Intrasubject variability in Potential Early Markers of Sensorineural Hearing Damage

Nele De Poortere [1], Sarineh Keshishzadeh[2], Hannah Keppler[1,4], Ingeborg Dhooge[3,4], Sarah Verhulst[2]

[1]Ghent University, Department of Rehabilitation Sciences – Audiology

[2] Ghent University, Dept. of Information Technology – Hearing Technology @ WAVES

[3] Ghent University Hospital, Department of Ear, Nose and Throat, Belgium

[4] Ghent University, Department of Head and Skin

Financial disclosure

Ghent University holds a patent on the RAM-EFR stimulation protocol (U.S. patent App. 17/791,985 (Inventors: Sarah Verhulst, Viacheslav Vasilkov). This work was supported by UGent BOF-IOP project: "Portable Hearing Diagnostics: Monitoring of Auditory-nerve Integrity after Noise Exposure (EarDiMon), European research counsil proof of concept grant CochSyn (899858) and EU Innovation counsil Grant EarDiTech (101058278).

Conflict of interest

The authors report there are no competing interests to declare.

Data Availability Statement

The datasets generated during and/or analyzed during the current study are not publicly available due to ethical restriction but are available from the corresponding author on reasonable request.

Acknowledgement:

The authors would like to thank Eef De Wilde and Ellen Sabau for their significant contribution in the data-collection

All correspondence should be addressed to:

Nele De Poortere, Faculty of Medicine and Health Sciences, Department of Rehabilitation Sciences, Corneel Heymanslaan 10 (2P1), Ghent University, 9000 Ghent, Belgium; nele.depoortere@ugent.be; +32 479 02 32 62.

ABSTRACT

The quest for noninvasive early markers for sensorineural hearing loss (SNHL) has yielded diverse measures of interest. However, comprehensive studies evaluating the test-retest reliability of multiple measures and stimuli within a single study are scarce, and a standardized clinical protocol for robust early markers of SNHL remains elusive. To address these gaps, this study explores the intra-subject variability of various potential EEG-biomarkers for cochlear synaptopathy (CS) and other SNHL-markers in the same individuals. Fifteen normal-hearing young adults underwent repeated measures of (extended high-frequency) pure-tone audiometry, speech-in-noise intelligibility, distortion-product otoacoustic emissions (DPOAEs), and auditory evoked potentials; comprising envelope following responses (EFR) and auditory brainstem responses (ABR). Results confirm high reliability in pure-tone audiometry, whereas the matrix sentence-test exhibited a significant learning effect. The reliability of DPOAEs varied across three evaluation methods, each employing distinct SNR-based criteria for DPOAE-datapoints. EFRs exhibited superior test-retest reliability compared to ABR-amplitudes. Our findings underscore the necessity for cautious interpretation of presumed noninvasive SNHL measures. While tonal-audiometry's robustness was corroborated, we observed a confounding learning effect in longitudinal speech audiometry. The variability observed in DPOAEs highlights the importance of consistent ear probe replacement and meticulous measurement techniques, suggesting that DP-thresholds may be less suitable for clinical application. As potential EEG-biomarkers of CS, EFRs are preferred over ABR-amplitudes based on the current study results.

Keywords: Intrasubject variability – noninvasive early markers of sensorineural hearing loss-cochlear synaptopathy – normal hearing young adults

INTRODUCTION

In the realm of clinical application, the standard procedure for assessing auditory thresholds relies on conventional pure-tone audiometry. Consequently, studies targeting the evaluation of hearing damage related to aging, ototoxicity and excessive noise exposure have primarily centered on identifying permanent hearing threshold changes within the frequency range of 250 to 8000 Hz (Cruickshanks et al., 2010; Rabinowitz et al., 2006). Temporary threshold shifts (TTS) resulting from noise exposure were historically regarded as less concerning markers for permanent hearing damage, as indicated by the National Institute of Occupational Safety and Health (NIOSH) in 1998. However, recent rodent studies have challenged this perspective, revealing that a noise-exposure-induced TTS may coincide with permanent deficits at the synaptic level (Kujawa & Liberman, 2009b), a phenomenon referred to as cochlear synaptopathy (CS) (Fernandez et al., 2020). CS predominantly affects the connections between type-I auditory nerve fiber terminals and inner hair cells (IHCs) (Furman et al., 2013; Kujawa & Liberman, 2009a) and can result in supra-threshold temporal coding deficits (Bharadwaj et al., 2015b). Unfortunately, CS mostly affects supra-threshold sound coding without affecting routine clinical hearing sensitivity measures such as the tonal audiogram or distortion-product otoacoustic emissions (DPOAEs) (Furman et al., 2013; Lobarinas et al., 2013). Nonetheless, CS is believed to contribute to symptoms such as tinnitus and hyperacusis (Guest et al., 2017; Paul et al., 2017; Schaette & McAlpine, 2011; Wojtczak et al., 2017), and is thought to underlie other perceptual challenges, including difficulties in discriminating sounds in complex acoustic environments and impaired temporal processing of sound and speech intelligibility (Bharadwaj et al., 2015a; Garrett, 2020; Guest, Munro, Prendergast, et al., 2018; Mepani et al., 2021; Oxenham, 2016; Prendergast et al., 2017; Smith et al., 2019).

Hence, researchers persist in their quest for precise and reliable measures of early SNHL markers in clinical settings. While numerous studies have investigated the test-retest reliability of various measures, there is a noticeable lack of research that administers a comprehensive battery of tests to the same individuals. Moreover, as novel SNHL treatments emerge, there is a pressing need for robust biomarkers to monitor treatment effectiveness. Consequently, this study aims to explore the intra-subject variability of potential EEG biomarkers for CS and other early SNHL indicators within the same individuals, with the objective of exploring their suitability for clinical use. By conducting multiple test on the same participants, we aim to provide a unique opportunity to compare intra-subject variability.

Early Indicators of Outer Hair Cell Loss

According to literature, extended high frequency (EHF) audiometry presents a more sensitive approach for the early detection of noise induced hearing loss (NIHL) compared to conventional frequencies (Singh et al., 2009; Wang et al., 2000). Additionally, EHF audiometry has proven to be a valuable predictive tool for identifying the risk of NIHL (Hunter et al., 2020; Mehrparvar et al., 2011), highlighting its clinical significance in the field of hearing health assessment.

An alternative, or complementary, approach to behavioral audiometry involves DPOAEs, which closely reflect the integrity of cochlear structures; particularly the outer hair cells (OHCs) (Jansen et al., 2009). Furthermore, DPOAEs are recognized for their sensitivity in detecting subtle cochlear damage before it manifests in pure-tone hearing threshold elevations (Coradini et al., 2007; Glavin et al., 2021; Knight et al., 2007; Reavis et al., 2015). Moreover, DPOAEs offer distinct advantages such as rapid acquisition, non-participatory nature, and suitability for measurement by non-specialist personnel trained in emission assessments (Reavis et al., 2015). However, despite the crucial role of DPOAEs in audiology, there remains

a noticeable scarcity of studies that concurrently investigate the most suitable DPOAE evaluation methodologies, along with the reliability of auditory thresholds, DP-amplitudes, and DP-thresholds within the same cohort of subjects.

Indicators of Cochlear Synaptopathy

Promising EEG biomarkers for CS include prominent suprathreshold neural potentials such as the Auditory Brainstem Response (ABR) wave I amplitude and the Envelope Following Response (EFR) strength, as highlighted in several studies (Guest et al., 2017; Liberman et al., 2016; Shaheen et al., 2015; Wilson et al., 2021). Animal models have emphasized the significance of ABR-amplitudes and EFR-strengths as clinical metrics for diagnosing age- and noise-induced CS (Kujawa & Liberman, 2009a; Shehabi et al., 2022; Skoe & Tufts, 2018). However, translating these measures as early biomarkers from rodents to humans faces challenges. Firstly, intersubject variabilities stemming from differences in head size and sex (Mitchell et al., 1989), electrode resistance, and various sources of electrical noise between individuals or sessions (Plack et al., 2016) act as confounding factors, limiting their diagnostic utility in humans. Secondly, while the early wave I of the ABR is thought to have diagnostic potential in individual listeners due to its clear link to CS in animal models, the OHC-loss aspect of sensorineural hearing damage also affects its amplitude (Verhulst et al., 2016). Additionally, the sensitivity of ABR measurements to low spontaneous rate (SR) ANFs, particularly vulnerable to CS, has been questioned due to the delayed onset response of these fibers (Bourien et al., 2014). ABRs are evoked by transient stimuli and primarily reflect onset responses, which tend to be relatively small in low SR fibers (Rhode & Smith, 1985; Taberner & Liberman, 2005). In contrast, EFR-strengths in response to sinusoidal amplitude modulated (SAM) tones are primarily driven by low SR-ANFs, especially when the modulation depths are shallow (Bharadwaj et al., 2014). High SR-ANFs also contribute to the EFR, when the modulation depth is maximal, as e.g. shown in the auditory simulation models of Encina-

Llamas et al. (2019), Vasilkov et al. (2021), and Van Der Biest et al. (2023). Additionally, it is important to acknowledge that there are multiple sources (including AN) that can contribute to the recorded EFRs. However, the contributing sources can be targeted towards more peripheral generators (AN, CN, IC) by using higher modulation frequencies above 80 Hz, as suggested by Purcell et al. (2004).

Given that CS has been assumed to impact suprathreshold hearing sensitivity and speech recognition abilities, particularly in challenging listening conditions (Kujawa & Liberman, 2009a; Lin et al., 2011; Parthasarathy & Kujawa, 2018; Skoe et al., 2019), various speech recognition in noise tests have been employed in human studies to explore CS (Garrett, 2020; Grinn et al., 2017; Guest, Munro, & Plack, 2018; Mepani et al., 2021; Vande Maele et al., 2021). Nonetheless, research into the within-subject variability of these tests has remained limited.

In sum, the reliability of proposed noninvasive metrics as early markers of SNHL in humans remains uncertain due to challenges posed by both intra- and inter-subject variabilities, impeding integration into clinical practice. This study explores the intra-subject variability of various potential EEG biomarkers for CS and other early SNHL indicators within the same individuals across three sessions in a cohort of normal hearing listeners. The aim is to explore and compare potential early indicators of SNHL that could enhance clinical diagnostics and monitoring of SNHL.

MATERIALS AND METHODS

PARTICIPANTS AND STUDY DESIGN

Fifteen young adults, nine men and six women, aged between 18 and 25 years (mean age 21.0 years ± 1.77 standard deviation; SD) participated at three test sessions. Participant selection involved administering a hearing evaluation questionnaire, followed by PTA and tympanometry

during the initial session. Individuals with known hearing disorders, a history of ear surgery, or tinnitus were excluded. The study encompassed three distinct sessions denoted as session 1, 2, and 3. Between each consecutive session, a time interval of two to three days was maintained, with the exception of two participants who had a 14- and 15-day interval between session 2 and 3. Throughout these intervals, participants were instructed to abstain from exposure to loud activities. During the first session, the best ear was selected based on PTA at both conventional and extended high frequencies. During each session, participants completed a comprehensive test battery consisting of (EHF) PTA, speech in quiet (SPiQ)- and speech in noise (SPiN)-tests, DPOAEs and AEP-measurements. The selection for the right ear was made for 10 participants, while the left ear was tested for five participants. As part of Covid-19 safety measures, subjects wore a face mask during the measurements. The test protocol had a maximum duration of three hours, and tests were administered in a consistent sequence for all subjects across all sessions. This study received approval from the UZ Gent ethical committee (BC-05214) and adhered to the ethical principles outlined in the Declaration of Helsinki. All participants were informed about the testing procedures and provided an informed consent.

OTOSCOPY AND TYMPANOMETRY

Otoscopy of the ear canal and the tympanic membrane was performed using a Heine Beta 200 LED otoscope (Dover, USA), and showed bilateral normal otoscopic in all subjects. Middle-ear admittance was bilaterally measured, followed by unilateral measurements (best ear) in the follow-up sessions, using a GSI TympStar (Grason-Stadler) tympanometer (Minneapolis, USA) with a 226 Hz, 85 dB sound pressure level (SPL) probe tone. All tympanograms were defined as a type-A according to the Liden-Jerger classification (Jerger, 1970; Lidén, 1969).

PURE-TONE AUDIOMETRY

Pure-tone thresholds were determined in a double-walled sound-attenuating booth by the use of an Equinox Interacoustics audiometer (Middelfart, Denmark). Stimuli were transmitted using

Interacoustics TDH-39 headphones (Middelfart, Denmark) and Sennheiser HDA-200 headphones (Wedemark, Germany) for conventional frequencies and EHFs, respectively. Air-conduction thresholds were measured using the modified Hughson-Westlake procedure at conventional octave frequencies 0.125, 0.250, 0.500, 1, 2, 4 and 8 kHz, half-octave frequencies 3 and 6 kHz, and EHF 10, 12.5, 14, 16, and 20 kHz. Both ears were tested at the first session to determine the test ear for the subsequent measurements, selecting the ear with superior thresholds on conventional frequencies. All participants were classified as having normal hearing according to the World Health Organization's (WHO) guidelines, which define normal hearing as a better-ear audiometric threshold averaged over 0.5, 1, 2, and 4 kHz equal or below 20 dB HL (Organization, 2021).

SPEECH INTELLIGIBILITY IN QUIET AND IN NOISE (SPiQ AND SPiN)

During each session, the SPiQ- and SPiN tests were administered in a quiet testing room, using the Flemish Matrix sentence test (Luts et al. 2014) and Apex 3 software (Francart et al., 2008). The sentences were presented to the best ear using a laptop connected to a Fireface UCX soundcard (RME) (Haimhausen, Germany) and HDA-300 (Sennheiser) headphones (Wedemark, Germany).

Speech performance was evaluated through four experimental tests lists, encompassing both broadband (BB) and high-pass filtered (HP) speech in quiet (SPiQ) and in noise (SPiN). In each session, participants were randomly assigned all four test lists, each containing 20 sentences in a randomized sequence. Due to protocol adjustments, BB-quiet was not executed for two participants. To counteract potential learning effects, two additional BB-noise training lists were administered (Luts et al., 2014).

BB-quiet involved presenting speech without filtering, while HP-quiet applied a zero-phase 1024th-order FIR HP-filter (cutoff 1650 Hz) to the speech signal. For HP-noise, both speech and

noise signals were filtered using the same cutoff values as HP-quiet, whereas BB-noise had no filtering on either the speech or noise signals.

The matrix-test consisted of a corpus of 50 words, categorized into 10 names, 10 verbs, 10 numerals, 10 adjectives, and 10 nouns. All sentences shared identical syntactical structures, and the semantic content remained unpredictable. The adaptive procedure outlined by Brand & Kollmeier (2002), was employed for all test lists, implementing a staircase paradigm to ascertain the speech-reception threshold. The speech level was adjusted by a maximum of 5 dB, progressively decreasing to a minimal step size of 0.1 dB. For SPiQ, the procedure commenced at a level of 50 dB SPL, while for SPiN, the noise was maintained at a fixed level of 70 dB SPL, starting at a SNR of -4 dB. In all tests lists, subjects were instructed to repeat the five-word sentences in a forced-choice setting, with 10 options provided for each word. The mean signal level or mean SNR from the six last reversals was utilized to determine the SPiQ and SPiN thresholds.

DISTORTION PRODUCT OTO-ACOUSTIC EMISSIONS (DPOAEs)

During each session, DPOAE measurements were carried out on the designated ear, encompassing DP-amplitudes and DP-thresholds. DPOAEs were collected in a quiet testing room, employing the Universal Smart Box (Intelligent Hearing Systems IHS) (Miami, United States). To ensure controlled conditions, both ears were shielded using earmuffs (Busters) (Kontich, Belgium) that were placed on top of a 10D IHS OAE-probe (Miami, United States). DPOAE responses and noise amplitudes were quantified using the simultaneous presentation of two primary tones, with f1 and f2 featuring a frequency ratio f2/f1 of 1.22. Noise artifact rejection was set at 10 dB SPL, and a total of 32 sweeps were recorded for each frequency or input-output level.

DP-amplitudes were obtained with a primary tone level combination of L1/L2 = 65/65 dB SPL and f2 ranging from 553 to 8837 Hz at two points per octave, and from 8837 to 11459 Hz at eight points per octave, resulting in twelve frequency bands with center frequencies 0.5, 0.7, 1.0, 1.4, 2.0, 2.8, 4.0, 5.7, 8.0, 8.7, 9.5, and 10.3 kHz.

DP-thresholds were obtained at octave frequencies between 0.5 to 8 kHz (i.e. $\sqrt{(f1*f2)}$ = 501, 1000, 2000, 3998, 8001, and 10376 Hz) with L2 ranging from 35 to 70 dB SPL in steps of 5 dB. L1/L2 varied across L2 intensities using the scissor paradigm of Kummer et al. (1998) whereby L1 = 0,4 L2 + 39 dB. Extrapolation and non-linear regression were used to estimate DP-thresholds in which a cubic function was fit to the I/O functions of DPOAE measurements of each frequency following the method of Verhulst et al. (2016). This way, DP-thresholds were determined as the level of L2 at which the curve reached the distortion component of -25 dB SPL.

The evaluation of DPOAEs was subdivided into three evaluation methods using commonly used inclusion criteria, i.e. response amplitude ≥ the noise floor; response amplitude ≥ 2SD above the noise floor; response amplitude ≥ 6 dB above the noise floor. When responses did not meet the inclusion criteria, the amplitudes were set to the noise floor levels. DP-thresholds outside the range -10 – 60 dB (Boege & Janssen, 2002) were excluded, since these responses are not considered as valid.

AUDITORY EVOKED POTENTIAL (AEP) MEASUREMENTS

AEP measurements, including EFRs and ABRs were conducted at the test ear using the IHS universal Smart box and SEPCAM software (Miami, United States). Recordings were performed in a quiet testing room, with subjects seated in a reclining chair, watching a muted video while resting their heads on a soft pillow. To minimize alpha-wave interference, subjects were instructed to relax without falling asleep. Controlled conditions within the hospital setting were maintained by shielding both ears with earmuffs (Busters) (Kontich, Belgium), turning off

extraneous lights and electronic devices, and applying NuPrep gel for skin preparation. Disposable Ambu Neuroline electrodes (Ballerup, Denmark) were placed on the vertex (inverting electrode), nasal flank on the non-test ear side (ground electrode), and bilateral mastoids (non-inverting electrodes). Electrode impedances were kept below 3 kΩ, and auditory stimuli were presented using etymotic ER-2 ear-probes (Chicago, USA).

EFRs were evoked using two stimulus types, distinguished by their modulation waveform, i.e. a sinusoidal amplitude modulated (SAM)-stimulus with a carrier frequency of 4 kHz, and rectangularly amplitude modulated (RAM)-stimuli, with carrier frequencies 4 and 6 kHz, and a duty cycle of 25% (Van Der Biest et al., 2023; Vasilkov et al., 2021). EFRs were evoked using 1000 alternating polarity sweeps. Stimuli had a modulation frequency of 110 Hz, a modulation depth of 100% and a duration of 500 ms which were presented at a rate of 2 Hz. The RAM stimuli with different carriers were calibrated in such a way to have the same peak-to-peak amplitude as a 70 dB SPL SAM-tone (carrier: 4 kHz, modulation frequency: 110 Hz, modulation depth: 100%). In this regard, the calibrated RAM stimuli with different carrier frequencies were presented at 68.24 dB SPL and had the same peak-to-peak amplitudes.

The EFR processing was performed in Matlab. Firstly, the recordings were filtered using a bandpass filter with low and high cutoff frequencies of 30 Hz and 1500 Hz, respectively. After filtering the EFRs, epoching and baseline correction was performed. Lastly, a bootstrapping approach according to Zhu et al. (2013) was adopted in the frequency domain to estimate the noise-floor and variability of the EFR, as detailed in Keshishzadeh et al. (2020). Subsequently, EFR-strengths represented the summation of the signal-to-noise spectral magnitude at the fundamental frequency and its following three harmonics, i.e. 110, 220, 330 and 440 Hz (Vasilkov et al., 2021).

ABRs were evoked using 4000 alternating polarity sweeps of six stimulus types, i.e. three broadband 80-µs clicks presented at levels of 70, 80 and 90 dBpeSPL and three narrowband toneburst (TB)-stimuli at 0.5 kHz, 1 kHz and 4 kHz with a stimulus duration of 5 ms, 4 ms and 2 ms, respectively. Clicks were presented at a rate of 11 Hz and TBs had a rate of 20 Hz. ABR recordings were filtered offline between 100 and 1500 Hz using a zero-phase filter. Afterwards, epoching and baseline correction was performed akin to the method described for EFR processing. After baseline correction, epochs were averaged to yield the ABR waveform. ABR waves I, III and V were manually peak-picked by audiologists to identify the respective ABR amplitudes (µV) and latencies (ms). ABR amplitudes were defined peak to baseline.

STATISTICAL ANALYSIS

All statistical analyses were conducted using IBM SPSS Statistics 27. The data-analysis employed a four-tiered methodology, including one-way repeated measures ANOVA, two-way random average measures intraclass correlation coefficient (ICC), standard errors of measurement (SEM), and calculation of individual 95% confidence intervals (95%CI). Firstly, a One-way repeated measures ANOVA was employed to assess variations in PTA, SPiQ and SPiN, DPOAEs and AEP outcomes across three consecutive measurements. Descriptive statistics were computed, and tests for normality, including the Shapiro-Wilk test, histograms, Q-Q plots, and box-and-whisker plots, were conducted to assess the assumptions of the one-way repeated measures ANOVA. If significance level ($p<0.05$) was reached, post-hoc analysis of least-square means was performed for intersession differences between the sessions. Two-tailed significance level was reached if $p < 0.017$ as adjusted by Bonferroni correction for multiple comparisons. Secondly, two-way random-average-measures intraclass correlation coefficients were computed to determine the relative consistency, i.e. the consistency of the position of individual scores relative to others. The interpretation of ICC-values followed the classification system proposed by Koo and Li (2016): excellent ICC (>0.90), good ICC (0.75 - 0.90), moderate ICC (0.50 – 0.75)

and poor ICC (< 0.50). Thirdly, SEM-scores were calculated to represent the reliability within repeated measures for an individual subject, reflecting absolute consistency. The latter is calculated as SEM = s*√(1-ICC), where 's' represents the standard deviation of all measurements. Finally, given that the substantial intersubject variability observed in each measure had an influence on the group-based test-retest 95% confidence intervals (CIs), we additionally computed 95%CIs of the repeated measures for each individual separately to visually assess the reliability of different hearing parameters in comparison to each other. This process entailed calculating 95% CIs across measurement sessions for each parameter and subject. The resulting distribution of individual CIs across subjects, is visualized using Kernel Density Estimation (KDE) plots, representing the upper and lower bounds of obtained test-retest CIs for each measure and each subject. These KDE-plots served to illustrate the variability of test-retest CIs across subjects and enhance the interpretation of test-retest variations within the data.

RESULTS

PURE-TONE AUDIOMETRY

At session 1, the mean pure-tone average at 3, 4, 6 and 8 kHz ($PTA_{3\text{-}8kHz}$) was 8.08 dB HL (SD 6.663, range -2.00 – 18.00) and the mean pure-tone average at 10, 12.5, 14, 16 and 20 kHz (PTA $EHF_{10\text{-}20\ kHz}$) was 4.40 dB HL (SD 8.382, range -8.00 – 18.00). One-way repeated measures ANOVA revealed no significant changes in pure-tone thresholds between measurements, except for the 0.25 kHz auditory thresholds [$F(2, 28) = 6.526$, $p = .005$]. Pairwise comparisons indicated a significant change of -4.00 dB from session 1 to session 3 ($p = .009$). Table I presents the averages per session and frequency, along with ICCs and SEMs for each tested frequency. In general, good to excellent ICCs with highly significant between subjects reliability ($p < .001$) were obtained and small SEMs were observed. However, at 6 kHz, 8 kHz and 20 kHz, lower ICCs were observed alongside wider corresponding 95%CIs and higher SEMs.

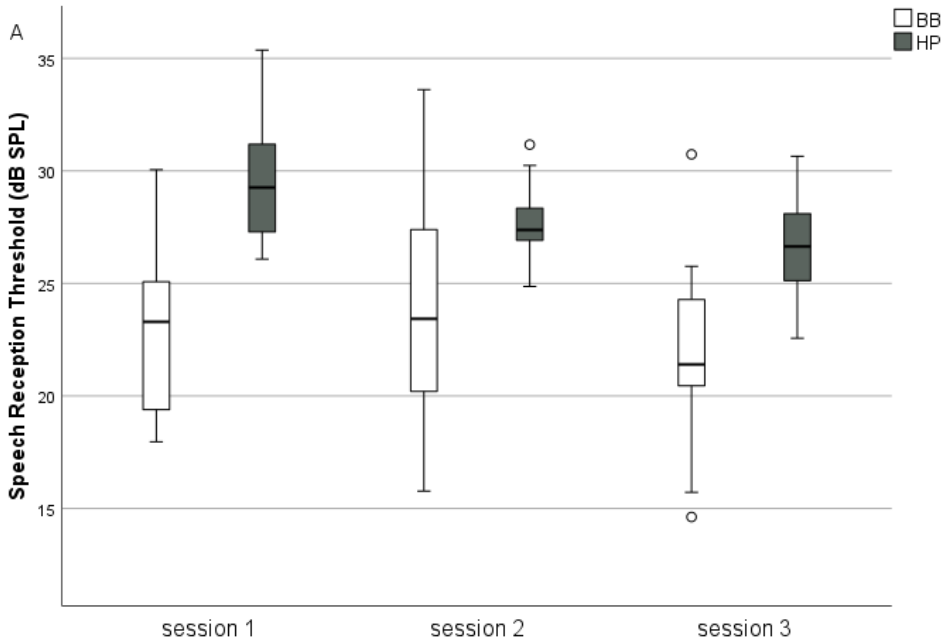
Table I. Summary of audiometric threshold averages (dB HL) per session and frequency, alongside intraclass correlation coefficients (ICCs) and corresponding 95% Confidence intervals (CIs), and standard error of measurements (SEMs) for all tested frequencies. P-values for between-subjects variability are reflected as *(0.05 > p > 0.01), **(0.01 > p > 0.001), and ***(p < 0.001).

Frequency (kHz)	**0.125**	**0.250**	**0.500**	**1.0**	**2.0**	**3.0**	**4.0**	**6.0**	**8.0**	**10.0**	**12.5.0**	**14.0**	**16.0**	**20.0**
Mean (SD) S1	8.33 (5.876)	4.67 (4.419)	1.67 (5.563)	1.00 (3.873)	3.00 (5.606)	6.43 (8.419)	3.67 (7.432)	13.67 (8.958)	9.00 (6.866)	2.00 (9.024)	4.00 (10.556)	3.33 (13.844)	7.67 (13.478)	5.00 (4.226)
Mean (SD) S2	6.33 (7.188)	4.00 (6.036)	2.00 (7.020)	0.00 (7.071)	2.67 (4.952)	5.00 (8.018)	2.33 (7.528)	14.33 (7.037)	10.33 (8.958)	4.00 (7.838)	3.00 (10.316)	2.33 (13.345)	8.00 (11.148)	3.33 (4.498)
Mean (SD) S3	5.00 (4.629)	0.67 (5.936)	0.67 (6.230)	-0.67 (5.627)	0.00 (4.629)	3.67 (6.673)	1.33 (9.348)	9.67 (8.550)	8.33 (7.480)	2.00 (8.619)	1.00 (9.856)	1.33 (13.157)	7.33 (11.629)	3.67 (7.898)
ICC	0.759***	0.805***	0.918***	0.851***	0.779***	0.936**	0.904***	0.636**	0.244	0.895***	0.911***	0.967***	0.963***	0.087
ICC 95%CI	0.448-0.911	0.517-0.930	0.809-0.970	0.650-0.945	0.488-0.918	0.845-0.978	0.776-0.965	0.180-0.864	-0.890-0.731	0.754-0.961	0.793-0.968	0.923-0.988	0.913-0.987	-1.293-0.676
SEM	2.952	2.504	1.770	2.157	2.418	1.192	2.485	5.008	6.681	2.711	3.010	2.392	2.280	5.447

S1, S2, S3 represent Session 1, Session 2, and Session 3, respectively

SPEECH IN QUIET AND SPEECH IN NOISE (SPiQ AND SPiN)

The distribution of SPiQ (dB SPL) and SPiN (dB SNR) thresholds across subjects is depicted in Figures 1A and 1B, respectively.



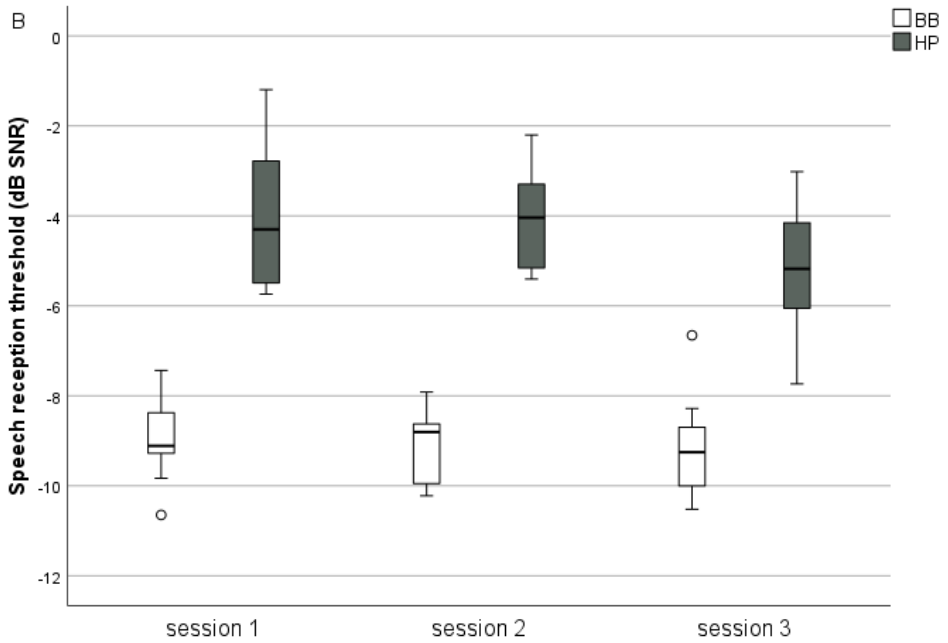


Figure 1: Boxplots illustrating the distribution of SPiQ (A) and SPiN (B) thresholds across sessions. White boxplots represent speech audiometry thresholds for BB-stimuli, while grey boxplots indicate thresholds for HP-stimuli.

One-way repeated measures ANOVA indicated no significant alterations in thresholds between measurement sessions for BB-quiet [$F(2, 24) = 1.549$, $p > 0.05$], nor BB-noise [$F(2, 28) = 0.690$, $p > 0.05$]. In contrast, significant threshold changes were found for HP-quiet [$F(2, 28) = 12.266$, $p < 0.001$], and HP-noise [$F(2, 28) = 7.788$, $p = 0.002$]. Pairwise comparisons unveiled threshold improvements for HP-quiet between each session with non-significant changes of 1.72 dB SPL from session 1 to session 2 ($p > 0.017$) and 1.12 dB SPL from session 2 to session 3 ($p > 0.017$), and a significant change of 2.84 dB SPL from session 1 to session 3 ($p = 0.004$). Furthermore, a significant change of 1.13 dB SNR between session 2 and 3 ($p = 0.002$), and a significant change of 1.21 dB SNR between session 1 and 3 ($p = 0.010$) were found for HP-noise. It should be noted that the initial training procedure before the start of the measurements only incorporated BB-speech in noise. Table II displays ICCs and SEMs for SPiQ and SPiN thresholds, indicating moderate ICCs with highly significant between-subject variability.

Table II. Summary of broadband (BB) and high-pass filtered (HP) speech audiometry threshold in quiet (dB SPL) and in noise (dB SNR) averages per session, along with Intraclass Correlation Coefficients (ICCs) and corresponding 95% Confidence Intervals (CIs), and Standard Error of Measurements (SEMs). P-values for between-subjects variability are reflected as *($0.05 > p > 0.01$), **($0.01 > p > 0.001$), and ***($p < 0.001$).

	Speech audiometry in quiet		Speech audiometry in noise	
	BB	**HP**	**BB**	**HP**
Mean (SD) S1	22.93 (4.011)	29.48 (2.906)	-8.93 (0.806)	-4.01 (1.522)
Mean (SD) S2	23.69 (4.794)	27.76 (1.763)	-9.15 (0.783)	-4.09 (1.091)
Mean (SD) S3	21.93 (4.036)	26.64 (2.166)	-9.22 (1.012)	-5.22 (1.424)
ICC	0.686**	0.666***	0.619**	0.694***
ICC 95%CI	0.204-0.900	0.196-0.878	0.096-0.862	0.286-0.887
SEM	2.395	1.481	0.532	0.797

S1, S2, S3 represent Session 1, Session 2, and Session 3, respectively

DISTORTION PRODUCT OTOACOUSTIC EMISSIONS (DPOAEs)

DP-AMPLITUDES

Per criterium, i.e. SNR ≥ 0, SNR ≥ 2 SD, and SNR ≥ 6 dB, one-way repeated measures ANOVA indicated no significant changes in DP-gram amplitudes ($p > 0.05$) for all tested frequencies. ICCs and SEMs are shown in Table III. Overall, the SNR ≥ 0-criterium showed the highest ICCs (moderate-to-good), followed by the SNR ≥ 6 dB-criterium and the SNR ≥ 2SD-criterium, respectively. The latter criterium is additionally characterized by greater variability among the different tested frequencies. Secondly, remarkably worse ICCs were found for the lower frequencies of 1 and 1.5 kHz. SEMs showed relatively large values overall, with the SNR ≥ 2SD-criterium showing the largest values relative to the other criteria. Figures 2 A, B, and C depict KDE-plots of the zero-criterion, illustrating the distribution of individual test-retest 95% CIs for different measures. A sharp peak in the KDE signifies a more concentrated distribution of test-retest CIs, indicating good overall reliability across the test population. Conversely, a broader peak implies increased variability of individual test-retest

CIs across individuals, reflecting a lower reliability for the corresponding parameter across the population.

Table III. Summary of DP-gram amplitude averages (μV) per session and frequency, alongside Intraclass correlation coefficients (ICCs) and corresponding 95% Confidence Intervals (CIs) ,and Standard error of measurements (SEMs). P-values for between-subjects variability are reflected as *(0.05 > p > 0.01), **(0.01 > p > 0.001), and ***(p < 0.001).

	500 Hz	**700 Hz**	**1 kHz**	**1.5 kHz**	**2 kHz**	**3 kHz**	**4 kHz**	**6 kHz**	**8 kHz**	**9 kHz**	**10 kHz**	**11 kHz**
SNR ≥ 0 criterium												
Mean (SD) S1	2.85	3.67	6.89	8.89	3.30	-0.29	-7.13	-6.80	-6.68	-10.20	-10.34	-6.47
	(5.384)	(7.238)	(7.883)	(6.987)	(8.157)	(8.008)	(6.720)	(5.809)	(6.245)	(5.058)	(6.556)	(6.459)
Mean (SD) S2	1.34	2.89	7.38	9.21	6.03	2.43	-6.03	-5.01	-6.67	-9.163	-6.27	-6.06
	(7.532)	(6.738)	(7.609)	(6.581)	(8.184)	(7.731)	(6.610)	(7.267)	(5.387)	(5.320)	(11.042)	(9.883)
Mean (SD) S3	2.01	1.68	7.82	10.46	5.55	3.61	-4.09	-3.93	-8.785	-9.54	-4.95	-3.79
	(6.291)	(4.524)	(3.779)	(6.724)	(6.498)	(5.543)	(7.522)	(8.206)	(6.371)	(6.816)	(10.011)	(9.353)
ICC	0.745**	0.774***	0.606*	0.507	0.700**	0.757***	0.800***	0.661**	0.788***	0.762**	0.659**	0.687**
ICC 95%CI	0.391-	0.469-	0.034-	-0.206-	0.297-	0.443-	0.538-	0.206-	0.508-	0.428-	0.226-	0.260-
	0.908	0.918	0.859	0.823	0.890	0.910	0.927	0.876	0.922	0.914	0.873	0.886
SEM	3.204	2.943	4.111	4.666	4.148	3.552	3.095	4.133	2.746	2.650	5.525	4.800
SNR ≥ 2SD criterium												
Mean (SD) S1	1.66	-1.16	5.96	8.53	2.68	-1.672	-8.05	-8.41	-8.09	-14.38	-13.77	-10.12
	(6.142)	(9.824)	(9.194)	(7.815)	(9.393)	(10.850)	(8.026)	(7.327)	(8.095)	(7.687)	(8.375)	(9.751)
Mean (SD) S2	-2.74	0.75	5.53	9.21	6.03	2.43	-7.39	-6.67	-10.57	-.12.93	-8.97	-7.969
	(8.221)	(8.565)	(10.348)	(6.581)	(8.184)	(7.731)	(8.537)	(9.513)	(9.196)	(7.663)	(12.646)	(11.359)
Mean (SD) S3	-1.01	-.48	7.25	10.46	4.91	3.61	-4.99	-6.65	-12.14	-12.38	-8.65	-6.2740
	(6.755)	(5.524)	(4.146)	(6.724)	(7.494)	(5.543)	(8.830)	(10.988)	(8.397)	(8.502)	(12.745)	(11.570)
ICC	0.614*	0.678**	0.582*	0.445	0.642*	0.713**	0.833***	0.550*	0.793***	0.622*	0.486	0.740**
ICC 95%CI	0.132-	0.226-	-0.019-	-0.357-	0.162-	0.350-	0.611-	-0.101-	0.521-	0.090-	-0.191-	0.390-
	0.856	0.883	0.850	0.801	0.869	0.893	0.939	0.838	0.924	0.864	0.811	0.905

SEM	4.453	4.557	5.291	5.175	4.978	4.530	3.425	6.184	3.888	4.810	8.190	5.502

<u>SNR ≥ 6dB criterium</u>

Mean (SD) S1	1.82	3.30	6.77	8.53	3.297	-0.29	-7.13	-7.04	-6.94	-11.26	-10.94	-6.94
	(5.524)	(7.746)	(8.073)	(7.815)	(8.157)	(8.008)	(6.720)	(6.205)	(6.574)	(6.273)	(7.386)	(6.977)
Mean (SD) S2	0.17	2.31	7.38	9.21	6.03	2.43	-6.03	-5.30	-6.93	-9.73	-6.83	-7.12
	(8.263)	(7.389)	(7.609)	(6.581)	(8.184)	(7.731)	(6.610)	(7.933)	(5.857)	(6.244)	(11.487)	(10.847)
Mean (SD) S3	1.55	0.89	7.82	10.46	5.55	3.61	-4.30	-4.67	-9.76	-10.557	-5.08	-4.25
	(6.508)	(5.166)	(3.780)	(6.724)	(6.498)	(5.5.43)	(7.835)	(9.030)	(7.350)	(7.150)	(10.152)	(9.866)
ICC	0.660**	0.811***	0.585*	0.445	0.700**	0.757***	0.800***	0.700**	0.741**	0.788***	0.640**	0.644*
ICC 95%CI	0.183-	0.558-	-0.019-	-0.357-	0.297-	0.443-	0.537-	0.289-	0.405-	0.495-	0.188-	0.156-
	0.877	0.931	0.851	0.801	0.890	0.910	0.927	0.891	0.904	0.923	0.866	0.871
SEM	3.931	2.947	4.268	5.175	4.148	3.552	3.136	4.216	3.363	2.969	5.949	5.522

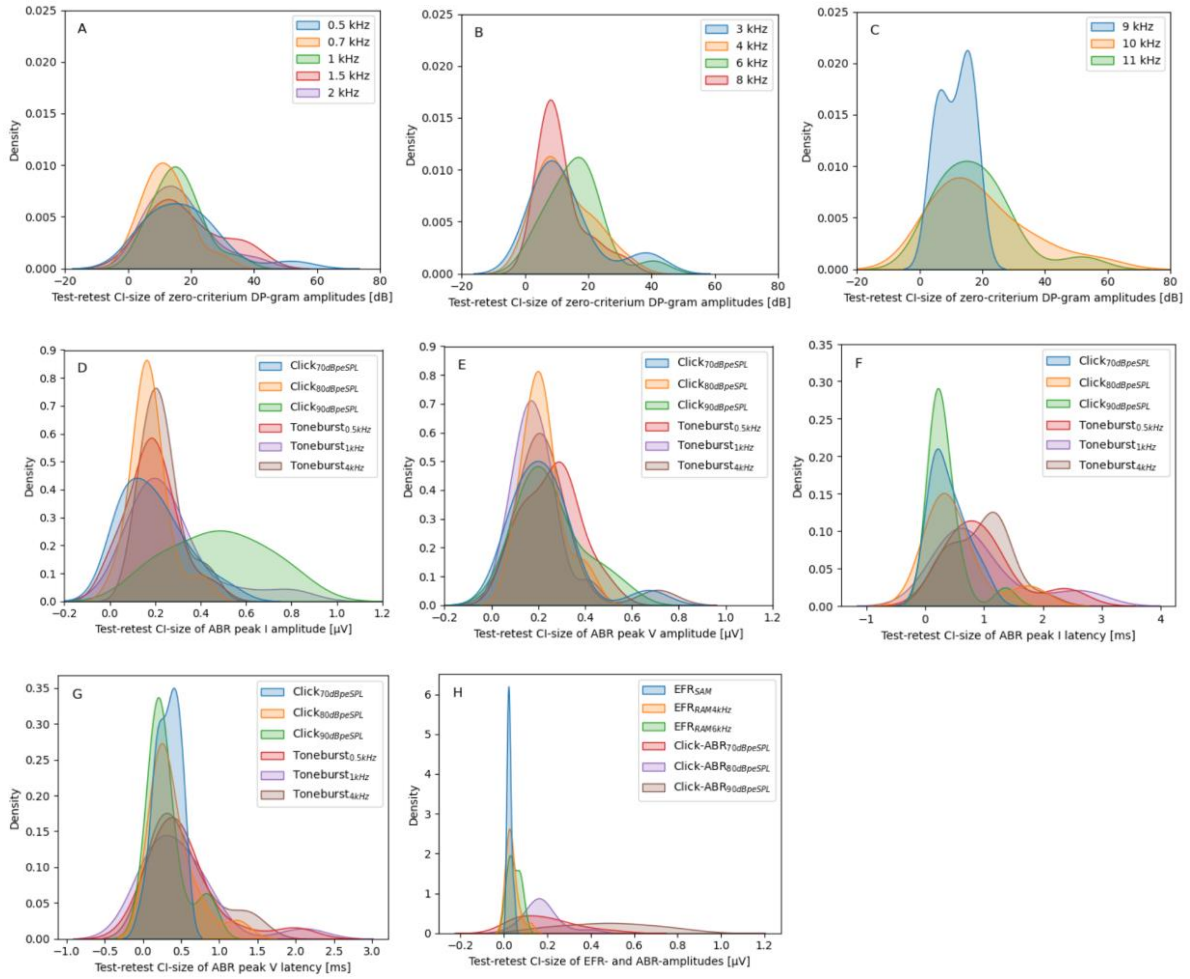


Figure 2: (color online) Kernel Density Estimate plots illustrating the distribution of individual test-retest 95% confidence interval (CIs) for zero-criterion DP-grams across different frequency ranges. Panel A represents the frequency range of 0.5-2 kHz, Panel B shows the range of 3-8 kHz, and Panel C displays the range 9-11. Subsequently, Panels D-G present ABR amplitudes for wave I (D) and wave V (E), along with ABR latencies for wave I (F) and wave V (G). Additionally, panel (H) contrasts the distribution of individual 95%CIs for EFR strengths with click-ABR amplitudes. A sharp peak observed in the KDE indicates a more concentrated distribution of test-retest CIs, suggesting good overall reliability across the tested population. Conversely, a broader peak implies signifies increased variability of individual test-retest CIs across individuals, reflecting a lower reliability for the corresponding parameter across the test population.

DP-THRESHOLDS

One-way repeated measures ANOVA indicated no significant changes ($p > 0.05$) across all tested frequencies, assessed per inclusion criterion. ICCs and SEMs are shown in Table IV. The SNR ≥ 6dB-criterium showed overall the highest ICCs, followed by SNR ≥ 2SD and SNR ≥ 0, retaining both very similar ICCs and SEMs. However, overall very poor ICCs with notably wide corresponding 95%CIs, alongside large SEMs, were observed across all six tested frequencies for DPOAE thresholds.

Table IV. Summary of DP-thresholds amplitude averages (dB SPL) per session and frequency, alongside Intraclass correlation coefficients (ICCs) and corresponding 95% Confidence Intervals (CIs), and Standard error of measurements (SEMs). P-values for between-subjects variability are reflected as *($0.05 > p > 0.01$), **($0.01 > p > 0.001$), and ***($p < 0.001$).

	500 Hz	1000 Hz	2000 Hz	4000 Hz	6000 Hz	8000 Hz
SNR ≥ 0 criterium						
Mean (SD) S1	12.83	13.69	17.83	29.45	25.74	25.65
	(12.635)	(9.107)	(11.308)	(6.256)	(6.868)	(8.190)
Mean (SD) S1	18.14	11.61	16.65	22.84	28.13	21.68
	(11.932)	(9.739)	(11.820)	(12.815)	(6.762)	(8.456)
Mean (SD) S1	13.98	14.41	16.074	28.89	27.78	28.16
	(12.893)	(7.766)	(8.831)	(9.571)	(9.747)	(9.917)
ICC	0.401	0.215	0.258	0.069	0.450	-0.028
ICC 95%CI	-1.265-0.889	-1.982-0.819	-0.1057-0.763	-1.104-0.663	-0.379-0.817	-1.304-0.626
SEM	9.529	7.806	9.089	9.706	5.762	9.186
SNR ≥ 2SD criterium						
Mean (SD) S1	12.83	13.69	19.86	29.45	25.74	25.65
	(12.635)	(12.635)	(8.445)	(6.256)	(6.868)	(8.190)
Mean (SD) S1	18.14	11.61	16.65	22.84	28.126	21.68
	(11.932)	(11.932)	(11.820)	(12.815)	(6.762)	(8.456)
Mean (SD) S1	13.80	14.36	16.22	28.93	28.06	29.21
	(9.917)	(7.748)	(8.785)	(9.590)	(9.677)	(9.841)
ICC	0.401	0.209	0.252	0.075	0.457	0.066
ICC 95%CI	-1.265-0.889	-2.017-0.818	-1.077-0.761	-1.089-0.665	-0.357-0.819	-0.995-0.652
SEM	9.499	7.830	9.113	9.683	5.712	8.856
SNR ≥ 6 dB criterium						
Mean (SD) S1	15.72	14.177	20.50	30.126	27.24	29.66
	(12.208)	(9.677)	(8.799)	(6.824)	(6.629)	(5.045)
Mean (SD) S2	18.36	13.81	20.71	29.12	28.40	25.96
	(11.061)	(11.078)	(11.442)	(7.696)	(7.048)	(9.345)
Mean (SD) S3	19.57	16.36	17.05	27.05	28.57	29.40
	(9.841)	(8.517)	(9.262)	(9.262)	(10.111)	(9.303)
ICC	0.782*	0.630*	0.653*	0.175	0.377	0.090
ICC 95%CI	0.110-0.960	-0.087-0.894	0.147-0.879	-1.119-0.718	-0.643-0.797	-1.190-0.681
SEM	5.163	5.877	6.309	6.313	6.223	7.740

S1, S2, S3 represent Session 1, Session 2, and Session 3, respectively

ICCs and SEMs of DP-grams (μV) and DP-thresholds (dB SPL) , in relation to pure tone thresholds (dB) are illustrated in Figures 3A, B, C, and D.

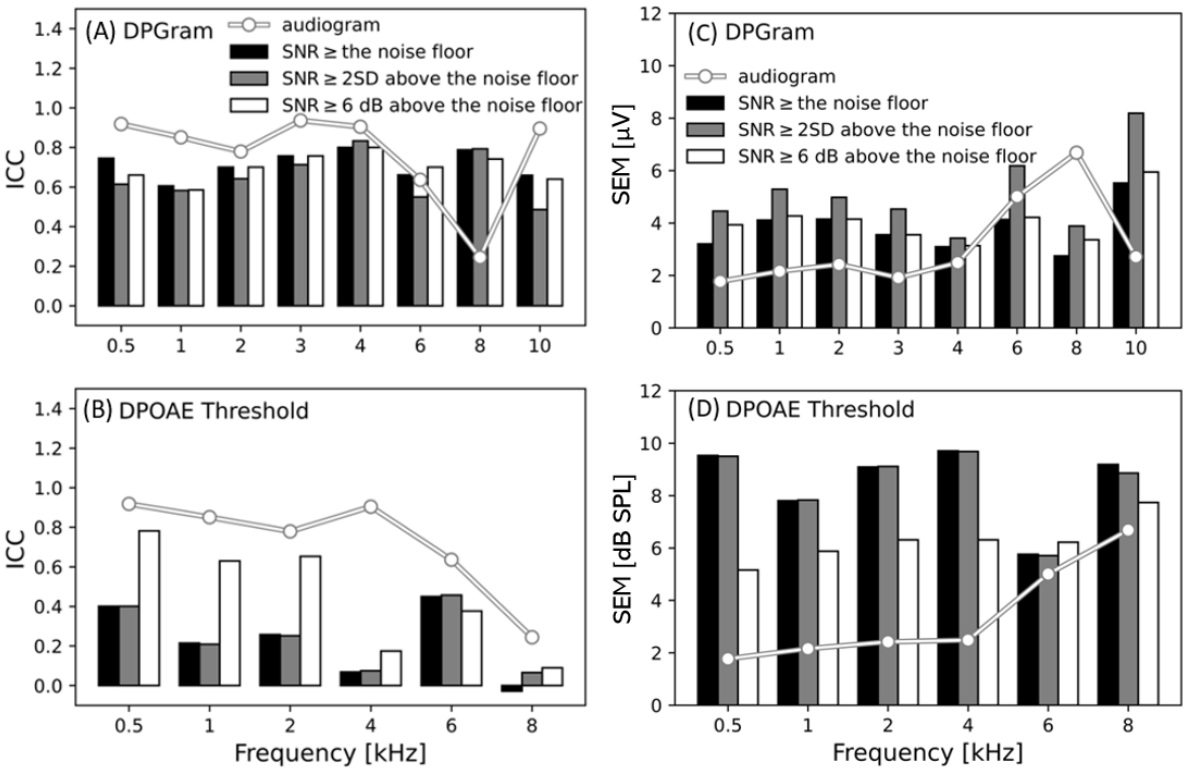


Figure 3A, 3B, 3C, and 3D: Intraclass correlation coefficients (ICCs) (A and B) and standard error of measurements (SEMs) (C and D) of DP-amplitudes (μV) and DP-thresholds (dB SPL) in relation to audiogram thresholds, are illustrated in panels A, B, C and D, respectively. ICCs and SEMs of the three different inclusion criteria regarding signal to noise ratio, i.e. SNR ≥ the noise floor, SNR ≥ 2SD above the noise floor, SNR ≥ 6 dB above the noise floor, are illustrated in black, grey and white, while ICCs and SEMs of pure-tone audiometry are illustrated by a line.

In terms of DP-amplitude ICCs (A), a pattern of generally lower but more consistent outcomes across different frequencies and evaluation criteria was observed compared to audiogram thresholds. Notably, exceptions were observed at 6 kHz and, predominantly, 8 kHz, where DP-amplitudes exhibited better ICCs than audiogram thresholds. This trend corresponded with the DP-amplitude SEMs (C). The DP-thresholds exhibited notably lower ICCs when juxtaposed

with audiogram thresholds (B). Moreover, increased variability across different evaluation criteria was evident, particularly with improved outcomes for the 6 dB criterion. SEMs confirmed less favorable results for DP-thresholds (D), emphasizing the 6 dB criterion as the most reliable in this study population.

AUDITORY EVOKED POTENTIALS (AEP)

AUDITORY BRAINSTEM RESPONSES (ABR)

A one-way repeated measures ANOVA revealed no significant differences in click-ABR amplitudes and latencies of waves I, III, and V at 70 dBpeSPL, 80 dBpeSPL, and 90 dBpeSPL, as well as for TB-stimuli, across measurement sessions ($p > 0.05$). Tables V and VI display Click- and TB-ABR amplitudes and latencies per session and stimulus level, alongside ICCs and SEMs. It is important to highlight that a correction of 1 ms should be applied when comparing the current latencies with those obtained using clinical-grade equipment. Generally, good-to-excellent ICCs with highly significant between-subject variances and small SEMs were observed for click-ABR wave I-, III- and V-latencies. These findings align with click-ABR wave V-amplitudes, showing good ICCs with highly significant between-subject variances, except for the click-ABR 70 dBpeSPL, retaining a moderate ICC. In contrast to wave I- and III-click latencies, wave I- and III-amplitudes showed moderate-to-very poor ICCs, with poor average measures.

Table V. Summary of Click-ABR amplitudes (μV) and latencies (ms) per session and stimulus level (dBpeSPL), alongside intraclass correlation coefficients (ICCs) and standard errors of measurements (SEMs). P-values for between-subjects variability are reflected as *(0.05 > p > 0.01), **(0.01 > p > 0.001), and ***(p < 0.001).

	70 dBpeSPL		**80 dBpeSPL**		**90 dBpeSPL**	
	Latency	Amplitude	Latency	Amplitude	Latency	Amplitude
Wave I						
Mean (SD) S1	3.37 (0.315)	0.06 (0.082)	3.02 (0.444)	0.05 (0.053)	2.75 (0.210)	0.06 (0.078)
Mean (SD) S2	3.40 (0.251)	0.04 (0.084)	3.09 (0.283)	0.07 (0.068)	2.69 (0.233)	0.08 (0.104)
Mean (SD) S3	3.35 (0.340)	0.04 (0.062)	2.99 (0.316)	0.03 (0.080)	2.76 (0.124)	0.05 (0.103)
ICC	0.640*	0.273	0.787***	0.518*	0.755**	0.100
ICC 95%CI	0.120-0.871	-0.788-0.740	0.497-0.922	-0.079-0.820	0.431-0.910	-1.244-0.680
SEM	0.179	0.064	0.161	0.048	0.096	0.090
Wave III						
Mean (SD) S1	5.63 (0.277)	0.11 (0.109)	5.27 (0.258)	0.06 (0.096)	5.00 (0.227)	0.13 (0.106)
Mean (SD) S2	5.68 (0.353)	0.12 (0.052)	5.25 (0.323)	0.11 (0.096)	4.96 (0.241)	0.12 (0.127)
Mean (SD) S3	5.66 (0.350)	0.07 (0.084)	5.31 (0.333)	0.09 (0.093)	4.89 (0.209)	0.09 (0.109)
ICC	0.967***	0.287	0.968***	0.647**	0.703**	0.552*
ICC 95%CI	0.923-0.988	-0.594-0.733	0.925-0.988	0.200-0.869	0.316-0.891	-0.060-0.837
SEM	0.058	0.073	0.054	0.057	0.123	0.076
Wave V						
Mean (SD) S1	7.43 (0.394)	0.26 (0.076)	7.17 (0.308)	0.31 (0.117)	6.86 (0.309)	0.32 (0.135)
Mean (SD) S2	7.46 (0.346)	0.24 (0.118)	7.05 (0.325)	0.27 (0.123)	6.93 (0.279)	0.27 (0.137)
Mean (SD) S3	7.47 (0.346)	0.30 (0.118)	7.13 (0.306)	0.32 (0.148)	6.83 (0.264)	0.37 (0.120)
ICC	0.969***	0.691**	0.916***	0.880***	0.937***	0.821***
ICC 95%CI	0.926-0.989	0.299-0.885	0.802-0.970	0.720-0.956	0.850-0.977	0.556-0.936
SEM	0.062	0.059	0.090	0.045	0.070	0.057

S1, S2, S3 represent Session 1, Session 2, and Session 3, respectively

Table VI. Summary of TB-ABR amplitudes (μV) and latencies (ms) per session and frequency (Hz), alongside intraclass correlation coefficients (ICCs) and standard errors of measurements (SEMs). P-values for between-subjects variability are reflected as *(0.05 > p > 0.01), **(0.01 > p > 0.001), and ***(p < 0.001).

	0.5 kHz		**1 kHz**		**4 kHz**	
	Latency	Amplitude	Latency	Amplitude	Latency	Amplitude
Wave I						
Mean (SD) S1	3.86 (0.620)	0.08 (0.064)	3.89 (0.597)	0.03 (0.074)	3.31 (0.572)	0.04 (0.069)
Mean (SD) S2	3.76 (0.595)	0.05 (0.098)	3.76 (0.781)	0.04 (0.056)	3.42 (0.502)	0.03 (0.056)
Mean (SD) S3	3.73 (0.564)	0.05 (0.065)	3.71 (0.581)	0.05 (0.083)	3.53 (0.448)	0.01 (0.084)
ICC	0.841***	0.524*	0.859***	0.241	0.824***	0.044
ICC CI	0.623-0.942	-0.134-0.827	0.659-0.951	-0.931-0.740	0.593-0.935	-0.968-0.626
SEM	0.232	0.053	0.244	0.061	0.214	0.072
Wave III						
Mean (SD) S1	6.53 (0.714)	0.06 (0.059)	6.42 (0.406)	0.08 (0.118)	5.91 (0.452)	0.09 (0.082)
Mean (SD) S2	6.35 (0.710)	0.05 (0.0778)	6.21 (0.461)	0.04 (0.081)	5.70 (0.431)	0.09 (0.104)
Mean (SD) S3	6.40 (0.666)	0.07 (0.067)	6.34 (0.397)	0.05 (0.109)	5.80 (0.501)	0.10 (0.103)
ICC	0.843***	0.603*	0.636*	0.331	0.876***	0.687**
ICC CI	0.631-0.943	0.122-0.851	0.160-0.870	-0.543-0.761	0.709-0.955	0.288-0.884
SEM	0.272	0.044	0.255	0.086	0.162	0.054
Wave V						
Mean (SD) S1	8.56 (0.451)	0.20 (0.086)	8.16 (0.402)	0.21 (0.081)	7.72 (0.273)	0.23 (0.101)
Mean (SD) S2	8.41 (0.439)	0.21 (0.096)	8.06 (0.401)	0.25 (0.100)	7.55 (0.327)	0.26 (0.100)
Mean (SD) S3	8.53 (0.459)	0.23 (0.072)	8.09 (0.339)	0.25 (0.091)	7.64 (0.180)	0.27 (0.129)
ICC	0.910*	0.601**	0.836***	0.778***	0.729***	0.679**
ICC CI	0.790-0.967	0.112-0.844	0.603-0.943	0.468-0.922	0.385-0.900	0.241-0.883
SEM	0.133	0.053	0.152	0.043	0.140	0.062

S1, S2, S3 represent Session 1, Session 2, and Session 3, respectively

TB-amplitudes generally exhibited slightly lower ICCs and higher SEMs compared to clicks. Similar to click-stimuli findings, wave I amplitudes exhibited moderate-to-very poor ICCs. For latencies, moderate-to-good ICCs with highly significant between-subject variances ($p < .001$) and small SEMs were observed for waves I- and III- and V-latencies. Figures 2 D, E, F and G display KDE-plots of wave I and V ABR-amplitudes, and -latencies, representing the distribution of individual test-retest 95%CIs for different measures.

ENVELOPE FOLLOWING RESPONSES (EFR)

Figure 4 depicts the EFR-strength distribution for SAM- and RAM-stimuli across the three consecutive sessions.

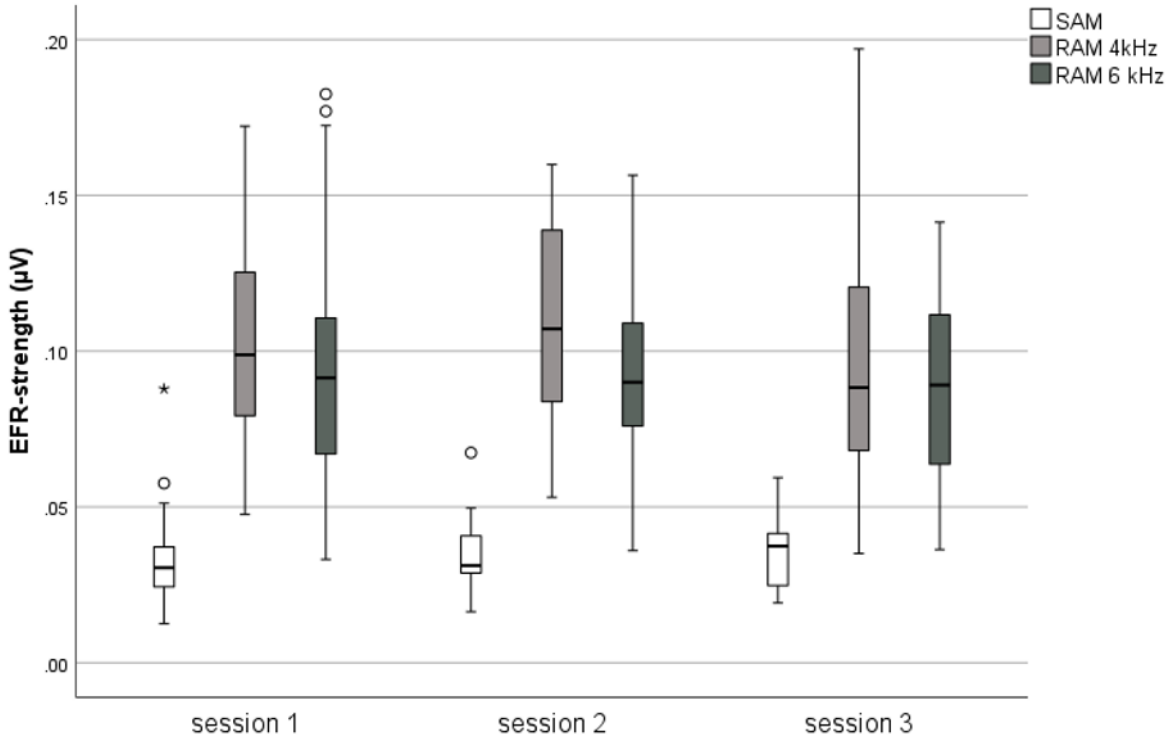


Figure 4: Boxplots presenting the distribution of EFR-strengths across sessions, with white, light grey, and dark grey denoting sinusoidally-amplitude modulated (SAM)-, 4 kHz Rectangularly-amplitude modulated (RAM)-, and 6 kHz RAM-EFR strengths, respectively.

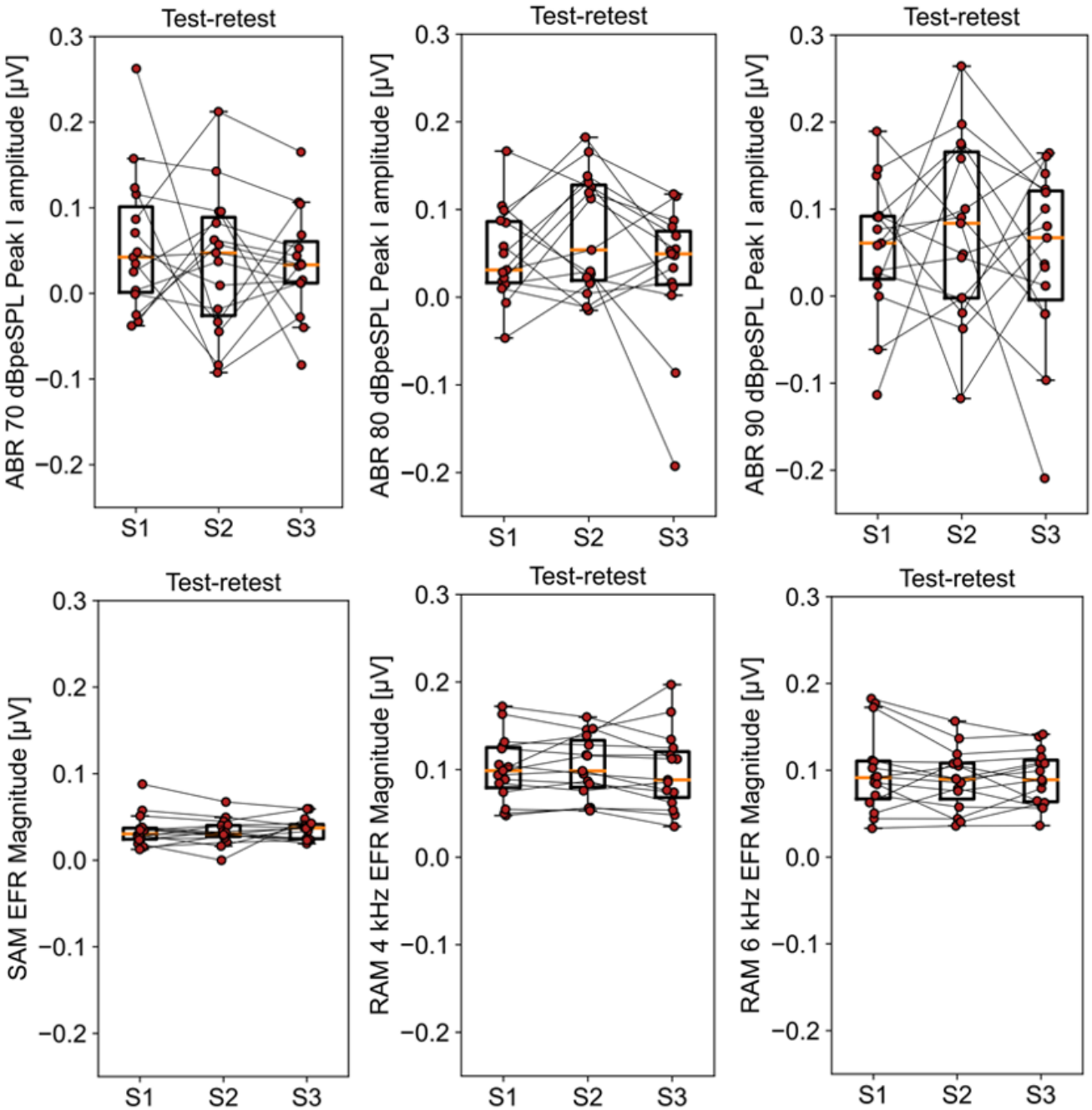
Consistent with prior findings (Vasilkov et al., 2021), EFRs were stronger for the RAM stimuli than the SAM stimulus. An outlier identified in the SAM-evoked response during session two was excluded due to potential data corruption caused by a 50 Hz noise from nearby electrical equipment. No significant changes in EFR-strength were found between measurements for the SAM-stimulus [$F(2, 26) = 0.066$, $p > 0.05$], and RAM-stimuli; i.e. RAM 4 kHz [$F(2, 28) = 0.383$, $p > 0.05$] and RAM 6 kHz [$F(2, 28) = 1.299$, $p > 0.05$]. Moreover, the ICC demonstrated excellent test-retest reliability for both the SAM and RAM stimuli at 4 kHz and 6 kHz, yielding average measures of 0.882 (95% BI [0.708;0.959]; $F(13, 26) = 7.975$, $p < 0.001$), 0.950 (95% BI [0.883;0.982]; $F(14, 28) = 19.355$, $p < 0.001$) and 0.930 (95% BI [0.837;0.974]; $F(14, 28) =14.553$, $p < 0.001$), respectively. Table 7 provides detailed information on ICCs and SEMs.

Table 7. Summary of EFR-strengths (μV) per session and stimulus-parameter along with intraclass correlation coefficients (ICCs), and corresponding 95% Confidence Intervals (CIs), and standard errors of measurements (SEMs). P-values for between-subjects variability are reflected as *(0.05 > p > 0.01), **(0.01 > p > 0.001), and ***(p < 0.001).

	SAM	**RAM 4 kHz**	**RAM 6 kHz**
Mean (SD) S1	0.03 (0.019)	0.10 (0.038)	0.10 (0.047)
Mean (SD) S2	0.03 (0.015)	0.10 (0.035)	0.09 (0.035)
Mean (SD) S3	0.04 (0.012)	0.010 (0.045)	0.09 (0.032)
ICC	0.882***	0.950***	0.930***
ICC 95%CI	0.644-0.946	0.883-0.982	0.837-0.974
SEM	0.005	0.009	0.010

S1, S2, S3 represent Session 1, Session 2, and Session 3, respectively

To provide a more comprehensive perspective on observed variations in ABR- and EFR-magnitudes, both considered as potential EEG-markers of CS, Figure 5 displays individual strengths and distribution boxplots.



S1, S2, S3 represent Session 1, Session 2, and Session 3, respectively.

Figure 5: Individual ABR-amplitudes and distribution plots for click-ABRs at 70 dBpeSPL, 80 dBpeSPL, and 90 dBpeSPL (row 1), along with individual EFR-magnitudes and distribution boxplots for EFR SAM, RAM 4 kHz, and RAM 6 kHz (row 2). Horizontal lines within the boxplots denote the median ABR-amplitudes and EFR-magnitudes.

Additionally, Figure 2H presents KDE-plots for both parameters, showing the distribution of individual 95% CIs computed across three sessions. Both figures highlight the superior reliability of EFR-magnitudes over ABR-amplitudes. 95%CIs of the ICCs further confirm the higher reliability of RAM-EFRs compared to ABR-amplitudes, as their CIs demonstrate non-overlapping ranges.

DISCUSSION

VALIDATING THE HIGH RELIABILITY OF PURE-TONE AUDIOMETRY WITH CONSIDERATION FOR FREQUENCY-SPECIFIC VARIATIONS

Prior research has recommended using a frequency range up to 14 kHz for monitoring purposes (Rodríguez Valiente et al., 2014), as auditory thresholds beyond this frequency show substantial intra-subject threshold variability (Frank, 2001; Schmuziger et al., 2007). While the current study validated good-to-excellent ICCs for both conventional and extended high frequencies, frequency-specific disparities should be taken into consideration within clinical practice since thresholds at 6 kHz, 8kHz, and 20 kHz retained moderate-to-poor ICCs and considerably larger corresponding 95%CIs. The higher variability at 6 and 8 kHz aligns with the findings of Schlauch and Carney (2011), and may be linked to suboptimal earphone positioning or calibration methods. The increased variability at 20 kHz is likely due to standing waves, suggesting that extending the frequency range up to 20 kHz is not advisable for monitoring purposes.

The good test-retest reliability of (high-frequency) audiometry is corroborated by studies conducted by Swanepoel et al. (2010) and Ishak et al. (2011), as well as by several other

investigations that employed diverse transducer models (Fausti et al., 1998; Frank, 1990, 2001; Frank & Dreisbach, 1991; Schmuziger et al., 2004).

The significant difference in 250 Hz thresholds revealed by the repeated measures ANOVA, prompts contemplation regarding the potential influence of mask-wearing on response accuracy. This is especially pertinent in lower frequency assessments, where participants reported conspicuous interference stemming from the act of breathing while wearing face masks, notably affecting their perception of very low bass tones. The observed enhancement in results during session three suggests a potential adaptation among subjects to the masking effects induced by mask-wearing.

NEED FOR A RELIABLE SPiQ-AND SPiN TEST.

SPiQ- and SPiN-tests showed no significant differences between sessions for the BB-lists, while significant threshold changes were found for the HP-lists in quiet and in noise. This learning effect was previously documented by Luts et al. (2014), highlighting a large threshold decrease occurring between the first and the second measurement which decreased to a value below 1 dB after the second list. Similar trends were observed for all language-specific tests covered in the review paper by Kollmeier et al. (2015), suggesting that the training effect might be associated more with the nature of the task and test structure rather than language-specific characteristics. However, in the current study, a training effect was observed despite the provision of two training lists. Firstly, the inclusion of two HP filtered training lists might have counteracted learning, as only BB training lists were intended. However, presenting multiple training lists extends the test duration, affecting subjects' attention span and potentially influencing outcomes. Nevertheless, Vande Maele et al. (2021) reported significant SNR-improvements in all tested conditions, including BB, even with two BB-training lists provided. Secondly, a closed-set test format might contribute to the learning effect, as subjects could more easily learn words when both heard and visualized. The intention of displaying the possible words was to mitigate potential performances

improvement across sessions, as subjects are aware of the words they may encounter from the outset. Nonetheless, the review paper of Kollmeier et al. (2015) reported a training effect in both open- and closed-set test formats for each language examined. Thirdly, within this study, participants were directed to respond in a forced-choice format to mitigate the learning effect. This was prompted by the observation that during the initial session, subjects frequently signaled non-detection of the word more swiftly, yet exhibited increasing confidence in subsequent measurements. This behavior might lead to speculation and potentially improved performance in subsequent sessions, which might indirectly contribute to better results. In sum, although further investigation is warranted, given the small study population and the potential for type I errors in the current study, caution is advised in using the matrix test in repeated measures or monitoring, due to the potential influence of learning effects.

DPOAEs EXHIBIT LARGE VARIABILITY.

DP-AMPLITUDES

One-way-repeated measures ANOVA indicated no significant changes in DP-amplitudes between the measurements for all tested criteria and frequencies. These findings algin with a study conducted in 2010, reporting no significant differences within time-intervals up to 60 minutes (Keppler et al., 2010). However, when the time-interval extended to 7 days, a significant difference was noted, suggesting decreased reliability of DP-grams with increased time intervals (Keppler et al., 2010). Engdahl et al. (1994) and Wagner et al. (2008) noted that prolonged time intervals lead to increased standard deviations due to greater variation in middle ear pressure, room- and biological noise. Probe refitting at each session on different days further contributes to variability, as indicated by research highlighting the impact of probe replacement on the level of background noise and acoustic leakage (Beattie & Bleech, 2000; Beattie et al., 2003; Franklin et al., 1992; Keppler et al., 2010; Mills et al., 2007; Wagner et al., 2008; Zhao & Stephens, 1999).

SEM values in this study generally surpassed those reported in other studies (Beattie et al., 2003; Franklin et al., 1992; Keppler et al., 2010; Ng & Mcpherson, 2005; Wagner et al., 2008). Various factors could account for this increased variability. Firstly, subject-generated noise has the potential to impact DP-gram amplitudes, introducing variability due to differences in patient cooperation and ear canal acoustics (Keppler et al., 2010; Wagner et al., 2008). Secondly, equipment-related noise (Keppler et al., 2010) and recording parameters could significantly influence DP-gram amplitudes. Franklin et al. (1992) demonstrated lower reliability for DPOAE amplitudes elicited using lower primary tone level combinations (L1/L2 = 65/55 dB SPL), as used in this study, compared to higher primary level combinations (L1/L2 = 75/70 dB SPL). Another investigation, assessing two stimulus protocols, consistently found larger absolute DP-amplitudes for higher primary intensities, with greater variability for lower primary level combinations relative to higher intensities (Hall, 2000). Thirdly, in this study, both the positioning of the probe at each session and the positioning of the earmuffs might have influenced the outcomes. Lastly, analysis strategies vary significantly across studies. In contrast to previous research, this study did not exclude responses that did not meet inclusion criteria. Instead, amplitudes were adjusted to the minimum level (i.e. the noise floor level), resulting in larger standard deviations, and consequently, larger SEMs. Additionally, the reliability varies with different inclusion criteria, as discussed earlier, and across frequencies, as depicted in Table III. The notably higher standard deviations observed at frequencies 1.0 and 1.5 kHz, compared to others, are likely attributed to low-frequency noise contaminations (Beattie et al., 2003; Keppler et al., 2010; Wagner et al., 2008; Zhao & Stephens, 1999). The increased variability in DP-gram amplitudes at higher frequencies is probably caused by ear-canal acoustics, particularly standing waves, amplifying intrinsic variability due to differences in sound pressure at the tympanic membrane and probe microphone (Keppler et al., 2010; Mills et al., 2007).

DP-THRESHOLDS

Prior research has indicated the potential of estimated DPOAE thresholds to predict pure-tone thresholds (Boege & Janssen, 2002; Goldman et al., 2006; Gorga et al., 2003). However, there is a scarcity of studies investigating test-retest reliability. In the present study, a one-way repeated measures ANOVA indicated no significant alterations in DP-thresholds across measurements for all examined criteria and frequencies. Despite this, ICCs and SEMs produced notably poor outcomes, indicating a low level of reliability in this study. The highly variable nature of the I/O function among subjects and even for different stimuli at different frequencies, as highlighted by Kimberley and Nelson (1989), Hall (2000), and Harris (1990), questions the clinical utility of this method. Additionally, Harris (1990) emphasized the need for strict minimum noise requirements for reliable responses, recommending measures such as conducting DPOAE measurements in a sound-attenuating booth, setting test protocol stopping criteria for a very low noise level, and employing continuous signal averaging until the minimum noise level is reached. Popelka et al. (1993) reported that achieving a noise floor of -40 dB for recording a single I/O function may take up to 45 minutes of testing. In conclusion, our study, in conjunction with the findings presented by Popelka et al. (1993), raises concerns regarding the clinical efficacy of DP-thresholds in their present state.

EFR MEASURES YIELD BETTER TEST-RETEST RELIABILITY, RELATIVE TO THE ABR.

The present study showed no significant changes for waves I, III and V between the three different test sessions. These results align with a 2018 study, showing no significant changes in click- and speech-evoked brainstem responses across test sessions (Bidelman et al., 2018). Additionally, Munjal et al. (2016), evaluating the reliability of the absolute latency of waves I, III and V and interpeak latencies, showed good test-retest reliability for all response parameters, except for the absolute latency of wave I.

The analyses of ICCs and SEMs unveiled several trends among the different waves and stimuli. Firstly, a higher within-subject reliability was found for wave V relative to wave I. These results are in line with the study of Lauter and Karzon (1990), who reported low level of consistency across subjects for wave I of ABR. Sininger and Cone-Wesson (2002) have shown that peripheral hearing and testing parameters, amongst others ambient noise and minimal wax in the external auditory canal, can affect the latency of wave I in ABR-measurements. Secondly, wave I may potentially be reduced due to CS and OHC-damage, while wave V may be enhanced due to central gain mechanisms (Auerbach et al., 2014; Gu et al., 2012; Schaette & McAlpine, 2011). Therefore, the interpretation of wave I-amplitudes and latencies requires some caution.

In addition to greater reliability for wave V relative to wave I, the present study also retained smaller ICCs and larger SEMs for click-amplitudes compared to click-latencies, consistent with Bidelman et al. (2018). Negligibly small intra-subject variability in ABR latencies are in addition in agreement with previous studies of Edwards et al. (1982) and Oyler et al. (1991). Firstly, evoked potential amplitudes are susceptible to nonbiological factors, such as electrode impedance and orientation relative to source generators. This suggest that the amplitude might be a poor metric for reliably assessing subtle changes in ABR-measurements with certain experimental manipulations, including noise exposure, ototoxicity, age, and training (Bidelman et al., 2018). The use of ear canal tiptrodes, as opposed to scalp mounted electrodes could result in higher reliability since the recording site has moved closer to the generator of wave I, specifically the auditory nerve. This assumption aligns with the study of Bauch and Olsen (1990), showing increased wave I-amplitudes with ear canal tiptrodes compared to mastoid electrodes, and Bieber et al. (2020), reporting good-to-excellent wave I and wave V amplitude ICCs when measured from the ear canal. However, the study of Prendergast et al. (2018) demonstrated only a small increase in reliability for waves I and V when using canal tiptrodes compared to mastoid electrodes. The benefits for the summation potential however, were greater. In sum, while wave

I has proved valuable, particularly in research studies, as a more direct measure of peripheral auditory function, our study revealed low amplitude ICCs, casting doubt on the feasibility of clinical waveform interpretation under the specified conditions. Prior investigations with good-to-excellent test-retest reliability often extended their test durations, incorporating up to 10000 sweeps and/or or automated peak- and trough-picking procedures (Bieber et al., 2020; Guest et al., 2019; Prendergast et al., 2018), suggesting considerable advantages of automated peak-picking algorithms in clinical practice. Moreover, stimulus levels were frequently elevated, reaching 115.5 dB peSPL in the study by Prendergast et al. (2018). Finally, it is essential to acknowledge that the relatively modest size of our study population and the testing environment within hospital settings could have influenced the observed results.

When comparing broadband clicks to toneburst stimuli in our study, it is evident that clicks evoke larger responses and serve as slightly more reliable biomarkers. This is likely because broadband clicks activate more auditory nerve fibers simultaneously, potentially resulting in larger and more robust responses. Subsequently, TBs are identified as less clearly detectable peaks, indicating higher interrater variability and therefore lower overall reliability. This hypothesis is supported by the generally smaller amplitudes and longer latencies in TB-responses, which likely stem from the narrower basilar-membrane stimulation (Gorga et al., 1988; Rasetshwane et al., 2013). Additionally, the impact of prolonged latencies and reduced amplitudes with narrower BM stimulation is more pronounced at lower frequencies compared to higher frequencies (500 Hz vs 4 kHz), primarily due to cochlear wave dispersion (Rasetshwane et al., 2013).

As noise-induced CS primarily targets AN fibers with high thresholds, and phase locking to temporal envelopes is in addition particularly strong in these fibers, the EFR-strength could potentially be a more robust measure, relative to ABR-amplitudes (Vasilkov et al., 2021). Additionally, phase information can be extracted from EFRs, and measures of phase-locking values might be less susceptible to anatomical variations in humans (Gorga et al., 1988),

generally interfering amplitude measures. These findings align with Bidelman et al. (2018), who reported that Frequency Following Responses generally showed higher test-retest reliability compared to conventional click-evoked ABRs. Furthermore, our study's results, which indicated favorable ICCs and non-overlapping corresponding 95%CIs between ABR-amplitudes and RAM-EFR strengths, align with this observed trend. However, Guest et al. (2019) reported only minor differences in reliability between ABR-amplitudes and EFR strengths. Although their experimental setup resembled ours in terms of the number of presentations (ranging from 5200-56000 versus 4000) and click levels (90, 96, and 102 dB peSPL versus 70, 80 and 90 dBpeSPL), they employed automated peak-picking algorithms to select peaks, which likely improved to reliability of the ABR results.

CONCLUSION

In the pursuit of identifying noninvasive early markers of noise-induced SNHL in humans, various measures have been explored. However, comprehensive studies evaluating the test-retest reliability of multiple measures and stimuli within a single study remain limited, and a standardized clinical protocol encompassing robust noninvasive early markers of SNHL has not yet been established. Addressing these gaps, this exploratory study aimed to explore the intra-subject variability of various potential noninvasive EEG-biomarkers of CS and other early indicators of SNHL within the same individuals. While pure-tone audiometry has confirmed good reliability in this study, caution is advised when extending the frequency range beyond 16 kHz. The observed learning effect in the speech-sentence test emphasizes the need for caution when employing the matrix sentence test in repeated measurements. The variability noted in DPOAEs highlights the importance of consistent ear probe replacement, meticulous measurement techniques, and optimal testing conditions to minimize variability in DP-amplitudes, and suggests that DP-thresholds may not be effective for clinical application. Regarding auditory evoked

potentials, the study found that EFRs exhibited greater reliability compared to ABRs when manually selecting the ABR-waveforms.

REFERENCES

Auerbach, B. D., Rodrigues, P. V., & Salvi, R. J. (2014). Central gain control in tinnitus and hyperacusis. *Frontiers in Neurology*, *5*, 206.

Bauch, C. D., & Olsen, W. O. (1990). Comparison of ABR amplitudes with TIPtrode™ and mastoid electrodes. *Ear and Hearing*, *11*(6), 463-467.

Beattie, R., & Bleech, J. (2000). Effects of sample size on the reliability of noise floor and DPOAE. *British Journal of Audiology*, *34*(5), 305-309.

Beattie, R. C., Kenworthy, O., & Luna, C. A. (2003). Immediate and short-term reliability of distortion-product otoacoustic emissions: Confiabilidad inmediata ya corto plazo de las emisiones otoacústicas por productos de distorsión. *International Journal of Audiology*, *42*(6), 348-354.

Bharadwaj, D., Verhulst, S., Shaheen, L., Liberman, M. C., & Shinn-Cunningham, B. G. (2014). Cochlear neuropathy and the coding of supra-threshold sound. *Frontiers in Systems Neuroscience*, *8*, 26. https://doi.org/10.3389/fnsys.2014.00026

Bharadwaj, H. M., Masud, S., Mehraei, G., Verhulst, S., & Shinn-Cunningham, B. G. (2015a). Individual differences reveal correlates of hidden hearing deficits. *Journal of Neuroscience*, *35*(5), 2161-2172.

Bharadwaj, H. M., Masud, S., Mehraei, G., Verhulst, S., & Shinn-Cunningham, B. G. (2015b). Individual differences reveal correlates of hidden hearing deficits. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *35*(5), 2161-2172. https://doi.org/10.1523/JNEUROSCI.3915-14.2015

Bidelman, G. M., Pousson, M., Dugas, C., & Fehrenbach, A. (2018). Test–retest reliability of dual-recorded brainstem versus cortical auditory-evoked potentials to speech. *Journal of the American Academy of Audiology*, *29*(02), 164-174.

Bieber, R. E., Fernandez, K., Zalewski, C., Cheng, H., & Brewer, C. C. (2020). Stability of early auditory evoked potential components over extended test-retest intervals in young adults. *Ear and Hearing*, *41*(6), 1461.

Boege, P., & Janssen, T. (2002). Pure-tone threshold estimation from extrapolated distortion product otoacoustic emission I/O-functions in normal and cochlear hearing loss ears. *The Journal of the Acoustical Society of America*, *111*(4), 1810-1818. https://www.ncbi.nlm.nih.gov/pubmed/12002865

Bourien, J., Tang, Y., Batrel, C., Huet, A., Lenoir, M., Ladrech, S., Desmadryl, G., Nouvian, R., Puel, J. L., & Wang, J. (2014). Contribution of auditory nerve fibers to compound action potential of the auditory nerve. *Journal of Neurophysiology*, *112*(5), 1025-1039. https://doi.org/10.1152/jn.00738.2013

Coradini, P. P., Cigana, L., Selistre, S. G., Rosito, L. S., & Brunetto, A. L. (2007). Ototoxicity from cisplatin therapy in childhood cancer. *Journal of Pediatric Hematology/Oncology*, *29*(6), 355-360.

Cruickshanks, K. J., Nondahl, D. M., Tweed, T. S., Wiley, T. L., Klein, B. E., Klein, R., Chappell, R., Dalton, D. S., & Nash, S. D. (2010). Education, occupation, noise exposure history and the 10-yr cumulative incidence of hearing impairment in older adults. *Hearing Research*, *264*(1-2), 3-9.

Edwards, R. M., Buchwald, J. S., Tanguay, P. E., & Schwafel, J. A. (1982). Sources of variability in auditory brain stem evoked potential measures over time. *Electroencephalography and Clinical Neurophysiology*, *53*(2), 125-132.

Encina-Llamas, G., Harte, J. M., Dau, T., Shinn-Cunningham, B., & Epp, B. (2019). Investigating the Effect of Cochlear Synaptopathy on Envelope Following Responses Using a Model of the Auditory Nerve. *Journal of the Association for Research in Otolaryngology : JARO*, *20*(4), 363-382. https://doi.org/10.1007/s10162-019-00721-7

Engdahl, B., Arnesen, A. R., & Mair, I. W. (1994). Reproducibility and short-term variability of transient evoked otoacoustic emissions. *Scandinavian Audiology*, *23*(2), 99-104.

Fausti, S. A., Henry, J. A., Hayden, D., Phillips, D. S., & Frey, R. H. (1998). Intrasubject reliability of high-frequency (9-14 kHz) thresholds: tested separately vs. following conventional-frequency testing. *Journal of the American Academy of Audiology*, *9*(2).

Fernandez, K. A., Guo, D., Micucci, S., De Gruttola, V., Liberman, M. C., & Kujawa, S. G. (2020). Noise-induced cochlear synaptopathy with and without sensory cell loss. *Neuroscience*, *427*, 43-57.

Francart, T., van Wieringen, A., & Wouters, J. (2008). APEX 3: a multi-purpose test platform for auditory psychophysical experiments. *Journal of Neuroscience Methods*, *172*(2), 283-293. https://doi.org/10.1016/j.jneumeth.2008.04.020

Frank, T. (1990). High-frequency hearing thresholds in young adults using a commercially available audiometer. *Ear and Hearing*, *11*(6), 450-454.

Frank, T. (2001). High-frequency (8 to 16 kHz) reference thresholds and intrasubject threshold variability relative to ototoxicity criteria using a Sennheiser HDA 200 earphone. *Ear and Hearing*, *22*(2), 161-168.

Frank, T., & Dreisbach, L. E. (1991). Repeatability of high-frequency thresholds. *Ear and Hearing*, *12*(4), 294-295.

Franklin, D. J., McCoy, M. J., Martin, G. K., & Lonsbury-Martin, B. L. (1992). Test/retest reliability of distortion-product and transiently evoked otoacoustic emissions. *Ear and Hearing*, *13*(6), 417-429.

Furman, A. C., Kujawa, S. G., & Liberman, M. C. (2013). Noise-induced cochlear neuropathy is selective for fibers with low spontaneous rates. *Journal of Neurophysiology*, *110*(3), 577-586. https://doi.org/10.1152/jn.00164.2013

Garrett, M. (2020). *Degradation of auditory processing and perception with age: The role of near and supra-threshold sensorineural hearing deficits* Carl von Ossietzky Universität Oldenburg. Medizinische Physik].

Glavin, C. C., Siegel, J., & Dhar, S. (2021). Distortion product otoacoustic emission (DPOAE) growth in aging ears with clinically normal behavioral thresholds. *Journal of the Association for Research in Otolaryngology*, *22*, 659-680.

Goldman, B., Sheppard, L., Kujawa, S. G., & Seixas, N. S. (2006). Modeling distortion product otoacoustic emission input/output functions using segmented regression. *The Journal of the Acoustical Society of America*, *120*(5), 2764-2776.

Gorga, M. P., Kaminski, J. R., Beauchaine, K. A., & Jesteadt, W. (1988). Auditory brainstem responses to tone bursts in normally hearing subjects. *Journal of Speech, Language, and Hearing Research*, *31*(1), 87-97.

Gorga, M. P., Neely, S. T., Dorn, P. A., & Hoover, B. M. (2003). Further efforts to predict pure-tone thresholds from distortion product otoacoustic emission input/output functions. *The Journal of the Acoustical Society of America*, *113*(6), 3275-3284.

Grinn, S. K., Wiseman, K. B., Baker, J. A., & Le Prell, C. G. (2017). Hidden Hearing Loss? No Effect of Common Recreational Noise Exposure on Cochlear Nerve Response Amplitude in Humans. *Frontiers in Neuroscience*, *11*, 465. https://doi.org/10.3389/fnins.2017.00465

Gu, J. W., Herrmann, B. S., Levine, R. A., & Melcher, J. R. (2012). Brainstem auditory evoked potentials suggest a role for the ventral cochlear nucleus in tinnitus. *Journal of the Association for Research in Otolaryngology*, *13*, 819-833.

Guest, H., Munro, K. J., & Plack, C. J. (2018). Acoustic Middle-Ear-Muscle-Reflex Thresholds in Humans with Normal Audiograms: No Relations to Tinnitus, Speech Perception in Noise, or Noise Exposure. *Neuroscience*. https://doi.org/10.1016/j.neuroscience.2018.12.019

Guest, H., Munro, K. J., Prendergast, G., Howe, S., & Plack, C. J. (2017). Tinnitus with a normal audiogram: Relation to noise exposure but no evidence for cochlear synaptopathy [Article]. *Hearing Research*, *344*, 265-274. https://doi.org/10.1016/j.heares.2016.12.002

Guest, H., Munro, K. J., Prendergast, G., Millman, R. E., & Plack, C. J. (2018). Impaired speech perception in noise with a normal audiogram: No evidence for cochlear synaptopathy and no relation to lifetime noise exposure. *Hearing Research*, *364*, 142-151.

Guest, H., Munro, K. J., Prendergast, G., & Plack, C. J. (2019). Reliability and interrelations of seven proxy measures of cochlear synaptopathy. *Hearing Research*. https://doi.org/10.1016/j.heares.2019.01.018

Hall, J. (2000). Distortion product and transient evoked OAEs: Nonpathologic factors influencing measurement. Handbook of Otoacoustic Emissions. In: Singular Publishing Group. Cengage Learning, San Diego.

Harris, F. (1990). Distortion-product otoacoustic emissions in humans with high frequency sensorineural hearing loss. *Journal of Speech, Language, and Hearing Research*, *33*(3), 594-600.

Hunter, L. L., Monson, B. B., Moore, D. R., Dhar, S., Wright, B. A., Munro, K. J., Zadeh, L. M., Blankenship, C. M., Stiepan, S. M., & Siegel, J. H. (2020). Extended high frequency hearing and speech perception implications in adults and children. *Hearing Research*, *397*, 107922.

Ishak, W. S., Zhao, F., Stephens, D., Culling, J., Bai, Z., & Meyer-Bisch, C. (2011). Test-retest reliability and validity of Audioscan and Békésy compared with pure tone audiometry. *Audiological Medicine*, *9*(1), 40-46.

Jansen, E., Helleman, H., Dreschler, W., & De Laat, J. (2009). Noise induced hearing loss and other hearing complaints among musicians of symphony orchestras. *International Archives of Occupational and Environmental Health*, *82*, 153-164.

Jerger, J. (1970). Clinical experience with impedance audiometry. *Archives of Otolaryngology*, *92*(4), 311-324.

Keppler, H., Dhooge, I., Maes, L., D'Haenens, W., Bockstael, A., Philips, B., Swinnen, F., & Vinck, B. (2010). Transient-evoked and distortion product otoacoustic emissions: A short-term test-retest reliability study. *International Journal of Audiology*, *49*(2), 99-109. https://doi.org/10.3109/14992020903300431

Keshishzadeh, S., Garrett, M., Vasilkov, V., & Verhulst, S. (2020). The derived-band envelope following response and its sensitivity to sensorineural hearing deficits. *Hearing Research*, *392*, 107979.

Knight, K. R., Kraemer, D. F., Winter, C., & Neuwelt, E. A. (2007). Early changes in auditory function as a result of platinum chemotherapy: use of extended high-frequency audiometry and evoked distortion product otoacoustic emissions. *Journal of Clinical Oncology*, *25*(10), 1190-1195.

Kollmeier, B., Warzybok, A., Hochmuth, S., Zokoll, M. A., Uslar, V., Brand, T., & Wagener, K. C. (2015). The multilingual matrix test: Principles, applications, and comparison across languages: A review. *International Journal of Audiology*, *54 Suppl 2*, 3-16. https://doi.org/10.3109/14992027.2015.1020971

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155-163.

Kujawa, S. G., & Liberman, M. C. (2009a). Adding insult to injury: cochlear nerve degeneration after "temporary" noise-induced hearing loss. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *29*(45), 14077-14085. https://doi.org/10.1523/JNEUROSCI.2845-09.2009

Kujawa, S. G., & Liberman, M. C. (2009b). Adding insult to injury: cochlear nerve degeneration after “temporary” noise-induced hearing loss. *Journal of Neuroscience*, *29*(45), 14077-14085.

Kummer, P., Janssen, T., & Arnold, W. (1998). The level and growth behavior of the 2 f1-f2 distortion product otoacoustic emission and its relationship to auditory sensitivity in normal hearing and cochlear hearing loss. *The Journal of the Acoustical Society of America*, *103*(6), 3431-3444.

Lauter, J. L., & Karzon, R. G. (1990). Individual Differences in Auditory Electric Responses: Comparisons of Between-Subject and Within-Subject Variability IV. Latency-variability Comparisons in Early, Middle, and Late Responses. *Scandinavian Audiology*, *19*(3), 175-182.

Liberman, M. C., Epstein, M. J., Cleveland, S. S., Wang, H. B., & Maison, S. F. (2016). Toward a Differential Diagnosis of Hidden Hearing Loss in Humans [Article]. *PloS One*, *11*(9), 15, Article e0162726. https://doi.org/10.1371/journal.pone.0162726

Lidén, G. (1969). The scope and application of current audiometric tests. *The Journal of Laryngology & Otology*, *83*(6), 507-520.

Lin, H. W., Furman, A. C., Kujawa, S. G., & Liberman, M. C. (2011). Primary neural degeneration in the Guinea pig cochlea after reversible noise-induced threshold shift. *Journal of the Association for Research in Otolaryngology : JARO*, *12*(5), 605-616. https://doi.org/10.1007/s10162-011-0277-0

Lobarinas, E., Salvi, R., & Ding, D. (2013). Insensitivity of the audiogram to carboplatin induced inner hair cell loss in chinchillas. *Hearing Research*, *302*, 113-120. https://doi.org/10.1016/j.heares.2013.03.012

Luts, H., Jansen, S., Dreschler, W., & Wouters, J. (2014). Development and normative data for the Flemish/Dutch Matrix test.

Mehrparvar, A. H., Mirmohammadi, S. J., Ghoreyshi, A., Mollasadeghi, A., & Loukzadeh, Z. (2011). High-frequency audiometry: a means for early diagnosis of noise-induced hearing loss. *Noise & health*, *13*(55), 402-406. https://doi.org/10.4103/1463-1741.90295

Mepani, A. M., Verhulst, S., Hancock, K. E., Garrett, M., Vasilkov, V., Bennett, K., de Gruttola, V., Liberman, M. C., & Maison, S. F. (2021). Envelope following responses predict speech-in-noise performance in normal-hearing listeners. *Journal of Neurophysiology*, *125*(4), 1213-1222.

Mills, D. M., Feeney, M. P., Drake, E. J., Folsom, R. C., Sheppard, L., & Seixas, N. S. (2007). Developing standards for distortion product otoacoustic emission measurements. *The Journal of the Acoustical Society of America*, *122*(4), 2203-2214.

Mitchell, C., Phillips, D. S., & Trune, D. R. (1989). Variables affecting the auditory brainstem response: audiogram, age, gender and head size. *Hearing Research*, *40*(1-2), 75-85.

Munjal, S., Panda, N., & Pathak, A. (2016). Long term test-retest reliability of auditory brainstem response (ABR) and middle latency response (MLR). *Glob J Oto*, *1*, 555-559.

Ng, I. H.-Y., & Mcpherson, B. (2005). Test-retest reliability of distortion product otoacoustic emissions in the 1 to 7 kHz range. *Audiological Medicine*, *3*(2), 108-115.

Organization, W. H. (2021). *World report on hearing*. World Health Organization.

Oxenham, A. (2016). Predicting the perceptual consequences of hidden hearing loss. Trends Hear. 20: 233121651668676. In.

Oyler, R. F., Lauter, J., & Matkin, N. (1991). Intrasubject variability in the absolute latency of the auditory brainstem response. *Journal of the American Academy of Audiology*, *2*(4), 206-213.

Parthasarathy, A., & Kujawa, S. G. (2018). Synaptopathy in the aging cochlea: Characterizing early-neural deficits in auditory temporal envelope processing [Article]. *Journal of Neuroscience*, *38*(32), 7108-7119. https://doi.org/10.1523/JNEUROSCI.3240-17.2018

Paul, B. T., Bruce, I. C., & Roberts, L. E. (2017). Evidence that hidden hearing loss underlies amplitude modulation encoding deficits in individuals with and without tinnitus. *Hearing Research*, *344*, 170-182.

Plack, C. J., Léger, A., Prendergast, G., Kluk, K., Guest, H., & Munro, K. J. (2016). Toward a diagnostic test for hidden hearing loss. *Trends in hearing*, *20*, 2331216516657466.

Popelka, G. R., Osterhammel, P. A., Nielsen, L. H., & Rasmussen, A. N. (1993). Growth of distortion product otoacoustic emissions with primary-tone level in humans. *Hearing Research*, *71*(1-2), 12-22.

Prendergast, G., Millman, R. E., Guest, H., Munro, K. J., Kluk, K., Dewey, R. S., Hall, D. A., Heinz, M. G., & Plack, C. J. (2017). Effects of noise exposure on young adults with normal audiograms II: Behavioral measures. *Hearing Research*, *356*, 74-86.

Prendergast, G., Tu, W., Guest, H., Millman, R. E., Kluk, K., Couth, S., Munro, K. J., & Plack, C. J. (2018). Supra-threshold auditory brainstem response amplitudes in humans: Test-retest

reliability, electrode montage and noise exposure [Article]. *Hearing Research*, *364*, 38-47. https://doi.org/10.1016/j.heares.2018.04.002

Purcell, D. W., John, S. M., Schneider, B. A., & Picton, T. W. (2004). Human temporal auditory acuity as assessed by envelope following responses. *The Journal of the Acoustical Society of America*, *116*(6), 3581-3593.

Rabinowitz, P., Taiwo, O., Sircar, K., Aliyu, O., & Slade, M. (2006). Physician hearing loss. *American Journal of Otolaryngology*, *27*(1), 18-23.

Rasetshwane, D. M., Argenyi, M., Neely, S. T., Kopun, J. G., & Gorga, M. P. (2013). Latency of tone-burst-evoked auditory brain stem responses and otoacoustic emissions: level, frequency, and rise-time effects. *The Journal of the Acoustical Society of America*, *133*(5), 2803-2817. https://doi.org/10.1121/1.4798666

Reavis, K. M., McMillan, G. P., Dille, M. F., & Konrad-Martin, D. (2015). Meta-analysis of distortion product otoacoustic emission retest variability for serial monitoring of cochlear function in adults. *Ear and Hearing*, *36*(5), e251.

Rhode, W. S., & Smith, P. H. (1985). Characteristics of tone-pip response patterns in relationship to spontaneous rate in cat auditory nerve fibers. *Hearing Research*, *18*(2), 159-168.

Rodríguez Valiente, A., Trinidad, A., García Berrocal, J., Górriz, C., & Ramírez Camacho, R. (2014). Extended high-frequency (9–20 kHz) audiometry reference thresholds in 645 healthy subjects. *International Journal of Audiology*, *53*(8), 531-545.

Schaette, R., & McAlpine, D. (2011). Tinnitus with a normal audiogram: physiological evidence for hidden hearing loss and computational model. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *31*(38), 13452-13457. https://doi.org/10.1523/JNEUROSCI.2156-11.2011

Schlauch, R. S., & Carney, E. (2011). Are false-positive rates leading to an overestimation of noise-induced hearing loss?

Schmuziger, N., Patscheke, J., & Probst, R. (2007). An assessment of threshold shifts in nonprofessional pop/rock musicians using conventional and extended high-frequency audiometry. *Ear and Hearing*, *28*(5), 643-648.

Schmuziger, N., Probst, R., & Smurzynski, J. (2004). Test-retest reliability of pure-tone thresholds from 0.5 to 16 kHz using Sennheiser HDA 200 and Etymotic Research ER-2 earphones. *Ear and Hearing*, *25*(2), 127-132.

Shaheen, L. A., Valero, M. D., & Liberman, M. C. (2015). Towards a Diagnosis of Cochlear Neuropathy with Envelope Following Responses. *Journal of the Association for Research in Otolaryngology : JARO*, *16*(6), 727-745. https://doi.org/10.1007/s10162-015-0539-3

Shehabi, A. M., Prendergast, G., & Plack, C. J. (2022). The relative and combined effects of noise exposure and aging on auditory peripheral neural deafferentation: A narrative review. *Frontiers in Aging Neuroscience*, *14*, 877588.

Singh, R., Saxena, R., & Varshney, S. (2009). Early detection of noise induced hearing loss by using ultra high frequency audiometry. *Int J Otorhinolaryngol*, *10*(2), 1-5.

Sininger, Y., & Cone-Wesson, B. (2002). Threshold prediction using ABR and SSEPs with infant and young children. *JackKatz. Handbook of clinical audiology*, 307-321.

Skoe, E., Camera, S., & Tufts, J. (2019). Noise exposure may diminish the musician advantage for perceiving speech in noise. *Ear and Hearing*, *40*(4), 782-793.

Skoe, E., & Tufts, J. (2018). Evidence of noise-induced subclinical hearing loss using auditory brainstem responses and objective measures of noise exposure in humans [Article]. *Hearing Research*, *361*, 80-91. https://doi.org/10.1016/j.heares.2018.01.005

Smith, S., Krizman, J., Liu, C., White-Schwoch, T., Nicol, T., & Kraus, N. (2019). Investigating peripheral sources of speech-in-noise variability in listeners with normal audiograms. *Hearing Research*, *371*, 66-74.

Swanepoel, D. W., Mngemane, S., Molemong, S., Mkwanazi, H., & Tutshini, S. (2010). Hearing assessment—reliability, accuracy, and efficiency of automated audiometry. *Telemedicine and e-Health*, *16*(5), 557-563.

Taberner, A. M., & Liberman, M. C. (2005). Response properties of single auditory nerve fibers in the mouse. *Journal of Neurophysiology*, *93*(1), 557-569.

Van Der Biest, H., Keshishzadeh, S., Keppler, H., Dhooge, I., & Verhulst, S. (2023). Envelope following responses for hearing diagnosis: Robustness and methodological considerations. *The Journal of the Acoustical Society of America*, *153*(1), 191-208.

Vande Maele, T. V., Keshishzadeh, S., Poortere, N. D., Dhooge, I., Keppler, H., & Verhulst, S. (2021). The variability in potential biomarkers for cochlear synaptopathy after recreational noise exposure. *Journal of Speech, Language, and Hearing Research*, *64*(12), 4964-4981.

Vasilkov, V., Garrett, M., Mauermann, M., & Verhulst, S. (2021). Enhancing the sensitivity of the envelope-following response for cochlear synaptopathy screening in humans: The role of stimulus envelope. *Hearing Research*, *400*, 108132.

Verhulst, S., Jagadeesh, A., Mauermann, M., & Ernst, F. (2016). Individual Differences in Auditory Brainstem Response Wave Characteristics: Relations to Different Aspects of Peripheral Hearing Loss [Article]. *Trends in hearing*, *20*. http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L621372574

Wagner, W., Heppelmann, G., Vonthein, R., & Zenner, H. P. (2008). Test–retest repeatability of distortion product otoacoustic emissions. *Ear and Hearing*, *29*(3), 378-391.

Wang, Y., Yang, B., Li, Y., Hou, L., Hu, Y., & Han, Y. (2000). Application of extended high frequency audiometry in the early diagnosis of noise--induced hearing loss. *Zhonghua Er Bi Yan Hou Ke Za Zhi*, *35*(1), 26-28.

Wilson, J. L., Abrams, K. S., & Henry, K. S. (2021). Effects of kainic acid-induced auditory nerve damage on envelope-following responses in the budgerigar (Melopsittacus undulatus). *Journal of the Association for Research in Otolaryngology*, *22*(1), 33-49.

Wojtczak, M., Beim, J. A., & Oxenham, A. J. (2017). Weak middle-ear-muscle reflex in humans with noise-induced tinnitus and normal hearing may reflect cochlear synaptopathy. *eNeuro*, *4*(6).

Zhao, F., & Stephens, D. (1999). Test-retest variability of distortion-product otoacoustic emissions in human ears with normal hearing. *Scandinavian Audiology*, *28*(3), 171-178.

Zhu, L., Bharadwaj, H., Xia, J., & Shinn-Cunningham, B. (2013). A comparison of spectral magnitude and phase-locking value analyses of the frequency-following response to complex tones. *The Journal of the Acoustical Society of America*, *134*(1), 384-395.