Received: 10 March 2023 | Revised: 17 October 2023 | Accepted: 23 October 2023

DOI: 10.1111/ejn.16187

RESEARCH REPORT

EJN European Journal of Neuroscience FENS

WILEY

Validity and reliability of self-reported and neural measures of listening effort

Yousef Mohammadi[1] | Jan Østergaard[2] | Carina Graversen[1,3] | Ole Kæseler Andersen[1,3] | José Biurrun Manresa[3,4]

[1]Integrative Neuroscience, Department of Health Science and Technology, Aalborg University, Aalborg, Denmark

[2]Department of Electronic Systems, Aalborg University, Aalborg, Denmark

[3]Center for Neuroplasticity and Pain (CNAP), Department of Health Science and Technology, Aalborg University, Aalborg, Denmark

[4]Institute for Research and Development in Bioengineering and Bioinformatics (IBB), CONICET-UNER, Oro Verde, Argentina

Correspondence
José Biurrun Manresa, Center for Neuroplasticity and Pain (CNAP), Department of Health Science and Technology, Aalborg University, Aalborg, Denmark.
Email: jbiurrun@hst.aau.dk

Funding information
Danmarks Frie Forskningsfond; Independent Research Fund Denmark, Technology and Production Sciences (FTP), Grant/Award Number: DNRF121

Edited by: Ali Mazaheri

Abstract

Listening effort can be defined as a measure of cognitive resources used by listeners to perform a listening task. Various methods have been proposed to measure this effort, yet their reliability remains unestablished, a crucial step before their application in research or clinical settings. This study encompassed 32 participants undertaking speech-in-noise tasks across two sessions, approximately a week apart. They listened to sentences and word lists at varying signal-to-noise ratios (SNRs) (−9, −6, −3 and 0 dB), then retaining them for roughly 3 s. We evaluated the test–retest reliability of self-reported effort ratings, theta (4–7 Hz) and alpha (8–13 Hz) oscillatory power, suggested previously as neural markers of listening effort. Additionally, we examined the reliability of correct word percentages. Both relative and absolute reliability were assessed using intraclass correlation coefficients (ICC) and Bland–Altman analysis. We also computed the standard error of measurement (SEM) and smallest detectable change (SDC). Our findings indicated heightened frontal midline theta power for word lists compared to sentences during the retention phase under high SNRs (0 dB, −3 dB), likely indicating a greater memory load for word lists. We observed SNR's impact on alpha power in the right central region during the listening phase and frontal theta power during the retention phase in sentences. Overall, the reliability analysis demonstrated satisfactory between-session variability for correct words and effort ratings. However, neural measures (frontal midline theta power and right central alpha power) displayed substantial variability, even though group-level outcomes appeared consistent across sessions.

KEYWORDS

EEG, frontal midline theta power, hearing impairment, repeatability, speech-in-noise perception

Abbreviations: ANOVA, analysis of variance; EEG, electroencephalography; ICC, intraclass correlation coefficient; LMM, linear mixed model; LoA, limits of agreement; SD, standard deviation; SDC, smallest detectable change; SEM, standard error of measurement; SNR, signal-to-noise ratio.

1 | INTRODUCTION

Comprehending speech in adverse situations is challenging and requires effort from the listener. Listening effort has been defined as the allocation of cognitive resources to overcome auditory challenges (Peelle, 2018; Pichora-Fuller et al., 2016). When speech is degraded, a greater use of cognitive resources is required, including selective attention to segregate and identify target speech from background interference and working memory to compensate for the reduction of target speech information (Edwards, 2016; Pichora-Fuller et al., 2016; Rönnberg et al., 2013). The assessment of listening contributes to our understanding of the real-life difficulties faced by individuals with hearing impairments (Alhanbali et al., 2017; Cañete et al., 2023).

Different subjective and objective methods have been used to assess listening effort (McGarrigle et al., 2014; Peelle, 2018). Among subjective measures, self-reported measures have been used extensively, which are simple and without cost. On the other hand, a variety of physiological measures are used for the objective assessment of listening effort, including electroencephalography (EEG), functional near-infrared spectroscopy, skin conductance and pupillometry (Alhanbali et al., 2017; Dimitrijevic et al., 2019; Mackersie et al., 2015; Ohlenforst et al., 2018; White & Langdon, 2021). In particular, changes in the amplitude of neural oscillations during the speech presentation and memory retention period measured by EEG have been shown to correlate with changes in listening effort. For instance, an increase in alpha band (8–13 Hz) power in the left inferior frontal gyrus and parietal cortex has been associated with the self-reported effort ratings in a speech-in-noise task (Ala et al., 2020; Dimitrijevic et al., 2019; Paul et al., 2021). An increase in theta band (4–7 Hz) power localised to frontal midline regions has also been reported to reflect subjective listening effort in a speech-in-noise task (Wisniewski et al., 2015).

While different neural correlates of listening effort have been proposed, reliability assessment of these measures is comparatively scarce. A reliability analysis examines the variability of response to a certain task across repeated measurements when experimental conditions remain unchanged (Downing, 2004). Variability in these responses may be caused by inherent large variation, problems with the equipment, improper understanding of the task, motivation or performance feedback, among other factors (Giuliani et al., 2021). Recently, a number of studies have attempted to assess the reliability of listening effort measures. For instance, Giuliani et al. (2021) studied the reliability of several measures of listening effort (self-reports, peak pupil diameter, skin conductance and reaction time) and reported fair-to-moderate reliability for the measures based on intraclass correlation coefficients (ICC) values in the speech-in-noise recognition task. Alhanbali et al. (2019) reported good reliability for the measures of listening effort (self-reported listening effort, pupil size, EEG alpha power and skin conductance) during a digit-span task. Specifically, they reported that alpha power in the parietal cortex is a reliable measure of listening effort.

In this study, we further assessed listening effort by measuring self-reported effort ratings and neural oscillations recorded by EEG while participants performing speech-in-noise tasks. Experiments were run in two sessions, in which coherent sentences (sentences) and words in random order (word lists) were used as speech material and presented in different signal-to-noise ratios (SNRs). We assessed the test–retest reliability of effort measures, including self-reported effort ratings, theta and alpha neural oscillations.

2 | MATERIALS AND METHODS

2.1 | Participants

The study included 32 healthy and normal-hearing participants (13 females, age = 24 ± 3 years; mean ± standard deviation, SD). In terms of precision, this sample size ensures that the 95% confidence intervals (95% CI) around the standard error of the measurement (standard error of measurement [SEM]) will not exceed 25% of its magnitude (Mokkink et al., 2023). There was no history of neurological or psychiatric illness or psychotropic medication use among the participants. Participants provided written informed consent and received financial compensation. The study was approved by the ethics committee of Northern Jutland, Denmark (N-20200061), and it was conducted at Aalborg University following the Declaration of Helsinki.

2.2 | Stimuli

Speech materials were obtained from the Dantale II database (Wagener et al., 2009). Dantale II contains 15 lists with 10 sentences each. Each sentence has the same syntactical structure consisting of a name, verb, number, adjective and object but is semantically unpredictable (e.g. in English: "Michael owns six nice houses"). The sentences in each list were generated by a random combination of the alternatives from a base list. Each base list consists of 10 sentences. The sentences were recorded at 44.1 kHz by a Danish female speaker. Duration of the sentences ranged from 1.85 to 2.52 s (2.22 ± 0.12 s; mean ± SD).

Random word lists with neither syntactic structure nor sentence-level semantic content were created. Each sentence of the base list was split into 5 words, yielding 50 different words. A natural pause after each word was kept by selecting the duration of individual words from the beginning of the given word to the beginning of the next word. Word lists were created by randomly combining 5 words from the list of 50 words (e.g. in English: "find won jackets nine new"). The duration of all word lists was between 1.58 and 2.71 s (2.20 s ± 0.16 s), comparable with that of the sentences.

The audio files were then masked by speech-shaped noise at SNRs of −9, −6, −3 and 0 dB by varying the intensity of the speech while keeping the background noise constant. Speech-shaped noise was created based on the long-term power spectrum of speech. This noise was also used during baseline and retention intervals (see below).

2.3 | Experimental design and stimulus presentation

The experiment was conducted in two sessions with identical conditions, separated by 6 ± 3 days. The experiment used a factorial design with speech type (sentences and word lists) and SNR (−9, −6, −3 and 0 dB) as independent variables. Combining two speech types and four SNR levels resulted in a total of eight conditions introduced in a block design with randomised order. Each block consists of 25 trials, for a total of 200 trials per session. Each trial consisted of four intervals and was started by a 'baseline' interval. During baseline, participants listened to background noise (speech-shaped noise) lasting 3 s plus a random interval of 0–1 s. This was followed by a 'listening' interval during which participants listened to sentences and word lists in the presence of background noise. Then, the trial was followed by a 'retention' interval in which speech had to be retained in memory for 3 s plus a random interval of 0.29–1.42 s depending on speech duration to ensure that all trials had the same post-stimulus (listening + retention) period of 6 s. Then, during the response interval, all corpus items of the base list appeared on the screen in front of the participants. Participants were asked to click on words verbatim, matching those they had heard. The percentage of correct words was calculated as a speech recognition performance. Participants were asked to rate their level of listening effort on a 1–10 scale using the NASA Task Load Index (Hart & Staveland, 1988) immediately after each block (25 trials) and then got a 3 min rest. Each session used different words/sentences from the same database.

The experiment was run using custom code written in MATLAB (R2021b, MathWorks Inc.). The audio signal was played via a soundcard (Scarlett 2i2 2nd Gen) and presented diotically through insert-earphones (a-JAYS Three). The presentation was controlled using the Psychophysics Toolbox (PTB-3). In preparation for the main experiment, participants heard some examples of speech in each condition and were familiar with all the procedures.

2.4 | EEG recording and preprocessing

EEG data were collected from 64 active sensors placed on a standard cap based on the 10–20 international system using a g.HIamp biosignal amplifier (g.tec medical engineering GmbH, Austria). The EEG was sampled continuously at 1200 Hz. The left earlobe (A1) was selected as a reference. During recordings, the impedance of all electrodes was kept below 5 kΩ. The experiment was carried out in an electromagnetically shielded room. The EEG data were processed using a MATLAB script and the EEGLAB toolbox (Delorme & Makeig, 2004). For each participant, EEG data of all conditions and sessions were concatenated. The data were then re-referenced to the average, band-passed between 0.5 and 40 Hz using a third-order zero-phase Butterworth filter and resampled to 128 Hz to reduce processing time. Portions of data contaminated by high-amplitude short-time artefacts produced by head and eye movement were detected and corrected automatically using the Artifact Subspace Reconstruction (ASR) algorithm (Mullen et al., 2013) in EEGLAB. Each trial was epoched to −2 to 6 s. Independent Component Analysis (ICA) was then performed to remove any remaining artefacts. The independent components derived by ICA were labelled using ICLabel (Pion-Tonachini et al., 2019) as implemented in EEGLAB. Components that belonged to each of the artefact classes (Muscle, Eye, Heart, Line Noise and Channel Noise) with a probability above 50% were visually examined for removal. Four EEG channels from one participant for all blocks were removed before ICA due to a very high muscle artefact level and were interpolated using spline interpolation after ICA in the EEGLAB toolbox. Data subsequently were exported in the BESA Research 7.1 (https://www.besa.de/) for time-frequency analysis. EEG data of two participants were excluded from further analysis: one due to an internal failure of the amplifier during recording and the other because of excessive artefacts.

2.5 | Analysis of EEG data

The power of neural oscillations for single trials was calculated using the complex demodulation method implemented in BESA Research. The complex demodulation uses two steps. First, the time-domain signal was multiplied by a complex exponential at the frequency of interest *f*. Then a low-pass finite impulse response (FIR) filter isolated the energy near frequency *f* (Hoechstetter et al., 2004). Data were processed for frequencies between 1 and 30 Hz with a time-frequency sampling rate of 1 Hz/50 ms. The time-frequency data were baseline corrected to the −2 to 0 s prestimulus interval. Power spectral changes relative to baseline were quantified as percent change.

2.6 | Statistical analysis

2.6.1 | Behavioural data

A two-way repeated measures analysis of variance (RM ANOVA) with the session (first and second) and SNR (levels: −9, −6, −3 and 0 dB) as independent variables was applied to the percentage of correct words values and listening effort scores. In case of violations of the sphericity assumption, the Greenhouse–Geisser correction was applied. IBM SPSS Statistics 27 was used for the analysis. Bonferroni–Holm corrected *p*-values were reported for multiple comparisons.

2.6.2 | Neural data

To test the differences between conditions in each session, we conducted cluster-based permutation tests. Specifically, we used a cluster-based permutation *t*-test (paired, two-tailed, with 5000 permutations, cluster entry criterion; $p = 0.05$) to compare word lists and sentences at each SNR level and for both sessions. We also used a cluster-based permutation repeated-measure ANOVA test (with 5000 permutations, cluster entry criterion of 0.05) to assess the effect of SNR on word lists and sentences in each session. When a significant SNR effect was observed, post-hoc cluster-based tests were conducted for pairwise comparisons. To account for multiple tests, we reported Bonferroni–Holm corrected *p*-values for each test.

For reliability analysis, power values were averaged over time intervals and electrodes of interest (for theta, frontal midline electrodes, and for the alpha, right central electrodes; Figures 4 and 5). On these data, a two-way RM ANOVA with the session (first and second) and SNR (levels: −9, −6, −3 and 0 dB) as independent variables was applied.

2.6.3 | Association between neural data and self-reported measures

To determine the association between the neural data (frontal theta and right central alpha) to self-reported effort ratings, we used a linear mixed model (LMM) with effort rating, SNR and correct word score as fixed effects and participants as random variables.

2.6.4 | Reliability analysis

Two types of reliability have been identified: relative and absolute reliability (Atkinson & Nevill, 1998; Biurrun Manresa et al., 2014; Bruton et al., 2000; Lamb, 2016). Relative reliability refers to the extent to which the individual's scores maintain their position over repeated measurement relative to others. Methods based on correlation coefficients such as the interclass correlation coefficient (ICC) provide an expression of relative reliability. Absolute reliability is the degree to which individuals' scores vary in repeated measurements and is expressed in actual units of measurement (Atkinson & Nevill, 1998). Absolute reliability can be assessed using the SEM and Bland–Altman (BA) analysis (Bland & Altman, 1999).

Reliability analyses were conducted on the percentage of correct words, listening effort scores, and theta and alpha power values for each condition. For the ICC, a two-way mixed model using absolute agreement was selected, and the ICC of single measurements was reported with its corresponding 95% CI. On the other hand, the SEM is calculated as $\mathrm{SD_{diff}}/\sqrt{2}$, where $\mathrm{SD_{diff}}$ is the standard deviation of the differences between repeated measurements. SEM can also be used to calculate the smallest detectable change (SDC), defined as the minimum amount of change in the score that can be interpreted as a real change for an individual rather than, potentially, the result of measurement error (Geerinck et al., 2019; Overend et al., 2010; Ries et al., 2009). A smaller SDC indicates a more reliable measure. The SDC was calculated as $1.96 \cdot \sqrt{2} \cdot \mathrm{SEM}$.

The BA analysis comprises plotting the average versus differences of paired measurements, from which the limits of agreement (LoA) are derived. The LoA are defined as the mean difference between repeated measurements (known as bias) $\pm\, 1.96 \cdot \mathrm{SD_{diff}}$. The LoA determine the range within which 95% of the differences between repeated measurements are expected to fall. The 95% CIs are reported for the SEM, the LoA and bias (Bland & Altman, 1999).

2.6.5 | Data and code availability statement

The EEG data, behavioural data and analysis code are available upon request to Ole Kæseler Andersen.

3 | RESULTS

3.1 | Behavioural outcomes

The individual average percentage of correct words and self-reported listening effort are shown in Figure 1. With regard to the number of correct words identified, a main effect of the session ($F(1,31) = 34.55$, $p < 0.001$, $\eta_p^2 = 0.52$), a main effect of SNR ($F(1.43,44.30) = 304$, $p < 0.001$, $\eta_p^2 = 0.90$) and a session × SNR interaction effect ($F(1.65,51.24) = 6.52$, $p = 0.005$, $\eta_p^2 = 0.17$) were found in the sentences condition. Post-hoc analysis revealed systematic bias between sessions (session 1 – session 2) at −9 dB (mean difference = −7.27%, $t(31) = -3.5$, $p = 0.005$), −6 dB (mean difference = −6%, $t(31) = -4.7$, $p < 0.001$ and −3 dB (mean difference = −2.45%, $t(31) = -3.11$, $p = 0.015$) but not at 0 dB (mean difference = −0.07%, $t(31) = -0.12$, $p = 0.90$). Furthermore, a main effect of the session ($F(1,31) = 29.22$, $p < 0.001$, $\eta_p^2 = 0.48$) and main effect of SNR ($F(2.15,66.74) = 590$, $p < 0.001$, $\eta_p^2 = 0.95$) were found for the word lists condition but no significant session × SNR interaction ($F(3,93) = 1.31$, $p = 0.27$, $\eta_p^2 = 0.01$). Post-hoc analysis on the effect of the session showed a mean difference with a magnitude of −4.92% between sessions ($t = -5.41$, $p < 0.001$). Post-hoc analysis on the effect of the SNR demonstrated differences between −9 and −6 dB (mean difference = −23.5, $t(31) = -20$, $p < 0.001$), between −9 and −3 dB (mean = −40.8, $t(31) = -28.4$, $p < 0.001$), between −9 and 0 dB (mean = −49.4, $t(31) = -33$, $p < 0.001$), between −6 and −3 dB (mean = −17.3, $t(31) = -13.2$, $p < 0.001$), between −6 and 0 dB (mean = −26, $t(31) = -19.3$, $p < 0.001$) and between −3 and 0 dB (mean = −8.6, $t(31) = -11.5$, $p < 0.001$). BA

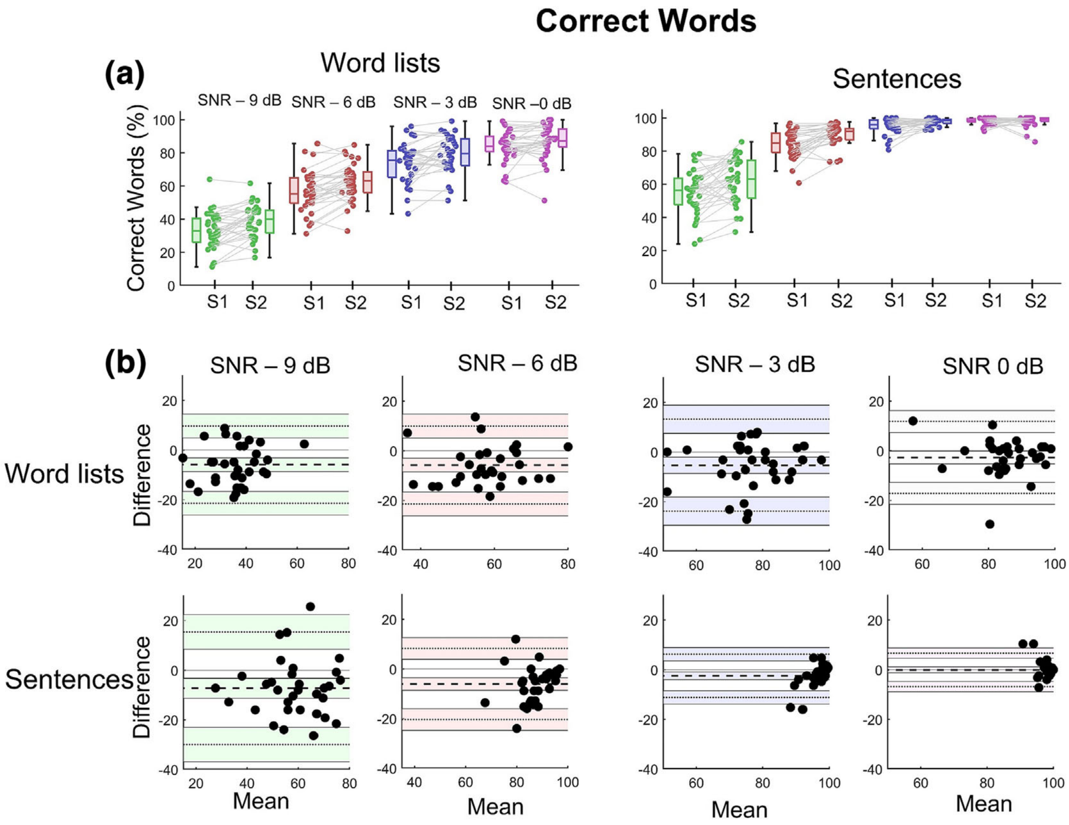

FIGURE 1 (a) Individual's percentage of correct words for each condition (sentences and word lists at a signal-to-noise ratio [SNR] of −9, −6, −3 and 0 dB) and in session 1 (S1) and session 2 (S2). Lines represent the mean (central dot) and standard deviation (SD) (whiskers). (b) Bland–Altman plots of the percentage of correct words for word lists and sentences. The dashed line indicates the bias between sessions, and the dotted lines are the limits of agreement (LoA), calculated as a bias ±1.96 times the standard deviation of the differences (SD_{diff}) in measurements between sessions. Shaded areas indicate the 95% confidence intervals of the bias and the LoA.

plots and ICC, SEM and SDC values and their 95% CI for each condition were listed in Figure 1 and Table 1, respectively.

In relation to listening effort, a main effect of the session (F(1,31) = 22.63, p < 0.001, $\eta_p^2 = 0.42$), a main effect of SNR (F(2.5,77.66) = 206, p < 0.001, $\eta_p^2 = 0.87$) and a session × SNR interaction effect (F(3,93) = 6.06, p = 0.001, $\eta_p^2 = 0.16$) were observed in the sentences condition. Indeed, differences in listening effort scores between sessions were observed at −6 dB (mean difference = 1.70, t(31) = 3.87, p = 0.003), at −3 dB (mean difference = 1.40, t(31) = 4.45, p < 0.001), but not at −9 dB (mean difference = 0.27, t(31) = 1.12, p = 0.63) or at 0 dB (mean difference = 0.40, t(31) = 1.68, p = 0.41). Concerning word lists, a main effect of the session (F (1,31) = 13.51, p < 0.001, $\eta_p^2 = 0.30$), a main effect of SNR (F(2.25,69.63) = 123, p < 0.001, $\eta_p^2 = 0.80$) and a session × SNR interaction effect (F(2.46,76.20) = 3.49, p = 0.027, $\eta_p^2 = 0.17$) were found. Differences between sessions were observed at −3 dB (mean difference = 1.34, t(31) = 3.92, p = 0.003) but not at −9 dB (mean difference = 0.17, t (31) = 1.88, p = 0.13), −6 dB (mean difference = 0.52, t (31) = 1.60, p = 0.11) or at 0 dB (mean difference = 0.82, t(31) = 1.88, p = 0.11). BA plots and ICC, SEM and SDC values and their 95% CI for each condition were listed in Figure 2 and Table 1, respectively.

TABLE 1 Intraclass correlation coefficient (ICC), standard error of measurement (SEM) and smallest detectable change (SDC) for the percentage of correct words, self-reported listening effort, theta power and alpha power at different signal-to-noise ratios (SNRs).

	SNR −9 dB		SNR −6 dB		SNR −3 dB		SNR 0 dB	
	Word lists	Sentences	Word lists	Sentences	Word lists	Sentences	Word lists	Sentences
Correct words (95% confidence interval)								
ICC	0.62 (0.19–0.82)	0.56 (0.2–0.77)	0.70 (0.30–0.87)	0.37 (−0.02–0.65)	0.62 (0.29–0.80)	0.21 (−0.09 to 0.49)	0.68 (0.44–0.83)	0.18 (−0.19–0.50)
SEM	5.6 (4.2–7)	8.2 (6.1–10.2)	5.6 (4.2–7)	5.2 (3.9–6.4)	6.7 (5.0–8.3)	3.1 (2.3–3.9)	5.2 (3.9–6.5)	2.4 (1.8–3)
SDC	15.5 (11.7–19.3)	22.7 (17.1–28.2)	15.6 (11.8–19.5)	14.3 (10.8–17.8)	18.5 (14.0–23)	8.7 (6.5–10.8)	14.4 (10.9–18)	6.7 (5.1–8.4)
Listening effort								
ICC	0.66 (0.41–0.82)	0.49 (0.18–0.71)	0.30 (−0.06–0.55)	0.26 (−0.04–0.54)	0.30 (−0.03–0.57)	0.42 (0.02–0.69)	0.36 (0.04–0.62)	0.50 (0.20–0.72)
SEM	0.37 (0.28–0.46)	0.98 (0.74–1.2)	1.28 (0.96–1.6)	1.76 (1.3–2.2)	1.39 (1.1–1.7)	1.26 (0.95–1.6)	1.51 (1.14–1.8)	0.96 (0.72–1.2)
SDC	1.02 (0.77–1.3)	2.7 (2–3.4)	3.5 (2.7–4.4)	4.9 (3.7–6.1)	3.8 (2.9–4.8)	3.5 (2.6–4.3)	4.2 (3.2–5.2)	2.7 (2–3.3)
Theta power								
ICC	0.69 (0.45–0.84)	0.68 (0.43–0.83)	0.51 (0.19–0.74)	0.46 (0.12–0.70)	0.65 (0.38–0.81)	0.35 (−0.01–0.63)	0.62 (0.34–0.80)	0.16 (−0.19–0.48)
SEM	5.3 (4.0–6.7)	5.5 (4.1–7.0)	6.5 (4.9–8.2)	5.9 (4.4–7.4)	6.0 (4.4–7.4)	6.3 (4.7–8.0)	5.7 (4.2–7.1)	6.3 (4.74–7.9)
SDC	14.7 (11.0–18.5)	15.4 (11.5–19.3)	18.2 (13.6–22.8)	16.4 (12.3–20.6)	16.3 (12.2–20.5)	17.6 (13.2–22.1)	15.8 (11.9–19.8)	17.6 (13.1–22.1)
Alpha power								
ICC	0.41 (0.07–0.66)	0.50 (0.17–0.73)	0.60 (0.30–0.79)	0.53 (0.22–0.74)	0.57 (0.28–0.77)	0.40 (0.05–0.66)	0.60 (0.32–0.80)	0.33 (−0.03–0.62)
SEM	7.6 (5.6–9.5)	7.2 (5.3 – 9.0)	7.3 (5.4–9.1)	7.8 (5.8–9.8)	6.5 (4.8–8.1)	8.8 (6.3–11.1)	6.7 (5.1–8.4)	7.7 (5.3–8.8)
SDC	21.1 (15.7–26.3)	19.9 (14.8 – 24.9)	20.2 (15.1–25.3)	21.6 (16.2–27.1)	17.9 (14.4–22.5)	24.6 (18.4–30.8)	18.6 (13.9–23.3)	19.6 (14.6–24.5)

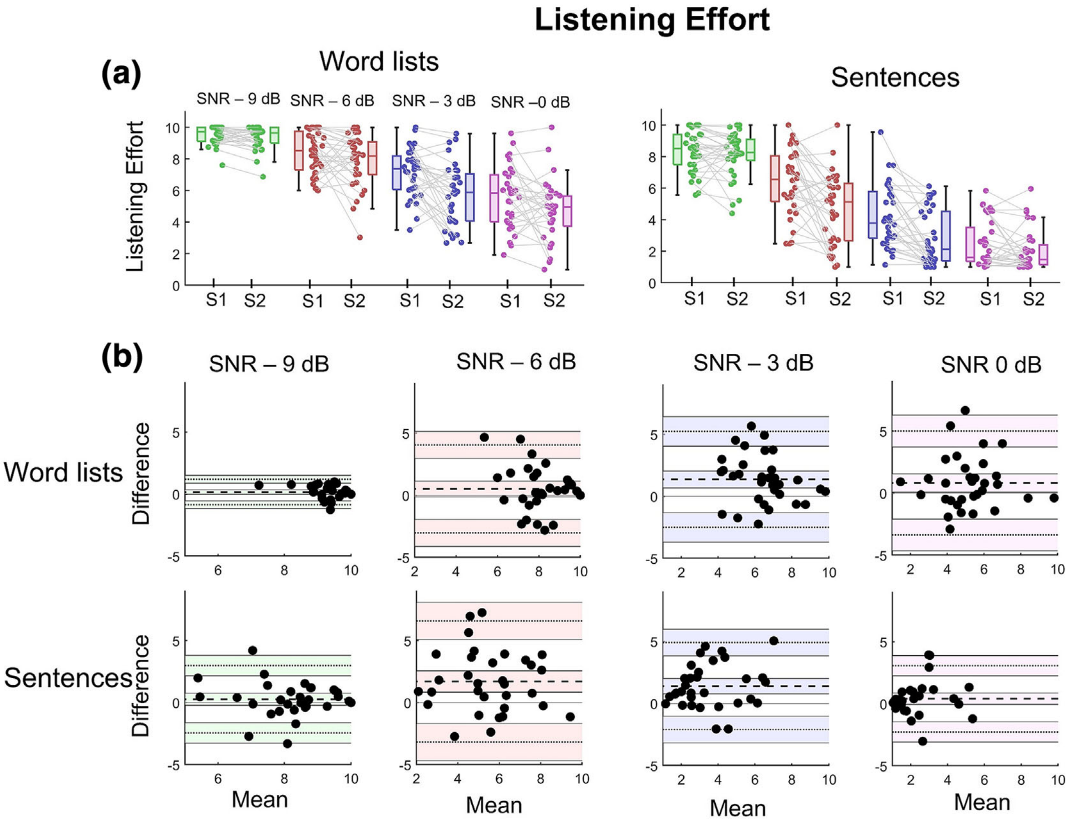

FIGURE 2 (a) Individual's listening effort scores for each condition (sentences and word lists at a signal-to-noise ratio [SNR] of −9, −6, −, and 0 dB) and in session 1 (S1) and session 2 (S2). Lines represent the mean (central dot) and standard deviation (SD) (whiskers). (b) Bland–Altman plots of the self-reported listening effort for word lists and sentences. The dashed line indicates the bias between sessions, and the dotted lines are the limits of agreement (LoA), calculated as a bias ±1.96 times the standard deviation of the differences (SD_{diff}) in measurements between sessions. Shaded areas indicate the 95% confidence intervals of the bias and the LoA.

3.2 | Neural results

Cluster-based permutation *t*-test showed a cluster of significant differences between word lists and sentences at SNR 0 dB (time interval: 3350–5850 ms; frequency: 2–8 Hz; channel at maximum: F1; $p = 0.028$) and SNR −3 dB (time interval: 2700–5600 ms; frequency: 2–8 Hz; channel at maximum: F3, $p = 0.019$) and not at SNR −6 dB and −9 dB ($p > 0.05$) for session one (Figure 3a). For session two, tests showed significant differences between word lists and sentences only at SNR 0 dB (time interval: 850–6000 ms; frequency: 2–10 Hz; channel at maximum Fz; $p = 0.024$) but not at SNR -3, SNR −6 and −9 dB (Figure 3b).

Cluster-based ANOVA test showed only an effect of SNR on sentences in session one (time interval: 500–3300 ms, channel at maximum CP4, $p < 0.001$). In the following, post-hoc tests showed a significant difference in power values between SNR −3 and 0 dB (time interval: 1400–2200 ms; frequency: 2–15 Hz; channel at maximum: C4; $p < 0.001$), a difference between SNR −3 and −6 dB (time interval: 1400–2600 ms; frequency: 7–17 Hz; channel at maximum: CP4; $p = 0.004$), a difference between SNR −3 dB and −9 dB (time interval: 1050–3050 ms; frequency: 4–24 Hz; channel at maximum: AF4; $p < 0.001$), a difference between SNR 0 and −9 dB (time interval: 2000–3350 ms; frequency: 9–21 Hz; channel at maximum: F4; $p < 0.001$) (Figure 3c). No relevant SNR effects were observed for word lists in both sessions and for sentences in session two ($p > 0.05$).

In the following permutation testing, time-frequency data were averaged across the frequency range of 4–7 Hz (theta band) and time interval of 3.5–5.5 s (retention interval) (Figure 4). To conduct further statistical analysis, the theta power was averaged over frontal midline electrodes, including AF3, AF4, F3, F1, FZ, F2, F4, FC1, FCZ and FC2, and the averaged values were shown in Figure 6a. These specific intervals and electrodes were chosen based on cluster-based test results, as presented above. Two-way RM ANOVA performed on frontal theta power for word lists showed a small main effect of SNR ($F(1.98,57.50) = 2.89$, $p = 0.04$, $\eta_p^2 = 0.09$) but no relevant

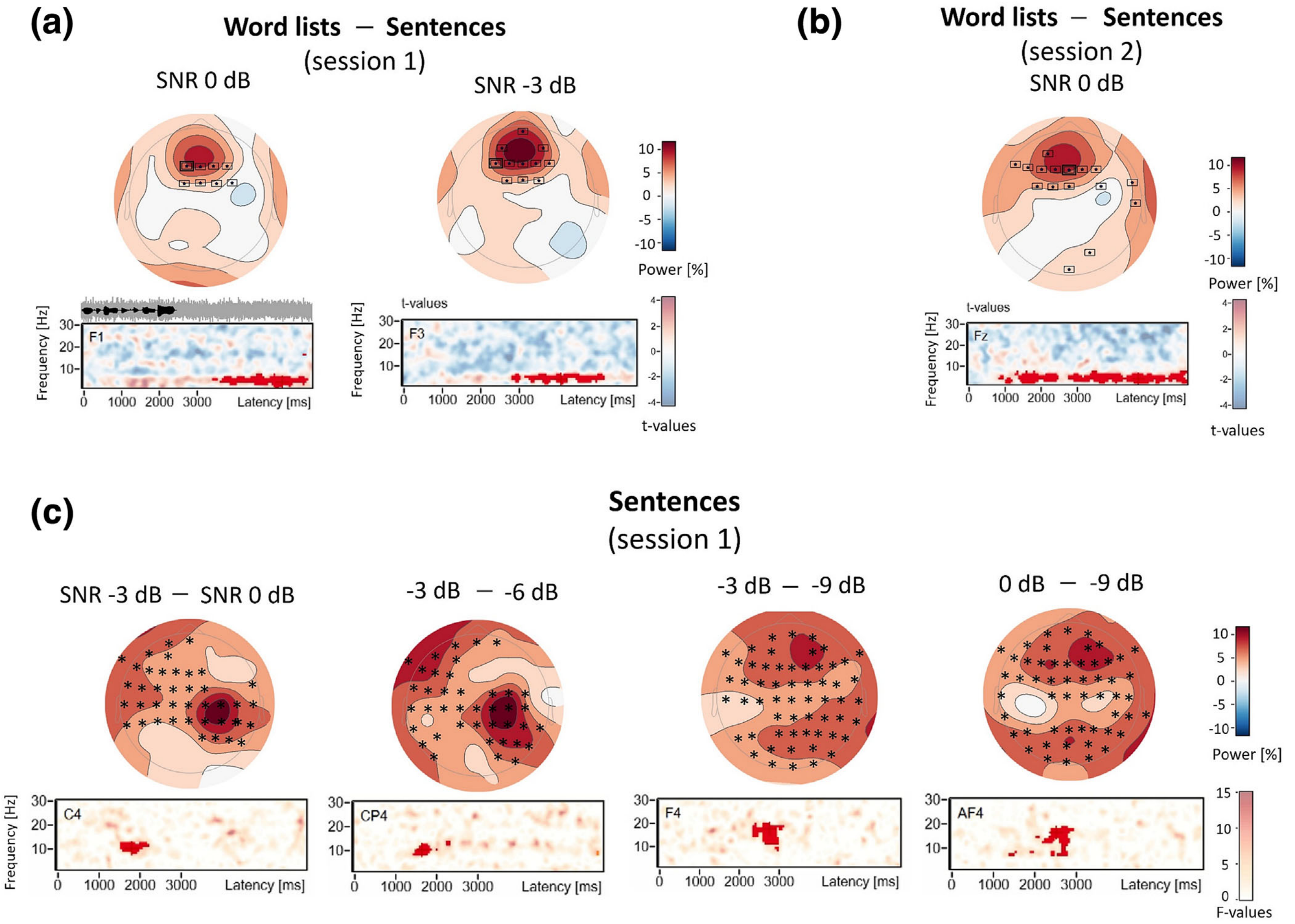

FIGURE 3 Cluster-based permutation test results. (a) Clusters of electrodes and time-frequency points that showed significant differences between word lists and sentences at a signal-to-noise ratio (SNR) of 0 dB ($p = 0.028$) and −3 dB ($p = 0.019$) in session 1. Power value is represented in percentage relative to the baseline. The speech in noise acoustic waveform is shown above time-frequency plot. (b) A cluster of electrodes and time-frequency points that showed a significant difference between word lists and sentences at SNR of 0 dB ($p = 0.024$) in session 2. (c) Clusters of electrodes and time-frequency points that showed significant differences in sentences between SNRs in session 1: −3–0 dB ($p < 0.001$), −3 to −6 dB ($p = 0.004$), −3 to −9 dB ($p < 0.001$) and 0 to −9 dB ($p < 0.001$). Electrodes belonging to significant clusters of differences are marked with asterisks. Images were generated in BESA statistics 7.1.

effect of session ($F(1,29) = 0.22$, $p = 0.64$, $\eta_p^2 = 0.008$) or session × SNR interaction($F(3,87) = 1.76$, $p = 0.16$, $\eta_p^2 = 0.05$). Post-hoc analysis of the SNR effect showed no relevant differences after the Bonferroni−Holm correction. Concerning sentences, the analysis showed a main effect of SNR ($F(3,87) = 3.55$, $p = 0.017$, $\eta_p^2 = 0.10$), no relevant effects of session ($F(1,29) = 1.86$, $p = 0.18$, $\eta_p^2 = 0.06$) and interaction ($F(3,87) = 0.66$, $p = 0.57$, $\eta_p^2 = 0.02$). The ICC, SEM and SDC values for each condition are listed in Table 1. LMM results showed no relevant relationship between the frontal midline theta power during retention and listening effort ratings ($p > 0.05$).

Figure 5 shows the topographical plot of alpha power (8–13 Hz) averaged over a time interval of interest of 0.5–2 s (listening interval). For further statistical analysis, the alpha power was then averaged over the time interval of interest and right central electrodes (FC2, FC4, FC6, C2, C4, C6, CP2, CP4, CP6) (Figure 6). The averaged alpha power values for word lists and sentences are indicated in Figure 7a. One these data, two-way RM ANOVA performed for word lists showed no relevant effects for the session ($F(1,29) = 3.68$, $p = 0.06$, $\eta_p^2 = 0.11$), SNR ($F(3,87) = 2.08$, $p = 0.11$, $\eta_p^2 = 0.06$) or for the session × SNR interaction effect ($F(3,87) = 0.82$, $p = 0.49$, $\eta_p^2 = 0.03$). RM ANOVA performed for sentences revealed a main effect of SNR ($F(3,87) = 6.35$, $p < 0.001$, $\eta_p^2 = 0.18$) and no relevant effects of session ($p = 0.32$) or session × SNR interaction ($p = 0.65$). Post-hoc analysis on the SNR effect showed differences between −9 and −3 dB (mean

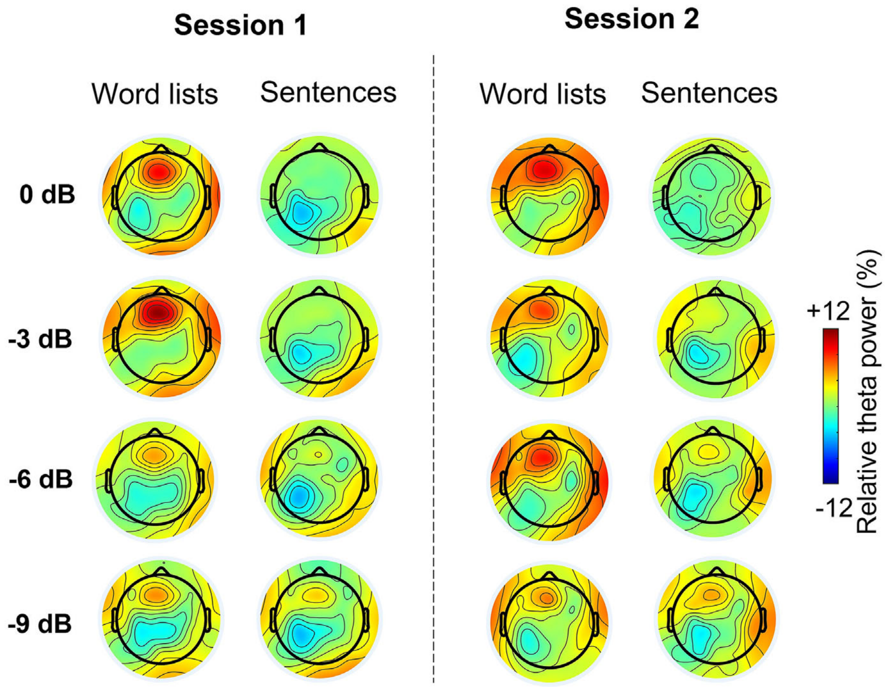

FIGURE 4 Average theta (4–6 Hz) power across participants during the retention interval (3.5–5.5 s) for word lists and sentences at a different signal-to-noise (SNR): −9, −6, −3 and 0 dB and sessions.

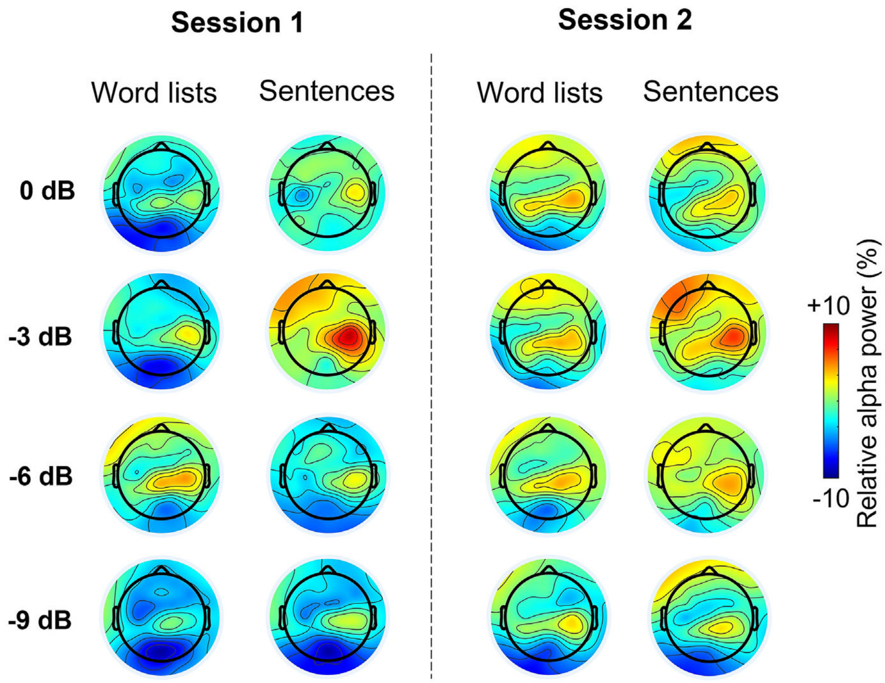

FIGURE 5 Average alpha (8–13 Hz) power across participants during the listening interval (0.5–2 s) for word lists and sentences at a different signal-to-noise (SNR): −9, −6, −3 and 0 dB and sessions.

difference = −5.12, $t(29) = -4.14$, $p = 0.001$). The ICC, SEM and SDC values for each condition are listed in Table 1.

LMM results showed a relationship between the right central alpha power and listening effort ratings in only sentences for session one ($p = 0.051$). It further showed no relevant relation for SNR ($p = 0.08$) and correct words ($p = 0.20$).

4 | DISCUSSION

4.1 | The reliability of the behavioural measures

Self-reported measures of listening effort typically refer to an answer on a scale to how effortful the task was (Johnson et al., 2015). Therefore, it is expected that the

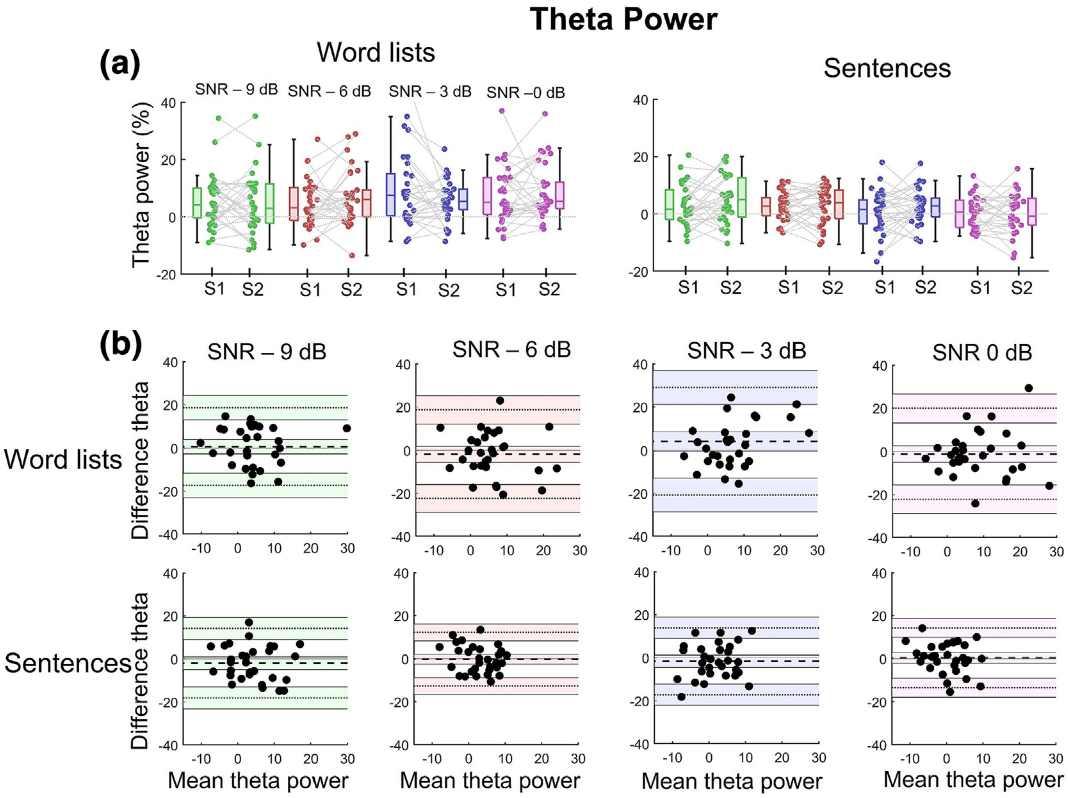

FIGURE 6 (a) Individual's frontal theta power for session 1 (S1) and session 2 (S2) at a signal-to-noise ratio (SNR) of −9, −6, −3 and 0 dB. Lines represent the mean (central dot) and standard deviation (SD) (whiskers). (b) Bland–Altman plots of frontal theta power at different SNRs for (top) word lists and (bottom) sentences. The dashed line indicates the bias between sessions, and the dotted lines are the limits of agreement (LoA), calculated as a bias ±1.96 times the standard deviation of the differences (SD_{diff}) in measurements between sessions. Shaded areas indicate the 95% confidence intervals of the bias and the LoA.

answer could be different between individuals, since people's understanding of the meaning of the question may differ. However, more important than that is how the individual's effort score varies between sessions. If between-session variability is acceptable, these measures could be used in a realistic setting to assess the daily life effort of hearing-impaired individuals (Cañete et al., 2023).

Previous studies used ICC to assess the reliability of self-reported listening efforts such as Alhanbali et al. (2019) which showed a 'good to excellent' (ICC = 0.83) reliability of effort rating. In this study, in addition to the ICC method, we used absolute reliability measures such as SEM, SDC and BA plots. As an example, for self-reported listening effort in sentences at 0 dB, we reported ICC of 0.50, SEM of 0.96 points and the SDC of 2.7 points. The SEM value indicates that the difference between a subject's measurement of effort rating and the hypothetical true score would be expected to be less than 1.96 · SEM, which equals 1.88 points for 95% of observations (Atkinson & Nevill, 2000; Bland & Altman, 1996). The SDC value means that the self-reported effort score of an individual would have to change by at least 2.7 points (on a scale of 1 to 10) before the observed change can be considered to be a real change in the effort rating of a subject, and not potentially a result of measurement error. Alternatively, from BA plot, the difference between the two measurements is expected to be less than 3 points (i.e. within the LoA) for 95% of the pair of observations within a week's time.

BA plots and ANOVA analysis for a percentage of correct words (Figure 2a) show a systematic bias, meaning participants recognised words better in the second session compared with the first one, most likely due to learning effects (Goldstone, 1998; Zhang et al., 2021). This systematic bias for sentences showed an interaction with SNR, indicating a smaller bias with increasing SNR. When the SNR was highest (0 dB), the bias was negligible. At this SNR, a ceiling effect is observed in the sentences condition, meaning that it was not possible to improve word recognition beyond the already high values, and differences between sessions are reduced accordingly. However, in word lists conditions, the average bias was almost the same (∼5%) across all SNRs.

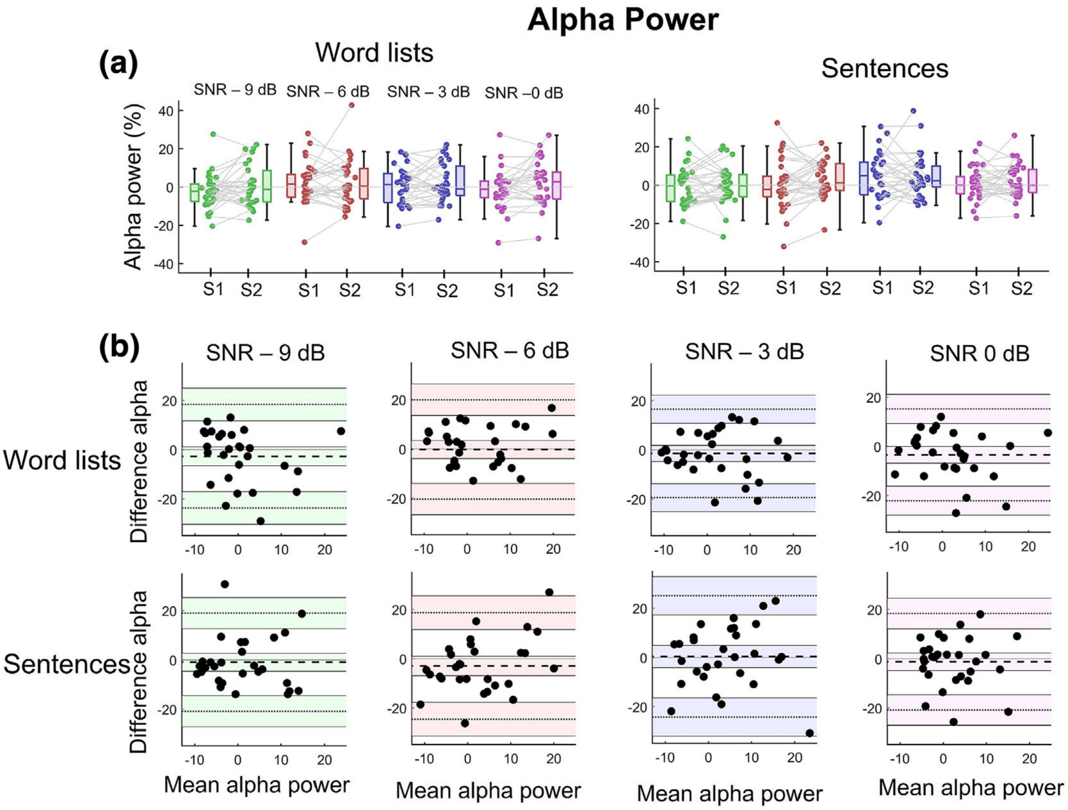

FIGURE 7 (a) Individual's alpha power for session 1 (S1) and session 2 (S2) at a signal-to-noise ratio (SNR) of −9, −6, −3 and 0 dB. Lines represent the mean (central dot) and standard deviation (SD) (whiskers). (b) Bland–Altman plots of alpha power at different SNRs for (top) word lists and (bottom) sentences. The dashed line indicates the bias between sessions, and the dotted lines are the limits of agreement (LoA), calculated as a bias ±1.96 times the standard deviation of the differences (SD_{diff}) in measurements between sessions. Shaded areas indicate the 95% confidence intervals of the bias and the LoA.

From BA plots, it is also evident that the measurement error diminishes with increasing SNR (i.e. the LoA get narrower) in sentences but not in word lists. This indicates that the source of between-session variation in word lists solely depends on the type of speech, which is the random combination of words, whereas, in sentences, measurement error depends on SNR.

It should be noted that our data emphasise the use of absolute measures of reliability compared with relative measures (e.g. ICC). For instance, the ICC value for correct words at 0 dB in sentences is 0.18, which would be classified as 'poor' reliability using current arbitrary scales. However, it is not appropriate to treat reliability results as *dichotomous* variables or as the outcome of a hypothesis test, and the measurement error should be considered (Biurrun Manresa et al., 2014; Bland & Altman, 1999; Völker et al., 2021). In this regard, the BA plot (Figure 2b) shows that the maximum difference for most of the observations (95%) between two sessions is approximately 7% (i.e. within the LoA) which is acceptable given the range of 0%–100% for correct words. This indicates that ICC is highly influenced by heterogeneity and the range of measured scores (Atkinson & Nevill, 1998; Martin Bland & Altman, 1986). Therefore, relative measures of reliability should be carefully interpreted, preferably along with absolute estimates of measurement error.

4.2 | The reliability of neural measures

The frontal midline theta oscillations have been observed in cognitive control (Cavanagh & Frank, 2014) and working memory tasks (Jensen & Tesche, 2002). Generally, frontal midline theta oscillations increase when participants actively engage in cognitive tasks and proportionally increase with task demand (Hsieh & Ranganath, 2014). In relation to listening effort, frontal midline theta oscillations have been proposed as a neural correlate of listening effort (Wisniewski, 2017; Wisniewski et al., 2015, 2018, 2021). For instance, Wisniewski et al. (2015) performed a speech-in-noise task where sentences were presented to participants in background noise at five different SNR levels (−12, −6, 0, 6 and 12 dB). They reported an effect of SNR on theta power ($\eta_p^2 = 0.29$), indicating that as SNR decreases frontal theta power increases which might be related to increased effort. Indeed, they also reported that increased frontal midline theta power was correlated to the self-reported listening effort.

In the present study, we used sentences and word lists and presented them in four SNR levels (−9, −6, −3 and 0 dB). We observed an increased frontal theta during retention interval for word lists than sentences at high SNRs (−3 and 0 dB), probably representing increased working memory load for word lists compared with sentences. We also observed a small effect of SNR on frontal theta power for sentences ($\eta_p^2 = 0.10$) and word lists ($\eta_p^2 = 0.09$). In addition, we observed an effect of SNR on right central alpha power in sentences $\left(\eta_p^2 = 0.18\right)$. Then, we assessed the reliability of frontal theta power and central alpha power for both speech types and observed high variability for both theta and alpha values despite differences observed in conditions (SNRs and speech types) at the group level. For example, the average theta power on the two sessions in word lists at 0 dB was 6.26%; therefore, it would be likely that a subsequent volunteer to be tested shows a frontal theta power magnitude of 7% in the first session. The LoA indicates a 95% probability that the retest theta power will be between −14% and 25.6%. The same reliability analysis for alpha power (for word lists at SNR 0 dB) showed that retest alpha power will be between −21% and 15.8% for a subsequent volunteer. These ranges are very wide and would be unacceptable for practical use.

Additionally, concerning association with behavioural data, in contrast to (Wisniewski et al., 2015), we observed no significant association between frontal theta power and self-reported listening effort for word lists and sentences. This is because we used LMM, considering SNR as a fixed effect and participants as random variables. However, in Wisniewski et al. (2015), the authors related averaged theta power, over participants, to the self-reported effort. This might be misleading since, in their paradigm, SNR is the main driver of theta power changes and should be involved in the model to access the real relation between theta power and effort rating.

While our study suggests that the neural data (frontal midline theta and right central alpha) may not be a reliable measure of listening effort, it is important to note that there were some limitations to our study that could have affected the reliability of the neural measures. We used 25 trials per condition, which is lower than the average number of trials used per condition in previous studies on the listening effort (Ala et al., 2020; Alhanbali et al., 2019; Dimitrijevic et al., 2019; Paul et al., 2021; Wisniewski et al., 2015) which is 53 (30–100, min and max). However, in our study, the number of trials was limited by the number of conditions and experiment duration, and increasing the number of trials would have resulted in fatigue and a decrease in signal quality. It is also possible that learning effects, as observed in behavioural data, could have contributed to the low reliability of neural measures, by adding more variability to neural responses. Therefore, future studies that aim to assess the reliability of listening effort measures are suggested to consider controlling for learning effects and increasing the number of trials.

5 | CONCLUSION

In this study, behavioural and neural data were recorded in speech-in-noise perception tasks. The behavioural measures (percentage of correct words and self-reported listening effort) showed low between-session variability, indicating a high level of reliability. An increased frontal theta power in word lists compared with sentences was observed during the retention interval which may indicate an increased working memory load for word lists. A pattern of increased right central alpha power was observed in response to listening to speech in noise that showed to be affected by SNR only in sentences. Neural data (frontal theta and right central alpha power) showed high between-session variability. The high variability in neural data might be due to a low number of trials (25) used or learning effects as observed in behavioural measures.

AUTHOR CONTRIBUTIONS

Yousef Mohammadi: Conceptualization; data curation; formal analysis; writing—original draft. **Jan Østergaard:** Conceptualization; funding acquisition; project administration; software; writing—review and editing. **Carina Graversen:** Conceptualization; writing—review and editing. **Ole Kæseler Andersen:** Conceptualization; funding acquisition; project administration; software; writing—review and editing. **José Biurrun Manresa:** Conceptualization; methodology; supervision; validation; writing—review and editing.

ACKNOWLEDGEMENTS

The study was supported by the Independent Research Fund Denmark, Technology and Production Sciences (FTP).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/ejn.16187.

ORCID

Yousef Mohammadi https://orcid.org/0000-0002-6211-9619

José Biurrun Manresa https://orcid.org/0000-0003-4060-9665

REFERENCES

Ala, T. S., Graversen, C., Wendt, D., Alickovic, E., Whitme, W. M., & Lunner, T. (2020). An exploratory study of EEG alpha oscillation and pupil dilation in hearing-aid users during effortful listening to continuous speech. *PLoS ONE*, *15*(7), e0235782. https://doi.org/10.1371/journal.pone.0235782

Alhanbali, S., Dawes, P., Lloyd, S., & Munro, K. J. (2017). Self-reported listening-related effort and fatigue in hearing-impaired adults. *Ear and Hearing*, *38*(1), e39–e48. https://doi.org/10.1097/AUD.0000000000000361

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

Atkinson, G., & Nevill, A. (2000). Typical error versus limits of agreement. *Sports Medicine (Auckland, N.Z.)*, *30*(5), 375–381. https://doi.org/10.2165/00007256-200030050-00005

Atkinson, G., & Nevill, A. M. (1998). Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Medicine (Auckland, N.Z.)*, *26*(4), 217–238. https://doi.org/10.2165/00007256-199826040-00002

Biurrun Manresa, J. A., Fritsche, R., Vuilleumier, P. H., Oehler, C., Mørch, C. D., Arendt-Nielsen, L., Andersen, O. K., & Curatolo, M. (2014). Is the conditioned pain modulation paradigm reliable? A test-retest assessment using the nociceptive withdrawal reflex. *PLoS ONE*, *9*(6), e100241. https://doi.org/10.1371/journal.pone.0100241

Bland, J. M., & Altman, D. G. (1996). Statistics notes: Measurement error. *BMJ*, *312*(7047), 1654. https://doi.org/10.1136/bmj.312.7047.1654

Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, *8*(2), 135–160. https://doi.org/10.1177/096228029900800204

Bruton, A., Conway, J. H., & Holgate, S. T. (2000). Reliability: What is it, and how is it measured? *Physiotherapy*, *86*(2), 94–99. https://doi.org/10.1016/S0031-9406(05)61211-4

Cañete, O. M., Nielsen, S. G., & Fuentes-López, E. (2023). Self-reported listening effort in adults with and without hearing loss: The Danish version of the effort assessment scale (D-EAS). *Disability and Rehabilitation*, *45*(1), 98–105. https://doi-org.zorac.aub.aau.dk/10.1080/09638288.2021.2022781

Cavanagh, J. F., & Frank, M. J. (2014). Frontal theta as a mechanism for cognitive control. *Trends in Cognitive Sciences*, *18*(8), 414–421. https://doi.org/10.1016/j.tics.2014.04.012

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Dimitrijevic, A., Smith, M. L., Kadis, D. S., & Moore, D. R. (2019). Neural indices of listening effort in noisy environments. *Scientific Reports*, *9*(1), 1–10. https://doi.org/10.1038/s41598-019-47643-1

Downing, S. M. (2004). Reliability: On the reproducibility of assessment data. *Medical Education*, *38*(9), 1006–1012. https://doi.org/10.1111/j.1365-2929.2004.01932.x

Edwards, B. (2016). A model of auditory-cognitive processing and relevance to clinical applicability. *Ear and Hearing*, *37* (Suppl 1), 85S–91S. https://doi.org/10.1097/AUD.0000000000000308

Geerinck, A., Alekna, V., Beaudart, C., Bautmans, I., Cooper, C., de Souza Orlandi, F., Konstantynowicz, J., Montero-Errasquín, B., Topinková, E., Tsekoura, M., Reginster, J. Y., & Bruyère, O. (2019). Standard error of measurement and smallest detectable change of the sarcopenia quality of life (SarQoL) questionnaire: An analysis of subjects from 9 validation studies. *PLoS ONE*, *14*(4), e0216065. https://doi.org/10.1371/journal.pone.0216065

Giuliani, N. P., Brown, C. J., & Wu, Y. H. (2021). Comparisons of the sensitivity and reliability of multiple measures of listening effort. *Ear and Hearing*, *42*(2), 465–474. https://doi.org/10.1097/AUD.0000000000000950

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology*, *49*, 585–612. https://doi.org/10.1146/annurev.psych.49.1.585

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

Hoechstetter, K., Bornfleth, H., Weckesser, D., Ille, N., Berg, P., & Scherg, M. (2004). BESA source coherence: A new method to study cortical oscillatory coupling. *Brain Topography*, *16*(4), 233–238. https://doi.org/10.1023/B:BRAT.0000032857.55223.5d

Hsieh, L. T., & Ranganath, C. (2014). Frontal midline theta oscillations during working memory maintenance and episodic encoding and retrieval. *NeuroImage*, *85 Pt 2*(02), 721–729. https://doi.org/10.1016/j.neuroimage.2013.08.003

Jensen, O., & Tesche, C. D. (2002). Frontal theta activity in humans increases with memory load in a working memory task. *The European Journal of Neuroscience*, *15*(8), 1395–1399. https://doi.org/10.1046/j.1460-9568.2002.01975.x

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

Lamb, K. (2016). Test-retest reliability in quantitative physical education research: A commentary. *European Physical Education Review*, *4*(2), 145–152. https://doi.org/10.1177/1356336X9800400205

Mackersie, C. L., Macphee, I. X., & Heldt, E. W. (2015). Effects of hearing loss on heart rate variability and skin conductance measured during sentence recognition in noise. *Ear and*

Hearing, *36*(1), 145–154. https://doi.org/10.1097/AUD.0000000000000091

Martin Bland, J., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, *327*(8476), 307–310. https://doi.org/10.1016/S0140-6736(86)90837-8

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'.

Mokkink, L. B., de Vet, H., Diemeer, S., & Eekhout, I. (2023). Sample size recommendations for studies on reliability and measurement error: An online application based on simulation studies. *Health Services & Outcomes Research Methodology*, *23*, 241–265.

Mullen, T., Kothe, C., Chi, Y. M., Ojeda, A., Kerth, T., Makeig, S., Cauwenberghs, G., & Jung, T. P. (2013). Real-time modeling and 3D visualization of source dynamics and connectivity using wearable EEG. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society*. Annual International Conference, *2013*, 2184–2187.

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. *Hearing Research*, *365*, 90–99. https://doi.org/10.1016/j.heares.2018.05.003

Overend, T., Anderson, C., Sawant, A., Perryman, B., & Locking-Cusolito, H. (2010). Relative and absolute reliability of physical function measures in people with end-stage renal disease. *Physiotherapy Canada. Physiotherapie Canada*, *62*(2), 122–128. https://doi.org/10.3138/physio.62.2.122

Paul, B. T., Chen, J., Le, T., Lin, V., & Dimitrijevic, A. (2021). Cortical alpha oscillations in cochlear implant users reflect subjective listening effort during speech-in-noise perception. *PLoS ONE*, *16*(7), e0254162. https://doi.org/10.1371/journal.pone.0254162

Peelle, J. E. (2018). Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior. *Ear and Hearing*, *39*(2), 204–214. https://doi.org/10.1097/AUD.0000000000000494

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W. Y., Humes, L. E., Lemke, U., Lunner, T., Matthen, M., Mackersie, C. L., Naylor, G., Phillips, N. A., Richter, M., Rudner, M., Sommers, M. S., Tremblay, K. L., & Wingfield, A. (2016). Hearing impairment and cognitive energy: The framework for understanding effortful listening (FUEL). *Ear and Hearing*, *37*(Suppl 1), 5S–27S. https://doi.org/10.1097/AUD.0000000000000312

Pion-Tonachini, L., Kreutz-Delgado, K., & Makeig, S. (2019). ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage*, *198*, 181–197. https://doi.org/10.1016/j.neuroimage.2019.05.026

Ries, J. D., Echternach, J. L., Nof, L., & Blodgett, M. G. (2009). Test-retest reliability and minimal detectable change scores for the timed "up & go" test, the six-minute walk test, and gait speed in people with Alzheimer disease. *Physical Therapy*, *89*(6), 569–579. https://doi.org/10.2522/ptj.20080258

Rönnberg, J., Lunner, T., Zekveld, A., Sörqvist, P., Danielsson, H., Lyxell, B., Dahlström, Ö., Signoret, C., Stenfelt, S., Pichora-Fuller, M. K., & Rudner, M. (2013). The ease of language understanding (ELU) model: Theoretical, empirical, and clinical advances. *Frontiers in Systems Neuroscience*, *7*(JUNE), 31. https://doi.org/10.3389/fnsys.2013.00031

Völker, J. M., Arguissain, F. G., Andersen, O. K., & Biurrun Manresa, J. (2021). Variability and effect sizes of intracranial current source density estimations during pain: Systematic review, experimental findings, and future perspectives. *Human Brain Mapping*, *42*(8), 2461–2476. https://doi.org/10.1002/hbm.25380

Wagener, K., Josvassen, J. L., & Ardenkjær, R. (2009). Design, optimization and evaluation of a Danish sentence test in noise: Diseño, optimización y evaluación de la prueba Danesa de frases en ruido. *International Journal of Audiology*, *42*(1), 10–17. https://doi.org/10.3109/14992020309056080

White, B. E., & Langdon, C. (2021). The cortical organization of listening effort: New insight from functional near-infrared spectroscopy. *NeuroImage*, *240*, 118324. https://doi.org/10.1016/j.neuroimage.2021.118324

Wisniewski, M. G. (2017). Indices of effortful listening can be mined from existing electroencephalographic data. *Ear and Hearing*, *38*(1), e69–e73. https://doi.org/10.1097/AUD.0000000000000354

Wisniewski, M. G., Iyer, N., Thompson, E. R., & Simpson, B. D. (2018). Sustained frontal midline theta enhancements during effortful listening track working memory demands. *Hearing Research*, *358*, 37–41. https://doi.org/10.1016/j.heares.2017.11.009

Wisniewski, M. G., Thompson, E. R., Iyer, N., Estepp, J. R., Goder-Reiser, M. N., & Sullivan, S. C. (2015). Frontal midline θ power as an index of listening effort. *Neuroreport*, *26*(2), 94–99. https://doi.org/10.1097/WNR.0000000000000306

Wisniewski, M. G., Zakrzewski, A. C., Bell, D. R., & Wheeler, M. (2021). EEG power spectral dynamics associated with listening in adverse conditions. *Psychophysiology*, *58*(9), e13877. https://doi.org/10.1111/psyp.13877

Zhang, L., Schlaghecken, F., Harte, J., & Roberts, K. L. (2021). The influence of the type of background noise on perceptual learning of speech in noise. *Frontiers in Neuroscience*, *15*, 484. https://doi.org/10.3389/fnins.2021.646137

How to cite this article: Mohammadi, Y., Østergaard, J., Graversen, C., Andersen, O. K., & Biurrun Manresa, J. (2023). Validity and reliability of self-reported and neural measures of listening effort. *European Journal of Neuroscience*, *58*(11), 4357–4370. https://doi.org/10.1111/ejn.16187