 of 26 infants and in 12 of 12 adults (experiment 1a average values: sensitivity = 0.63, specificity = 0.65, PPV = 0.63, AUC = 0.67; experiment 1b average values: sensitivity = 0.63, specificity = 0.61, PPV = 0.58, AUC = 0.64). However, with this type of analysis, we observed no significant differences between the tested intensity levels. Furthermore, in 37% of infant and 23% of adult analyses, the above-chance accuracy measures did not correspond to an overall increase in pupil size in target trials (Table 1 in Supplemental Digital Content 1, http://links.lww.com/EANDH/B617). Average sensitivity, specificity, and PPV scores at each intensity level are presented in Figure 3A (red lines). Sensitivity, specificity, and PPV scores in randomly selected single participants are presented in Figure 3B. The model fit for the full data models of one of the participants in each group (TS 38 in the infant and PA12 in the adult group) is presented in Figure 4A.

The Amount of Data Necessary to Obtain a Reliable Response on the Individual Level • To address the question of the amount of data necessary for a reliable estimate of the individual pupil response, we compared the results of the analyses when using the full and the reduced dataset, that is, the first six trials at each intensity block. In first six trials, at least two filler trials were present and at least one target sound.

Proportion of participants with increased pupil size during target sounds (column 6, Table 1 in Supplemental Digital Content 1, http://links.lww.com/EANDH/B617) did not significantly differ when first six trials were analyzed, nor did the classification accuracy when using the z score-based linear classification.

However, when applying the logistic GAMM classification model, the model accuracy with reduced data in experiment 1a significantly outperformed the full data model (Welch two-sample t test: $t[182.52] = -7.72$, p < 0.001; average values: sensitivity = 0.75, specificity = 0.77, PPV = 0.86, AUC = 0.81). Similarly, in experiment 1b, the model accuracy with reduced data significantly outperformed the full data model (Welch two-sample t test: $t[81, 21] = -7.61$, p < 0.001; average values: sensitivity = 0.71, specificity = 0.71, PPV = 0.70, AUC = 0.75). Average and individual sensitivity, specificity, and PPV scores at each intensity level are presented in Figure 3 with blue lines. The model fits for the data models for the same participants as earlier are presented in Figure 4B, with model summaries available in Supplemental Digital Content 3, http://links.lww.com/EANDH/B619.

Discussion

The analysis of variance showed that target sounds elicited an increased PDR at 70 dB in both infants and adults and at 30 and 50 dB in adults, thus confirming previously reported studies on pupillometry as an index of auditory orienting response elicited by a rare and unexpected occurrence of a sound deviating from the auditory background (Wetzel et al. 2016; Marois et al. 2018; Bala et al. 2019). While the PDR amplitude increased across intensity levels in the infant group, there was no such trend in the adult group. It therefore remains an open question whether a reliable group threshold can be estimated using pupillometry.

The alternative method, using various classification models in individual participants, showed that it is possible to classify pupil response according to the type of the stimulus (target versus standard sound). The accuracy of the simple linear model was lower than that of the generalized additive model, which was above chance in the vast majority of participants in both groups. The proportion further increased when only the first six trials were analyzed in each intensity block. Nonetheless, the classification success did not correlate with stimulus intensity, and it was not possible to use classification results to create a psychometric curve and estimate the individual auditory thresholds (Fig. 3B). Note, furthermore, that in some participants, the pupil response was systematically different from that in the majority of the participants, such that standard sounds elicited a higher pupil dilation response than the target ones. In some of these participants, too, the classification algorithm correctly classified their PDR according to the type of the stimulus, indicating that the algorithm may be overly sensitive to regularities in the PDR at the individual level, unrelated to auditory attention effects.

In experiment 2, we modified the experiment 1b to test whether the type of the auditory input (speech versus tones) and the type of the task (discrimination versus detection paradigm) affect how the pupil reacts to sounds. Two groups of young adults were tested and the results were compared with the results in experiment 1b.

EXPERIMENT 2

In experiment 2a, we tested whether the detection test yields reliable PDR responses to audible target stimuli, and whether there are systematic differences between the detection and discrimination tests. In experiment 2b, we tested whether listening to speech elicits a comparable response as to tones. The testing procedure and analysis of the data matched those in experiment 1.

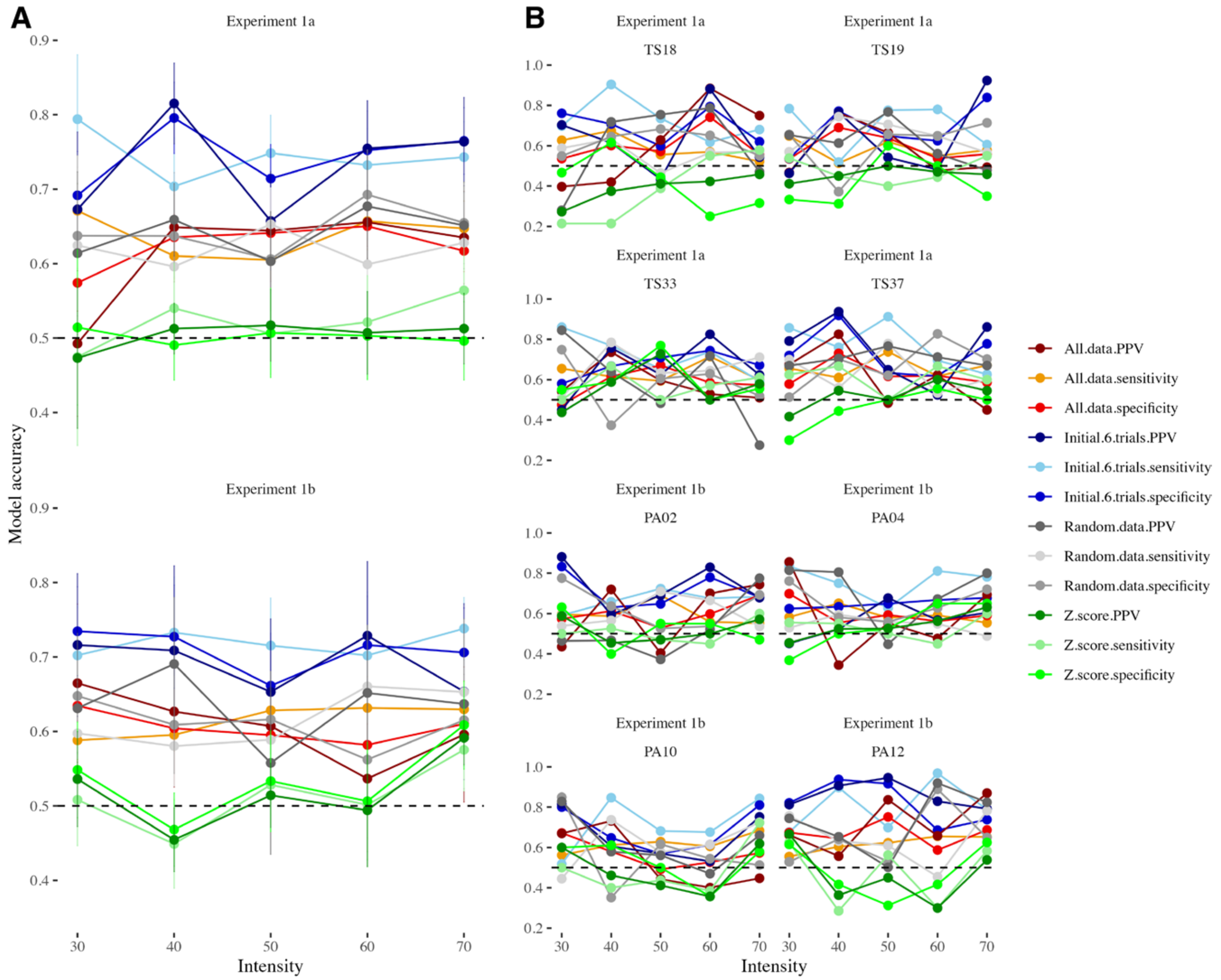

Fig. 3. A, Average sensitivity, specificity, and PPV scores at each intensity level in experiment 1. Red-to-orange lines represent the results of the models computed with all data, navy-to-skyblue lines represent the results of the models computed with first six trials at each intensity level, while gray lines represent the results of the models computed with randomized data. Error bars represent the 95% CI. B, Sensitivity, specificity, and PPV scores at each intensity level for randomly selected participants in both groups, using the same color scheme. CIs indicate confidence intervals; PPV, positive predictive value.

Methods

Participants • Young adults (n = 12, 11 women, mean age: 24 years, SD = 2.2) from various regions of Slovenia were tested in experiment 2a. Young adults (n = 12, 7 women, mean age: 29 years, SD = 3.6) from various regions of Italy were tested in experiment 2b. In both groups, their education level was a university degree or higher, obtained in the fields related to audiology or otorhinolaryngology, speech therapy, or audiometry. Written consent to participate in the study was obtained from all participants before the beginning of the testing. They were unaware of the specific scope of the study and were subsequently informed about it. All participants had normal hearing, normal or corrected visual acuity, no history of psychiatric diseases or neurological problems. All methods were performed in accordance with the relevant guidelines and regulations (e.g., the 1964 WMA Declaration of Helsinki and its later amendments). The testing took place in 2022.

Stimuli and Apparatus • In experiment 2a, target auditory stimuli were the same warble tones as in experiment 1, but with silent periods instead of the 250 Hz tone, thus creating a detection task instead of the oddball discrimination task.

In experiment 2b, the auditory stimuli were Ling-6 sounds ([m], [u], [a], [i], [ʃ], and [s]), which are commonly used to assess access to speech sounds across the speech frequency range. The live voice method of administering Ling-6 sounds was originally proposed by Daniel Ling (Scollie et al. 2012) and is often used to assess whether a child can detect sounds that lie within the speech spectrum of hearing. More recently, Ling-6 sounds have been used within a bracketed threshold measurement procedure in hearing aid research studies. Such studies have demonstrated that calibrated Ling-6 sounds can be used to reliably bracket speech sound detection thresholds, and are sensitive to change across different hearing aid treatments (Glista et al. 2009; Wolfe et al. 2011). Continuous repetitions of the Ling-6 sounds were recorded by a female Italian speaker, after which one instance of each sound was segmented and normalized for intensity. Here, the standard stimulus was the Ling-6 sound/a/ while other sounds served as the target stimuli in the procedure.

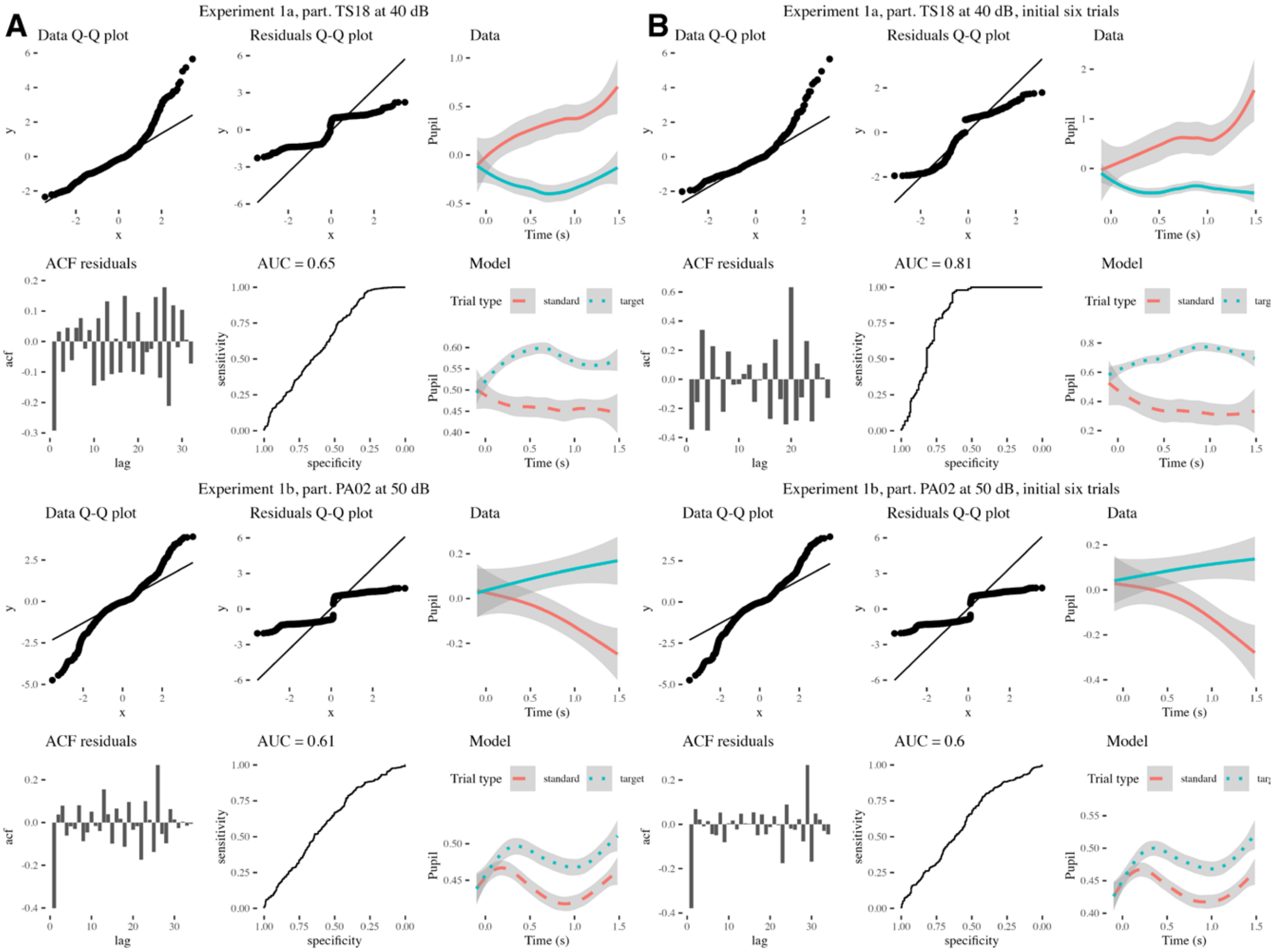

Fig. 4. A, The model fit computed for GAMM models based on the full dataset in the last of the randomly selected single participants from experiments 1a and 1b (Fig. 3B). B, The model fit computed for GAMM models based on the first six trials in the same participants. For each of the two participants and each of the two models, data and residual Q-Q plots to show the quantiles of the data and model residuals distribution, the ACF plot of the model residuals as a metric of model fit, the AUC as a metric of the model performance, and finally the smoothed plots of the actual and modeled data. ACF indicates autocorrelation function; AUC, area under the ROC curve; GAMM, general additive mixed models.

The apparatus in experiment 2 was comparable to the one in experiment 1, except that participants were tested either in the audiological chamber or in a quiet testing room. The ambient sound (background noise) level near the subject's head was in the same range as that of the audiological chamber (between 28 and 38 dB SPLZ), with a 60 dB reverberation time (RT60) of 200 msec matching the one in the audiological room. The brightness level in the testing room was slightly higher and reached around 7950 lumens.

Results

The PDR to Unexpected Stimuli at Different Levels of Intensity • The permutation tests revealed the 500 to 1200 msec time window in experiment 2a and the 500 to 1500 msec time window in experiment 2b. The average trial time course of the responses and the relative difference between the PDR in standard and target trials are presented in Figure 5A. The logistic GLM models showed a significant change in PDR between target and standard trials in both groups (analysis of variance, experiment 2a: deviandce [1, 237] = 12.13, resid = 320.58, $p < 0.001$; experiment 2a: deviandce [1, 217] = 22.96, resid = 282.02, $p < 0.001$), and the interaction between Intensity and the PDR in experiment 2a (deviandce [1, 236] = 4.23, resid = 316.35, $p = 0.04$).

As for experiment 1, the average response to target sounds was plotted as a function of intensity, and the average trial timeline of the PDR to individual stimuli are presented in Figures 5B, C. The difference between target and standard trials was significant at the 30 (Welch two-sample t test: $t[21.99] = -3.15$, $p = 0.005$, Cohen's $d = 1.29$) and 70 dB intensity levels (Welch two-sample t test: $t[21.99] = -2.38$, $p = 0.02$, Cohen's $d = 0.97$) in experiment 2a, and at all levels except 50 dB in experiment 2b (Welch two-sample t test: $t[\geq 19.98] \geq -2.47$, $p \leq 0.02$; see Table 1 in Supplemental Digital Content 2, http://links.lww.com/EANDH/B618).

We verified whether the PDR was equally elicited by all target tones and linguistic sounds. For experiment 2a, the model showed a significant effect of the target stimuli at 1000 (estimate = 0.15, SE = 0.06, $t = 2.33$, $p = 0.02$) and 4000 (estimate = 0.14, SE = 0.06, $t = 2.10$, $p = 0.04$) Hz, but the model as a whole did not explain a significant amount of variance [$F(4,51) = 1.92$, $p = 0.12$]. In experiment 2b, the model explained a significant amount of variance [$F(5,66) = 2.96$, $p = 0.02$], with a significant effect of the sounds [s] (estimate = 0.30, SE = 0.11,

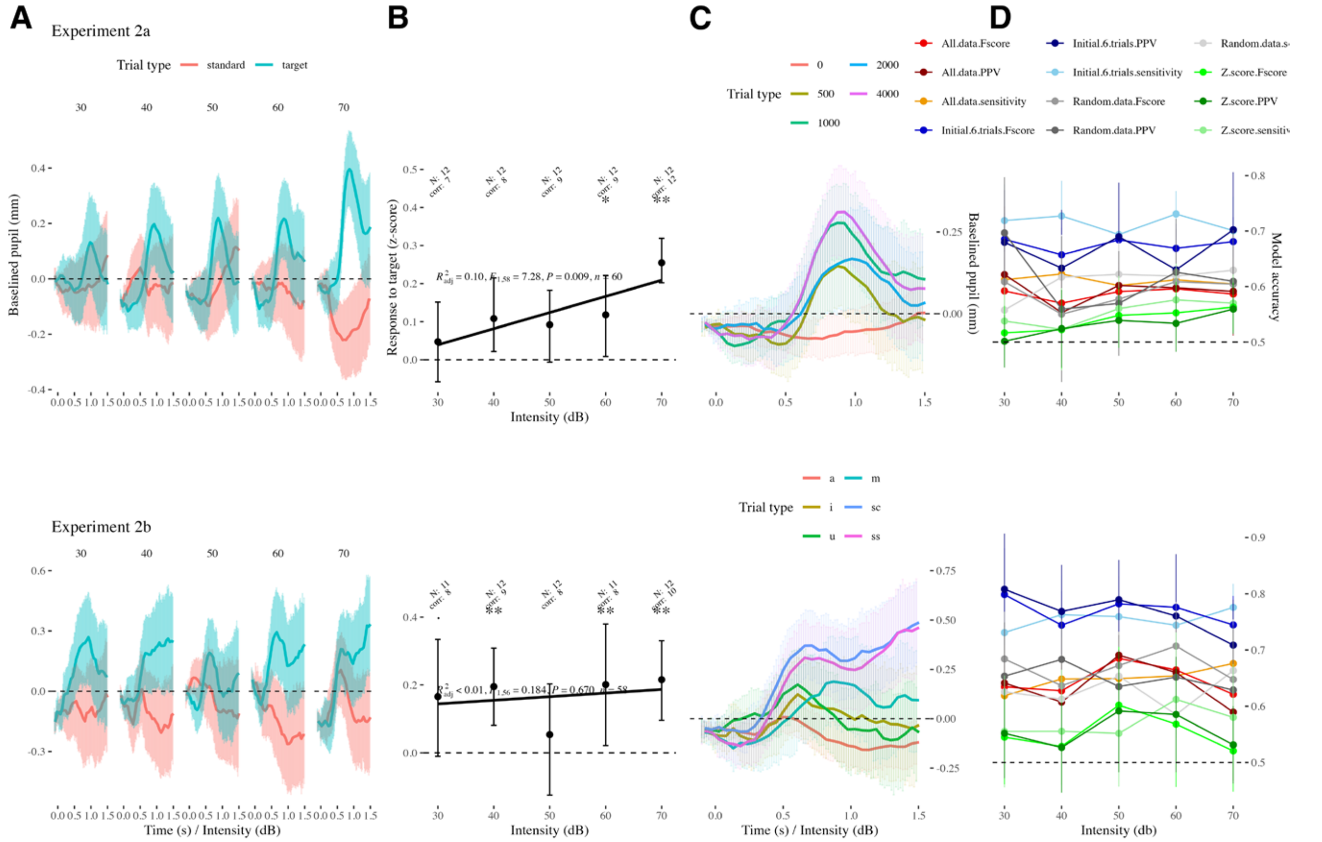

Fig. 5. A, The average trial timeline of the PDR during the 1500 msec post-stimulus onset at each tested intensity level for both experiments. Shades around the lines represent 95% CIs. B, The average response amplitude (responses to target subtracted by the responses to standard trials), plotted as a function of intensity. Error bars represent 95% CIs. C, The average (averaged across participants and stimulus intensity levels) trial timeline of the PDR during the 1500 msec post-stimulus onset to target frequencies. D, Average sensitivity, specificity, and PPV scores at each intensity level in experiment 2. Red-to-orange lines represent the results of the models computed with all data, navy-to-skyblue lines represent the results of the model computed with first six trials at each intensity level, while gray lines represent the results of the models computed with randomized data. Error bars represent the 95% CI. CIs indicate confidence intervals; PDR, pupillary dilation response.

t = 2.84, p = 0.01) and [ʃ] (estimate = 0.33, SE = 0.11, t = 3.10, p = 0.002) compared with the standard sound/a/. Model details are presented in Supplemental Digital Content 2, http://links.lww.com/EANDH/B618.

Within-Participant Classification at Different Intensity Levels and With First Six Trials Analyzed • When classification algorithms were applied on individual participants' data at each intensity level, GAMM again outperformed the linear z score model and the GLM and LDA models (gamm – glm: z = 4.37, p < 0.001, gamm – lda: z = 6.74, p < 0.001 [Kruskal–Wallis multiple comparison using the Dunn test, Bonferroni correction]), with sensitivity and specificity above-chance level in all but one participant in experiment 2b (experiment 2a average values: sensitivity = 0.60, specificity = 0.63, PPV = 0.62, AUC = 0.64; experiment 2b average values: sensitivity = 0.64, specificity = 0.67, PPV = 0.67, AUC = 0.69). As in experiment 1, we observed no significant differences between the tested intensity levels.

We also applied the logistic GAMM classification model to only the first six trials at each intensity level. The reduced model accuracy significantly outperformed the full data model both in experiment 2a (Welch two-sample t test: t[98.89] = −5.10, p < 0.001; average values: sensitivity = 0.74, specificity = 0.72, PPV = 0.69, AUC = 0.76) and in experiment 2b (Welch two-sample t test: t[94.93] = −6.29, p < 0.001; average values: sensitivity = 0.78, specificity = 0.78, PPV = 0.77, AUC = 0.84). Average sensitivity, specificity, and PPV scores at each intensity level are presented in Figure 5D. Sensitivity, specificity, and PPV scores in randomly selected single participants, along with the model fit in both experiments for the full data model in one of the participants in each experiment, are presented in Figure 3 in Supplemental Digital Content 3, http://links.lww.com/EANDH/B619.

The Effect of Task and Stimulus Type • Last, the accuracy measures of the GAMM—sensitivity, specificity, and PPV—were compared across the testing groups in experiments 1 and 2 using the Kruskal–Wallis multiple comparison tests with p values adjusted with the Bonferroni method. Results in experiment 2b, where participants listened to linguistic sounds, were significantly higher than those in other groups (sensitivity: experiment 1b to experiment 2b: z = −2.72, p = 0.03, confidence interval [CI] = −0.06 to −0.002; experiment 2a to experiment 2b: z = −2.65, p = 0.045, CI = [−0.06 to −0.009]; specificity: experiment 1b to experiment 2b: z = −3.06, p = 0.01, CI = [−0.07 to −0.012]; F score: experiment 2a to experiment 2b: z = −3.23, p = 0.01, CI = [−0.077 to −0.007]). No significant differences were measured between the results in the infant group (experiment 1a) and any of the adult groups in the study. Model details are presented in Supplemental Digital Content 4, http://links.lww.com/EANDH/B620.

Explorative Randomized Trial Type Analysis • Given the surprisingly accurate GAMM model predictions at the within-participant level, as compared with the simple linear classification of the standardized average pupil sizes (*z* scores), we conducted an additional GAMM analysis that included the same data and the same model structure as in the original model analysis but with trials randomly assigned as target or standard trials. This analysis allowed us to observe potential over-fit of the nonlinear models at the within-participant level. The sensitivity and specificity of the within-participant models using randomized trial types were compared with the measures from the models run on the actual data. We found no significant differences between the sensitivity, specificity, or the PPV of the models run on actual and those run on the randomized data. Average sensitivity, specificity, and PPV scores at each intensity level of the random models, compared with the accuracy measures of the model based on actual data, are presented in Figures 3A, 5D (gray lines). Nonetheless, the number of participants in whom both model sensitivity and model specificity exceeded the chance level (0.5) decreased significantly when models were run on the randomized data, from 71 to 58% when all data were analyzed (*t*[541.55] = 3.25, *p* value = 0.001), and from 91 to 74 when the first six trials were analyzed. The same trend was visible when we compared model performances on the first six trials of the randomized and actual datasets. Model details are available in Supplemental Digital Content 4, http://links.lww.com/EANDH/B620.

Discussion

In experiment 2b, where adults listened to linguistic sounds in an oddball (discrimination) task, participants' pupils responded to target sounds more consistently than in experiments 1b and 2a, where young adults listened to tones. Although the difference was not large, it may indicate an overall better attentional response to linguistic sounds compared with nonlinguistic and acoustically simpler warble tones (Kuhl 1991), which may at the same time correspond to distinct (attentional) neural networks that are activated for linguistic and nonlinguistic sounds (Quillen et al. 2021; Jurov et al. 2023).

It is interesting that in experiment 2b, sounds [s] and [ʃ] elicited the most consistent PDR. It is possible that this response is due to the acoustic properties of these sounds. Being fricatives, they are aperiodic, with most of their acoustic energy residing in the higher part of the spectrum (roughly, above 4 kHz for [s] and above 2.5 kHz for [ʃ]), while other Ling-6 sounds are all sonorants and thus periodic, with the acoustic energy concentrated in the lower part of the spectrum. This high-frequency energy can make them more easily distinguishable from other phonemes, especially in a clean, quiet environment, and it has been shown that pupillary responds are, on average, bigger for highly distinguishable sounds (Liao et al. 2016; Wetzel et al. 2016).

In experiment 2a, where young adults listened to warble tones in a detection task, the PDR to target trials as compared with standard trials showed marginally significant linear increase across the intensity levels, as we had hypothesized based on previous reports (Liao et al. 2016; Bala et al. 2019; Schilling et al. 2019), but the responses did not differ significantly from (adult) results in the discrimination tasks in experiments 1b and 2b. The elicited PDR response is, however, more clearly visible in this task in the form of a consistent increase of the pupil size around one second after the stimulus onset (Liao et al. 2016; Wetzel et al. 2016). The detection task may thus, despite the lack of a significant advantage over the discrimination task, nonetheless represent a more easily interpretable version of the test.

GENERAL DISCUSSION

Here we explore (1) whether the PDR to audible sounds can be reliably measured in normally hearing infants within their average attention span, and in normally hearing adults, (2) how accurate within-participant models are in classifying PDR based on the stimulus type at various intensity levels, (3) whether the amount of analyzed data affects the model reliability, and (4) whether we can observe systematic differences in the PDR between speech and nonspeech sounds, and between the discrimination and detection paradigms.

The idea to use pupillometry, that is, the pupil diameter response (PDR), to measure individual's response to unexpected sounds was conceived above all for the populations that might be, for various reasons, unable to respond behaviorally, most notably infants and young children. Participants belonging to such populations also impose time limitations as to how long they are willing to attend to a task and how many (physiological) measures can be obtained from them. Whereas this idea has been explored and even patented for infants (US 2020/0253526 A1 2020) and hearing-impaired adults as a method for adjusting hearing aids (EP 3 481 086 B1 2019; US 10,609,493 B2 2020), the actual accuracy of any model to predict the individual response to sounds with little and noisy data could represent an obstacle for the applicability of pupillometry in real-life diagnostic procedures.

The present study explores whether the relative increase of pupil diameter in response to rare or unexpected events (PDR) can serve as a reliable measure of auditory attention in infants and adults. The results show, across groups and across tasks, a consistent categorical response—an increase of the pupil size to target auditory stimuli for most participants. The PDR to target stimuli was the most consistent at the highest tested intensity level, 70 dB SPL, and the correlation between the intensity of the stimulus and the pupil size response was visible in the infant group (experiment 1a) and for the detection task (experiment 2a) while the trend was missing in the other two groups.

The response to the question regarding how accurate within-participant models are in classifying PDR based on the stimulus type at various intensity levels remains unanswered. In some participants, unexpected events do not elicit an increase in pupil size, and we can only speculate about the possible individual differences in the attentional capture, or uncontrolled factors that may have influenced the pupil dilation (cf., Johnson et al. 2014; Zekveld & Kramer 2014). When the nonlinear classification models were applied at the individual level, model accuracy measures were relatively high, in some cases even when the PDR to target stimuli was not in the expected direction, and the correspondence between the models and the test's ability to capture auditory attention remained unclear. The exploratory analysis reveals that the models, by being able to discriminate any, even randomly attributed categories, are not reliably detecting the effect of the sound to the pupil size. The reliability of the within-participant models in comparison with the existing behavioral methods is further discussed in Is pupillometry applicable as an additional objective measure of hearing threshold?, Amount of data and model accuracies, and PDR to stimulus type address the last two questions of this study, the amount of data and the impact of stimulus type to the results.

An additional point that emerged from the data analysis was that the age did not play an important role in participants' pupil response. No significant effect of age to the average PDR to target sounds indicates that it may be possible to use the method also in younger groups of infants. The lack of significant differences in the general time-course of the pupil response between infants and adults is in line with the previous studies that directly compared pupil responses in infants and adults (Wetzel et al. 2016; Zhang et al. 2019). Pupil response, as an index of information processing, may undergo a similar developmental trajectory as other indices of information processing, where the acceleration is most obvious in the first 8 months of life and becomes less visible after the second year of life (Hepach & Westermann 2016; Hochmann and Kouider 2022). In our infants' sample, the average age was 21 months and a comparable latency of the response was therefore expected.

Is Pupillometry Applicable as an Additional Objective Measure of Hearing Threshold?

The methods that proposed to measure hearing with pupillometry assume that the auditory threshold can be defined somewhere between the quietest sound that elicits a statistically significant PDR and the amplitude of a quiet sound that does not yield such a dilation (Liao et al. 2016; Bala et al. 2019). As such, pupillometry may be comparable to other indexes of auditory information processing (Hepach & Westermann 2016; Schilling et al. 2019). In our results, we observed an increase of the PDR amplitude across intensity levels in infants (experiment 1a) and in adults who were exposed to the detection task. While the individual thresholds, using the within-participant classification methods, could not be reliably estimated, the trend in these two groups is comparable with the previously reported data (Liao et al. 2016).

To estimate the reliability of the method to assess hearing, these results might benefit from a comparison with the results obtained through behavioral tests of auditory threshold detection including the visual reinforcement audiology procedure. These behavioral tests often use different types of adaptive staircase procedures adapted to infants, such as parameter estimation by sequential testing for infants or the observer-based psychoacoustic procedure, and their accuracy has been reported up to around 80% for loud stimuli at 60 to 80 dB and up to around 65% for stimuli at 30 to 40 dB (Sinnott et al. 1983; Sinnott & Aslin 1985; Trehub et al. 1986; Olsho et al. 1987). The results of the present study show similar accuracy levels for loud stimuli (76% in infants), while for low-intensity stimuli, we obtain somewhat lower results (52% in infants; see Figs. 2B, 5B). The proportion of participants with the accurate nonlinear model (GAMM) predictions was higher than in the behavioral tests; nonetheless, the value of these predictions can be questioned given the results based on the randomized Trial type.

Thus, pupillometry appears to have accuracy levels comparable to behavioral psychoacoustic procedures beyond the already known advantages, such as being suitable for a wider array of participants, less time-consuming, and requiring fewer human resources than traditional behavioral methods. Nonetheless, the present results do not offer a definite answer as to which task (detection or discrimination), which stimuli (speech or nonspeech), and which classification algorithm (a simple linear classification of standardized scores, linear, or a nonlinear logistic regression model) would yield the most reliable and systematic way of using pupillometry as an index of auditory attention and, consequently, hearing.

Simultaneously, the potential of the pupillometric test to assess either the threshold of the audibility or the comfortable listening level threshold could be further explored (Moore et al. 2011). For example, the ambient noise levels during testing were relatively high and they impacted the decision regarding the lowest measured intensity level (30 dB SPL). More controlled testing conditions and/or non-free-field presentation of the stimuli may allow the test of the PDR at lower intensities, possibly using an adaptive procedure, and compare them to the results in the previous studies (Liao et al. 2016; Bala et al. 2019).

Amount of Data and Model Accuracies

On the one side, pupillary response may be more reliable and robust in initial exposures to stimuli, potentially due to the novelty and heightened attentional capture of the first few sounds (Burg et al. 2021; Mathôt & Vilotijević 2023). Analyzing fewer trials reduces the overall testing time, making the procedure more practical and less burdensome, especially for populations with limited attention span such as infants and young children. Additionally, shorter testing sessions minimize the risk of data degradation due to participant fatigue, boredom, or habituation to the stimuli (Hochmann 2013).

On the other side, during the first trials, the effect of the habituation to filler sounds may not occur yet. In addition, throughout the analysis, pupil response also included the response to filler standard sounds which in some trials occurred when the event-related pupil dilation has not yet reached its maximum (sometimes as early as 700 msec after the target sound). The reason for inserting the filler standard sounds was exactly to enhance habituation effect in the oddball procedure while at the same time avoid additional lengthening of the experimental procedure (given the limited attention span in infants). Such filler sounds may have influenced the habituation process, especially at the beginning of each testing block, and alter the time-course of pupil dilation in response to target sounds—as indirectly observable from the visual inspection of the differences in average time courses of pupil dilation in experiments 1b and 2a where filler sounds have been replaced by silence (Fig. 2C versus Fig. 5C).

It was thus promising to see that, when nonlinear models were applied on the individual level, classification accuracy of the PDR to auditory stimuli improved when fewer trials (the first six trials at each intensity level) were analyzed. However, no difference in the model accuracy measures was found when the models were applied to the data with fewer trials in which trials were randomly attributed to the trial type (target or standard), indicating once more that the algorithm may be overly sensitive to the regularities in the PDR at the individual level, unrelated to the auditory attention effects. The results are, therefore, not conclusive regarding the optimal amount of data needed to reliably estimate pupil response.

The comparison of the accuracies of the simple *z* scores-based linear models when full and reduced datasets were analyzed confirms this inconclusive result. No significant differences were found between the analysis of the first six trials and the full dataset. Nonetheless, this might indicate that a reliable estimate of individual auditory attention is possible with a relatively small amount of PDR data.

The lack of differences might also indicate that the discrimination between low- and high-frequency sounds elicited the increased PDR, independently of the habituation (Montes-Lourido et al. 2021). Disentangling the effects that these two cognitive processes—discrimination between sounds and the rapid detection of frequency deviants (unexpected events)—may have on the pupil dilation process, is unfortunately beyond the scope of the present study. However, understanding of the event-related events in the pupil response in more detail could be of both clinical and research interest in the future.

PDR to Stimulus Type

In experiment 2b, where participants listened to linguistic sounds, the PDR was slightly more consistent compared with responses to nonspeech warble tones in experiments 1b and 2a. This suggests that speech sounds may elicit stronger or more reliable attentional responses than simpler nonspeech stimuli. The increased consistency in responses to linguistic sounds may be attributed to the inherent complexity and salience of speech, which is a biologically and socially significant signal for humans (Liberman & Mattingly 1985; Holt & Lotto 2010). The sounds [s[and [ʃ] in particular elicited the most consistent PDRs, likely due to their distinct acoustic properties such as aperiodicity and high spectrum, which make them easier to recognize (Tamási et al. 2017).

Because speech sounds are more ecologically valid and relevant in everyday communication, using them in auditory assessments could enhance the diagnostic utility of pupillometry, especially in evaluating speech perception and processing in individuals with hearing impairments (Scollie et al. 2012; Wolfe et al. 2015). Moreover, the ability to detect and respond to speech sounds is critical for language development in infants and young children, making it essential to incorporate speech stimuli in pediatric auditory assessments (Tamási et al. 2016).

Limitations

This study, while providing valuable insights into the potential of pupillometry as an auditory assessment tool, has at least two limitations that warrant consideration. First, the sample sizes were relatively small, consisting of 36 infants and 36 adults divided into three groups, with comparable results across groups. The relatively small sample sizes and specific demographics of the participants may limit the generalizability of the findings. However, while studies including individual (within-participants) models were not found, group-based analyses of pupil size as a function of auditory or audiovisual attention have been using samples of similar sizes (Wetzel et al. 2016; Bala et al. 2019; Zhao et al. 2019; Calignano et al. 2023).

Second, the high ambient noise levels during testing could have influenced the pupillary responses. The ambient noise in the testing rooms was relatively high compared with the standards for the audiological testing, and we acknowledge that this might have compromised our results to some extent. More specifically, the background noise may have affected the results at the lower intensity levels by hindering the response. The difference between the PDR amplitude at the lowest and at the highest intensity levels was visible, especially in infants, who reportedly tolerate lower levels of background noise (McMillan & Saffran 2016; Saksida et al. 2022). While testing conditions could not be changed retroactively, we acknowledge that for a clearer interpretation of the results, the more controlled testing environments are necessary in the future research.

Third, given that it differed for participants in experiments 1 and 2, the brightness level as a possible factor overlapped with the group factor in the analysis. By adding luminosity level, a redundant factor would be added in the analysis. To assess the possible effect of the luminosity level changes, a different experimental setup might be needed.

Conclusions

While measuring the PDR to unexpected sounds is a promising tool, providing accuracy levels comparable to traditional behavioral psychoacoustic procedures, further research is needed to define the most efficient protocol that would be suitable for diverse populations, with stimuli that would reliably elicit the PDR, and that would minimize the testing time while controlling for the possible sources of noise. Furthermore, the potential of the pupillometric test to assess either the audibility or the comfortable listening level threshold needs to be further explored. Addressing these points will help validate the utility of pupillometry in auditory assessments, potentially establishing it as a tool in both clinical and research settings.

ACKNOWLEDGMENTS

The authors thank the participants, parents, and infants and children for their participation. The authors also thank Eleonora Tonon, Andrea Martina Santoro, Anita Kastelic, Mateja Božič, and Associazione Villaggio per Crescere for their help with data acquisition. Special thanks to Associazione Ciao ci sentiamo Onlus for additional support with travel. The authors also thank Giampietro Ricci, Valeria Gambacorta, and Michele Orzan for help with participant recruitment, as well as Massimo Grassi and Francesco Pavani for discussions on the earlier versions of the manuscript.

This study has received funding from the Italian Health Ministry through the Institute for Maternal and Child Health—IRCCS "Burlo Garofolo"—Trieste (RC 15/21), and Associazione Italiana Liberi di Sentire.

A.S. and S.Ž. (sharing the first authorship) conceptualized and designed the study, drafted the manuscript, carried out the initial analyses, and critically reviewed and revised the manuscript. S.B. and A.S. coordinated and supervised data collection, carried out the initial analyses, and critically reviewed and revised the manuscript. A.S. and E.O. conceptualized and designed the study. E.O. coordinated and supervised data collection, and critically reviewed and revised the manuscript for important intellectual content. All authors have reviewed the content of the article and have approved the final version of the manuscript submitted for publication.

The study was approved by the regional ethical committee (Comitato Etico Unico Regionale, Agenzia Regionale di Coordinamento per la Salute, Friuli-Venezia-Giulia, Approval no. CEUR-2021-Os-97) in 2021, under the framework of the research project 15/21, nominated by the Italian Ministry of Health (Ufficio per la Ricerca Clinica IRCCS Burlo Garofolo). All methods were performed in accordance with the relevant guidelines and regulations (e.g., 1964 WMA Declaration of Helsinki and its later amendments). The testing took place in 2021 and 2022. Parents of participants and participants were informed about the study and signed the informed consent to collect, analyze, and publish results from the study.

Anonymized anamnestic data, the analysis script, and the anonymized raw dataset are available in the OSF data repository at https://osf.io/cywq7/?view_only=5af730403fce48faae5565e8f21c1524.

The authors have no conflicts of interest to disclose.

Address for correspondence: Amanda Saksida, Pediatric Audiology and Otolaryngology Unit, Institute for Maternal and Child Health - IRCCS "Burlo Garofolo" - Trieste, via dell'Istria 65/1, 34137, Trieste, Italy. Email: amanda.saksida@burlo.trieste.it

Received May 31, 2024; accepted January 24, 2025

REFERENCES

Ahn, J. H., Lee, H. S., Kim, Y. J., Yoon, T. H., Chung, J. W. (2007). Comparing pure-tone audiometry and auditory steady state response for the measurement of hearing loss. *Otolaryngol Head Neck Surg*, *136*, 966–971.

Bala, A. D. S, & Takahashi, T. T. (2020). *US 2020/0253526 A1*. United States.

Bala, A. D. S., Whitchurch, E. A., Takahashi, T. T. (2019). Human auditory detection and discrimination measured with the pupil dilation response. *J Assoc Res Otolaryngol*, *21*, 43–59.

Burg, E. A., Thakkar, T., Fields, T., Misurelli, S. M., Kuchinsky, S. E., Roche, J., Litovsky, R. Y., Litovsky, R. Y. (2021). Systematic comparison of trial exclusion criteria for pupillometry data analysis in individuals with single-sided deafness and normal hearing. *Trends Hear*, *25*, 23312165211013256.

Calignano, G., Dispaldro, M., Russo, S., Valenza, E. (2021). Attentional engagement during syllable discrimination: The role of salient prosodic cues in 6- to 8-month-old infants. *Infant Behav Dev*, *62*, 101504.

Calignano, G., Girardi, P., Altoè, G. (2023). First steps into the pupillometry multiverse of developmental science. *Behav Res Methods*, *56*, 3346–3365.

Fink, L., Simola, J., Tavano, A., Lange, E., Wallot, S., Laeng, B. (2023). From pre-processing to advanced dynamic modeling of pupil data. *Behav Res Methods*, *56*, 1376–1412.

Glista, D., Scollie, S., Bagatto, M., Seewald, R., Parsa, V., & Johnson, A. (2009). Evaluation of nonlinear frequency compression: Clinical outcomes. *Intern J Audiol*, *48*, 632–644.

Gordon, K. A., Papsin, B. C., Harrison, R. V. (2004). Toward a battery of behavioral and objective measures to achieve optimal cochlear implant stimulation levels in children. *Ear Hear*, *25*, 447–463.

Hepach, R., & Westermann, G. (2016). Pupillometry in infancy research. *J Cogn Dev*, *17*, 359–377.

Hochmann, J. R. (2013). Pupillometry in six-month-old infants. In Proceedings of the 37th annual conference on language development. Cascadilla Press.

Hochmann, J.-R., & Kouider, S. (2022). Acceleration of information processing en route to perceptual awareness in infancy. *Current Biology*, 1–5.

Hochmann, J.-R., & Papeo, L. (2014). The invariance problem in infancy: A pupillometry study. *Psychol Sci*, *25*, 2038–2046.

Holt, L. L., & Lotto, A. J. (2010). Speech perception as categorization. *Atten Percept Psychophys*, *72*, 1218–1227.

Jackson, I., & Sirois, S. (2009). Infant cognition: Going full factorial with pupil dilation. *Dev Sci*, *12*, 670–679.

Johnson, E. L., Miller Singley, A. T., Peckham, A. D., Johnson, S. L., Bunge, S. A. (2014). Task-evoked pupillometry provides a window into the development of short-term memory capacity. *Front Psychol*, *5*, 1–8.

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

Jurov, N., Idsardi, W., Feldman, N. H. (2023). *A neural architecture for selective attention to speech features*. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2023-August, 1778–1782.

Kret, M. E., & Sjak-Shie, E. E. (2019). Preprocessing pupil size data: Guidelines and code. *Behav Res Methods*, *51*, 1336–1342.

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Percept Psychophys*, *50*, 93–107. https://link.springer.com/content/pdf/10.3758/BF03212211.pdf.

Lee, O. E., Braun, T. M., Arbor, A. (2012). Permutation tests for random effects in linear mixed models. *Biometrics*, *68*, 486–493. https://doi.org/10.1111/j.1541-0420.2011.01675.x.

Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Percept Psychophys*, *63*, 1279–1292.

Liao, H., Kidani, S., Yoneya, M., Kashino, M., Furukawa, S. (2016a). Correspondences among pupillary dilation response, subjective salience of sounds, and loudness. *Psychon Bull Rev*, *23*, 412–425.

Liao, H. I., Yoneya, M., Kidani, S., Kashino, M., Furukawa, S. (2016b). Human pupillary dilation response to deviant auditory stimuli: Effects of stimulus properties and voluntary attention. *Front Neurosci*, *10*, 43.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revisited. *Cognition*, *21*, 1–36.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *J Neurosci Methods*, *164*, 177–190.

Marois, A., Labonté, K., Parent, M., Vachon, F. (2018). Eyes have ears: Indexing the orienting response to sound using pupillometry. *Int J Psychophysiol*, *123*, 152–162.

Marois, A., & Vachon, F. (2018). Can pupillometry index auditory attentional capture in contexts of active visual processing? *J Cogn Psychol*, *30*, 484–502.

Mathôt, S., & Vilotijević, A. (2023). Methods in cognitive pupillometry: Design, preprocessing, and statistical analysis. *Behav Res Methods*, *55*, 3055–3077.

McMillan, B. T. M., & Saffran, J. R. (2016). Learning in complex environments: The effects of background speech on early word learning. *Child Dev*, *87*, 1841–1855.

Montes-Lourido, P., Kar, M., Kumbam, I., Sadagopan, S. (2021). Pupillometry as a reliable metric of auditory detection and discrimination across diverse stimulus paradigms in animal models. *Sci Rep*, *11*, 1–15.

Moore, R., Gordon-Hickey, S., Jones, A. (2011). Most comfortable listening levels, background noise levels, and acceptable noise levels for children and adults with normal hearing. *J Am Acad Audiol*, *22*, 286–293.

Olsho, L. W., Koch, E. G., Halpin, C. F., Carter, E. A. (1987). An observer-based psychoacoustic procedure for use with young infants. *Dev Psychol*, *23*, 627–640.

Porretta, V., & Tucker, B. V. (2019). Eyes wide open: Pupillary response to a foreign accent varying in intelligibility. *Front Commun*, *4*, 1–12.

Quillen, I. A., Yen, M., Wilson, S. M. (2021). Distinct neural correlates of linguistic and non-linguistic demand. *Neurobiol Lang (Camb)*, *2*, 202–225.

Quirins, M., Marois, C., Valente, M., Seassau, M., Weiss, N., El Karoui, I., Naccache, L., Naccache, L. (2018). Conscious processing of auditory regularities induces a pupil dilation. *Sci Rep*, *8*, 1–11.

Saksida, A., Ghiselli, S., Picinali, L., Pintonello, S., Battelino, S., Orzan, E. (2022). Attention to speech and music in young children with bilateral cochlear implants: A pupillometry study. *J Clin Med*, *11*, 1745.

Schilling, A., Gerum, R., Krauss, P., Metzner, C., Tziridis, K., Schulze, H. (2019). Objective estimation of sensory thresholds based on neurophysiological parameters. *Front Neurosci*, *13*, 1–12.

Scollie, S., Glista, D., Tenhaaf, J., Dunn, A., Malandrino, A., Keene, K., Folkeard, P. (2012). Stimuli and normative data for detection of ling-6 sounds in hearing level. *Am J Audiol*, *21*, 232–241.

Sinnott, J. M., & Aslin, R. N. (1985). Frequency and intensity discrimination in human infants and adults. *J Acoust Soc Am*, *78*, 1986–1992.

Sinnott, J. M., Pisoni, D. B., Aslin, R. N. (1983). A comparison of pure tone auditory thresholds in human infants and adults. *Infant Behav Dev*, *6*, 3–17.

Strauch, C., Wang, C. A., Einhäuser, W., Van der Stigchel, S., Naber, M. (2022). Pupillometry as an integrated readout of distinct attentional networks. *Trends Neurosci*, *45*, 635–647.

Tamási, K., Mckean, C., Gafos, A., Fritzsche, T., Höhle, B. (2017). Pupillometry registers toddlers' sensitivity to degrees of mispronunciation. *J Exp Child Psychol*, *153*, 140–148.

Tamási, K., Wewalaarachchi, T. D., Höhle, B, Singh, L. (2016). Measuring sensitivity to phonological detail in monolingual and bilingual infants using pupillometry. In *Proceedings of the 16th Speech Science and Technology Conference*.

Trehub, S. E., Bull, D., Schneider, B. A., Morrongiello, B. A. (1986). PESTI: A procedure for estimating individual thresholds in infant listeners. *Infant Behav Dev*, *9*, 107–118.

van Rij, J., Hendriks, P., van Rijn, H., Baayen, R. H., Wood, S. N. (2019). Analyzing the time course of pupillometric data. *Trends Hear*, *23*, 1–22.

Visram, A. S., Innes-Brown, H., El-Deredy, W., McKay, C. M. (2015). Cortical auditory evoked potentials as an objective measure of behavioral thresholds in cochlear implant users. *Hear Res*, *327*, 35–42.

Voeten, C. C. (2018). *Permutes: Permutation Tests for Time Series Data. R package version 0.1*. Retrieved from https://cran.r-project.org/package=permutes.

Wendt, D., Lunner, T., Ksiazek, P., Alickovic, E. (2019). *EP 3 481 086 B1*.

Wendt, D., Lunner, T., Książek, P., Alickovic, E. (2020). *US 10,609,493 B2*. United States of America.

Wetzel, N., Buttelmann, D., Schieler, A., Widmann, A. (2016). Infant and adult pupil dilation in response to unexpected sounds. *Dev Psychobiol*, *58*, 382–392.

Widen, J. E., Folsom, R. C., Cone-Wesson, B., Carty, L., Dunnell, J. J., Koebsell, K., … Norton, S. J. (2000). Identification of Neonatal Hearing Impairment: Hearing status at 8 to 12 months corrected age using a visual reinforcement audiometry protocol. *Ear Hear*, *21*, 471–487.

Wolfe, J., John, A., Schafer, E., Nyffeler, M., Boretzki, M., Caraway, T., & Hudson, M. (2011). Long-term effects of non-linear frequency compression for children with moderate hearing loss. *Intern J Audiol*, *50*, 396–404.

Wolfe, J., Schafer, E., Mills, E., John, A., Hudson, M., Anderson, S. (2015). Evaluation of the benefits of binaural hearing on the telephone for children with hearing loss. *J Am Acad Audiol*, *26*, 93–100.

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

Zhang, F., Jaffe-Dax, S., Wilson, R. C., Emberson, L. L. (2019). Prediction in infants and adults: A pupillometry study. *Dev Sci*, *22*, 1–9.

Zhao, S., Bury, G., Milne, A., Chait, M. (2019). Pupillometry as an objective measure of sustained attention in young and older listeners. *Trends in hearing*, *23*, 1–21. https://doi.org/10.1177/2331216519887815.