Heliyon 8 (2022) e09631

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Comparing methods of analysis in pupillometry: application to the assessment of listening effort in hearing-impaired patients

Lou Seropian [a], Mathieu Ferschneider [a,b], Fanny Cholvy [a,b], Christophe Micheyl [a,c], Aurélie Bidet-Caulet [a], Annie Moulin [a,*]

[a] Lyon Neuroscience Research Center (CRNL), INSERM U1028, CNRS UMR 5292, Université Claude Bernard Lyon 1, Université de Lyon, Lyon, France
[b] Audition Conseil, Lyon, France
[c] Starkey, France

HIGHLIGHTS

- Different normalization techniques of the pupil dilation response result in similar outcomes.
- The choice of the baseline period is crucial to assess attention mobilization, in particular anticipation during listening effort, using the pupil dilation response.
- Pupil dilation response can reveal subtle changes in cognitive demands even in the case of perfect performances.

ARTICLE INFO

Keywords:
Pupil dilation response
Baseline
Normalization
Anticipation
Attention
Hearing-aid

ABSTRACT

Numerous studies showed that task-evoked pupil dilation is an objective marker of cognitive activity and listening effort. However, these studies differ in their experimental and analysis methods. Whereas most studies focus on a single method, the present study sought to compare different pupil-dilation data analysis methods, including different normalization techniques, baseline periods, and baseline durations, in order to assess their influence on the outcomes of pupillometry results obtained in an auditory task. To that purpose, we used pupillometry data recorded in response to words in noise in hearing-impaired individuals. The start-time of the baseline relative to stimulus timing turned out to have a significant influence on conclusions. In particular, a significant interaction in the effects of signal-to-noise ratio and hearing-aid use on pupil dilation was observed when the baseline period used started early relative to the word—an effect likely related to anticipatory, pre-stimulus cognitive processes, such as attention mobilization. This was the case even with only correct-response trials included in analyses, so that any confounding effect of performance in the word-repetition task was eliminated. Different normalization methods and baseline durations showed similar results, however the use of z-score transformation homogenized variability across conditions without affecting the qualitative aspect of the results. The consistency of results regardless of normalization methods, and the fact that differences in pupil dilation and subjective measures of listening effort could be observed despite perfect performance in the task, underlines the relevance of pupillometry as an objective measure of listening effort.

1. Introduction

Studies investigating cognitive effort have used different methods, such as questionnaires and behavioral indices, or objective measures, such as pupillometry. Pupillometry is a well-known neurophysiological investigation technique, based on measuring the size of the pupil. From silver photography to high sampling rate eye-tracking systems, pupillometry has drastically evolved over the years, and it has been used in a wide range of research fields. Changes in pupil size are now known to be linked to the activation of the locus-coeruleus-norepinephrine (LC-NE) system (Aston-Jones and Cohen, 2005) and have been associated with several neurophysiological and cognitive phenomena, such as adaptation to brightness (Barlow, 1972; Reeves, 1920), arousal and emotion processing (Bradley et al., 2008; Partala and Surakka, 2003), or increases in cognitive demands (Beatty, 1982; Hess and Polt, 1964; Payne et al., 1968).

Over the past twenty years or so, numerous studies have used pupillometry to investigate cognitive effort, using different methods of analysis (see Table 1). While some earlier studies used changes in raw

* Corresponding author.
E-mail address: annie.moulin@inserm.fr (A. Moulin).

https://doi.org/10.1016/j.heliyon.2022.e09631
Received 26 August 2021; Received in revised form 26 December 2021; Accepted 26 May 2022

Table 1. Studies investigating auditory cognitive demands and their respective method and parameters used in analysis.

Study	Method	Experiment design	Baseline period	Baseline duration	Task	Cue
(Zekveld et al., 2019)	Baseline correction	Event-related	Beginning of stimulus presentation	1000 ms	Active	A, V
Zekveld et al. (2014a)	Baseline correction	Blocked	Beginning of trial	1000 ms	Active	A
Winn et al. (2015)	Baseline correction	Blocked	Beginning of trial	2000 ms	Active	No
Kramer et al. (1997)	Baseline correction	Blocked	Prior to noise onset	1000 ms	Active	A
Lewis and Bidelman (2020)	Baseline correction	Blocked	Prior to stimulus onset	100 ms	Active	No
Zellin et al. (2011)	Baseline correction	Event-related	Prior to stimulus onset	200 ms	Active	V
Wetzel et al. (2016) Widmann et al. (2018)	Baseline correction	Event-related	Prior to stimulus onset	200 ms	Passive	No
Zekveld et al. (2010) Zekveld et al. (2011a, b) Koelewijn et al. (2012) Koelewijn et al. (2014a) Koelewijn et al. (2014b) Zekveld and Kramer (2014) Koelewijn et al. (2015) Koelewijn et al. (2017) Wendt et al. (2017) Francis et al. (2018) Jensen et al. (2018) Ohlenforst et al. (2018) Wang et al., 2018a) Wang et al., (2018b) Wendt et al. (2018)	Baseline correction	Blocked	Prior to stimulus onset	1000 ms	Active	A
Zekveld et al. (2014b) Ohlenforst et al. (2017)	Baseline correction	Event-related	Prior to stimulus onset	1000 ms	Active	A
Kramer et al. (2013)	Baseline correction	Blocked	Prior to stimulus onset	1500 ms	Active	A
McMahon et al. (2016)	Baseline correction	Event-related	Prior to stimulus onset	2000 ms	Active	A
Borghini and Hazan (2018)	Baseline correction	Blocked	Prior to stimulus onset	2000 ms	Active	A, V
McCloy et al. (2017) McCloy et al. (2018)	Deconvolution	Event-related	Prior to trial onset	500 ms	Active	A
Engelhardt et al. (2010)	Grand-mean scaling	Event-related			Active	No
Winn and Moore (2018)	Individual dynamic range scaling	Blocked			Active	A
Ayasse et al. (2017)	Individual dynamic range scaling	Event-related			Active	V
Wagner et al. (2016b)	Proportional change from baseline	Event-related	Prior to stimulus onset	200 ms	Active	No
Wagner et al. (2016a)	Proportional change from baseline	Blocked	Prior to stimulus onset	200 ms	Active	V
Wagner et al. (2019)	Proportional change from baseline	Event-related	Prior to stimulus onset	500 ms	Active	V
Miles et al. (2017)	Proportional change from baseline	Event-related	Prior to stimulus onset	1000 ms	Active	A
Winn and Moore (2018)	Proportional change from baseline	Blocked	Prior to stimulus onset	1000 ms	Active	A
Hyönä et al. (1995)	Raw pupil size	Blocked			Active	No
Wendt et al. (2016)	Within-trial mean scaling	Event-related	Prior to stimulus onset	1000 ms	Active	A, V
Mcgarrigle et al. (2016)	Within-trial mean scaling	Blocked	Prior to stimulus onset	1000 ms	Active	A
Kuchinsky et al. (2013) Kuchinsky et al. (2014) Kuchinsky et al. (2016)	Within-trial mean scaling	Blocked	Prior to stimulus onset	1000 ms	Active	A, V
Ayasse et al. (2017)	Within-trial mean scaling	Event-related	Prior to stimulus onset	2000 ms	Active	V
Korn and Bach (2016)	Z-score transformation	Event-related	Beginning of trial	First data point	Passive	No
McCloy et al. (2016)	Z-score transformation	Event-related	Prior to stimulus onset	500 ms	Active	A

pupil size as measures of cognitive activity (Hyönä et al., 1995; Just and Carpenter, 1993; Kahneman and Beatty, 1966), in more recent years, it has become customary to measure pupil-size changes relative to a baseline period using subtractive baseline correction (e.g., Laeng and Alnaes, 2019). Mathôt et al. (2018) specifically compared, on simulated data, the absence of baseline correction and the two most used methods of baseline correction: the subtraction method and the divisive method (similar to percent change), and concluded in favor of the subtractive method, the divisive method being very sensitive to artefacts in the baseline and to very low baseline values. Reilly, Kelly, Kim, Jett and Zuckerman (2019) have shown that, even with different absolute values of baselines (obtained by modulating brightness in the testing room), similar results could be obtained by simple baseline correction, without scaling. However, the dynamic range of pupil size, i.e., the difference between baseline and peak pupil size, can vary widely across individuals. Therefore, in two different individuals performing the same task, the same amount of pupil dilation may in fact correspond to different amounts of cognitive effort. Given such inter-individual variability, some authors, e.g., Winn et al. (2018), have advised that normalization methods should be applied, in addition to baseline correction (some examples are described in Table 1). Furthermore, analyses of pupil-dilation data may also vary with respect to the selected baseline period (see Table 1). Although it is customary in pupillometry studies to use, as baseline, a period preceding stimulus onset (e.g., Laeng and Alnaes, 2019), the duration of that period, and how close in time it is to stimulus onset, can vary widely across studies. While most studies have used a 1000-ms baseline, baseline durations can actually vary from as short as 100 ms to as long as 2000 ms (see Table 1). Given substantial differences in pupil-data analysis methods across studies, it seems important to ask whether different combinations of baseline corrections and analysis methods can lead to different conclusions. This issue has recently been investigated by Attard-Johnson et al. (2019) regarding

pupillary responses recorded to measure arousal elicited by viewing pictures of men and women. The use of different analysis methods resulted in similar outcomes on their specific data set, but this methodological matter needs to be addressed for the study of other cognitive processes (such as listening effort) and using different type of stimuli. Indeed, Attard-Johnson et al. (2019) compared different analysis methods on pupil responses to visual stimulation, whereas listening effort studies use auditory stimulation. The two types of stimuli differ in their modality, but also in their temporality. As Attard-Johnson et al. (2019) used pictures, which content is fixed, auditory stimulation is changing over time. Besides this aspect, pupil responses recorded during exclusive auditory tasks are free from luminance effects or image-based factors from the stimuli, that can contaminate the pupillometric signal (Barlow, 1972; Ellis, 1981). Finally, Attard-Johnson et al. (2019) compared data recorded during a free-viewing paradigm. In listening effort studies, participants are generally asked to fixate some cue during the experiment (Koelewijn et al., 2012, 2014a, b; Zekveld et al., 2019), preventing eye movements that can affect the estimation of the pupil size (Brisson et al., 2013; Gagl et al., 2011).

In recent years, pupillometry has been applied to measuring listening effort. The latter has been defined by McGarrigle et al. (2014) as "*the mental exertion required to attend to, and understand, an auditory message*". In those applications, pupil diameter is recorded during various speech-in-noise recognition tasks (Koelewijn et al., 2014; Kuchinsky et al., 2013; Ohlenforst et al., 2018; Zekveld et al., 2010, 2011a, b). In most of those tasks, participants are asked to repeat as correctly as possible sentences heard in adverse conditions, while the diameter of their pupils is being recorded. In those studies, a common working hypothesis is that changes in cognitive load are associated to changes in pupil size: the more challenging the listening condition, the larger the pupil dilation. However, in some cases, the listening situation can be so challenging that some participants may stop trying to perform the task correctly; consequently, no modification, or even a decrease in pupil dilation, may be observed in these test conditions. Another complication in the interpretation of the results of these studies, stems from the multiplicity of the cognitive processes involved in task performance. Being able to correctly repeat a sentence involves attention, short-term memory, and a host of auditory and linguistic processes. Thus, pupil dilation may reflect more complex cognitive processes than purely auditory processing of the stimulus (Zekveld et al., 2019). Lastly, it is conceivable that in tasks with a predictable (e.g., repeating) temporal structure, participants start mobilizing cognitive resources, such as attention, *in advance* of the stimulus; such anticipatory effects underline the importance of selecting an adequate baseline in analyses.

Following a review of analysis methods used in pupillometry studies involving listening tasks (Table 1), the present study systematically compares several normalization methods, and several ways of taking into account baselines. The different analysis methods were applied to data from a listening study in hearing-impaired patients. An additional point, which this study sought to address, is whether pupil-dilation data collected during a word-in-noise recognition test could be used to assess listening effort induced by a single-word identification task. This question has been addressed in the literature for young normal-hearing people (Kramer et al., 2013), and older hearing-impaired individuals who were not familiar with the use of hearing-aids, using a forced-choice paradigm (Kuchinsky et al., 2013). The present study sought to address this particular point in hearing-impaired individuals, regular hearing-aid wearers, using a word-in-noise recognition task as close as possible to the task widely used in clinical audiology to evaluate speech-perception performance. Test conditions varying in the amount of effort were included into the experimental design, and differences in pupil dilation between these conditions were tested. Importantly, variations in listening effort across conditions can exist, despite similar performance scores across the same conditions. Here, the words were presented at high signal-to-noise ratios, so that recognition performance was generally high. Another hypothesis, which was tested in this study, is that listening effort, reflected in pupil-dilation changes, is a combination anticipatory and stimulus-processing efforts. To test this hypothesis, we systematically analyzed the pupillometry data using four different baselines (with different timings relative to stimulus onset), and five normalization techniques (subtractive baseline correction, proportional change from baseline, within-trial mean scaling, grand mean scaling and z-score transformation).

Table 1. The cue column indicates whether participants are given indices regarding stimulus presentation or not. Cues might be explicit (a visual (V) or auditory (A) indication before the presentation of the stimulus) or implicit (the beginning of the noise before the presentation of a sentence in adverse conditions for instance). The experiment design column indicates whether conditions are presented in a block ('Blocked') or trial ('Event-related') fashion. Computation for each method: ***baseline correction****: the average pupil size in the baseline period is subtracted from data points.* ***Proportional change from baseline****: pupil data are normalized by first applying a baseline correction then dividing all data point by the baseline (it can be expressed in percentage by multiplying by 100).* ***Within-trial mean scaling****: pupil data are normalized by calculating the mean of all data points in each trial and then dividing each data point by the mean. Baseline correction is then applied.* ***Grand-mean scaling****: pupil data are normalized by calculating the grand mean of the complete trace and then dividing each data point by the grand mean.* ***Z-score transformation****: pupil diameters are expressed as z-scores then baseline corrected.* ***Individual dynamic range scaling****: pupil dilation is expressed as a proportion of each participant's dynamic range.* ***Deconvolution****: pupil data are first baseline corrected on each trial then data points are dividing by the standard deviation of pupil sizes across all trials. Pupil data are then deconvolved with pupil impulse response kernel.*

2. Material and methods

2.1. Participants

Thirty participants (15 men) aged between 29 and 91 years (mean: 70 years) were recruited among hearing-impaired patients of a hearing-aid dispenser. All participants provided written informed consent in accordance with the Declaration of Helsinki (General Assembly of the World Medical Association, 2014). All data were processed anonymously and in agreement with the French MR003 regulation and the European General Data Protection Regulation (GDPR). This study was approved by an ethics committee for research on humans ("comité de protection des personnes" de Ile de France II, Paris, France", n°21.01.08.67105//4314802). Participants were screened for major cognitive alterations and dementia using the MoCA test (7.2) (http://www.mocatest.org/) (Nasreddine and Patel, 2016). As the participants were all experienced hearing-aid users, the MOCA test was performed in the usual situation for them, i.e. when they were wearing their hearing aids (Saunders et al., 2018). To alleviate any problem of audibility, the instructions for the MOCA test were presented to the patients both by live voice and as large print material, in the lines of the MOCA test adapted for hearing impaired patients (Lin et al., 2017). The participants also completed a listening effort questionnaire adapted from the Effort Assessment Scale (EAS) (Alhanbali et al., 2017). Twenty participants were included in analysis (inclusion criteria are detailed in the *Task-Evoked Pupil Response (TEPR): baseline-related analysis* section): 13 men, 7 women, aged between 29 and 85 years (mean: 66.5 years, SD: 15.3). Included participants had a mean hearing loss (Pure Tone Average at 500, 1000, 2000 and 4000 Hz) of 41 dB HL in the better ear (SD = 13.7 dB HL) and 50.1 dB HL in the worse ear (SD = 12.8 dB HL). Based on participants' answers to a hearing-aid use questionnaire, they had been wearing hearing-aids for 3.4 years on average (SD = 5.8 years), for an average of approximately 10-hour per day.

2.2. Pupillometry session

Prior to the main tests, the sound level of the dissyllabic words was adjusted individually to yield at least 90% word-correct scores unaided, i.e., without hearing-aids, in "quiet". In the context of this study, "quiet" test conditions refer to conditions in which words were presented in a background of speech-shaped noise (i.e., noise having the same long-term average power spectrum as the words), with a high (30 dB) signal-to-noise ratio (SNR). The reason background noise was included even in "quiet" conditions is that the room in which participants were tested was not entirely sound-proof, and the computer used to control the test procedure could not be moved outside of the room. The individually adjusted sound level was then kept constant for the entire duration of the tests. For the "noisy" conditions, the SNR was adjusted individually prior to the start of the main test, in such a way that scores measured without hearing-aids in the presence of background noise would also equal around 75%. SNR ranged from +3 dB to +18 dB (mean: +10.2 dB, SD: +5.1 dB) for included participants. Following these preliminary individual adjustments of signal level and SNR, four different conditions were tested in a block design:

- two "quiet" conditions, one with hearing-aids (Quiet Aided) and one without hearing-aids (Quiet Unaided);
- two "noisy" conditions, one with hearing-aids (Noise Aided) and the other without the hearing-aids (Noise Unaided);

For the tests with hearing-aids, each participant used his/her own hearing-aids, with their habitual gain settings. The hearing-aids included different brands (Phonak, Starkey) as well as different models. To facilitate analyses and interpretations of the results, except for the feedback-canceller, the signal-processing algorithms on these hearing-aids, such as noise-reduction and directional processing, were turned off for the duration of the tests. A subset of participants also performed a fifth condition, in which they were tested with their hearing-aids on, but with the hearing-aid noise-reduction algorithm turned on. However, because this condition could only be tested in a subset of participants, while a complete dataset was needed for the current study, results of this fifth condition were not included into this study.

For each condition, two lists of 12 dissyllabic words were presented to the participant (leading to a total of 8 different 12-word lists). Participants were asked to repeat the word or what they had heard, even if it was a single syllable or a single phoneme, after an auditory cue (see Stimuli section). There was a short training block of five words (not included in the 12-word lists), for participants to familiarize themselves with the task; in particular, participants had to learn to wait for the cue before repeating the word – an important methodological feature of the present study.

Word order within a list, 12-word lists and test conditions were randomized across participants. At the end of each 12-word list, participants rated the perceived difficulty of the task using a 10-point visual analogue scale (1: very easy task, 10: very difficult task). Each 12-word sequence lasted approximately 2 min 30 s and this pupillometry session lasted approximately 1 h, including breaks and calibration procedure.

2.3. Apparatus

The walls in the room were covered with acoustic panels and the room was illuminated by a lamp placed above and behind the participant. The participant was comfortably seated in front of three loudspeakers (Yamaha, HS 50M) (see Figure 1). The first one, directly in front of the participant, was used to play the target words – consistent with the most common conversation-listening situation, with an interlocutor in front. The two other loudspeakers were used to play the noise, and were positioned at -60° and +60° angles relative to the midline. Each loudspeaker was placed at a distance of 1.20 m from the participant. A microphone installed next to the participant was used to record answers

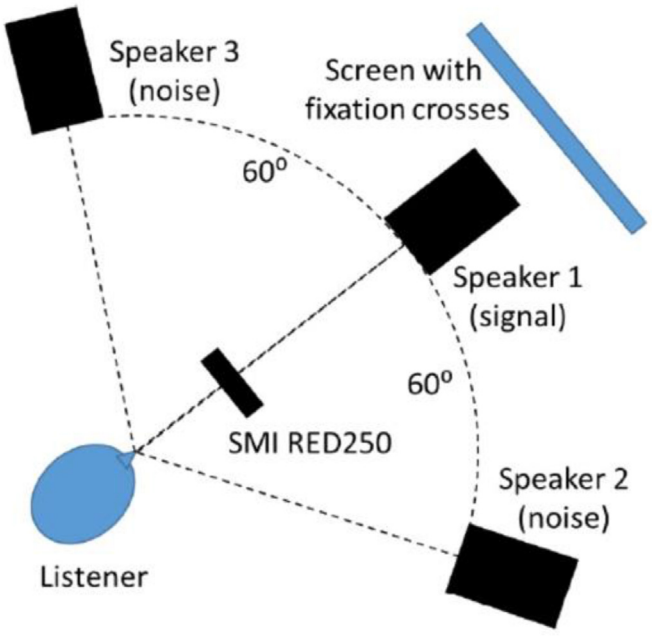

Figure 1. Experimental setup. Schematic depiction of the experimental setup. The participant was seated in front of three loudspeakers. The first one directly placed in front of them and was used to play the target signals (words). The two other loudspeakers were at a 60° angle from the front one and were used to play the noise. The eye-tracking system was placed in front of the participant. During the tests, participants had to fixate a visual cue on the loudspeaker in front of them.

during the task. The eye-tracking system (SensoMotoric Instrument, RED250mobile (Berlin, Germany)) was placed at a maximum of 70 cm in front of the participant. Data were sampled at 250 Hz with a 0.4° spatial resolution. Eye positions and pupil data were recorded for both eyes. To avoid color/luminance variation effect, no computer screen was used. A five-point calibration was performed using colored pieces of paper on each corner of a virtual rectangle, placed around the front loudspeaker. The same setup was used during the task, and the participants were asked to fixate a white paper cue on the loudspeaker placed in front of them.

2.4. Stimuli

Words used as stimuli were disyllabic words, with high frequency occurrence in the French language and pronounced by a female voice. Each word lasted 500 ms on average. Word lists were balanced with respect to their linguistic properties (e.g. occurrence frequency above 2 occurrences per million words and more than 2 phonological neighbors in the French language), acoustic characteristics (frequency spectra), and psychometric curves (50% word recognition threshold and slope) (Moulin and Fourcaud-Trocmé, 2019) so that each word list, presented at the same level, gave similar speech perception performance, as comparable word lists commonly used in audiology laboratories typically do (Moulin et al., 2017).

Each trial was designed as follows (see Figure 2):

- first second: stationary noise;
- auditory cue ("Beep": 500 Hz pure-tone lasting 400 ms), announcing the occurrence of the next word;
- 1.5 s of stationary noise preceding the word;
- 4 s (including the word) to process the word;
- auditory cue ("Blop": two 400-Hz pure tones followed by a 450-Hz pure-tone, each 100 ms in duration, separated by a 50-ms inter-stimulus interval), announcing the start of the temporal window for repeating the word;

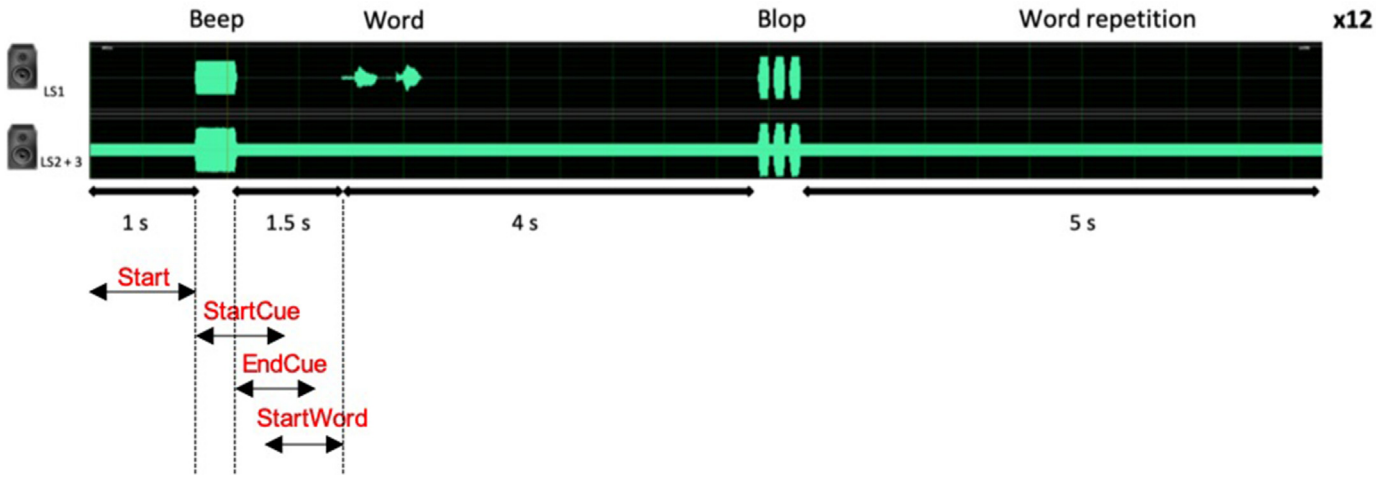

Figure 2. Design of one trial with baseline periods used in analyses. Noise is first played, followed by a first auditory cue announcing the beginning of the word ("Beep"). After the word has been processed, a second auditory cue allows participant to repeat the word heard ("Blop"). Time is fixed for all trials. Each block contains 12 trials.

- 5 s temporal window, to allow the participant to repeat the word that he/she just heard.

Each trial lasted 12.25 s.

3. Analyses

3.1. Subjective data

3.1.1. Speech perception scores

A correct answer was recorded when participants repeated the whole word correctly after the auditory cue. Scores are averaged across conditions for each participant and expressed as percentages.

3.1.2. Subjective rating

Subjective ratings given at the end of each sequence in the pupillometry session are also averaged across conditions for each participant (10 = maximum difficulty).

3.2. Pupillometric data

To allow for adaptation of the participant and of the hearing-aids to changes in the signal and/or noise levels, only data corresponding to the 10 last words of each 12-word block were included in the analyses. Trials for which participants repeated the word before the auditory cue were also excluded from analysis.

3.2.1. Pre-processing of pupil data

Pupil diameters were recorded binocularly and continuously using an SMI° RED eye-tracking system. Data were sampled at 250 Hz. All data processing and analysis were performed using Python (Python Software Foundation. Python Language Reference, version 3.7. Available at http://www.python.org/, the "cili" library (Acland and Braver, 2014) was adapted for data pre-processing. Zero values and pupil diameters below two standard errors of the mean were coded respectively as blinks and "half-blinks". Blink events were linearly interpolated from 8 ms before the starting point until 16 ms after the ending point of the blink. "Half-blink" events were interpolated from 40 ms before the starting point until 40 ms after the ending point of the "half-blink". The data were then corrected for artefacts due to eye movements using linear interpolation. Each step of this pre-processing was performed for both eyes. Pupil data were then smoothed using a low-pass 7^{th} order Butterworth filter (cut-off frequency: 25 Hz) to remove high frequency artefacts. Trials containing more than 60% of interpolated data were rejected. Finally, traces were visually inspected to manually include or exclude trials. Following these criteria, 3 participants were excluded from analysis due to insufficient recording of pupil data and too much interpolated data. Around 10% of trials were excluded from the remaining data of 27 participants, leading to 474, 501, 471, 473 trials in the Quiet Aided, Quiet Unaided, Noise Aided, Noise Unaided conditions respectively. For each participant, we chose to analyze pupil traces from the eye in which we found least interpolations and missing data across all conditions.

3.2.2. Task-Evoked Pupil Response (TEPR): baseline-related analysis

Overall, task performance was high, with a mean percent score of 90.3 (SEM = 2.7) in the Quiet Aided condition, 89.3 (SEM = 2.9) in the Quiet Unaided condition, 73.5 (SEM = 3.2) in the Noise Aided condition and 78.8 (SEM = 1.9) in the Noise Unaided condition. To identify changes in pupil dilation related to listening effort unbiased by response correctness, main analyses focused on pupil responses measured during correct-response trials ("hits") only. However, analyses including data from all trials (correct and incorrect) were also performed, and their results were compared with those of the main analyses using only correct-response trials (Supplementary Figures S3 to S6). Therefore, participants who repeated correctly at least 8 entire words out of 20 (2 × 10 words per condition) for each condition were included in the analyses (N = 20 participants).

3.2.2.1. Baseline correction. All included trials were baseline-corrected. Baseline correction is a common operation in pupillometry, whereby the average pupil size measured over some time period (epoch) prior to stimulus onset is computed, then subtracted from pupil-dilation measurements collected over a subsequent epoch –typically, post stimulus onset. To investigate the influence of the choice of baseline epoch, analyses were performed using four different baseline epochs (see Figure 2), with durations of 500 ms (Supplementary Figures S1 to S4) or 1 s: *Baseline Start:* The first baseline period used in these analyses coincided with the trial onset, defined as the onset of the background noise. This baseline period is the most anterior, relative to the onset of the word. *Baseline StartCue*: This baseline period started with the onset of the cue announcing the word. *Baseline EndCue:* This baseline period started with the offset of the cue announcing the word. *Baseline StartWord:* This baseline period *ended* with the onset of the word.

3.2.2.2. Normalization. To take inter-individual differences in physiological pupil dynamic range into account, several normalization methods (Winn et al., 2018) were applied on the data. This was done separately, for each baseline correction described above. Results obtained using these normalization methods were then compared with results obtained

using baseline-corrected data (Koelewijn et al., 2014; Kramer et al., 2010; Zekveld et al., 2011b) to observe potential differences.

Proportional change from baseline: One way to express local deviation from baseline is to present the percentage of pupillary dilation (Hess and Polt, 1964; Payne et al., 1968) or to express data as a proportional change from baseline (Johnson et al., 2014; Wierda et al., 2012). In our analysis, task-evoked pupil dilation was expressed in this way for each data point x: $x_{norm} = (x_{data} - baseline)/baseline$ (Wierda et al., 2012).

Within-trial mean scaling: This method has been used in previous studies to ensure a consistent scaling of pupil across trials and participants (Kuchinsky et al., 2013, 2014, 2016). Each trial data point was divided by the mean of the relevant period of the trial (from the start of the trial to the cue before the repetition of the word), then, a baseline correction was performed.

Grand mean scaling: Grand mean scaling has also been used to assess change in pupil size (Engelhardt et al., 2010). The grand mean of the whole time series (periods from the start of each trial to the cue before the repetition of the word of the entire block (10 words)) was computed and each data point was divided by the grand mean. Then, we applied baseline correction on the normalized data.

Z-score transformation: Finally, a z-score transformation was applied (McCloy et al., 2016; Nassar et al., 2012). To do so, the mean of the relevant period of the trial (from the start of the trial to the cue before the repetition of the word) was subtracted from each trial data point, then each data point was divided by the standard deviation of the relevant period of the trial. The z-score transformation was followed by a baseline correction.

3.3. Statistical analyses

3.3.1. Behavioral data

Performance during the task was analyzed using a Generalized Linear Mixed-effects Model (GLMM) at the trial-level, following a binomial distribution (0/1 for correct/incorrect trial). The model contained fixed-effects for the hearing-aid factor (Aided vs Unaided), the noise factor (Quiet vs Noise) and interaction between these two factors. Subject was modeled as a random factor. To confirm the need for a mixed nested model, we used a likelihood ratio analysis to test each model fit, before and after sequential addition of random effects (in particular, random intercepts and slopes for the hearing-aid and noise factors). We used the Akaike information criterion and the Bayesian information criterion as estimators of the quality of the statistical models generated (Matuschek et al., 2017). The best model contained fixed-effects for the hearing-aid factor (Aided vs Unaided), the noise factor (Quiet vs Noise) and interaction between these two factors and Subjects as a random factor with by-subject slopes for the factor noise. A type-II ANOVA (using type-II Wald X^2 test) was then applied to the model. Analysis were conducted with R (R Development Core Team, 2020), the "lme4" package (Bates et al., 2015) was used to compute the model, the "car" package (Fox and Weisberg, 2019) was used to compute the type-II ANOVA.

Participants' subjective ratings expressed as ordinal data (from 1 to 10) were analyzed at the block level using a Cumulative Link Mixed Models (CLMM) to test fixed within-subject effects of hearing-aid (Aided vs Unaided) and background noise (Quiet vs Noise). Subject was modeled as a random factor. To confirm the need for a mixed nested model, we used a likelihood ratio analysis to test each model fit, before and after sequential addition of random effects (in particular, random intercepts and slopes for the hearing-aid and noise factors). We used the Akaike information criterion and the Bayesian information criterion as estimators of the quality of the statistical models generated (Matuschek et al., 2017). The best model contained fixed-effects for the hearing-aid factor (Aided vs Unaided), the noise factor (Quiet vs Noise) and interaction between these two factors and Subjects as a random factor with by-subject slopes for the factor noise. A type-II ANOVA (using type-II Wald X^2 test) was then applied to the model. Post-hoc pairwise comparisons were conducted using least-squares means testing. The "emmeans" package was used to compute least-squares means testing (Searle et al., 1980).

3.3.2. Pupillometric data

Since we are interested in anticipation processes and how different conditions could affect them, statistical analyses were conducted on two time-windows: anticipation and stimulus processing. The anticipation window is defined between the cue announcing the word and the beginning of the word (i.e., 1900 ms long). The stimulus processing window is defined between the beginning of the word and the cue before the repetition of the word (i.e., 4000 ms).

We investigated the impact of the baseline period on the mean dilation of the pupil obtained in the anticipation window, using a Linear Mixed-effects Model (LMM) with "baseline period" (Start, StartCue, EndCue, StartWord) and "condition" (QuietUnaided, QuietAided, NoiseUnaided, NoiseAided) as fixed factors. Subject was modeled as a random factor. Data were averaged across normalization methods. We used the Akaike information criterion and the Bayesian information criterion (Matuschek et al., 2017) as estimators of the quality of the statistical models generated, before and after sequential addition of random effects (in particular, random intercepts and slopes for the 'baseline period' and 'condition' factors). The best model contained fixed effects for the "baseline period" factor and "condition" factor and interaction between these two factors, and Subjects as a random factor. A type-II ANOVA (using type-II Wald X^2 test) was then applied to the model. Post-hoc pairwise comparisons were conducted using least-squares means testing. The "emmeans" package was used to compute least-squares means testing (Searle et al., 1980).

Between-subject variability between conditions over time was inspected for each normalization method. To do so, a ratio was computed for each time point. This ratio was computed per normalization method as follows: $h(t)/H$, with h(t) the width of the confidence interval of the group mean computed for each time point per condition and H the width of the confidence interval of the group mean computed on the entire time window per condition. H is then constant for each time window and condition.

Further statistics were conducted in order to find within-subject effects of the different factors on the amplitude dilation of the pupil during the task. Listening effort is usually assessed using the maximum dilation value (peak pupil dilation) in a defined window (Koelewijn et al., 2014; Ohlenforst et al., 2018; Y. Wang, Kramer, et al., 2018; Zekveld et al., 2014a; Zekveld and Kramer, 2014; Zekveld et al., 2014b). In our study a peak value would hardly be definable in the anticipation window, we will then use the mean dilation of the pupil in the two time-windows to assess cognitive demands during the task (Koelewijn et al., 2012; Kramer et al., 2013; Zekveld et al., 2010, 2011a). Therefore, dependent variables were defined as the mean dilation in the anticipation window and the mean dilation around the peak pupil dilation (maximum dilation) in the stimulus processing window. Since the stimulus processing window is larger than the anticipation window, the mean dilation in the stimulus processing window was computed on the same length as in the anticipation window (i.e., 1900 ms) centered around peak pupil dilation. Mean dilations of the pupil were analyzed using a Linear Mixed-effects Model (LMM) with the fixed factors hearing-aid (Aided vs Unaided) and noise (Quiet vs Noise). Subject was modeled as a random factor. We used the Akaike information criterion and the Bayesian information criterion (Matuschek et al., 2017) as estimators of the quality of the statistical models generated, before and after sequential addition of random effects (in particular, random intercepts and slopes for the hearing-aid and noise factors). For each window separately (Anticipation and Stimulus processing), the best model contained fixed-effects for the hearing-aid factor, the noise factor and interaction between these two factors, and Subjects as a random factor. We then performed a type-II ANOVA (using type-II Wald X^2 test) on the model. Post-hoc pairwise comparisons were conducted using least-squares means testing.

4. Results

4.1. Behavioral data

As mentioned earlier, the task was very well completed with high speech perception scores (see Material and Methods section). The type-II ANOVA showed a main effect of the noise factor (X^2 (1, N = 20) = 19.34, p = 1.1e-5), indicating better performances when words were presented in Quiet. The subjective difficulty of the task was rated as follows (the higher the more difficult): 1.98 (SEM = 0.47) in the Quiet Aided condition, 1.65 (SEM = 0.39) in the Quiet Unaided condition, 5.28 (SEM = 0.59) in the Noise Aided condition and 4.05 (SEM = 0.59) in the Noise Unaided condition. The type-II ANOVA indicated significant effect of noise on subjective ratings (X^2 (1, N = 20) = 19.45, p = 1*10e-5) showing that the task was perceived as easier when words were presented with a high SNR. A significant effect of aid was also found on subjective ratings (X^2 (1, N = 20) = 12.38, p = 0.00043) indicating that participants rated the task as easier when they were not wearing their hearing aids. A significant interaction between the noise and aid factors was also found (X^2 (1, N = 20) = 4.02, p = 0.045). Post-hoc pairwise comparisons indicated that all conditions were significantly different from each other, except the Quiet Aided and the Quiet Unaided conditions, suggesting that the task was perceived as easier without hearing aid in case of high SNR only.

4.2. Pupillometric data

Different patterns were observed in participants pupil traces. In more challenging conditions, some participants showed a dilation peak when the word is presented and when the word is repeated (see Figure 3: left), but other showed a continuous increase in dilation from the beginning of the trial to the presentation of the word reflecting preparation to process the word (see Figure 3: right). In this particular case, the choice of the baseline period can have a critical outcome regarding the measure of the listening effort: if the chosen baseline period was too close to the word, this anticipation would not be taken into account. Following this idea, several baseline periods were tested (see Figure 4a to d). As well as different patterns of dilation, differences in terms of variability and pupil size were observed in collected data. In light of such inter-individual variability, several normalization methods were computed as well (see Figure 4.1 to 4.5).

When considering all participants (N = 20), the choice of the baseline period seemed to influence dilation values obtained in the anticipation window, between the first auditory cue and the presentation of the word (i.e. 1900 ms–2900 ms from trial start) (see Figures 5 and 6). The type-II ANOVA (using type-II Wald X^2 test) showed a main effect of the "baseline period" factor (X^2 (3, N = 20) = 12.81, p = 0.0051) and a main interaction between the "baseline period" and "condition" factor (X^2 (9, N = 20) = 22.02, p = 0.0088). Post-hoc pairwise comparisons of the baseline period main effect showed larger pupil dilation in the anticipation window when using the baseline period Start compared to the baseline period StartWord (p = 0.0061). Post-hoc pairwise comparisons of the baseline period by condition interaction indicated significant differences in mean dilation between conditions during the anticipation window while using the Start baseline period only, with larger pupil dilations in the NoiseAided condition compared to the QuietAided condition (p = 0.0038) and to the NoiseUnaided condition (p = 0.0009) (see Figure 7). This finding suggests that the choice of the baseline period can have a strong impact on the potential differences between conditions in anticipation of the relevant sounds. In particular, anticipation processes could be observed using a baseline period sufficiently anterior to the stimulation, but not when the baseline period is defined just before the presentation of the stimuli. Further analyses were then conducted on traces obtained with baseline correction using the baseline *Start*.

Applied to all participants (N = 20), the different normalization techniques showed similar patterns (see Figures 5 and 6). However, the z-score transformation seemed to better homogenize the variability between conditions (see Figures 5.5 and 6.5), we looked at the variability between conditions over time in the anticipation and processing window for each normalization method (using the baseline period Start). As observed in Figure 8, only the z-score transformation provided homogenous variability between conditions. This normalization method was then used in further analysis.

As well as different baseline periods and normalization methods, different baseline lengths were tested (500 ms and 1000 ms). Similar results were obtained with both baseline lengths (see supplementary figures S1 to S4).

Results obtained with all traces included (words repeated correctly or not) and those obtained with only hit trials (word repeated correctly only) showed similar results (see supplementary figures S3 to S6).

In the following, we applied z-score transformation and 1000 ms baseline period from start to our pupil responses, in order to compare pupil dilation between the different conditions, and between the different time windows, using hit trials only.

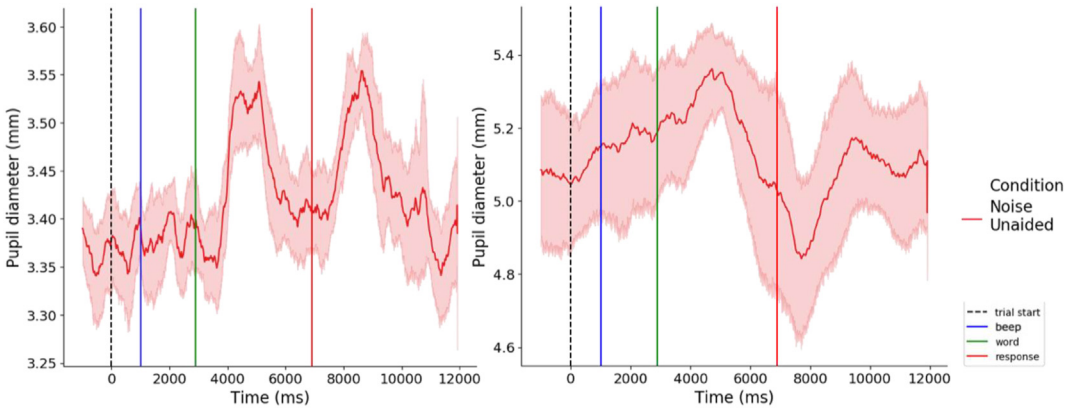

Figure 3. Different individual dilation patterns. Left: Pupil trace from a participant showing no dilation increase before the word presentation (green line) in a more challenging condition (words presented in noise) (N = 1). Right: Pupil trace from a participant showing a continuous increase in dilation before the word presentation (green line) in a more challenging condition (words presented in noise) (N = 1). Traces were obtained by averaging the correctly answered trials of the 10 last trials of each block (the 2 first trials are excluded) in the Noise Unaided condition. Traces represent raw pupil diameter during the task. Shaded areas represent 95% confidence intervals.

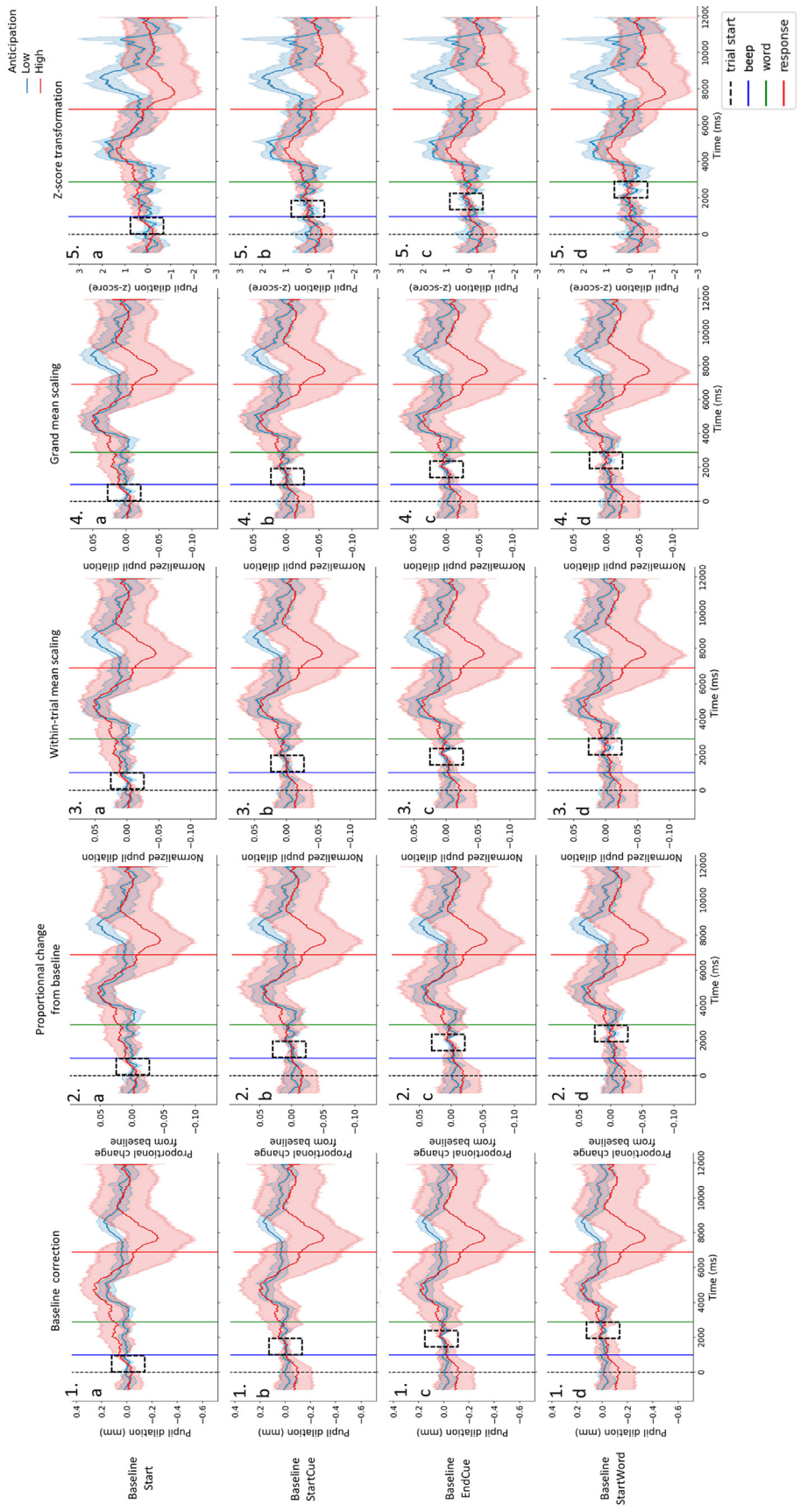

Figure 4. Individual pupil-dilation patterns computed using different normalization methods and baseline periods. Pupil traces obtained by averaging the correctly answered trials in the Noise Unaided condition. Blue traces were obtained from a participant presenting no increase in pupil dilation before the word presentation (green line). Red traces were obtained from a participant showing a continuous increase in pupil dilation before the word presentation (green line). Traces were computed using several normalization methods (1: baseline correction only, 2: proportional change from baseline, 3: within-trial mean scaling, 4: grand mean scaling, 5: z-score transformation) and several baseline periods (dashed rectangles) (a: Start: from start (dashed black line) to the auditory cue before word presentation (blue line), b: StartCue: starting at the beginning of the auditory cue before word presentation (blue line), c: EndCue: starting at the end of the auditory cue before word presentation, d: StartWord: ending at the beginning of the word (green line)). Baseline lengths were set at 1000 ms. Shaded areas represent 95% confidence intervals.

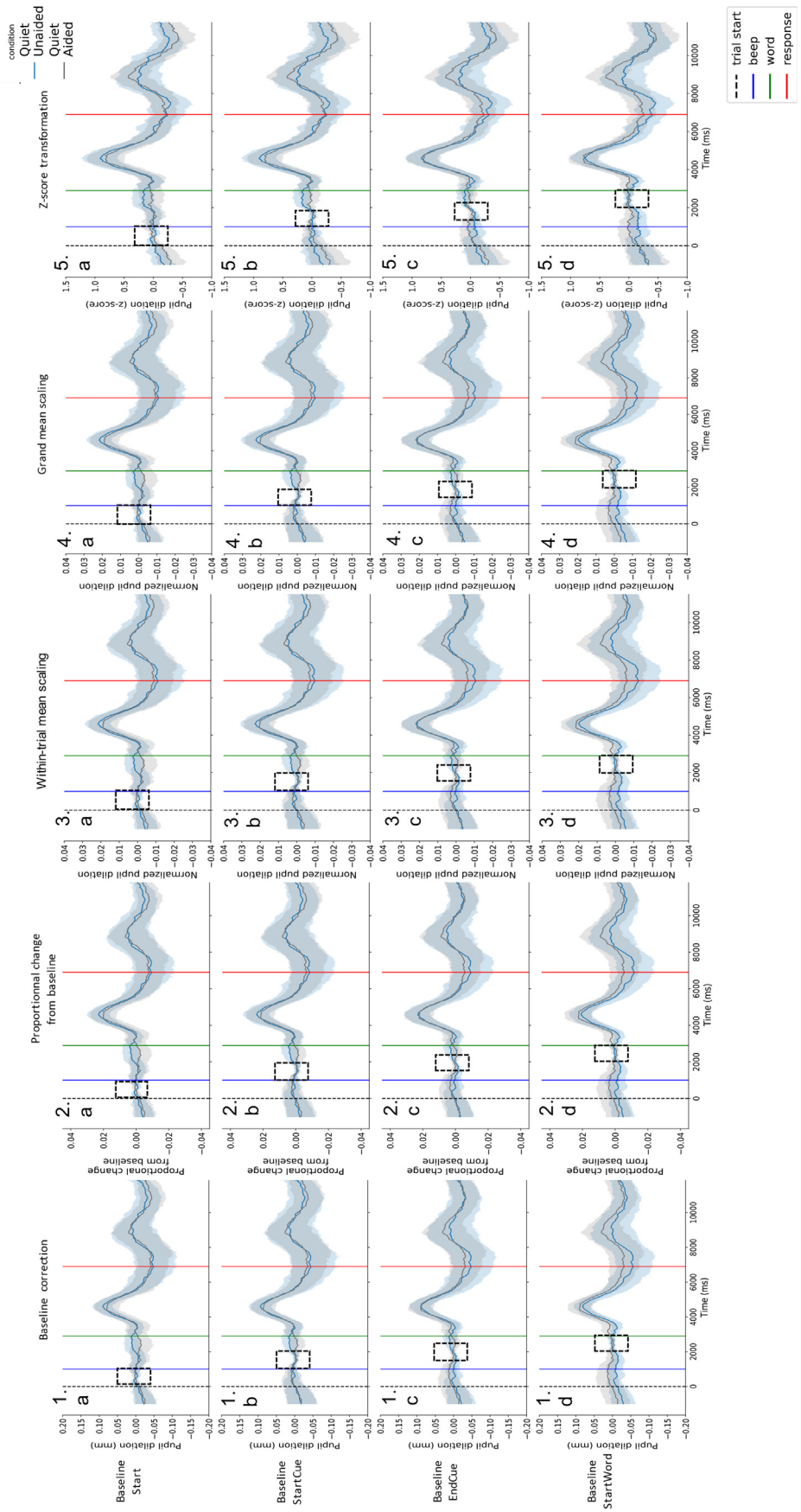

Figure 5. Group pupil dilation in the Quiet condition according to different normalization methods and baseline periods. Pupil traces obtained by averaging the correctly answered trials in the Quiet conditions (N = 20). Grey traces were obtained in the Quiet Aided condition, blue traces were obtained in the Quiet Unaided condition. Traces were computed using several normalization methods (1: baseline correction only, 2: proportional change from baseline, 3: within-trial mean scaling, 4: grand mean scaling, 5: z-score transformation) and several baseline periods (dashed rectangles) (a: Start: from start (dashed black line) to the auditory cue before word presentation (blue line), b: StartCue: starting at the beginning of the auditory cue before word presentation (blue line), c: EndCue: starting at the end of the auditory cue before word presentation, d: StartWord: ending at the beginning of the word (green line)). Baseline lengths were set at 1000 ms. Shaded areas represent 95% confidence intervals.

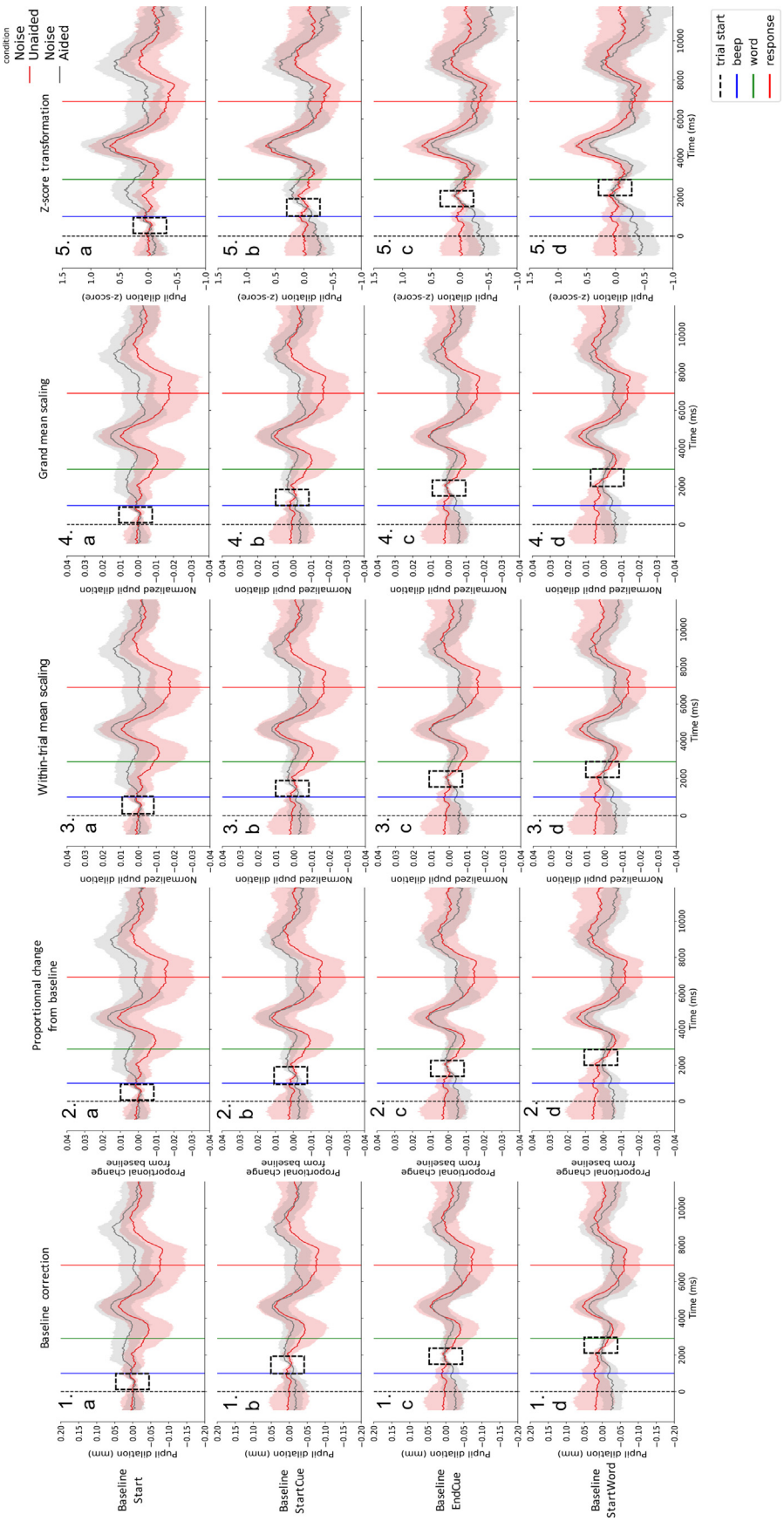

Figure 6. Group pupil dilation in the Noise condition according to different normalization methods and baseline periods. Pupil traces obtained by averaging the correctly answered trials in the Noise conditions (N = 20). Grey traces were obtained in the Noise Aided condition, red traces were obtained in the Noise Unaided condition. Traces were computed using several normalization methods (1: baseline correction only, 2: proportional change from baseline, 3: within-trial mean scaling, 4: grand mean scaling, 5: z-score transformation) and several baseline periods (dashed rectangles) (a: Start: from start (dashed black line) to the auditory cue before word presentation (blue line), b: StartCue: starting at the beginning of the auditory cue before word presentation (blue line), c: EndCue: starting at the end of the auditory cue before word representation, d: StartWord: ending at the beginning of the word (green line)). Baseline lengths were set at 1000 ms. Shaded area represent 95% confidence intervals.

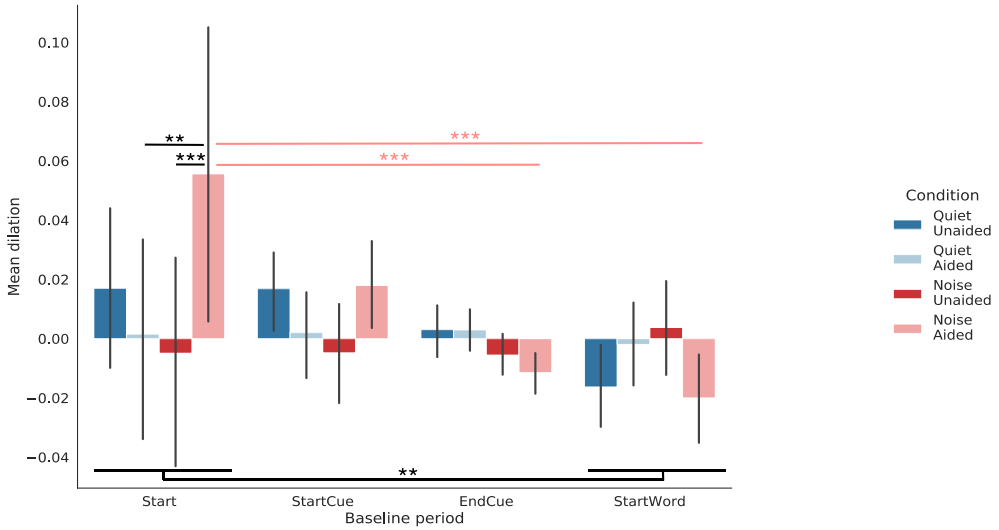

Figure 7. Mean pupil dilation in the Anticipation window using different baseline periods. Mean pupil dilation in the Anticipation window for each condition using different baseline periods (N = 20). **: p < 0.01, ***: p < 0.001.

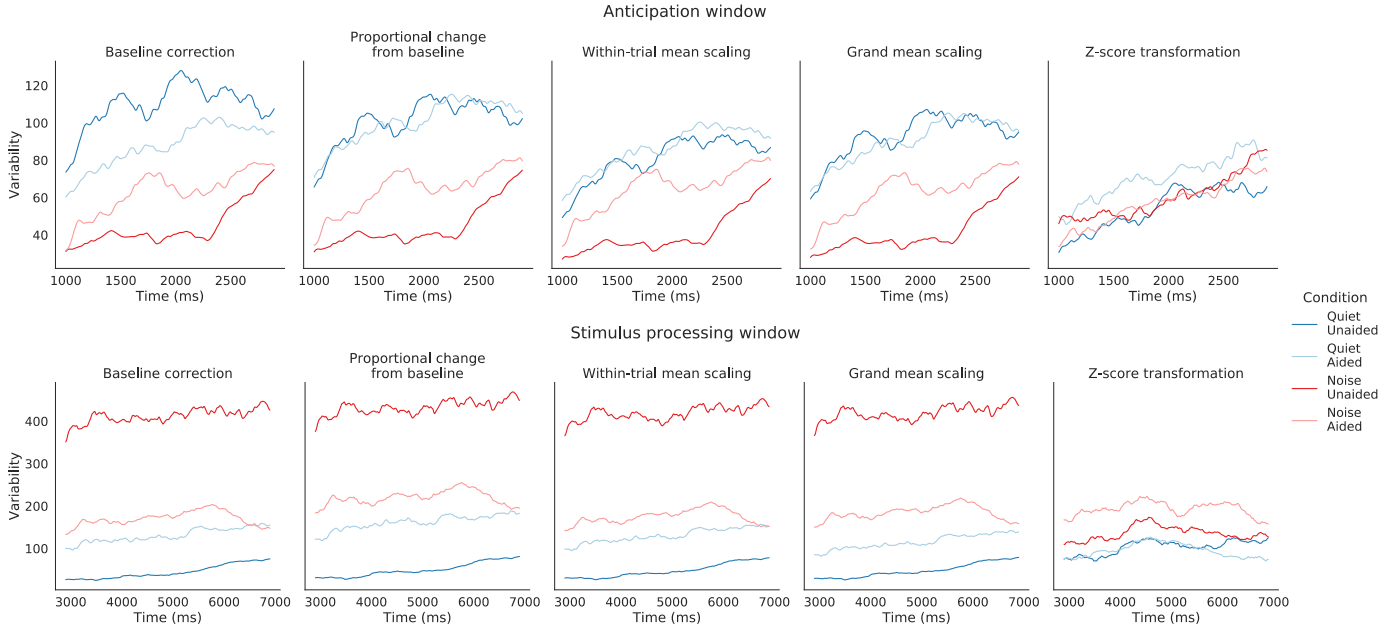

Figure 8. Inter-subject variability according to different normalization methods. Variability over time for each condition per normalization method in the anticipation window (Top) and in the processing window (Bottom). The 1000 ms baseline period Start was used and only correctly answered trials were included.

4.3. Hearing in noise

In addition to the methodological investigation, our work aimed at highlighting pupillometric measures contribution to listening effort measures. Therefore, statistics were conducted in order to find within-subject effects of noise and hearing-aid on objective pupillometric responses elicited by participants. The type-II ANOVA was performed on z-score transformed pupil dilation amplitude using a 1000 ms baseline *Start* for baseline correction only when participants repeated the word correctly (hits only) (see Figure 9: left). Statistical analyses in the anticipation window showed a main interaction between the aid and noise factor (X^2 (1, N = 20) = 6.87, p = 0.0088). Post-hoc pairwise comparisons showed an effect of the aid in the Noise condition only: mean pupil dilation in the anticipation window was significantly larger when participants wore their hearing-aids rather than not (p = 0.0039). They also showed an effect of the noise in the Aided condition only: the mean pupil dilation in the anticipation window was significantly larger when words were presented in a low SNR condition rather than in a high SNR condition (p = 0.0106) (see Figure 9: middle). No significant effects of factors on mean pupil dilation were found in the stimulus processing window (see Figure 9: right) .

5. Discussion

Pupillary responses have been widely used to assess cognitive effort and various ways of analyzing them are described in the literature (see Table 1). In audiology, pupil responses are mostly processed using subtractive baseline correction with a 1-s pre-stimulus baseline (see Table 1). Other analysis methods have been used as well, but the aforementioned processing seems to prevail in this research area. While Attard-Johnson et al. (2019) compared different pupillometric analysis methods to measure arousal in response to images, to our knowledge, such systematic comparisons had not been undertaken in the context of listening-effort studies. In the present study, we compared several

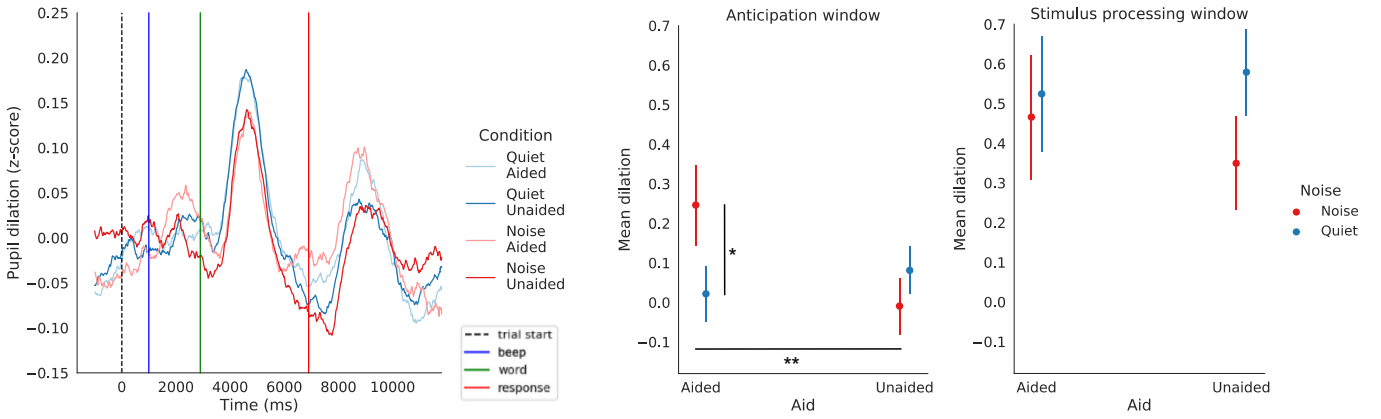

Figure 9. Pupillary measures using z-score transformation and Start baseline period. Left: Pupil traces obtained by averaging the correctly answered trials of the 10 last trials of each block in the different conditions (N = 20). Traces were computed using z-score transformation and 1000 ms baseline Start period (from start (dashed black line) to the auditory cue before word presentation (blue line)). Shaded areas represent 95% confidence intervals. Middle: Mean dilation in the anticipation window (from the auditory cue before word presentation (blue line) to the word presentation (green line). Right: Mean dilation in the stimulus processing window (centered around maximum dilation between the word presentation (green line) and the cue before the repetition of the word (red line), with the same length as the anticipation window). Error bars represent standard errors of the mean. **: $p < 0.01$.

methods of analysis using different normalization techniques, baseline periods and baseline durations, using data collected during a speech-in-noise task to assess listening effort in hearing-impaired listeners. Five normalization techniques were compared: subtractive baseline correction, proportional change from baseline, within-trial mean scaling, grand mean scaling and z-score transformation; as well as four baseline periods: anterior to just before stimulus onset; and two baseline durations: 500 ms and 1 s.

The different normalization methods yielded similar results. Attard-Johnson et al. (2019) reached similar conclusions when comparing raw pupil sizes, z-scored pupillary data, percentage changes in pupil size and subtractive baseline corrected data recorded to measure arousal elicited by viewing images. Three of the normalization methods included in the current study were also considered in this previous study, but the two studies used different stimuli (images vs uttered words), different populations (healthy vs hearing-impaired), different recording systems (SR EyeLink 1000 recording left eye at 1000Hz vs SMI RED-mobile250 recording both eyes at 250Hz), and they investigated very different cognitive processes (sexual arousal vs listening effort). This makes more detailed comparisons between these studies difficult. However, the consistent conclusions of these two studies regarding the lack of impact of the normalization method on the pupil dilation pattern suggests that this conclusion may hold across different applications of pupillometry, including different types of stimuli, populations, and cognitive processes. Although pupillary traces observed with the different normalization methods were highly similar, the z-score transformation elicited a uniformed inter-subject variability between conditions, while greater differences in terms of variance were observed between easy and most adverse conditions with other methods (see Figure 7). The standardization of the data therefore homogenized variability across conditions without affecting the qualitative aspect of the results. This method though, calls for some caution, as it is more sensible to differences in pupil tonic activity. Reilly et al. (2019) modulated the pupil dilation baseline level of their participants by using different levels of brightness. They showed that, after simple baseline correction, the different baseline levels didn't change the task evoked pupillary response. However, their different baseline levels were elicited by an external stimulus, mediated first by the retinal photoreceptors, i.e. modulating the parasympathic system. Different baseline levels mediated by arousal, or tonic activity (i.e., sympathic activity) might have a different influence on the task evoked pupillary peak. Also, it is important to keep in mind that divisive baseline correction can be highly sensitive to artefacts present during the baseline period such as blinks, eyelid closure and especially very low baseline values (Mathôt et al., 2018).

Indeed, the results reveal that the choice of baseline period can have a major impact on the conclusions of studies involving pupillometry to evaluate listening effort. This conclusion may apply, in particular, to studies in which participants may anticipate the occurrence of perceptual or motor demands, which require effort for correct task performance. In the current study, an attempt was made to control for such anticipatory processes, by including into the stimulus sequence explicit cues to announce the occurrence of target stimuli and response periods. However, even with such precautions, some participants only showed a dilation peak during stimulus presentation, while for others, pupil dilation increased continuously, starting as soon as cue presentation. The latter pattern was observed for words in high-level noise, but not for words in low-level noise, suggesting greater cognitive preparation when listening conditions are challenging than when they are less so. The comparison of several baseline periods showed that the early increase in pupil dilation observed in some participants was taken into account in the results, when the chosen period was the most anterior to stimulus presentation. Therefore, it appears that using a baseline period well before cue onset allowed us to measure both anticipatory processes and stimulus processing. While a few pupil-dilation studies have explicitly limited or controlled for the influence of anticipatory processes on pupil dilation (Lewis and Bidelman, 2020; Winn et al., 2015), most studies of listening effort using implicit or explicit cues have focused exclusively on pupil dilation during stimulus processing. Anticipation has also been observed with pupillometry, when participants were engaged in a difficult attention task (McCloy et al., 2016, 2017) or preparing to answer questions in a task involving linguistic challenge (Vogelzang et al., 2016). In particular, McCloy observed that participants' pupil started to show greater dilation as soon as they were informed that they would have to switch their attention from one talker to another, instead of just maintaining their attention on one talker. Likewise, Vogelzang et al. (2016) showed that pupil dilation remained larger during the presentation of a story with a complex structure until participants had to answer a question about the story, while pupil dilation slowly decreased before a question about a simpler story. As well as in the auditory modality, anticipatory behavior has been observed through pupillary activity in visual attention tests. Indeed, larger pupil dilations were elicited in anti-saccade tasks before difficult trials (Wang et al., 2015), especially when participants reported being "on-task" rather than during mind wandering (Hutchison et al., 2020), reflecting preparation and anticipation processes during the

most challenging conditions. Previous fMRI studies (Kurniawan et al., 2013; Vassena et al., 2014) showed that preparation to cognitively demanding tasks seemed to rely on the same brain system than the one implied in attentional control (Krebs et al., 2012; Voisin et al., 2006) and working memory (Engström et al., 2013). These findings suggest that anticipatory effects in speech-in-noise listening tasks should be considered for inclusion into the overall measure of listening effort. In this context, it may be important to use baseline periods that precede explicit or implicit cues to stimulus and/or response periods.

Finally, results obtained using two different baseline durations (500 and 1000 ms) were similar. These findings are in accordance with Winn et al. (2018), who compared baseline-corrected pupil data using different baseline lengths (ranging from 100 ms to 3 s) and found negligible differences between the resulting curves. Although, the baseline durations being compared in the present study only differed by a factor of two, in the literature, baseline-period durations typically range from 100 ms (Lewis and Bidelman, 2020) to 2 s (Ayasse et al., 2017; Winn et al., 2015) (see Table 1). This variability may be related, in part, to differences in recording devices and techniques. For instance, if the recording sampling rate is relatively low, longer baseline durations (>1000 ms) may be needed to prevent the occurrence of blink (during the baseline period) to lead to the necessary elimination of the trial (Winn et al., 2018); with a higher sampling rate, longer baseline periods are more affected by pupil fluctuations (Mathôt et al., 2018). Using short baseline periods may reduce the risk that baseline estimates be contaminated by after-effects from the previous trial (Winn et al., 2018), but it increases the risk that precise baseline estimation be precluded due to a blink or some other source of physiological "noise" in the recording (Mathôt et al., 2018).

Similarly to studies investigating listening effort with speech-in-noise task requiring participants to repeat sentences (Koelewijn et al., 2014; Koelewijn et al., 2018; Ohlenforst et al., 2018; Y. Wang, Kramer, et al., 2018; Zekveld and Kramer, 2014; Zekveld et al., 2010; Zekveld et al., 2011a), we managed to obtain clear pupillary responses to shorter stimuli such as disyllabic words, similarly to Kramer et al. (2013) in young normally hearing participants. Contrary to previous studies assessing listening effort at different intelligibility levels (Koelewijn et al., 2012; Koelewijn et al., 2014; Zekveld et al., 2014a; Zekveld and Kramer, 2014; Zekveld et al., 2010; Zekveld et al., 2011a), our aim was to investigate whether pupillary responses can provide information about listening effort even when speech intelligibility is perfect, i.e., when only correct-response trials are included in the analysis. Sounds levels were adjusted individually to yield scores above 90% in quiet without hearing-aids, in order to have a sufficient number of hit responses per condition per patient. One limitation of this methodological choice is that the resulting sound levels were relatively high (above the typical conversational level for speech), so that when participants wore their hearing-aids, the level of these sounds may have been unusually high (compared to conversational speech). This may partly explain the counter-intuitive outcome, whereby participants rated the listening task as more difficult with hearing-aid than without, in the low SNR conditions (i.e. higher noise). A more likely explanation is that this effect was caused by the background noise being amplified by the hearing-aids compared to the high SNR conditions. Although only correct-response trials were used in the main data analyses, the greater listening effort in the Noise Aided than in the Noise Unaided condition is consistent with the results of these analyses: an interaction between the factors, "noise" and "hearing-aid" was observed for pupil dilation measured during the anticipation window, and post-hoc comparisons revealed that participants' pupil dilation was greater in the Noise Aided than in the Noise Unaided condition. This is consistent with the fact that patients reported greater difficulties with their hearing-aid, than without, in the noise condition, and that they anticipated that difficulty.

However, no effect of hearing-aid or noise on pupil dilation was found in the stimulus processing window. Even though hearing-aid or noise effects are usually observed during the processing period, Zekveld, Kramer, et al. (2018) did not observe any effect of signal-to-noise ratio on pupil responses during this specific time window in a concurrent memory and auditory perception task. This result indicates that differences in listening effort are not systematically observed during the processing of the target stimulus, especially when maximum intelligibility is reached, but could be observed at a different period during the trial. Indeed, in the present study, differences in provided effort during the task were observed before stimulus processing, during the anticipation window. Therefore, this time-window played a key role to highlight differences in listening effort during this task. As Kuchinsky et al. (2014, 2013) did in an orthographic Visual Word Paradigm, we managed to show subtle changes in cognitive load during a speech in noise recognition task via pupil dilation with short uttered words in older adults and exclusively in correct trials. Such differences could not be observed with subjective measures, hence proving the relevance of the use of an objective measure, such as pupillary responses, to effectively assess listening effort.

To conclude, this study compared several methods of analyzing pupillary responses during a speech-in-noise task in hearing-impaired patients. In agreement with previous findings, different normalization methods and baseline durations result in similar pupil dilation patterns. Importantly, the present results highlight the strong influence of the choice of baseline period, with regards to anticipatory processes during effortful listening. A significant impact of baseline period on conclusions was observed, even though only correct-response trials were included in the analyses, so that task performance and intelligibility were near perfect. The findings confirm the feasibility, and potential usefulness, of pupil-dilation as an objective measure to investigate listening effort in older hearing-impaired individuals, with or without hearing-aids.

Open practices statement

Pupillometric data, stimuli material and data processing script (Python language) are available at https://hal.archives-ouvertes.fr/hal-03185825. However, we cannot make available indirectly identifying data. Indeed, according to the French MR003 regulation, indirectly identifying data of persons participating in research and directly or indirectly identifying data of professionals involved in research may be transferred outside the European Union when the transfer is strictly necessary for implementation of the research or the exploitation of its results and under the conditions provided for in Chapter V of the General Data Protection Regulation (GDPR), which is the new legislation that governs the processing of personal data in Europe.

Declarations

Author contribution statement

Lou Seropian: Analyzed and interpreted the data; Contributed analysis tools or data; Wrote the paper.

Mathieu Ferschneider: Conceived and designed the experiments; Performed the experiments; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Fanny Cholvy: Performed the experiments; Contributed analysis tools or data; Wrote the paper.

Christophe Micheyl: Conceived and designed the experiments; Analyzed and interpreted the data; Wrote the paper.

Aurélie Bidet-Caulet: Contributed analysis tools or data; Wrote the paper.

Annie Moulin: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This work was supported in part by the "Auvergne-Rhône Alpes" region ("pack ambition recherche", "Effecbruit"); and the LABEX CELYA (ANR-11-LABX-0060) of Université de Lyon, France, within the program

"Investissements d'Avenir" (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR).

C. Micheyl was supported by Starkey Hearing Technologies, a private entity and manufacturer of hearing technology. Other than through funding of this coauthors' salaries, the sponsors for this study had no involvement in the design of the study, the data analysis, or the writing of the manuscript.

Data availability statement

Data associated with this study is available at https://hal.archives-ouvertes.fr/hal-03185825.

Declaration of interests statement

The authors declare no conflict of interest.

Additional information

Supplementary content related to this article has been published online at https://doi.org/10.1016/j.heliyon.2022.e09631.

Acknowledgements

The authors wish to thank Prof. Stéphane Gallego, Audition Conseil°, for useful advice, continuous support and help with patient's recruitment, and Roxane Hoyer for useful advice concerning statistical analysis.

References

Acland, B., Braver, T., 2014. Cili (v0.5.4 *[Software]*.
Alhanbali, S., Dawes, P., Lloyd, S., Munro, K.J., 2017. Self-reported listening-related effort and fatigue in hearing-impaired adults. Ear Hear. 38 (1), e39–e48.
Aston-Jones, G., Cohen, J.D., 2005. An integrative theory of locus function : adaptive gain and optimal performance. Annu. Rev. Neurosci. 28, 403–450.
Attard-Johnson, J., Ciardha, C.Ó., Bindemann, M., 2019. Comparing methods for the analysis of pupillary response. Behav. Res. Methods 51, 83–95.
Ayasse, N.D., Lash, A., Wingfield, A., 2017. Effort not speed characterizes comprehension of spoken sentences by older adults with mild hearing impairment. Front. Aging Neurosci. 8 (329).
Barlow, H.B., 1972. Dark and light adaptation: psychophysics. In: Jameson, D., Hurvich, L.M. (Eds.), Visual Psychophysics, pp. 1–28.
Bates, D., Mächler, M., Bolker, B.M., Walker, S.C., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Software 67 (1).
Beatty, J., 1982. Task-evoked pupillary responses, processing load, and the structure of processing resources. Psychol. Bull. 91 (2), 276–292.
Borghini, G., Hazan, V., 2018. Listening effort during sentence processing is increased for non-native listeners: a pupillometry study. Front. Neurosci. 12 (152).
Bradley, M.M., Miccoli, L., Escrig, M.A., Lang, P.J., 2008. The pupil as a measure of emotional arousal and autonomic activation. Psychophysiology 45 (4), 602–607.
Brisson, J., Mainville, M., Mailloux, D., Beaulieu, C., Serres, J., Sirois, S., 2013. Pupil diameter measurement errors as a function of gaze direction in corneal reflection eyetrackers. Behav. Res. Methods 45 (4), 1322–1331.
Ellis, C.J.K., 1981. The pupillary light reflex in normal subjects. Br. J. Ophthalmol. 65 (11), 754–759.
Engelhardt, P.E., Ferreira, F., Patsenko, E.G., 2010. Pupillometry reveals processing load during spoken language comprehension. Q. J. Exp. Psychol. 63 (4), 639–645.
Engström, M., Landtblom, A.M., Karlsson, T., 2013. Brain and effort: brain activation and effort-related working memory in healthy participants and patients with working memory deficits. Front. Hum. Neurosci. 7 (140).
Fox, J., Weisberg, S., 2019. An R Companion to Applied Regression (Third Edit). Retrieved from. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.
Francis, A.L., Tigchelaar, L.J., Zhang, R., Zekveld, A., 2018. Effects of second language proficiency and linguistic uncertainty on recognition of speech in native and nonnative competing speech. J. Speech Lang. Hear. Res. 61 (7), 1–16.
Gagl, B., Hawelka, S., Hutzler, F., 2011. Systematic influence of gaze position on pupil size measurement: analysis and correction. Behav. Res. Methods 43 (4), 1171–1181.
General Assembly of the World Medical Association, 2014. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. J. Am. Coll. Dent. 81 (3), 14–18. Retrieved from. http://www.ncbi.nlm.nih.gov/pubmed/25951678.
Hess, E.H., Polt, J.M., 1964. Pupil size in relation to mental activity during simple problem-solving. Science 140 (3611), 1190–1192.
Hutchison, K.A., Moffitt, C.C., Hart, K., Hood, A.V.B., Watson, J.M., Marchak, F.M., 2020. Measuring task set preparation versus mind wandering using pupillometry. J. Exp. Psychol. Learn. Mem. Cogn. 46 (2), 280–295.
Hyönä, J., Tommola, J., Alaja, A., 1995. Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. Q. J. Exp. Psychol. 48A (3), 598–612.
Jensen, J.J., Callaway, S.L., Lunner, T., Wendt, D., 2018. Measuring the impact of tinnitus on aided listening effort using pupillary response. Trends Hear. 22.
Johnson, E.L., Singley, A.T.M., Peckham, A.D., Johnson, S.L., Bunge, S.A., Chiew, K.S., 2014. Task-evoked pupillometry provides a window into the development of short-term memory capacity. Front. Psychol. 5, 1–8.
Just, M.A., Carpenter, P.A., 1993. The intensity dimension of thought : pupillometric indices of sentence processing. Can. J. Exp. Psychol. 47 (2), 310–339.
Kahneman, D., Beatty, J., 1966. Pupil diameter and load on memory. Science 154 (3756), 1583–1586.
Koelewijn, T., Kluiver, H. De, Shinn-cunningham, B.G., Zekveld, A.A., Kramer, S.E., 2015. The pupil response reveals increased listening effort when it is difficult to focus attention. Hear. Res. 323, 81–90.
Koelewijn, T., Shinn-Cunningham, B.G., Zekveld, A.A., Kramer, S.E., 2014a. The pupil response is sensitive to divided attention during speech processing. Hear. Res. 312, 114–120.
Koelewijn, T., Versfeld, N.J., Kramer, S.E., 2017. Effects of attention on the speech reception threshold and pupil response of people with impaired and normal hearing. Hear. Res. 354, 56–63.
Koelewijn, T., Zekveld, A.A., Festen, J.M., Kramer, S.E., 2012. Pupil dilation uncovers extra listening effort in the presence of a single-talker masker. Ear Hear. 33 (2), 291–300.
Koelewijn, T., Zekveld, A.A., Festen, J.M., Kramer, S.E., 2014b. The influence of informational masking on speech perception and pupil response in adults with hearing impairment. J. Acoust. Soc. Am. 135 (3), 1596–1606.
Koelewijn, T., Zekveld, A.A., Lunner, T., Kramer, S.E., 2018. The effect of reward on listening effort as reflected by the pupil dilation response. Hear. Res. 367, 106–112.
Korn, C.W., Bach, D.R., 2016. A solid frame for the window on cognition: modelling event- related pupil responses. J. Vis. 16 (3).
Kramer, S.E., Kapteyn, T.S., Festen, J.M., Kuik, D.J., 1997. Assessing aspects of auditory handicap by means of pupil dilation. Audiology 36 (3), 155–164.
Kramer, S.E., Lorens, A., Coninx, F., Zekveld, A.A., Piotrowska, A., Skarzynski, H., 2013. Processing load during listening: the influence of task characteristics on the pupil response. Lang. Cognit. Process. 28 (4), 426–442.
Kramer, S.E., Zekveld, A. a, Koelewijn, T., Beek, H. Van., 2010. Increased Interest in Listening Effort Pupillometry .. Study 1 Data Selection & Preprocessing Study, 2, pp. 4–7.
Krebs, R.M., Boehler, C.N., Roberts, K.C., Song, A.W., Woldorff, M.G., 2012. The involvement of the dopaminergic midbrain and cortico-striatal-thalamic circuits in the integration of reward prospect and attentional task demands. Cerebr. Cortex 22 (3), 607–615.
Kuchinsky, S.E., Ahlstrom, J.B., Cute, S.L., Humes, L.E., Judy, R., Eckert, M.A., 2014. Speech-perception training for older adults with hearing loss impacts word recognition and effort. Psychophysiology 51 (10), 1046–1057.
Kuchinsky, S.E., Ahlstrom, J.B., Jr, K.I.V., 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.
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. Exp. Aging Res. 42, 50–66.
Kurniawan, I.T., Guitart-Masip, M., Dayan, P., Dolan, R.J., 2013. Effort and valuation in the brain: the effects of anticipation and execution. J. Neurosci. 33 (14), 6160–6169.
Laeng, B., Alnaes, D., 2019. Pupillometry. Chapter 11. In: Klein, C., Ettinger, U. (Eds.), Eye Movement Research. Springer, pp. 449–502.
Lewis, G.A., Bidelman, G.M., 2020. Autonomic nervous system correlates of speech categorization revealed through pupillometry. Front. Neurosci. 13 (1418).
Lin, V.Y.W., Chung, J., Callahan, B.L., Smith, L., Gritters, N., Chen, J.M., Masellis, M., 2017. Development of cognitive screening test for the severely hearing impaired: hearing-impaired MoCA. Laryngoscope 127 (S1), S4–S11.
Mathôt, S., Fabius, J., Heusden, E. Van, Stigchel, S. Van Der., 2018. Safe and sensible preprocessing and baseline correction of pupil-size data. Behav. Res. Methods 94–106.
Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., Bates, D., 2017. Balancing Type I error and power in linear mixed models. J. Mem. Lang. 94, 305–315.
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.
McCloy, D.R., Larson, E., Lee, A.K.C., 2018. Auditory attention switching with listening difficulty: behavioral and pupillometric measures. J. Acoust. Soc. Am. 144 (5), 2764–2771.
McCloy, D.R., Lau, B.K., Larson, E., Pratt, K.A.I., Lee, A.K.C., 2017. Pupillometry shows the effort of auditory attention switching. J. Acoust. Soc. Am. 141 (4), 2440–2451.
Mcgarrigle, R., Dawes, P., Stewart, A.J., Kuchinsky, S.E., 2016. Pupillometry reveals changes in physiological arousal during a sustained listening task. Psychophysiology, 00.
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. Int. J. Audiol. 53 (7), 433–440.
McMahon, C.M., Boisvert, I., Lissa, P. De, Granger, L., Ibrahim, R., Lo, C.Y., Graham, P.L., 2016. Monitoring alpha oscillations and pupil dilation across a performance-intensity function. Front. Psychol. 7 (745).
Miles, K., McMahon, C., Boisvert, I., Ibrahim, R., de Lissa, P., Graham, P., Lyxell, B., 2017. Objective assessment of listening effort: coregistration of pupillometry and EEG. Trends Hear. 21, 1–13.

Moulin, A., Fourcaud-Trocmé, N., 2019. Open-set word lists development, using acoustic, psycholinguistics, and psychometric factors. J. Hear. Sci. 9 (1), 03881. Supplement: 99: ID.
Moulin, A., Bernard, A., Tordella, L., Vergne, J., Gisbert, A., Martin, C., Richard, C., 2017. Variability of word discrimination scores in clinical practice and consequences on their sensitivity to hearing loss. Eur. Arch. Oto-Rhino-Laryngol. 274 (5), 2117–2124.
Nasreddine, Z.S., Patel, B.B., 2016. Validation of montreal cognitive assessment, MoCA, alternate French versions. Can. J. Neurol. Sci. 43 (5), 665–671.
Nassar, M.R., Rumsey, K.M., Wilson, R.C., Parikh, K., Heasly, B., Gold, J.I., 2012. Rational regulation of learning dynamics by pupil–linked arousal systems. Nat. Neurosci. 15 (7), 1040–1046.
Ohlenforst, B., Wendt, D., Kramer, S.E., Naylor, G., Zekveld, A.A., Lunner, T., 2018. Impact of SNR , masker type and noise reduction processing on sentence recognition performance and listening effort as indicated by the pupil dilation response. Hear. Res. 365, 90–99.
Ohlenforst, B., Zekveld, A.A., Lunner, T., Wendt, D., Naylor, G., Wang, Y., Kramer, S.E., 2017. Impact of stimulus-related factors and hearing impairment on listening effort as indicated by pupil dilation. Hear. Res. 351, 68–79.
Partala, T., Surakka, V., 2003. Pupil size variation as an indication of affective processing. Int. J. Hum. Comput. Stud. 59, 185–198.
Payne, D.T., Parry, M.E., Harasymiw, S.J., 1968. Percentage of pupillary dilation as a measure of item difficulty. Percept. Psychophys. 4 (3), 139–143.
R Development Core Team, 2020. R: A Language and Environment for Statistical Computing. Retrieved from. http://www.r-project.org.
Reeves, P., 1920. The response of the average pupil to various intensities of light. J. Opt. Soc. Am. 4 (2), 35–43.
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. Behav. Res. Methods 51, 865–878.
Saunders, G.H., Odgear, I., Cosgrove, A., Frederick, M.T., 2018. Impact of hearing loss and amplification on performance on a cognitive screening test. J. Am. Acad. Audiol. 29 (7), 648–655.
Searle, S.R., Speed, F.M., Milliken, G.A., 1980. Population marginal means in the linear model: an alternative to least squares means. Am. Statistician 34 (4), 216.
Vassena, E., Silvetti, M., Boehler, C.N., Achten, E., Fias, W., Verguts, T., 2014. Overlapping neural systems represent cognitive effort and reward anticipation. PLoS One 9 (3), e91008.
Vogelzang, M., Hendriks, P., van Rijn, H., 2016. Pupillary responses reflect ambiguity resolution in pronoun processing. Lang. Cogn. Neurosci. 31 (7), 876–885.
Voisin, J., Bidet-Caulet, A., Bertrand, O., Fonlupt, P., 2006. Listening in silence activates auditory areas: a functional magnetic resonance imaging study. J. Neurosci. 26 (1), 273–278.
Wagner, A.E., Nagels, L., Toffanin, P., Opie, J.M., Başkent, D., 2019. Individual variations in effort : assessing pupillometry for the hearing impaired. Trends Hear. 23, 1–18.
Wagner, A.E., Toffanin, P., Başkent, D., 2016a. The timing and effort of lexical access in natural and degraded speech. Front. Psychol. 7 (398).
Wagner, A., Pals, C., De Blecourt, C.M., Sarampalis, A., Başkent, D., 2016b. Does signal degradation affect top–down processing of speech? Adv. Exp. Med. Biol. 894, 297–306.
Wang, C.-A., Brien, D.C., Munoz, D.P., 2015. Pupil size reveals preparatory processes in the generation of pro-saccades and anti-saccades. Eur. J. Neurosci. 41 (8), 1102–1110.
Wang, Y., Kramer, S.E., Wendt, D., Naylor, G., Lunner, T., Zekveld, A.A., 2018a. The pupil dilation response during speech perception in dark and light : the involvement of the parasympathetic nervous system in listening effort. Trends Hear. 22, 1–11.
Wang, Y., Naylor, G., Kramer, S.E., Zekveld, A.A., Wendt, D., Ohlenforst, B., Lunner, T., 2018b. Relations between self-reported daily-life fatigue, hearing status, and pupil dilation during a speech perception in noise task. Ear Hear. 39 (3), 573–582.
Wendt, D., Dau, T., Hjortkjær, J., 2016. Impact of background noise and sentence complexity on processing demands during sentence comprehension. Front. Psychol. 7 (345).
Wendt, D., Hietkamp, R.K., Lunner, T., 2017. Impact of noise and noise reduction on processing effort : a pupillometry study. Ear Hear. 38 (6), 690–700.
Wendt, D., Koelewijn, T., Książek, P., Kramer, S.E., 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. Hear. Res. 369, 67–78.
Wetzel, N., Buttelmann, D., Schieler, A., Widmann, A., 2016. Infant and adult pupil dilation in response to unexpected sounds. Dev. Psychobiol. 58 (3), 382–392.
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. Biol. Psychol. 133, 10–17.
Wierda, S.M., Rijn, H. Van, Taatgen, N.A., Martens, S., 2012. Pupil dilation deconvolution reveals the dynamics of attention at high temporal resolution. Proc. Natl. Acad. Sci. Unit. States Am. 109 (22), 8456–8460.
Winn, M.B., Edwards, J.R., Litovsky, R.Y., 2015. The impact of auditory spectral resolution on listening effort revealed by pupil dilation. Ear Hear. 36 (4), e153–e165.
Winn, M.B., Moore, A.N., 2018. Pupillometry reveals that context benefit in speech perception can Be disrupted by later-occurring sounds , especially in listeners with cochlear implants. Trends Hear. 22, 1–22.
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 Hear. 22, 1–32.
Zekveld, A.A., Heslenfeld, D.J., Johnsrude, I.S., Versfeld, N.J., Kramer, S.E., 2014a. The eye as a window to the listening brain: neural correlates of pupil size as a measure of cognitive listening load. Neuroimage 101, 76–86.
Zekveld, A.A., Kramer, S.E., 2014. Cognitive processing load across a wide range of listening conditions : insights from pupillometry. Psychophysiology 51, 277–284.
Zekveld, A.A., Kramer, S.E., Festen, J.M., 2010. Pupil response as an indication of effortful listening: the influence of sentence intelligibility. Ear Hear. 31 (4), 480–490.
Zekveld, A.A., Kramer, S.E., Festen, J.M., 2011a. Cognitive load during speech perception in noise: the influence of age, hearing loss, and cognition on the pupil response. Ear Hear. 32 (4), 498–510.
Zekveld, A.A., Kramer, S.E., Rönnberg, J., Rudner, M., 2019. In a concurrent memory and auditory perception task, the pupil dilation response is more sensitive to memory load than to auditory stimulus characteristics. Ear Hear. 40 (2), 272–286.
Zekveld, A.A., Rudner, M., Johnsrude, I.S., Festen, J.M., van Beek, J.H.M., Rönnberg, J., 2011b. The influence of semantically related and unrelated text cues on the intelligibility of sentences in noise. Ear Hear. 32 (6), e16–e25.
Zekveld, A.A., Rudner, M., Kramer, S.E., Lyzenga, J., Rönnberg, J., 2014b. Cognitive processing load during listening is reduced more by decreasing voice similarity than by increasing spatial separation between target and masker speech. Front. Neurosci. 8 (88).
Zellin, M., Pannekamp, A., Toepel, U., Meer, E. Van Der., 2011. In the eye of the listener : pupil dilation elucidates discourse processing. Int. J. Psychophysiol. 81, 133–141.