Disentangling listening effort and memory load beyond behavioural evidence: Pupillary response to listening effort during a concurrent memory task

Yue ZHANG[1,2,3,4*◑], Alexandre LEHMANN[1,2,3,4◑], Mickael DEROCHE[1,2,3,4,5◑],

1 Department of Otolaryngology, McGill University, Montreal, Canada
2 Centre for Research on Brain, Language and Music, Montreal, Canada
3 Laboratory for Brain, Music and Sound Research, Montreal, Canada
4 Centre for Interdisciplinary Research in Music Media and Technology, Montreal, Canada
5 Department of Psychology, Concordia University, Montreal, Canada

◑These authors contributed equally to this work.
* yue.zhang7@mail.mcgill.ca (YZ)

Abstract

Recent research has demonstrated that pupillometry is a robust measure for quantifying listening effort. However, pupillary responses in listening situations where multiple cognitive functions are engaged and sustained over a period of time remain hard to interpret. This limits our conceptualisation and understanding of listening effort in realistic situations, because rarely in everyday life are people challenged by one task at a time. Therefore, the purpose of this experiment was to reveal the dynamics of listening effort in a sustained listening condition using a word repeat and recall task.

Words were presented in quiet and speech-shaped noise at different signal-to-noise ratios (SNR). Participants were presented with lists of 10 words, and required to repeat each word after its presentation. At the end of the list, participants either recalled as many words as possible or moved on to the next list. Simultaneously, their pupil dilation was recorded throughout the whole experiment.

When only word repeating was required, peak pupil dilation (PPD) was bigger in 0dB versus other conditions; whereas when recall was required, PPD showed no difference among SNR levels and PPD in 0dB was smaller than repeat-only condition.

Baseline pupil diameter and PPD followed different growth patterns across the 10 serial positions in conditions requiring recall: baseline pupil diameter built up progressively and plateaued in the later positions (but shot up at the onset of recall, i.e. the end of the list); PPD decreased at a pace quicker than in repeat-only condition.

The current findings concur with the recent literature in showing that additional cognitive load during a speech intelligibility task could disturb the well-established relation between pupillary response and listening effort. Both the magnitude and temporal pattern of task-evoked pupillary response differ greatly in complex listening conditions, urging for more listening effort studies in complex and realistic listening situations.

Introduction

Effortless as it seems, everyday communication is cognitively demanding. Degraded speech input induced by adverse listening conditions (e.g., background noise, reverberation etc.) and peripheral hearing loss introduces mismatch between perceived acoustic signals and their canonical forms [1–3]. Resolving this mismatch demands more resources from the finite pool of cognitive resources, leading to fewer resources for other cognitive tasks and eventually overload [4, 5]. Populations facing long-term auditory challenges are specifically at risk. For instance, people with hearing impairment and particularly those using cochlear implants (CI) often experience high and sustained effort, even when speech recognition performance is similar [6–10]. CI listeners have to engage and deploy more cognitive resources to achieve a satisfactory level of speech communication due to electric hearing. Such elevated and sustained listening effort is associated with detrimental psychosocial consequences including greater need for recovery after work, increased incidence of sick leave and social interaction withdrawal [6, 11–13]. Therefore, there is a growing interest in the field of hearing science to conceptualise and quantify listening effort during speech perception for different populations.

Pupillometry (the continuous recording of changes in pupil diameter) has been one of most widely used methods for assessing listening effort. Its popularity can be attributed to its sensitivity to a wide range of cognitive tasks and processing that relate to the concept of listening effort [3, 14, 15]. Past studies have shown that pupil size varies with different speech intelligibility, hearing impairment, lexical manipulation, masker type, spectral resolution, memory load and divided/focused attention [4, 16–23]. Typically, when task demands increase, for instance, with lower SNR, degraded spectral resolution or more digits to remember, pupil size increases. However, when the task becomes so demanding that it exceeds the capacity limit, pupil size stops increasing and/or starts decreasing, forming a relation similar to inverse-U shape between task demands and listening effort [14, 24–30].

Because pupil size variation is the result of a complex interplay between the parasympathetic and sympathetic system, pupillometry can also reveal aspects of listening effort relating to fatigue, motivation and arousal [31–34]. For instance, Wang et al. [34] showed a negative correlation between the need for recovery and peak pupil dilation relative the baseline (PPD), supporting the assumption that high fatigue could be related to a reduced state of arousal (hence smaller pupillary response) [35]. Furthermore, pupillometry has a reasonable temporal locking to cognitive events, with some delay due to the slow locus coeruleus (LC)-norepinephrine (NE) response. Typically, the peak of event-evoked pupillary dilation arrives within the time window from 0.7 to 1.5 sec following the target stimuli [21, 36, 37]. This allows pupillary response to show trial-by-trial and within-trial variation in listening effort, which can reveal the underlying cognitive processing and allocation policy that are hardly measurable via behavioural outcomes. For instance, pupil size typically decreases with increasing trial/block numbers within one condition, suggesting fatigue or habituation with similar stimuli and task [17,38–40]; it also varies with the level of engagement that changes from one trial to the next [41].

Due to these multiple influences on pupillary responses, there is only limited understanding of how pupil size varies in complex situations, where multiple cognitive functions are engaged and effort sustained over a period of time. Rarely in everyday life are people challenged by one task at a time. Even in a simple conversation, one needs to decode the incoming speech input embedded in various types of background noise, retain some information for mental processing, pondering over the best choices of words and articulating a verbal response (potentially monitoring the feedback of one's own voice), all of which require sustained cognitive processing over time. Understanding pupillary response to speech understanding in those situations is essential to conceptualise and quantify listening effort in ecological conditions, especially in the case of hearing aid or cochlear implant users.

Specifically, the relation between single task demand and pupil dilation has been shown and well-replicated in studies manipulating speech intelligibility and memory load [24–29]. However, there are only a handful of pupillometry studies involving multiple and sustained tasks within hearing science. For instance, Karatekin et al. [16] found that pupil diameter increased progressively with more digits to remember during a digit span task and a dual task (digit span with visual response time task), but the rate of increase was shallower in the dual task than the single task. McGarrigle et al. [42] asked NH participants to listen to one short passage per trial, presented with multi-talker babble noise, and at the end of each passage judge whether images presented on the screen were mentioned in the previous passage. A steeper decrease in (baseline corrected) pupil size was found for difficult than easy SNR, but only in the second paragraph. This was interpreted as an index of the onset of fatigue in listening conditions requiring sustained effort. However, paragraphs were between 13-18s and the target word was periodically varied inside the paragraph, making it difficult to measure directly the pupillary response evoked by recognising and encoding the target item. In Zekveld et al. [43], participants had to recall the four-word cues (either related or unrelated to the following sentence) presented visually before the onset of the sentence embedded in a speech masker. The 7dB SNR difference between two sentence-in-noise conditions (-17dB and -10dB) elicited a difference in intelligibility, but not in peak and mean pupil dilation. This contrasted with the well-established effect of auditory task demands on the pupillary response, suggesting that an external cognitive load (i.e., memory) during speech processing could nullify the intelligibility effects on pupil dilation response. Overall, these studies point to a complicated, but under-investigated, relation between the speech task and the pupil dilation response, when other cognitive task load is present.

Therefore, the current experiment starts addressing the lack of systematic investigation on the dynamics of pupillary response in complex and sustained listening situations. To do so, we designed a behavioural paradigm including two TASKs with different demands in cognitive resources: a repeat-only condition where participants listen and repeat one word consecutively for ten words, and a repeat-with-recall condition where after listening and repeating each of the ten words, they need to recall as many words as possible at the end of the tenth word. Using words instead of digits or paragraphs, the paradigm utilises natural speech, yet still provides precise time-locking to the canonical task-evoked pupil response. The recall task poses a substantial and sustained requirement of cognitive resources (attention and working memory) that are also essential for speech understanding: participants had to complete both word recognition and memorising tasks within the same time window, and keep retaining more words in the memory until the end of the list. The task difficulty was further manipulated by embedding words in different levels of speech-shaped noise to compare pupillary responses under high and low listening effort (LISTENING condition). The effect of SNR on pupil dilation response during speech-in-noise tasks has been well-established in past studies, but remains unclear when the memory load is present. Simultaneously, pupil size variations were recorded. By comparing pupil traces of words recalled and forgotten, we could potentially identify the time window and sequence for word recognition and memory encoding. Participants' subjective ratings on effortfulness were also collected, and results were correlated with individuals' behavioural and pupillary responses. This analysis helps to disentangle further pupil responses corresponding to word recognition and memory, by identifying pupillary metrics that are significantly related to word recognition, recall and self-rating performance.

According to past studies, the main hypotheses were:

- Fewer words correctly repeated in difficult versus easy SNR conditions due to more degraded acoustic input, and fewer stated words recalled with more adverse SNRs due to limited cognitive capacity to prioritise the word recognition task.
- Bigger pupillary response in difficult versus easy SNR conditions, due to more degraded acoustic input. Bigger pupillary response in repeat-with-recall versus repeat-only condition: bigger baseline pupil diameter due to accumulating memory load and bigger PPD due to greater cognitive demands. This difference might also depend on the serial position.
- Quick and large increase in pupil diameter at the time of recall (similar to Cabestrero et al. [26]), and possibly bigger increase in difficult versus easy SNR conditions.
- Higher self-report effort in difficult SNR and repeat-with-recall conditions, reflecting the increased subjective experience of effort for conditions with more degraded acoustic input and sustained effort.

1 Methods

1.1 *Participants*

Data were collected from 25 adults (age range:18-49 years; average: 29 years). A pure tone audiometry was administered to ensure that all participants had binaural thresholds at or better than 25 dB HL at 0.25, 0.5, 1, 2, 4, 8 kHz. All participants were native speakers of either French or North American English (the study being always run in their native language).

1.2 *Stimuli*

Stimuli were standard CNC words recorded from a male American English Speaker and monosyllabic Fournier words recorded from a male French Speaker (mean duration = 0.62s, SD = 0.09s). Words were fully randomised, grouped into lists of 10 and occurred only once in each list. They were then masked by speech-shaped noise (filtered on the long-term excitation pattern of the entire material, respectively in English or French) at three SNR levels of 0 dB, 7 dB and 14 dB. A quiet condition was also included, making a total of 4 LISTENING conditions.

Each LISTENING condition was paired with TASK condition (repeat-only and repeat-with-recall) and was repeated three times (using different word lists), making a total of 4x2x3=24 test blocks. Condition sequences and word lists were fully randomised for each participant.

1.3 *Procedure*

Participants sat on a chair in a soundproof room, 2m in front of a 35-inch screen monitor and wearing an infrared binocular eyetracker (Tobii Glasses Pro2, 100 Hz sampling rate). The room and screen luminance levels were adjusted to reach 75lx (measured using a luxometer with the sensor positioned at the same height as the participants' left eye and facing the screen). The luminance levels were fixed throughout the experiment, to avoid changes in light level inducing task-unrelated pupillary response. All audio stimuli were presented through a Beyer Dynamics DT 990 Pro headphone via an external soundcard (Edirol UA), calibrated at 65 dB SPL. Experiments were run in Matlab 2016b, using Psychtoolbox and custom software.

After demonstrating the task and explaining the procedure, participants practised with one repeat-with-recall condition at 14 dB to familiarise themselves with the test sequence and requirements for pupil recording.

Before each test block, participants were notified by words on the screen to either recall (printed in red) or not recall (printed in black) at the end of the ten words. 3s after the notification, a black fixation cross appeared and stayed for another 1s, to indicate the start of the first trial and eliminate any carry-over effect from reading the coloured words in the pre-block notification. In each trial within the block, the presentation of speech-shaped noise masker (or quiet in the quiet condition) started 1.5s before the onset of the word. This was to provide time for the pupils to recover from the previous trial to temper carry-over effect (0.5s) and to measure baseline pupil diameter (1s). SNR was varied by fixing the masker level and adjusting the target level. In this way, listeners could not estimate the upcoming task difficulty based on the noise level [30]. Participants were instructed to fixate on the black fixation cross displayed at the centre of the screen. After 1.5s, the word was played, and the presentation of the masker noise (or quiet in the quiet condition) was turned off 1s after the word offset. Upon the masker offset, the fixation cross turned into a circle, and this prompted participants to repeat back the word. They were instructed to fixate on the black circle during the verbal response. The experimenter then typed down the repeated word and pressed ENTER to proceed to the next trial. Words were scored automatically based on whether the characters typed matched the transcripts. No fixed time was enforced on the participants and experimenter to repeat back and type down the correct word. Both the participants and the experimenter were instructed to take time. This was to avoid extra mental stress and ensure the correct scoring of word recognition and recall performance. On average, it took 2.11s (SD=1.08s) from the onset of the prompt cue to the onset of the next trial.

In blocks requiring recall, at the end of the 10th word, the word RECALL appeared on the screen followed by a black circle to prompt the participants to recall as many words as possible from the previous 10 words in any order. Participants were instructed to fixate on the black circle during recall. Their responses were typed down by the experimenter and scored automatically based on character matching with the response typed during word repeat. Therefore, correctly recalled words would include words that were correctly recalled misperceptions (similar to [44]), dissociating the impact of intelligibility from recall performance.

At the end of each block regardless of the TASK condition, participants were asked verbally to rate *How effortful the last block was* from 1 to 10, 10 being most effortful. Their subjective ratings were typed down by the experimenter. An illustration of the test sequence is shown in Fig 1.

Fig 1. Test sequence in a block. Before each block, participants were presented with either words ‘please listen, repeat and recall ’ in red or words ‘please listen, repeat and no recall ’ in black against a white screen, indicating whether the incoming block was repeat-only or repeat-with-recall condition. 3s after the words notification, a black fixation cross appeared and stayed for another 1s, to signal the start of the first trial. The trial started with acoustic presentation of 0.5s speech-shaped noise (or quiet in the quiet condition) and visual presentation of a black fixation cross (‘intertrial ’). Another 1s of baseline measurement followed, with the same acoustic and visual presentation (‘baseline’). The word was then played at 1.5s into the trial, followed by noise presentation (or quiet in the quiet condition) for 1s (‘waitpeak ’), with the same visual presentation. Upon the offset of ‘waitpeak ’, the black fixation cross turned into a black circle to prompt listeners to repeat back the word ‘repeat ’. If the block was a repeat-with-recall condition, at the end of the 10th word, participants were prompted by the word RECALL followed by a black circle on the screen to start recalling previously repeated words. At the end of the block, participants were verbally reminded to rate *How effortful was the last block* from 1 to 10, 10 being most effortful.

The experiment lasted for 1 hour.

1.4 *Data processing and analysis*

There were no differences between the French-speaking and English-speaking listeners in word recognition ($t = 0.44, df = 20.45, p = 0.63$), word recall ($t = 0.09, df = 20.68, p = 0.92$) and subjective rating ($t = 0.68, df = 22.57, p = 0.50$), using between-subjects two-tailed t-tests. Therefore, data were firstly aggregated over language (as this played no role and was not a factor of interest in our study).

1.4.1 *Word recognition performance*

To examine the effect of LISTENING and TASK conditions on word recognition, a logistic mixed-effect model was fitted on listeners' word recognition, using LISTENING and TASK conditions as fixed effect factors and LISTENER as random effect factor. Mixed effect models allow for controlling the variance associated with random factors without data aggregation. Therefore, by using LISTENER as random effect factor in the model, we controlled for the variance in overall performance (random intercept) and dependency on other fixed factors (random slope) that were associated with LISTENER. Models were constructed using the lme4 package [45] in R [46], and figures were produced using the ggplot2 package [47]. Fixed and random effect factors entered the model, and retained in the model only if they significantly improved the model fitting, using Chi-squared tests based on changes in deviance ($p < 0.05$). Differences between levels of each factor and interactions were examined with post-hoc Wald test. p values were estimated using the z distribution in the test as an approximation for the t distribution [48].

1.4.2 *Word recall performance*

To examine the effect of background noise on stated word recall performance, a logistic mixed-effect model was fitted on the number of words correctly recalled, with LISTENING condition as fixed effect factor and LISTENER as random effect factor, and following the same procedure reported above. Note that the recall performance was counted as stated word correct, and as such a word could be misunderstood and yet corrected recalled.

1.4.3 *Pupil data preprocessing*

Baseline pupil diameter in each trial was calculated as averaged pupil trace 1s before each word onset. The pupil diameter measured from the word onset to the end of the trial was subtracted from that baseline level to obtain relative changes in pupil diameter elicited by the task. Sample points were coded as blinks when pupil diameter values were below 3 standard deviation (SD) of the mean of the unprocessed trace or when gazing positions were 3 SD away from the centre of the fixation. Traces between 10 data points (0.1s) before the start and after the end of blink were interpolated cubically in Matlab, to further decrease the impact of the obscured pupil from blinks. Trials that had over 20% of the data points coded as blinks from the start of baseline to the start of the next trial were excluded. Trials containing blinks longer than 0.4s were also excluded, because they were more likely to be artefacts than normal blinks [49]. Three participants had more than 20% of the overall trials discarded and were excluded from the pupillometry analysis (but kept for behavioural and subjective rating analysis).

All valid traces were low-pass filtered at 10 Hz with a first-order Butterworth filter to preserve only cognitively related pupil size modulation [50]. Processed traces were then aligned by the onset of the response prompt (the display of circle to signal participants to repeat back the word) and aggregated per listener, by each WORD

POSITION in the 10-word list, TASK and LISTENING conditions.

1.4.4 *Pupil data analysis*

Three indices of pupil response (baseline pupil diameter, peak pupil dilation PPD and peak latency) were obtained from processed traces, consistent with the method in [17]. PPD was the maximum diameter of pupil measurements from word onset to response prompt (time window 1), relative to the baseline pupil diameter. Note that we used the averaged pupil trace 1s before each word as the baseline during baseline correction, therefore, PPD corresponded to the phasic pupillary response evoked by word recognition. This method was in line with the aim of our experiment to investigate pupillary response to listening effort when another cognitive load was present. (For comparison, supplementary material S1 File showed an alternative method to calculate PPD, i.e. baseline corrected by the averaged pupil trace 1s before the first word in the list, and its impact on understanding the results. To summarise, this alternative method could not disentangle the compound impact of listening effort and memory load on pupillary response.) Peak latency response was the time between word onset to the peak dilation. During this time window, listeners were predominantly listening and decoding the acoustic signals. There were also no significant differences in baseline pupil diameter ($t = 0.75, df = 19.7, p = 0.46$), PPD ($t = -0.49, df = 18.53, p = 0.63$) and peak latency ($t = 1.02, df = 17.04, p = 0.32$) between native English and French speakers, so data were aggregated over language.

To investigate how the experimental manipulations on listening effort and memory load affected the dynamics of pupillary response, three mixed effect models were then fitted on baseline diameter, PPD and peak latency respectively. LISTENING and TASK conditions were entered as fixed effect factors to investigate the impact of experimental conditions on the pupillary response averaged over the ten-word list. WORD POSITION was coded as from 1 to 10, corresponding to the serial position of each word in the list. Entering this variable as another fixed factor enabled us to examine the temporal variations of different pupil metrics. Also, the interaction between WORD POSITION and other fixed effect factors showed how the pupil dynamics differed in the conditions with and without memory load, and under high and low listening effort. LISTENER was entered in the model as a random effect factor. Model buildings followed the same procedure above.

To further explore the sequence of different cognitive processing stages, pupil traces of words correctly versus incorrectly recognised, and pupil traces of words forgotten versus recalled were compared. For words correctly and incorrectly recognised, two logistic mixed effect models were fitted on the word recognition correct, using PPD and peak latency (calculated in time window 1 from word onset to response prompt) as fixed effect factors and LISTENER as random effect factor. For words recalled and forgotten, a new time window was added into analysis. New PPD and peak latency were calculated at the time window from the response prompt to 1.5s after the response prompt (time window 2). This time window corresponded to when participants were probably rehearsing and encoding the perceived word to working memory storage. The inclusion of extra 1.5s after the response prompt in the analysis was to include the time for rehearsing and encoding the perceived word to working memory storage. Logistic mixed effect models were fitted on the word recall, using PPD and peak latency in two time windows as fixed effect factors. Note that in this particular analysis pupillary parameters were used as independent variables to assess behavioural outcomes, to understand how the strategy of cognitive resources allocation affected word recognition and recall. In other words, it was examined as a predictive tool: predict whether a given word would be correctly understood or not, and recalled or forgotten, from the particular shape of a pupil trace.

Finally, to explore the impact of LISTENING condition on the pupillary response during recall, pupil traces from recall onset cue to 15s after the cue was firstly baseline-corrected by subtracting the average diameter of all previous word trials in the block. They were then de-blinked and low-pass filtered using the same parameters as above. Processed traces were then aggregated per listener by LISTENING condition. The mean of the trace during word recall was calculated. A mixed effect model was fitted on the mean pupil diameter during recall, with LISTENING condition as fixed effect factor and LISTENER as random effect factor.

1.4.5 *Subjective listening effort rating and individual differences*

To examine the effect of LISTENING and TASK conditions on subjective rating, a logistic mixed-effect model was fitted on ratings, with LISTENING and TASK conditions as fixed effect factors and LISTENER as random effect factor, and following the same procedure reported above.

In a final attempt to delineate different components of the pupillary dynamics, each participant's pupillary responses (baseline diameter and PPD) were correlated with their age, word recognition, word recall and subjective rating performance.

All best fitting models and summary output were reported in the Supplementary Materials S1 Table.

2 Results

2.1 *Word recognition performance*

There was a significant main effect of LISTENING condition ($\chi^2 = 684.11, df = 3, p < 0.001$) and interaction between LISTENING and TASK conditions($\chi^2 = 10.64, df = 3, p = 0.01$), but no main effect of TASK ($\chi^2 = 1.49, df = 1, p = 0.22$). Post-hoc Wald test showed that word recognition at 0dB was lower that at 7dB ($\beta = -1.8, se = 0.13, p < 0.001$), 14dB

($\beta = -2.61, se = 0.18, p < 0.001$) and quiet ($\beta = -3.72, se = 0.33, p < 0.001$); 7dB was lower than 14dB ($\beta = -0.82, se = 0.2, p < 0.001$) and quiet ($\beta = -1.92, se = 0.34, p < 0.001$); 14dB was lower than quiet ($\beta = -1.1, se = 0.36, p < 0.001$). At 0dB, word recognition was higher in repeat-with-recall than in repeat-only condition ($\beta = 0.27, se = 0.12, p = 0.03$). Surprisingly, in quiet, word recognition was lower in repeat-with-recall than in repeat-only condition ($\beta = -1.5, se = 0.64, p = 0.02$) (Fig 2). Recognition performance did not vary across ten word positions within each block ($\chi^2 = 15.14, df = 9, p = 0.09$).

Fig 2. Behavioural performance. All data are averaged across 25 listeners. The error bars and shaded width denote 1 standard error of the mean. (a) shows word recognition performance as a function of LISTENING and TASK conditions, and (b) shows free recall performance as a function of the LISTENING condition.

2.2 *Word recall performance*

There was a significant main effect of LISTENING condition ($\chi^2 = 18.46, df = 3, p < 0.001$), and post-hoc Wald test showed that fewer stated words were recalled at 0dB than 7dB ($\beta = 0.38, se = 0.11, p < 0.001$), 14dB ($\beta = 0.34, se = 0.11, p = 0.003$) and quiet ($\beta = 0.45, se = 0.11, p < 0.001$), with no other significant differences (Fig 2b).

2.3 *The effect of noise and memory load on pupillary response*

Fig 3a and Fig 4a show the pupil diameter variation from the onset of baseline to 1.5s after the response cue.

Fig 3. Pupillometry results as a function of LISTENING and TASK conditions. All data are aggregated across 22 listeners, and WORD POSITION, LISTENING, TASK conditions. The error bars and shaded width denote 1 standard error of the mean. (a) shows changes in pupil size as a function of time during each trial, for each LISTENING and TASK conditions. (b) and (c) plot baseline pupil diameter and PPD as a function of LISTENING and TASK conditions respectively.

Fig 4. Pupillometry results as a function of TASK and WORD POSITION. All data are aggregated across 22 listeners, and WORD POSITION, LISTENING, TASK conditions. The error bars and shaded width denote 1 standard error of the mean. (a) shows changes in pupil size as a function of time at each WORD POSITION for each TASK condition. (b) and (c) plot baseline pupil diameter and PPD as a function of WORD POSITION and TASK condition respectively.

For baseline pupil diameter, there was a significant main effect of LISTENING condition ($\chi^2 = 11.21, df = 3, p = 0.01$), TASK ($\chi^2 = 283.49, df = 1, p < 0.001$) and WORD POSITION ($\chi^2 = 24.85, df = 9, p = 0.003$), and significant interaction between TASK:WORD POSITION ($\chi^2 = 82.99, df = 9, p < 0.001$). Post-hoc tests showed that baseline pupil diameter at 0dB was not different from 7dB ($\beta = 0.004, se = 0.01, p = 0.68$), but both were bigger than 14dB ($\beta = 0.04, se = 0.01, p = 0.002; \beta = 0.03, se = 0.01, p = 0.007$) and quiet ($\beta = 0.04, se = 0.01, p = 0.04$; $\beta = 0.03, se = 0.01, p = 0.04$); 14dB was not different from quiet ($\beta = 0.01, se = 0.01, p = 0.32$). Overall, baseline pupil diameter at repeat-with-recall condition was significantly bigger (about 0.2 mm) than that at repeat-only condition ($\beta = 0.18, se = 0.01, p < 0.001$) (Fig 3b). A trend analysis on WORD POSITION showed that from the 1st to 10th word, repeat-only condition had a linearly decreasing trend ($\beta = -0.18, se = 0.01, p < 0.001$), whereas repeat-with-recall condition had a linearly increasing trend ($\beta = 0.18, se = 0.01, p < 0.001$) (Fig 4b). Baseline diameter in repeat-with-recall condition also showed a significant quadratic trend ($\beta = -0.09, se = 0.03, p < 0.001$), suggesting that the greatest increase in baseline diameter occurred in the mid-section of the word list. No significant cubic trend was detected.

For PPD, there was a significant main effect of WORD POSITION ($\chi^2 = 104.39, df = 9, p < 0.001$), and no significant main effect of LISTENING ($\chi^2 = 2.55, df = 3, p = 0.47$) and TASK conditions ($\chi^2 = 1.85, df = 1, p = 0.17$). Interactions between LISTENING:TASK ($\chi^2 = 13.15, df = 3, p = 0.004$) and

TASK:WORD POSITION ($\chi^2 = 22.98, df = 9, p = 0.006$) were significant, and no significant three-way interaction ($\chi^2 = 31.05, df = 27, p = 0.27$). Post-hoc tests showed that at 0dB, repeat-only condition evoked bigger PPD than repeat-with-recall condition ($\beta = 0.03, se = 0.01, p = 0.04$), and no difference between two tasks at other SNR levels (Fig 3c). Examining the same interaction differently: SNR only affected the repeat-only condition, showing a bigger PPD at 0 dB than at other SNR levels. A trend analysis on WORD POSITION showed that from the 1st to the 10th word, there was a decrease in PPD ($\chi^2 = 55.73, df = 1, p < 0.001, \beta = -0.08, se = 0.01, p < 0.001$), and this decrease was steeper in the repeat-with-recall condition than repeat-only condition ($\beta = -0.07, se = 0.007, p < 0.001$) (Fig 4c). No further significant quadratic or cubic trend.

For peak latency, there was a significant main effect of LISTENING condition ($\chi^2 = 8.67, df = 3, p = 0.03$) and WORD POSITION ($\chi^2 = 66.98, df = 9, p < 0.001$), and significant interaction between TASK:WORD POSITION($\chi^2 = 21.93, df = 9, p = 0.009$). Post-hoc test showed that at 0dB pupil size peaked significantly later than at 7dB ($\beta = 0.07, se = 0.03, p = 0.008$), 14dB ($\beta = 0.06, se = 0.02, p = 0.01$), and quiet ($\beta = 0.05, se = 0.03, p = 0.05$). From the 1st to the 10th word, there was an increase in repeat-only condition ($\beta = -0.11, se = 0.04, p = 0.007$), and also an increase ($\beta = -0.3, se = 0.04, p < 0.001$) in repeat-with-recall condition, but steeper than repeat-only condition ($\beta = 0.2, se = 0.05, p = 0.001$). No further significant quadratic or cubic trend.

2.4 *Pupillary response: incorrectly versus correctly repeated words*

For the pupillary responses of words that were correctly and incorrectly recognised, no difference in baseline diameter was found ($\chi^2 = 0.001, df = 1, p = 0.94$), suggesting that there was no differential arousal that could explain the word intelligibility. There was a main effect of PPD ($\chi^2 = 12.59, df = 1, p < 0.001$) and a significant interaction of TASK:PPD ($\chi^2 = 13.9, df = 1, p < 0.001$). No significant effect of peak latency ($\chi^2 = 1.96, df = 1, p = 0.16$) was found. Post-hoc tests showed that at repeat-only condition, bigger PPD was associated with incorrectly repeated words ($\beta = -1.8, se = 0.35, p < 0.001$), and no such relation at repeat-with-recall task (Fig 5a).

Fig 5. Comparing pupil traces for words correctly and incorrecty repeated, recalled and forgotton. All data are averaged across 22 listeners. The shaded width denotes 1 standard error of the mean. (a) compares the pupil traces for words correctly and incorrectly repeated in each TASK condition. (b) compares the pupil traces for words that are successfully recalled and forgotten. Traces in two time windows are analysed: first analysis window is from the onset of word to the onset of the response prompt, and the second analysis window is from the onset of the response prompt to 1.5s after the prompt.

2.5 *Pupillary response: recalled versus forgotten words*

Comparing the pupillary responses of words that were later recalled or forgotten, no difference in baseline size was found ($\chi^2 = 0.001, df = 1, p = 0.9$). At the first time window, there was no significant main effect of PPD ($\chi^2 = 1.76, df = 1, p = 0.18$) and latency ($\chi^2 = 1.49, df = 1, p = 0.22$). At the second time window, there was a significant main effect of peak pupil diameter ($\chi^2 = 4.87, df = 1, p = 0.03$). Post-hoc Wald test showed that bigger peak dilation at the second time window was associated with the successful recall of the word ($\beta = 3.18, se = 1.47, p = 0.03$) (Fig 5b).

2.6 *The effect of noise on pupillary response during word recall at the end of a block*

For the mean pupil diameter during the listeners' word recall, there was no difference among SNRs ($\chi^2 = 0.67, df = 3, p = 0.88$) (Fig 6); and the mean pupil diameter jumped from about 4.0 to 4.3-4.4 mm (just short of 10%). However, across the individuals, we observed an interesting relationship to the memory performance: in quiet condition, bigger mean pupil diameter during recall was associated with more stated words correctly recalled ($\beta = 0.65, se = 0.26, p = 0.01$).

Fig 6. Pupil traces from 10s before the recall onset to 15s after the recall onset. Each panel shows the averaged traces in each LISTENING condition. All data are aggregated across 22 listeners. The shaded width denotes 1 standard error of the mean.

2.7 *Subjective listening effort rating*

There was a significant main effect of LISTENING ($\chi^2 = 2278.51, df = 3, p < 0.001$) and TASK conditions ($\chi^2 = 7137.01, df = 1, p < 0.001$), and a significant interaction of LISTENING:TASK ($\chi^2 = 239.78, df = 3, p < 0.001$) on subjective rating. Subjective rating at 0dB was higher than at 7dB ($\beta = 0.85, se = 0.04, p < 0.001$), 14dB ($\beta = 0.89, se = 0.04, p < 0.001$) and quiet ($\beta = 1.29, se = 0.05, p < 0.001$); 7dB was higher than quiet ($\beta = 0.44, se = 0.05, p < 0.001$) but not 14dB ($\beta = 0.04, se = 0.05, p = 0.38$); and 14dB was higher than quiet ($\beta = 0.4, se = 0.05, p < 0.001$). Overall, subjective rating at repeat-with-recall condition was higher than that at repeat-only condition ($\beta = 1.56, se = 0.03, p < 0.001$), and the difference was smaller at 0dB than other SNR levels ($\beta = -1.13, se = 0.06, p < 0.001$) (Fig 7a).

Fig 7. Individual differences Each data point corresponds to one participant. The error bars denote 1 standard error of the mean. (a) plots subjective rating as a function of LISTENING and TASK conditions. (b) to (d) show the significant correlations ($p < 0.05$) between behavioural and pupillary measures.

2.8 *Individual differences*

On an individual level, baseline diameter (within word lists) positively correlated with word recall performance ($r = 0.45, p = 0.04$, Fig 7b), and negatively correlated with subjective rating ($r = -0.45, p = 0.04$, Fig 7c). PPD negatively correlated with word recognition performance ($r = -0.48, p = 0.02$, Fig 7d), but this was only true when no memory requirement was involved: in repeat-with-recall condition, there was no significant correlation between PPD and word recognition performance ($r = 0.08, p = 0.21$). These relations were modulated by participants' age: word recall performance worsened with age ($r = -0.5, p = 0.01$); baseline diameter shrunk with age ($r = -0.52, p = 0.01$); and subjective rating shifted up with age ($r = 0.5, p = 0.01$). Note that these correlations should be considered with caution due to no corrections.

Discussion

The current experiment used a word recall paradigm to elicit sustained and concurrent memory load on word recognition in noise. Pupil diameters were recorded simultaneously to investigate the dynamics of pupillary response in complex listening situations. A number of our findings can be contrasted with the literature, advancing current debates on 1) interferences between concurrent tasks, 2) the nature of pupil dynamics in dual versus single tasks, 3) the predictive power of pupillometry for intelligibility and memory, and 4) individual differences.

2.9 Word recall task interfering with the word recognition task

Consistent with our first hypothesis, results showed that noise impaired both word recognition and recall. Fewer stated words were recalled at 0dB than 7dB, 14dB and quiet conditions. Note that to dissociate the impact of word recognition from recall performance, word recall scoring was based on whether the recalled words matched the words repeated by participants, rather than the transcripts (similar to [44]). Past studies using the recall paradigm reported similar results. McCoy et al. [7] showed that even when word recognition was near perfect (>98%), listeners with mild-to-moderate hearing loss had worse word recall performance than NH listeners in a running memory task. In [51], NH participants repeated the final word of each of 8 sentences embedded in babble-speech noise, and at the end of the 8th sentence recalled as many of the previously reported words as possible. Results showed that challenging signal-to-noise (SNR) condition impaired both word recognition and recall of the stated words performance. When a noise reduction algorithm [52] was turned on, participants' word recognition performance did not change, but their word recall performance improved (at least for sentences with high contextual information). Particularly, the recall of items at the beginning of the lists was most affected (suggesting a benefit in the primacy effect). Ng et al. [53] tested moderate to severe hearing loss participants using a similar memory recall paradigm referred to as the sentence-final word identification and recall (SWIR). Results showed that even under similar intelligibility, babble-speech noise impaired word recall performance more than speech-shaped noise. And with the assistance of a noise reduction algorithm, participants with better working memory capacity recalled more words in babble-speech noise, particularly in the recency position. Lunner et al. [44] also replicated the benefit of using noise reduction algorithm on word recall performance using a Danish version of SWIR for native Danish-speaking hearing-aid users. In line with the interpretation in previous studies, we believe that this SNR effect on recall reflects that higher listening effort during word recognition evoked at lower SNR leaves fewer cognitive resources for encoding and retrieving words, leading to the decreased performance in the word recall task [4, 8, 54–56].

Surprisingly, we found a possible interference from the recall task on the word recognition task. At 0dB, word recognition performance was better when participants expected word recall at the end of the list; and in quiet, word recognition was worse when participants expected word recall task at the end. Although word recognition was essentially the same task in repeat-only condition and repeat-with-recall condition, participants might evaluate and anticipate the amount of cognitive resources differently.

At 0dB, listeners might be more attentive and ready to engage overall more cognitive resources when they were notified at the beginning of the block that they should recall at the end of 10th word because they anticipated the incoming block to be demanding. When no recall was required, they might have judged beforehand that the incoming block was not worthwhile to mobilise too many resources, hence worse recognition performance. Furthermore, in quiet with repeat-with-recall condition, listeners should have sufficient capacity to reach a better primary task performance (as shown by a higher word recognition in repeat-only condition), but instead, they performed worse in the word recognition task compared to in the repeat-only condition. This might suggest that they did not prioritise the word recognition task (although they were instructed explicitly to do so by the experimenter), and may have shifted some resources to the recall task probably because it was more interesting and rewarding [57–60].

This interference warrants further investigation, because it concerns the validity of using a dual-task paradigm in measuring listening effort. In order to interpret safely the difference in secondary task performance as a result of listening effort, implicit assumptions of the dual-task paradigm need to be reviewed [61]. Firstly, the paradigm assumes that participants have a limited pool of cognitive resources, but The Framework for Understanding Effortful Listening (FUEL) model also notes that resources that are available to be allocated are fluctuating with other factors besides overall task demands [3, 4]. In other words, the relationship between task difficulty and effort is not linear, but modulated by factors like fatigue, motivation and (dis)pleasure [33, 62–67]. Secondly, the paradigm assumes that listeners, under explicit instructions, will prioritise the primary task by investing as many resources as possible, and only leaves whatever left of the resources for the secondary task. However, individual differences and task characteristics might affect listeners' actual strategy [3]. For instance, older adults may differ from younger adults in the extent to which they prioritise one task over another [57–59]. And when the primary task is too complex or secondary task more novel, participants may consciously or unconsciously shift more resources to the secondary task relative to the primary task [68–70]. Although the recall paradigm from previous studies is sensitive to the relative allocation of cognitive resources, there is no direct method to gauge the total amount of resources deployed and how they are allocated [61]. As illustrated in the current experiment, listeners might not mobilise and/or allocate the same amount of cognitive resources for the speech recognition task when a secondary recall task was anticipated, even under explicit instruction. This makes it unclear whether the difference in the recall performance is due to differences in the listening effort, or prior mobilisation of overall cognitive resources, or internal shift of resources between primary and secondary task. Previous studies using SWIR paradigm have typically fixed the SNR levels at or close to ceiling performance, to ensure no substantial differences in sentence intelligibility. But this still does not exclude the possibilities mentioned above, because even at ceiling performance level (similar to the quiet condition in the current experiment), interferences could occur. This might be of particular concern when applying the test to listener groups who are susceptible to fatigue and task interference, for instance hearing impaired populations and children, because they might either give up or not fully engaged in the first place even when the available capacity can meet the processing demand [3, 66, 68–70].

2.10 Pupillary response to intelligibility during a concurrent and sustained memory load

Consistent with our second hypothesis, pupil diameter was larger in repeat-with-recall than repeat-only condition. In this respect, the present design has the advantage of dissecting how this difference arises, thanks to the trial-by-trial sensitivity of pupillometry. The difference arises from a progressive decrease in pupil diameter within the repeat-only condition, and a progressive increase in baseline diameter within the repeat-with-recall condition from the 1st to the 10th word. Although past studies have reported similar trends, they were using different materials and test designs, making it hard to demonstrate clearly the impact of additional memory task on listening effort in both magnitude and dynamics. For instance, within one speech perception task, pupil diameter gradually decreased with increasing trial numbers, due to task/stimuli habituation [17, 38–40]. However, when listeners needed to remember the digits or pseudo-words presented auditorily, pupil diameter increased progressively, until the memory span was exceeded [16, 24, 26, 71]. Note that in the current experiment, listeners needed to continuously decode words embedded in noise, which was more effortful than listening to digits or pseudo-words in quiet. The more demanding primary speech recognition task led to more accumulated and sustained effort over time. This might explain earlier plateau in baseline diameter in our experiment than observed in those studies. We observed a quadratic trend of baseline pupil diameter from the 1st to the 10th word within a list. [26] reported the plateau at the 9th digit for young adults, and [72] reported the plateau at 6th digits for children and 8th digit for adults. Our results are in good agreement with such estimates, and confirm that additional memory task places a heavier and sustained load on cognitive effort. More specifically, baseline diameter could reveal the impact on cognitive effort from the additional task, and the rate of increase in baseline diameter could be suggestive of the magnitude of sustained effort in a test paradigm with multiple sources of cognitive effort.

However, the steeper decrease of PPD in repeat-with-recall condition compared to repeat-only condition was unexpected. PPD has been shown to be sensitive to memory load, therefore, with more words to be remembered, we expected PPD to increase accordingly over time [16, 25, 26]. Decrease in PPD was reported when listeners tended to give up in the tasks that were impossibly difficult [27, 29]. In those cases, performance level was typically low (around 0%). But we did not observe a decrease in recognition and recall performance for words in the later part of the list in our results, or a worse word recognition performance in repeat-with-recall condition at difficult 0dB condition (in fact, word recognition was higher in repeat-with-recall than repeat-only condition). This suggests that listeners did not give up at the later part of the word list, or at 0dB. Similarly, a smaller PPD at 0dB in repeat-with-recall than repeat-only condition was surprising. Additional recall task with difficult SNR is certainly more demanding than a single task, therefore, we expected PPD to be larger in the repeat-with-recall condition and at difficult SNR level. But we observed the opposite: PPD actually decreased in the repeat-with-recall condition. We do not believe that these are spurious results. This huge contrast with the well-established effect of task demands on the pupillary response was also observed in Zekveld et al. [43]. In Zekveld et al. [43], participants had to recall the four-word cues (either related or unrelated to the following sentence) presented visually before the onset of the sentence embedded in speech masker. The 7dB SNR difference between two sentence-in-noise conditions (-17dB and -10dB) elicited a difference in intelligibility, but not in peak and mean pupil dilation. Zekveld et al. [43] interpreted the absence of pupillary difference between two SNRs as participants prioritising the central factors (memory task) than peripheral factors (sentence recognition task). There are a few characteristics that distinguish our design from Zekveld et al. [43]. Firstly, the memory and sentence recognition tasks in Zekveld et al. [43] were more independent: participants read the cue words for 5s before the auditory stimulus onset; after the auditory stimulus offset, participants either repeated the sentence or the cue words. This separation between two tasks could facilitate intentional prioritisation of the memory over the speech recognition task. Secondly, participants in Zekveld et al. [43] only needed to memorise a four-word cue at the start of each trial, with no accumulation of memory load over time. In comparison, the memory task in our paradigm was more imposing on the limited cognitive resources: participants had to complete both word recognition and memorising tasks within the same time window, and they needed to keep retaining more words in the memory from the 1st to the 10th word. Therefore, it is not surprising that we observed not only a lack of correlation between task demands and pupillary response at easier SNR levels, but also a reversal of that relation at the most cognitively demanding condition (0dB and repeat-with-recall).

One explanation for the steep decrease of PPD in sustained listening condition could be due to fatigue. In a similar sustained listening condition, McGarrigle et al. [42] asked NH participants to listen to two short passages of text with multi-talker babble noise at either -8 dB and 15 dB, and at the end of each passage judge whether images presented on the screen were mentioned in the previous passage. A steeper decrease in (normalised and baseline corrected) pupil size during listening was found for difficult SNR than easy SNR, but only in the second half of the trial block. This was interpreted as fatigue kicking in at the second section of the test. It is likely that in our study, the steeper decrease of PPD in repeat-with-recall condition could also be the sign of overload and fatigue with continuing effort to recognise, encode and rehearse isolated words. However, the decreasing trend reported in McGarrigle et al. [42] was not found in McGarrigle et al. [73] when using a similar test for school-aged children, so it is still unclear how reliably and accurately this metric is related to fatigue.

Yet another possible explanation to the steeper decrease of PPD in repeat-with-recall condition is that the dynamic range of pupillary could be constrained by baseline diameter. Critically, for the first word in the list, PPD was bigger in repeat-with-recall than repeat-only condition but the baseline diameter was similar. As the baseline diameter grew bigger and plateaued in repeat-with-recall condition, PPD did not have much space to grow, so it decreased faster than repeat-only condition.

Similarly, at repeat-with-recall condition, baseline diameter was already bigger than repeat-only condition for all SNR levels to start with, leaving little room for PPD to increase further during the task. It looks as if under sustained listening condition, there is a limit on the magnitude of pupil dilation, beyond which no further increase is possible. This interpretation is tempting in its logic. However, this limit must not be imposed by physiological constraint of the iris muscles, because at the onset of the recall, pupil diameter increased dramatically, on average by 0.3mm or equivalent to an effect six times bigger than the average PPD at the 10th word (also seen in Cabestrero et al. [26] and discussed in Zekveld et al. [43]). Instead, this limit might be of a cognitive origin. Puma et al. [74] reported a similar ceiling in EEG alpha and theta band power when participants were overloaded with multiple concurrent tasks. This limit might be associated with the saturation in cognitive resources allocation. In order to ensure successful retrieval of words from long- and short-term memory storage at the recall stage, some cognitive resources should be preserved and held until the later part of the test. Therefore, as memory load accumulated (increase in baseline diameter) and approached the limit allocated for the recognition and encoding stage, fewer new resources would be assigned (decrease in PPD), so that enough resources were reserved for the recall stage. The reserved cognitive resources were finally put to use at the onset of recall, leading to a big ‘jump ’ in pupil diameter. This could be a phenomenal illustration of how cognitive resources are managed in a highly flexible and goal-directed manner. In Cabestrero et al. [26], the biggest ‘jump ’ at the onset of recall was when 5 digits were to be recalled (low load), and the smallest ‘jump ’ was when 11 digits were to be recalled (overload), suggesting that this sharp increase in pupil diameter is proportionate to the cognitive resources left for the recall task. Arguably, how cognitive resources are allocated to different tasks could also depend on individual cognitive capacity and cognitive abilities. Listeners with bigger cognitive capacity and better abilities to process speech in noise, might allocate fewer resources (lower limit) to word recognition and encoding, because they will be more efficient in completing the task [75,76]. Therefore, to fully test this hypothesis, future studies need to include more individual cognitive ability measurements and different types of manipulations on cognitive load.

2.11 Pupillary response to word recognition and memory

Baseline pupil diameter held a lot of predictive power in showing the accumulation of memory load from one serial position to the next. On an individual level, baseline diameter was also responsive for recall performance, as shown by their significant correlation.

Bigger PPD and more delayed dilation for incorrectly than correctly repeated words in repeat-only condition is also observed in other studies using sentence stimuli [17,20,27]. But in the condition requiring heavy and sustained effort (repeat-with-recall), PPD saturated too quickly, especially later in the word list, to support the correlation with word recognition. It seemed that the dynamic range of pupillary response was constrained by the baseline diameter. This further highlights the issue aforementioned, namely that the saturation in pupillary response under sustained load might make PPD problematic for quantifying the actual effort.

Nevertheless, PPD remains a reliable index of cognitive effort and explanatory factor of some behavioural performance. Typically, when comparing the recall performance, we found words that were successfully recalled had bigger pupillary response than those forgotten. Papesh et al. [77] suggested a similar relation between PPD and memory encoding success. Participants first listened to 80 words and nonwords spoken by two speakers; then during the test session, they listened to 160 items and judged, along a 6-point scale, how confident they were that the words were old/new. Words that were remembered with higher degree of confidence showed bigger PPD, relative to words that were remembered with less confidence or forgotten.

Taken as a whole, these results picture a complex story of the allocation and dynamics of cognitive resources during speech perception and memory task. Failure to recognise the word is associated with more effortful processing, possibly because more lexical competitors are activated for explicit decision when listeners fail to decode the acoustic signals without ambiguity. This might also initiate retroactive corrective processing that would keep the effort elevated post-stimulus [21]. When words need to be remembered for the recall task, the memory encoding probably becomes a priority after completing the word recognition. If more cognitive resources are expended at this stage to encode the word in the working memory storage, there is a higher chance that it will be retrieved successfully later.

2.12 Individual differences

Behavioural performance was correlated with pupillary response, but in different manners: better word recognition performance was related with smaller PPD; better stated word recall performance was related with bigger baseline diameter; bigger baseline diameter was related with easier subjective rating; better word recall performance was related with easier subjective rating. Consistent with the results discussed above, these suggest that different metrics of pupillary responses might relate to different cognitive processing. PPD was an indicator of transient effort expended for decoding the words presented in noise, hence correlated with the word recognition performance. Listeners' subjective feeling is affected both by external task demands (SNR levels and TASK), and one's evaluation of recall success. Note that all three measures (pupillary response, word recall performance and subjective rating) also significantly correlated with age, making it possible that the correlations observed were due to a latent variable, for instance individual cognitive capacity [23, 27, 44, 53, 78, 79].

To summarise, while behavioural performance (i.e., recall) and subjective rating indicate the final outcome of a series of cognitive processes, pupillometry can reveal the difference in listening effort between conditions, the temporal dynamics of different stages of cognitive processing, as well as the allocation policy of cognitive resources. However, only a handful of studies have looked into the dynamics of pupillary response in realistic conditions, where listening is not the only task demanding cognitive resources. The current experiment is a good example showing the importance of looking at pupillary metrics (time-series variations, baseline diameter) other than PPD when investigating listening effort under sustained memory or other cognitive loads. PPD might be constrained by the baseline diameter induced by concurrent tasks, making it less related to actual listening effort. Accordingly, new pupillary metrics and analysis pipeline should be developed to quantify the dynamic aspect of listening effort.

2.13 Limitation

Pupil recordings during word repeat and recall were inevitably contaminated by movements during speech production and involuntary eye movement. No algorithm has been developed yet to reliably adjust pupil diameter for these factors. Special care was taken during the experiment and data preprocessing: participants were instructed to keep fixating at the fixation circle during verbal responses; we extrapolated points in the pupil traces where the centre of gazing was beyond 3SD from the centre and excluded trials where over 20% of the traces were either blinks or erratic gazing. Although this lead to loss of data, we ensured that the data left for analysis was valid.

Nevertheless, speech production following the response cue could potentially interfere with the pupillary response corresponding to memory encoding. Individual differences in the timing of responding could also interfere with the correspondence between memory encoding and pupillary response. However, this artefact was present for every word because participants needed to repeat words in all conditions. Therefore, the difference in pupil trace observed within this time window could not be entirely due to production confounds.

2.14 Conclusion

As one of the first few studies to investigate pupillary responses under sustained and complex listening condition, the present study serves as a bridge between established listening effort research and future direction of understanding and quantifying listening effort in real-life communication in various populations. The concurrent recall task did not allow listeners to process just one item, shake off the load once finished and start afresh for the next item. Instead, they needed to be constantly attentive and allocating cognitive resources to process new items while holding other information in (working) memory. This is similar to a real-life communication scenario where multiple tasks compete for a limited pool of cognitive resources over a period of time. Results suggest that both the magnitude and temporal pattern of pupillary response differ greatly in sustained listening condition from those in a single task. Accordingly, parameters of pupillary responses used for indexing listening effort need to be reviewed in the light of the more ecological listening conditions.

Although real-life speech communication is even more complex and dynamic, the present study serves as a good starting point by choosing a paradigm that could provide enough approximation to cognitive processing in speech communication, yet sufficient time locking to a given type of cognitive processing to ensure the interpretability of the results. A better understanding of listening effort in ecological environments is also important for developing clinical measurement, especially for CI users and HI listeners. It is possible that prior motivational, emotional, cognitive factors and social pressure could disturb the relation between pupillary response and listening effort that is well-established in research settings.

Supporting information

S1 File. Alternative method to calculate PPD Results and discussions on the alternative method to perform baseline correction using the averaged pupil trace 1s before the first word in the list.

S1 Table. Model summary outputs. Model parameter estimates and model comparison statistics for the best fitting models. The reference level for the categorical factor LISTENING is 0dB, for the factor TASK is repeat-only.

Acknowledgement

This research was supported by MITACS and Oticon Medical. We also thank Florian Malaval and Arthur Delage for assistance with running the experiment with native French-speaking participants.

3 References

References

1. Rönnberg J, Rudner M, Foo C, Lunner T. Cognition counts: A working memory system for ease of language understanding (ELU). International Journal of Audiology. 2008;47(sup2):S99–S105.

2. Mattys SL, Davis MH, Bradlow AR, Scott SK. Speech recognition in adverse conditions: A review. Language and Cognitive Processes. 2012;27(7-8):953–978.

3. Pichora-Fuller MK, Kramer SE, Eckert MA, Edwards B, Hornsby BW, Humes LE, et al. Hearing impairment and cognitive energy: The framework for understanding effortful listening (FUEL). Ear and Hearing. 2016;37:5S–27S.

4. Kahneman D. Attention and effort. vol. 1063. Citeseer; 1973.

5. Rudner M. Cognitive spare capacity as an index of listening effort. Ear and hearing. 2016;37:69S–76S.

6. Kramer SE, Kapteyn TS, Festen JM, Kuik DJ. Assessing aspects of auditory handicap by means of pupil dilatation. Audiology. 1997;36(3):155–164.

7. McCoy SL, Tun PA, Cox LC, Colangelo M, Stewart RA, Wingfield A. Hearing loss and perceptual effort: Downstream effects on older adults memory for speech. The Quarterly Journal of Experimental Psychology Section A. 2005;58(1):22–33.

8. Gosselin PA, Gagné JP. Use of a Dual-Task Paradigm to Measure Listening Effort Utilisation dun paradigme de double tâche pour mesurer lattention auditive. Inscription au Répertoire. 2010;34(1):43.

9. Rönnberg J, Lunner T, Zekveld A, Sörqvist P, Danielsson H, Lyxell B, et al. The Ease of Language Understanding (ELU) Model: theoretical, empirical, and clinical advances. Frontiers in systems neuroscience. 2013;7:31.

10. McGarrigle R, Munro KJ, Dawes P, Stewart AJ, Moore DR, Barry JG, et al. Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group white paper. International journal of audiology. 2014;53(7):433–445.

11. Nachtegaal J, Kuik DJ, Anema JR, Goverts ST, Festen JM, Kramer SE. Hearing status, need for recovery after work, and psychosocial work characteristics:

Results from an internet-based national survey on hearing. International journal of audiology. 2009;48(10):684–691.

12. Grimby A, Ringdahl A. Does having a job improve the quality of life among post-lingually deafened Swedish adults with severe-profound hearing impairment? British Journal of Audiology. 2000;34(3):187–195.

13. Kramer SE, Kapteyn TS, Houtgast T. Occupational performance: Comparing normally-hearing and hearing-impaired employees using the Amsterdam Checklist for Hearing and Work: Desempeño laboral: Comparación de empleados con audición normal o alterada usando el Listado Amsterdam para Audición y Trabajo. International journal of audiology. 2006;45(9):503–512.

14. Ohlenforst B, Zekveld AA, Jansma EP, Wang Y, Naylor G, Lorens A, et al. Effects of hearing impairment and hearing aid amplification on listening effort: A systematic review. Ear and hearing. 2017a;38(3):267.

15. Zekveld AA, Koelewijn T, Kramer SE. The pupil dilation response to auditory stimuli: Current state of knowledge. Trends in hearing. 2018;22:2331216518777174.

16. Karatekin C, Couperus JW, Marcus DJ. Attention allocation in the dual-task paradigm as measured through behavioral and psychophysiological responses. Psychophysiology. 2004;41(2):175–185.

17. Zekveld AA, Kramer SE, Festen JM. Pupil response as an indication of effortful listening: The influence of sentence intelligibility. Ear and hearing. 2010;31(4):480–490.

18. Koelewijn T, Zekveld AA, Festen JM, Rönnberg J, Kramer SE. Processing load induced by informational masking is related to linguistic abilities. International journal of otolaryngology. 2012;2012.

19. Koelewijn T, de Kluiver H, Shinn-Cunningham BG, Zekveld AA, Kramer SE. The pupil response reveals increased listening effort when it is difficult to focus attention. Hearing research. 2015;323:81–90.

20. Winn MB, Edwards JR, Litovsky RY. The impact of auditory spectral resolution on listening effort revealed by pupil dilation. Ear and hearing. 2015;36(4):e153.

21. Winn MB. Rapid release from listening effort resulting from semantic context, and effects of spectral degradation and cochlear implants. Trends in Hearing. 2016;20:2331216516669723.

22. McMurray B, Farris-Trimble A, Rigler H. Waiting for lexical access: Cochlear implants or severely degraded input lead listeners to process speech less incrementally. Cognition. 2017;169:147–164.

23. Wendt D, Hietkamp RK, Lunner T. Impact of noise and noise reduction on processing effort: A pupillometry study. Ear and hearing. 2017;38(6):690–700.

24. Peavler WS. Pupil size, information overload, and performance differences. Psychophysiology. 1974;11(5):559–566.

25. Granholm E, Asarnow RF, Sarkin AJ, Dykes KL. Pupillary responses index cognitive resource limitations. Psychophysiology. 1996;33(4):457–461.

26. Cabestrero R, Crespo A, Quirós P. Pupillary dilation as an index of task demands. Perceptual and motor skills. 2009;109(3):664–678.

27. Zekveld AA, Kramer SE. Cognitive processing load across a wide range of listening conditions: Insights from pupillometry. Psychophysiology. 2014;51(3):277–284.

28. Kramer SE, Teunissen CE, Zekveld AA. Cortisol, chromogranin A, and pupillary responses evoked by speech recognition tasks in normally hearing and hard-of-hearing listeners: a pilot study. Ear and hearing. 2016;37:126S–135S.

29. Ohlenforst B, Zekveld AA, Lunner T, Wendt D, Naylor G, Wang Y, et al. Impact of stimulus-related factors and hearing impairment on listening effort as indicated by pupil dilation. Hearing Research. 2017b;351:68–79.

30. Ohlenforst B, Wendt D, Kramer SE, Naylor G, Zekveld AA, Lunner T. Impact of SNR, masker type and noise reduction processing on sentence recognition performance and listening effort as indicated by the pupil dilation response. Hearing research. 2018;365:90–99.

31. Aston-Jones G, Cohen JD. An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. Annu Rev Neurosci. 2005;28:403–450.

32. Murphy PR, O'connell RG, O'sullivan M, Robertson IH, Balsters JH. Pupil diameter covaries with BOLD activity in human locus coeruleus. Human brain mapping. 2014;35(8):4140–4154.

33. Koelewijn T, Zekveld AA, Lunner T, Kramer SE. The effect of reward on listening effort as reflected by the pupil dilation response. Hearing research. 2018;367:106–112.

34. Wang Y, Naylor G, Kramer SE, Zekveld AA, Wendt D, Ohlenforst B, et al. Relations between self-reported daily-life fatigue, hearing status, and pupil dilation during a speech perception in noise task. Ear and Hearing. 2018;39(3):573–582.

35. Hockey R. The psychology of fatigue: Work, effort and control. Cambridge University Press; 2013.

36. Verney SP, Granholm E, Marshall SP. Pupillary responses on the visual backward masking task reflect general cognitive ability. International Journal of Psychophysiology. 2004;52(1):23–36.

37. Winn MB, Wendt D, Koelewijn T, Kuchinsky SE. Best practices and advice for using pupillometry to measure listening effort: An introduction for those who want to get started. Trends in hearing. 2018;22:2331216518800869.

38. Beatty J. Task-evoked pupillary responses, processing load, and the structure of processing resources. Psychological bulletin. 1982;91(2):276.

39. Damsma A, van Rijn H. Pupillary response indexes the metrical hierarchy of unattended rhythmic violations. Brain and cognition. 2017;111:95–103.

40. Marois A, Labonté K, Parent M, Vachon F. Eyes have ears: Indexing the orienting response to sound using pupillometry. International Journal of Psychophysiology. 2018;123:152–162.

41. Gilzenrat MS, Nieuwenhuis S, Jepma M, Cohen JD. Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. Cognitive, Affective, & Behavioral Neuroscience. 2010;10(2):252–269.

42. McGarrigle R, Dawes P, Stewart AJ, Kuchinsky SE, Munro KJ. Pupillometry reveals changes in physiological arousal during a sustained listening task. Psychophysiology. 2017a;54(2):193–203.

43. Zekveld AA, Kramer SE, Rönnberg J, Rudner M. In a concurrent memory and auditory perception task, the pupil dilation response is more sensitive to memory load than to auditory stimulus characteristics. Ear and hearing. 2019;40(2):272.

44. Lunner T, Rudner M, Rosenbom T, Ågren J, Ng EHN. Using speech recall in hearing aid fitting and outcome evaluation under ecological test conditions. Ear and hearing. 2016;37:145S–154S.

45. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software. 2015;67(1):1–48. doi:10.18637/jss.v067.i01.

46. R Core Team. R: A Language and Environment for Statistical Computing; 2019. Available from: https://www.R-project.org/.

47. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York; 2016. Available from: https://ggplot2.tidyverse.org.

48. Mirman D. Growth curve analysis and visualization using R. CRC Press Boca Raton, FL; 2014.

49. Bristow D, Frith C, Rees G. Two distinct neural effects of blinking on human visual processing. Neuroimage. 2005;27(1):136–145.

50. Klingner J, Kumar R, Hanrahan P. Measuring the task-evoked pupillary response with a remote eye tracker. In: Proceedings of the 2008 symposium on Eye tracking research & applications. ACM; 2008. p. 69–72.

51. Sarampalis A, Kalluri S, Edwards B, Hafter E. Objective measures of listening effort: Effects of background noise and noise reduction. Journal of Speech, Language, and Hearing Research. 2009;.

52. Ephraim Y, Malah D. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Transactions on acoustics, speech, and signal processing. 1984;32(6):1109–1121.

53. Ng EHN, Rudner M, Lunner T, Pedersen MS, Rönnberg J. Effects of noise and working memory capacity on memory processing of speech for hearing-aid users. International Journal of Audiology. 2013;52(7):433–441.

54. Downs DW. Effects of hearing aid use on speech discrimination and listening effort. Journal of Speech and Hearing Disorders. 1982;47(2):189–193.

55. Wingfield A. Evolution of models of working memory and cognitive resources. Ear and hearing. 2016;37:35S–43S.

56. Edwards B. A model of auditory-cognitive processing and relevance to clinical applicability. Ear and hearing. 2016;37:85S–91S.

57. Li KZ, Lindenberger U, Freund AM, Baltes PB. Walking while memorizing: Age-related differences in compensatory behavior. Psychological science. 2001;12(3):230–237.

58. Madden DJ, Langley LK. Age-related changes in selective attention and perceptual load during visual search. Psychology and aging. 2003;18(1):54.

59. Hein G, Schubert T. Aging and input processing in dual-task situations. Psychology and Aging. 2004;19(3):416.

60. Plummer P, Eskes G. Measuring treatment effects on dual-task performance: a framework for research and clinical practice. Frontiers in human neuroscience. 2015;9:225.

61. Gagne JP, Besser J, Lemke U. Behavioral assessment of listening effort using a dual-task paradigm: A review. Trends in hearing. 2017;21:2331216516687287.

62. Brehm JW, Self EA. The intensity of motivation. Annual review of psychology. 1989;40(1):109–131.

63. Eckert MA, Teubner-Rhodes S, Vaden Jr KI. Is listening in noise worth it? The neurobiology of speech recognition in challenging listening conditions. Ear and hearing. 2016;37(Suppl 1):101S.

64. Hornsby BW, Naylor G, Bess FH. A taxonomy of fatigue concepts and their relation to hearing loss. Ear and hearing. 2016;37(Suppl 1):136S.

65. Matthen M. Effort and displeasure in people who are hard of hearing. Ear and hearing. 2016;37:28S–34S.

66. Richter M. The moderating effect of success importance on the relationship between listening demand and listening effort. Ear and Hearing. 2016;37:111S–117S.

67. Peelle JE. Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior. Ear and Hearing. 2018;39(2):204.

68. Paas F, Renkl A, Sweller J. Cognitive load theory and instructional design: Recent developments. Educational psychologist. 2003;38(1):1–4.

69. Choi S, Lotto A, Lewis D, Hoover B, Stelmachowicz P. Attentional modulation of word recognition by children in a dual-task paradigm. Journal of Speech, Language, and Hearing Research. 2008;.

70. McFadden B, Pittman A. Effect of minimal hearing loss on childrens ability to multitask in quiet and in noise. Language, speech, and hearing services in schools. 2008;.

71. López-Ornat S, Karousou A, Gallego C, Martín L, Camero R. Pupillary measures of the cognitive effort in auditory novel word processing and short-term retention. Frontiers in psychology. 2018;9.

72. Johnson EL, Miller Singley AT, Peckham AD, Johnson SL, Bunge SA. Task-evoked pupillometry provides a window into the development of short-term memory capacity. Frontiers in psychology. 2014;5:218.

73. McGarrigle R, Dawes P, Stewart AJ, Kuchinsky SE, Munro KJ. Measuring listening-related effort and fatigue in school-aged children using pupillometry. Journal of experimental child psychology. 2017b;161:95–112.

74. Puma S, Matton N, Paubel PV, Raufaste É, El-Yagoubi R. Using theta and alpha band power to assess cognitive workload in multitasking environments. International Journal of Psychophysiology. 2018;123:111–120.

75. Unsworth N, Engle RW. The nature of individual differences in working memory capacity: active maintenance in primary memory and controlled search from secondary memory. Psychological review. 2007;114(1):104.

76. Ng EHN, Rudner M, Lunner T, Rönnberg J. Noise reduction improves memory for target language speech in competing native but not foreign language speech. Ear and Hearing. 2015;36(1):82–91.

77. Papesh MH, Goldinger SD, Hout MC. Memory strength and specificity revealed by pupillometry. International Journal of Psychophysiology. 2012;83(1):56–64.

78. Kuchinsky SE, Vaden Jr KI, Ahlstrom JB, Cute SL, Humes LE, Dubno JR, et al. Task-related vigilance during word recognition in noise for older adults with hearing loss. Experimental aging research. 2016;42(1):50–66.

79. Tsukahara JS, Harrison TL, Engle RW. The relationship between baseline pupil size and intelligence. Cognitive psychology. 2016;91:109–123.

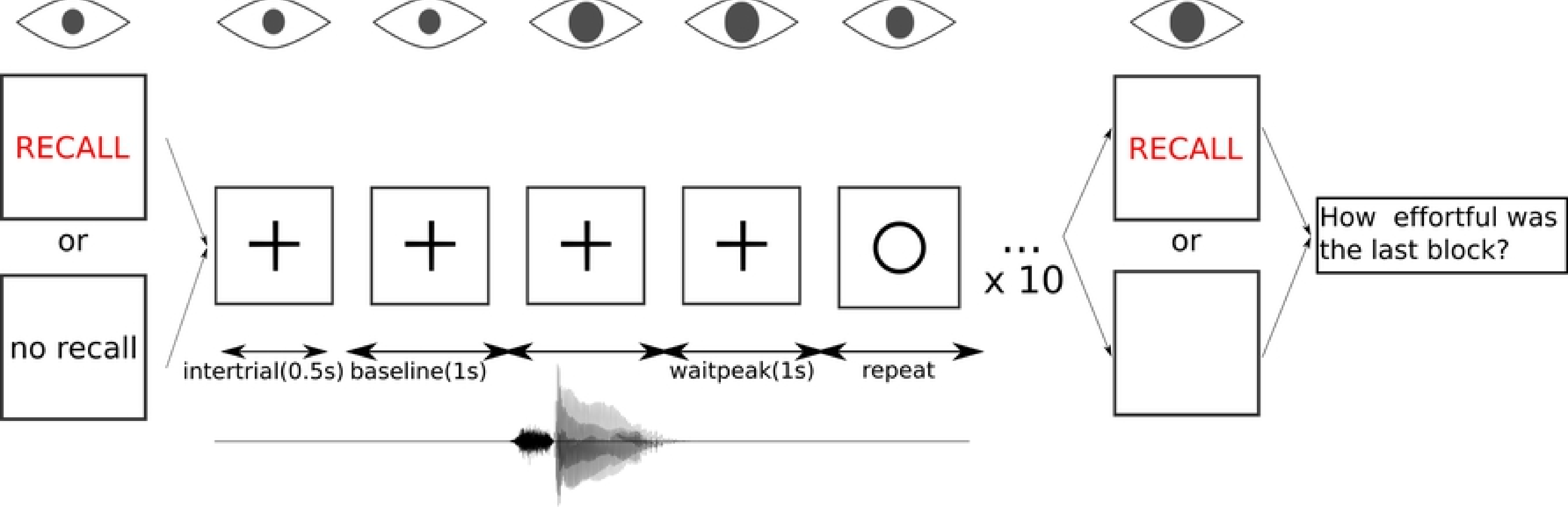
RECALL
or
no recall
intertrial(0.5s)
baseline(1s)
waitpeak(1s)
repeat
...
x 10
RECALL
or
How effortful was the last block?

fig1

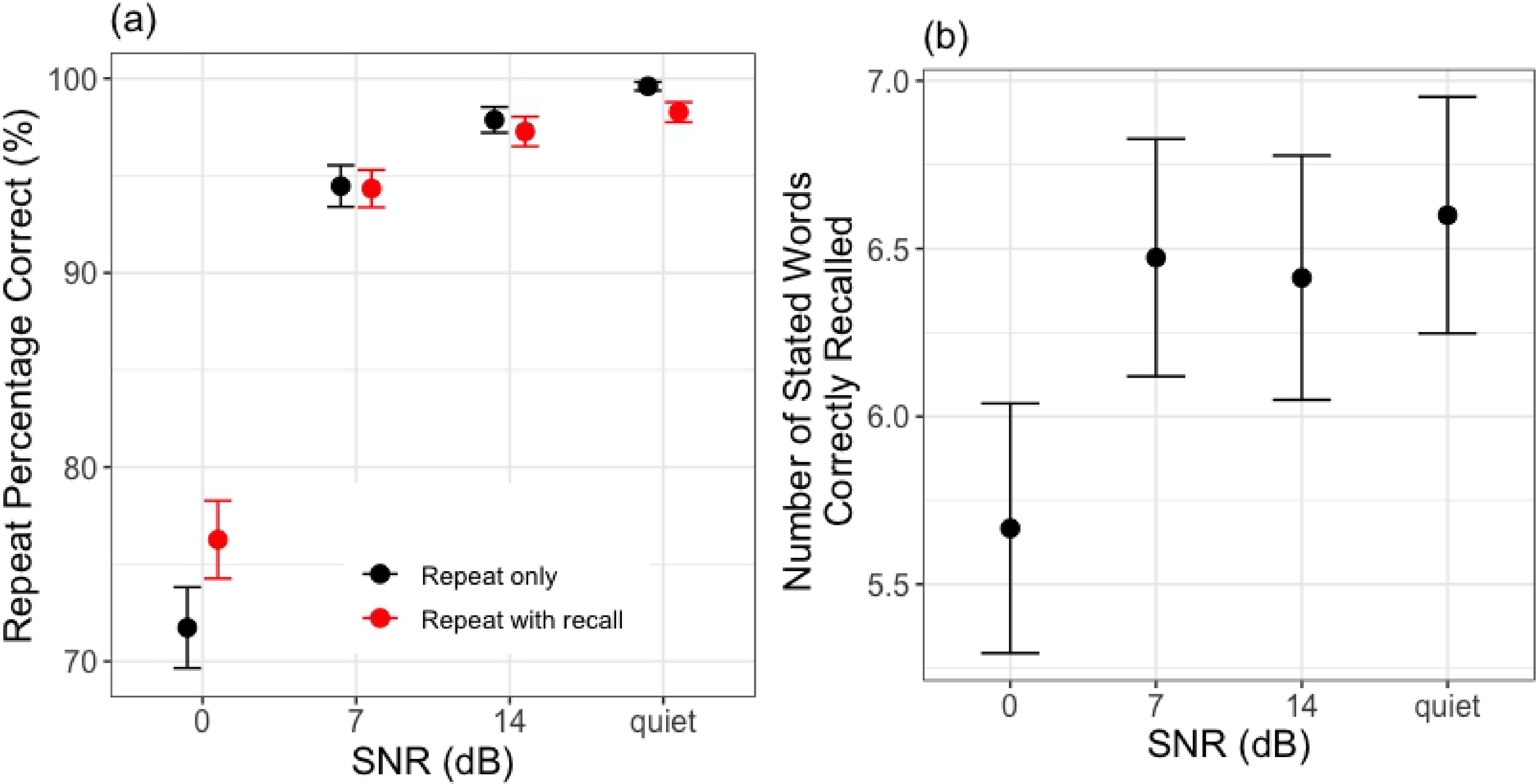
(a)
Repeat Percentage Correct (%)
100
90
80
70
0
7
14
quiet
SNR (dB)
Repeat only
Repeat with recall
(b)
Number of Stated Words
Correctly Recalled
7.0
6.5
6.0
5.5
0
7
14
quiet
SNR (dB)

fig2

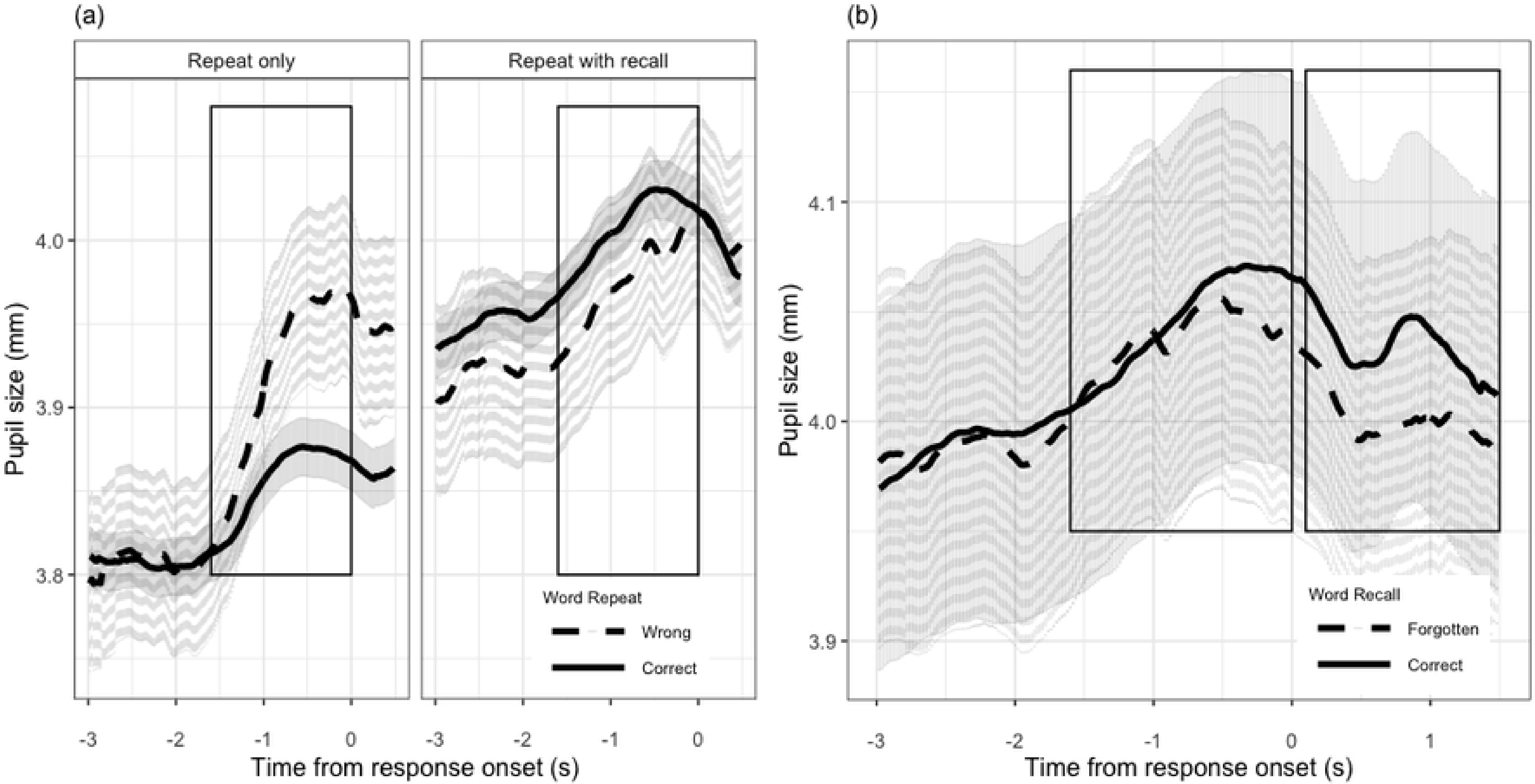
(a)
Repeat only
Repeat with recall
Pupil size (mm)
4.0
3.9
3.8
-3
-2
-1
0
Time from response onset (s)
Word Repeat
Wrong
Correct
(b)
Pupil size (mm)
4.1
4.0
3.9
-3
-2
-1
0
1
Time from response onset (s)
Word Recall
Forgotten
Correct

fig5

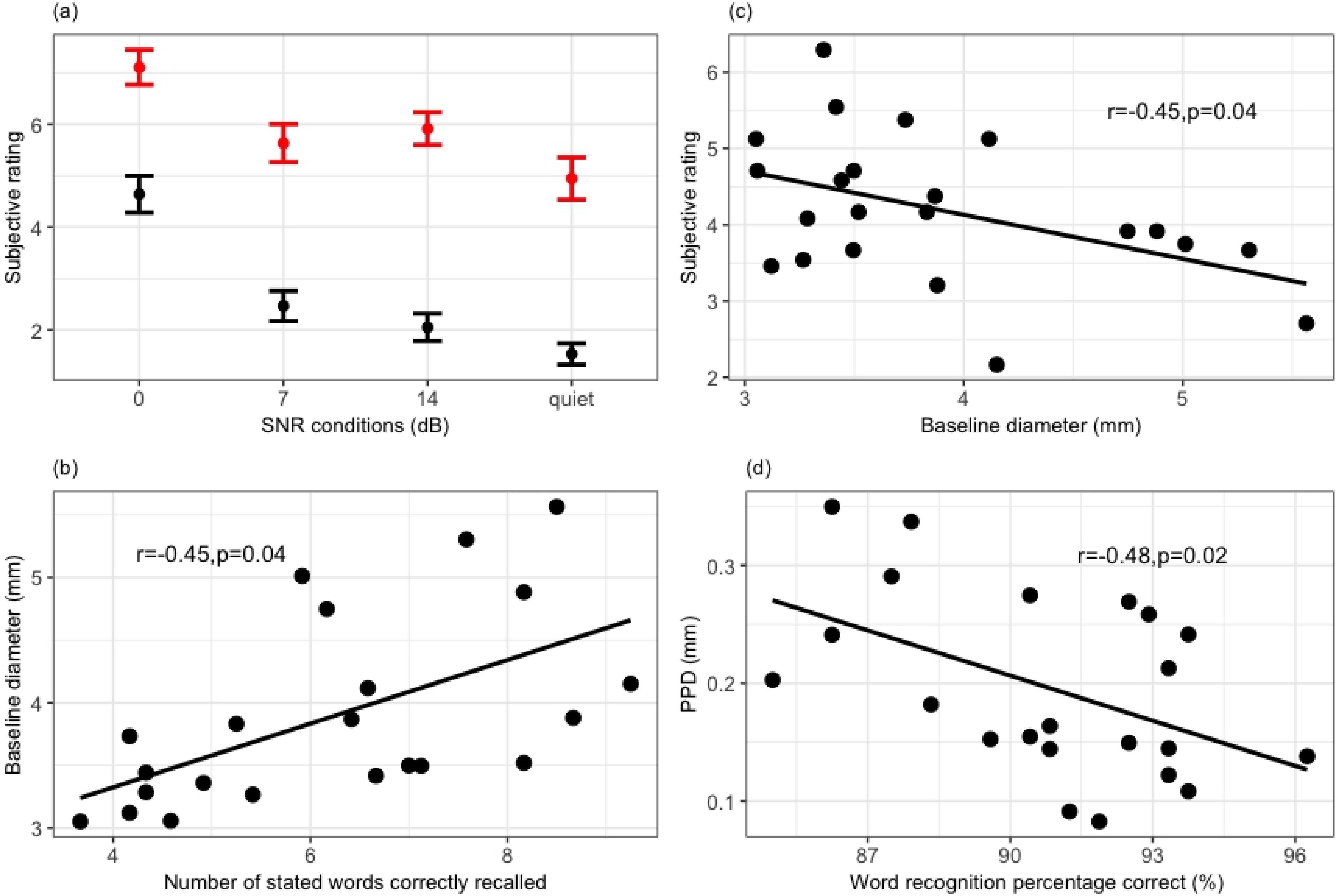
(a)
Subjective rating
2
4
6
0
7
14
quiet
SNR conditions (dB)
(b)
Baseline diameter (mm)
3
4
5
r=-0.45,p=0.04
4
6
8
Number of stated words correctly recalled
(c)
Subjective rating
2
3
4
5
6
r=-0.45,p=0.04
3
4
5
Baseline diameter (mm)
(d)
PPD (mm)
0.1
0.2
0.3
r=-0.48,p=0.02
87
90
93
96
Word recognition percentage correct (%)

fig7

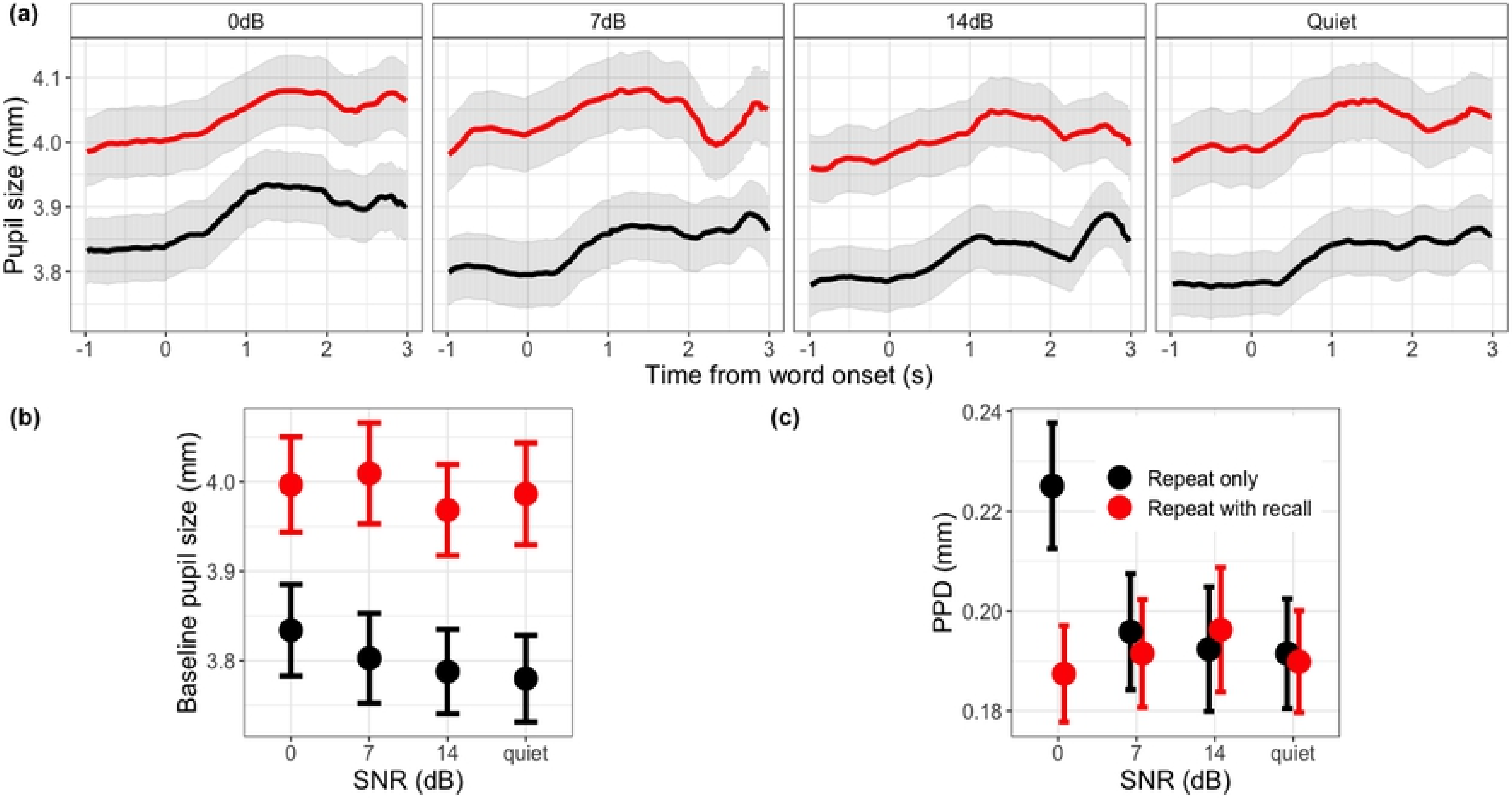
(a)
0dB
7dB
14dB
Quiet
Pupil size (mm)
4.1
4.0
3.9
3.8
-1
0
1
2
3
Time from word onset (s)
(b)
Baseline pupil size (mm)
4.0
3.9
3.8
0
7
14
quiet
SNR (dB)
(c)
PPD (mm)
0.24
0.22
0.20
0.18
● Repeat only
● Repeat with recall
0
7
14
quiet
SNR (dB)

fig3

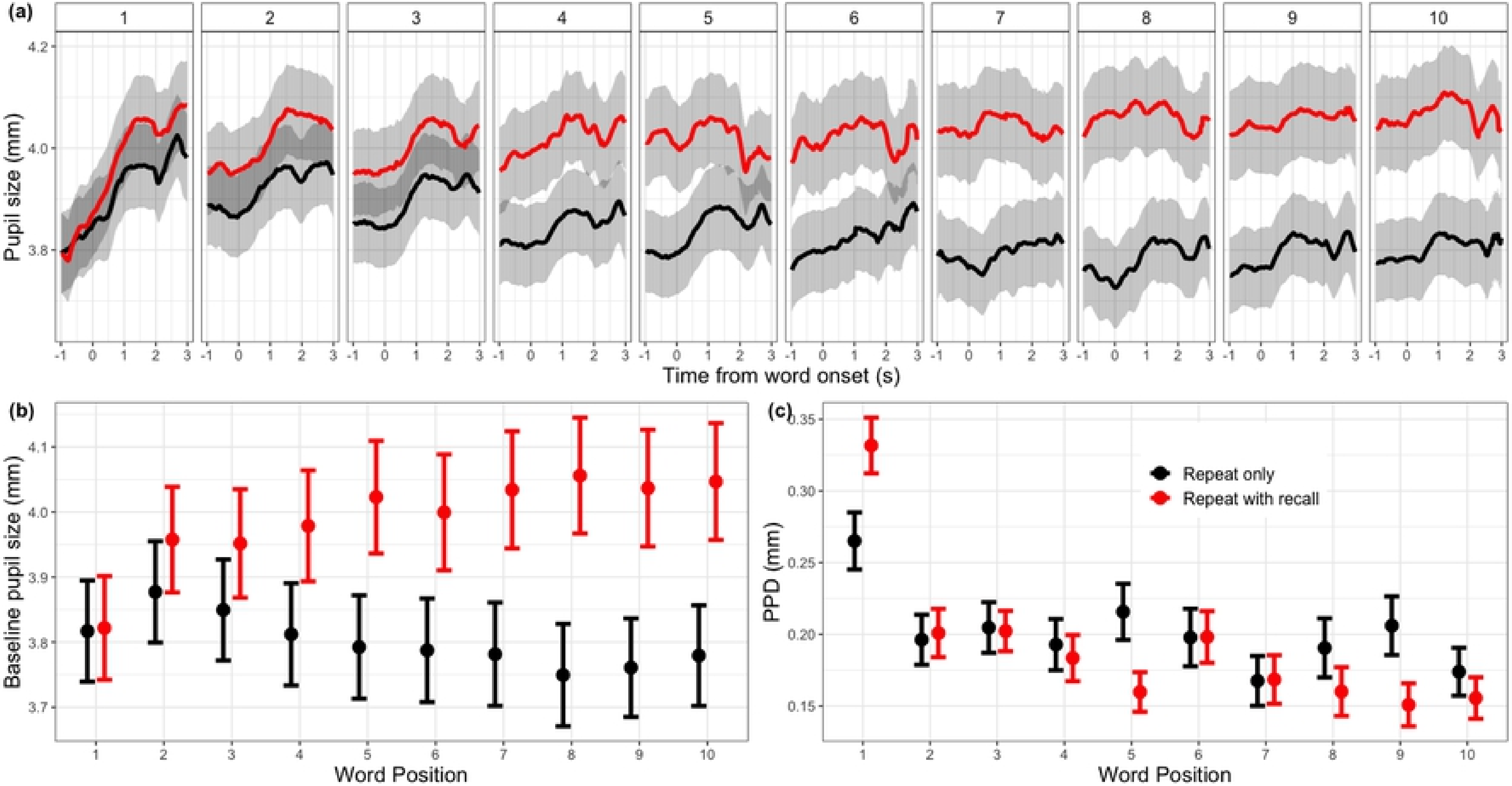
(a)
1
2
3
4
5
6
7
8
9
10
Pupil size (mm)
4.2
4.0
3.8
-1 0 1 2 3
Time from word onset (s)
(b)
Baseline pupil size (mm)
4.1
4.0
3.9
3.8
3.7
1 2 3 4 5 6 7 8 9 10
Word Position
(c)
PPD (mm)
0.35
0.30
0.25
0.20
0.15
● Repeat only
● Repeat with recall
1 2 3 4 5 6 7 8 9 10
Word Position

fig4

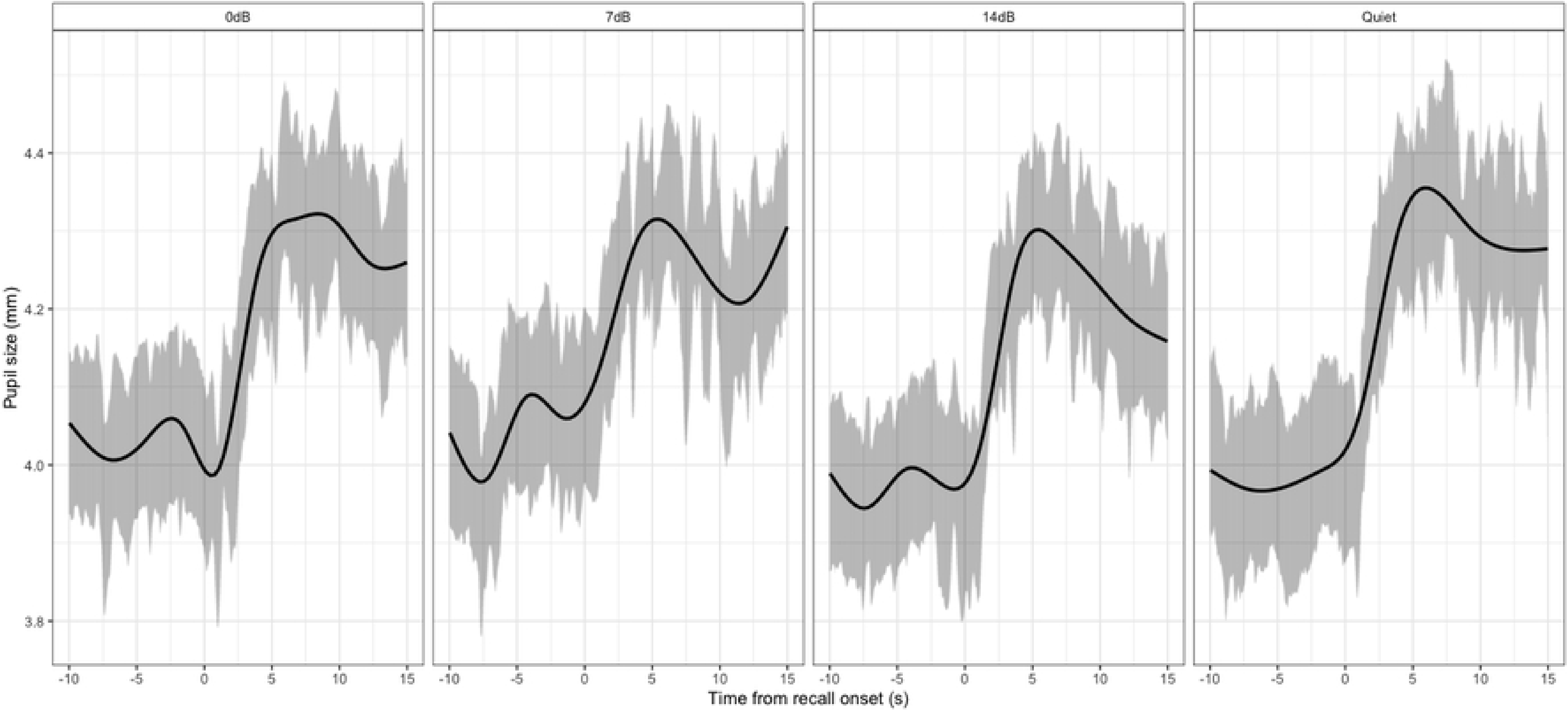
0dB
7dB
14dB
Quiet
Pupil size (mm)
4.4
4.2
4.0
3.8
-10
-5
0
5
10
15
Time from recall onset (s)

fig6