Supplementary material of the article: "Auditory tests for characterizing hearing deficits: The BEAR test battery"

Raul Sanchez-Lopez[a,*], Silje Grini Nielsen[a], Mouhamad El-Haj-Ali[b], Federica Bianchi, Michal Fereczkowski[a,b,c], Oscar M Cañete[a], Mengfan Wu[b,c], Tobias Neher[b,c], Torsten Dau[a,] and Sébastien Santurette[a,d]

[a] *Hearing Systems Section, Department of Health Technology, Technical University of Denmark, Kgs. Lyngby, Denmark.*

[b] *Institute of Clinical Research, University of Southern Denmark, Odense, Denmark.*

[c] *Research Unit for Oto-Rhino-Laryngology, Odense University Hospital, Odense, Denmark*

[d] *Centre for Applied Audiology Research, Oticon A/S, Smørum, Denmark*

Contents

A Detailed test battery methods

Participants

Seventy-five listeners (38 females) participated in the study. The normal-hearing (NH) group consisted of five participants (PTA ≤ 25 dB HL). The age for the NH listeners ranged between 59 and 76 (median 69) years. The hearing-impaired (HI) group consisted of 70 participants with symmetric sensorineural hearing loss aged between 59 and 82 (median 71) years. Symmetric sensorineural hearing loss was defined as an interaural difference (ID) ≤ 15 dB HL at frequencies below 8 kHz and ID ≤ 25 dB HL at 8 kHz and air-bone gap < 10 dB HL. The participants were recruited from the BEAR database (Wolff et al., 2020) in Odense University Hospital (OUH) and from Bispebjerg Hospitalet (BBH) and Hearing Systems Section at the Technical University of Denmark (DTU) databases. None of the participants had a history of any neurological diseases, and they all had a self-reported normal or corrected-to-normal vision. None of the HI participants reported tinnitus as their major hearing problem. The study was approved by the Science-Ethics Committee for the Capital Region of Denmark H-16036391 and all participants gave written informed consent, compensation was provided for their participation.

The participants eligible for the present study had audiometric thresholds ≤ 55 dB HL (pure-tone audiometry not older than 1 year) in the range between 125 and 1000 Hz. Participants with a pure tone threshold ≥ 75 dB HL at 2 kHz were excluded from the study as it is unlikely that it will be feasible to perform all of the tests due to audibility issues.

Equipment and set-up

The basic audiological assessment consisted of pure-tone audiometry, wideband tympanometry (Rosowski, Stenfelt, & Lilly, 2013) and middle ear muscle reflex, and was conducted in the facilities of OUH, BBH and DTU. The rest of the tests were performed via PC in a double-walled sound-insulated booth (BBH and DTU) or in a small anechoic chamber (OUH). The tests were implemented in Matlab with a graphical user interface (GUI) that the examiner could operate without programming experience. Most of the tests were implemented using a modular framework for psychoacoustic experiments (AFC; Ewert, 2013). The participants were seated in the room and the stimuli were presented through headphones (Sennheiser HDA200) connected to a headphone-amplifier (SPL phonic) and an audio interface (RME Surface 24-bit). The equipment was calibrated using an artificial ear according to IEC 60318-1:2009.

Test procedure

The tests were conducted by three examiners with a background in audiology and hearing research. An interface containing all the tests was implemented in MATLAB. The MATLAB GUI enabled each examiner to perform a demonstration and a short training before each listening test, so the listener can get familiar with the procedure before starting measurements. For monaural conditions, the right ear was tested first in all listeners. If the standard deviation was higher than the one defined for each test or the listener was not able to perform the procedure, the examiner was able to administer an additional measurement. The tests consisting of threshold estimation using the AFC framework (Ewert, 2013) were repeated at least three times and the mean of the three measurements was considered as the final value. A repetition was considered as an outlier if it was greater than three scaled median absolute deviations. The time needed to

complete the entire test battery did not take longer than 3 hours, distributed over two sessions, for any of the participants.

Test Protocol

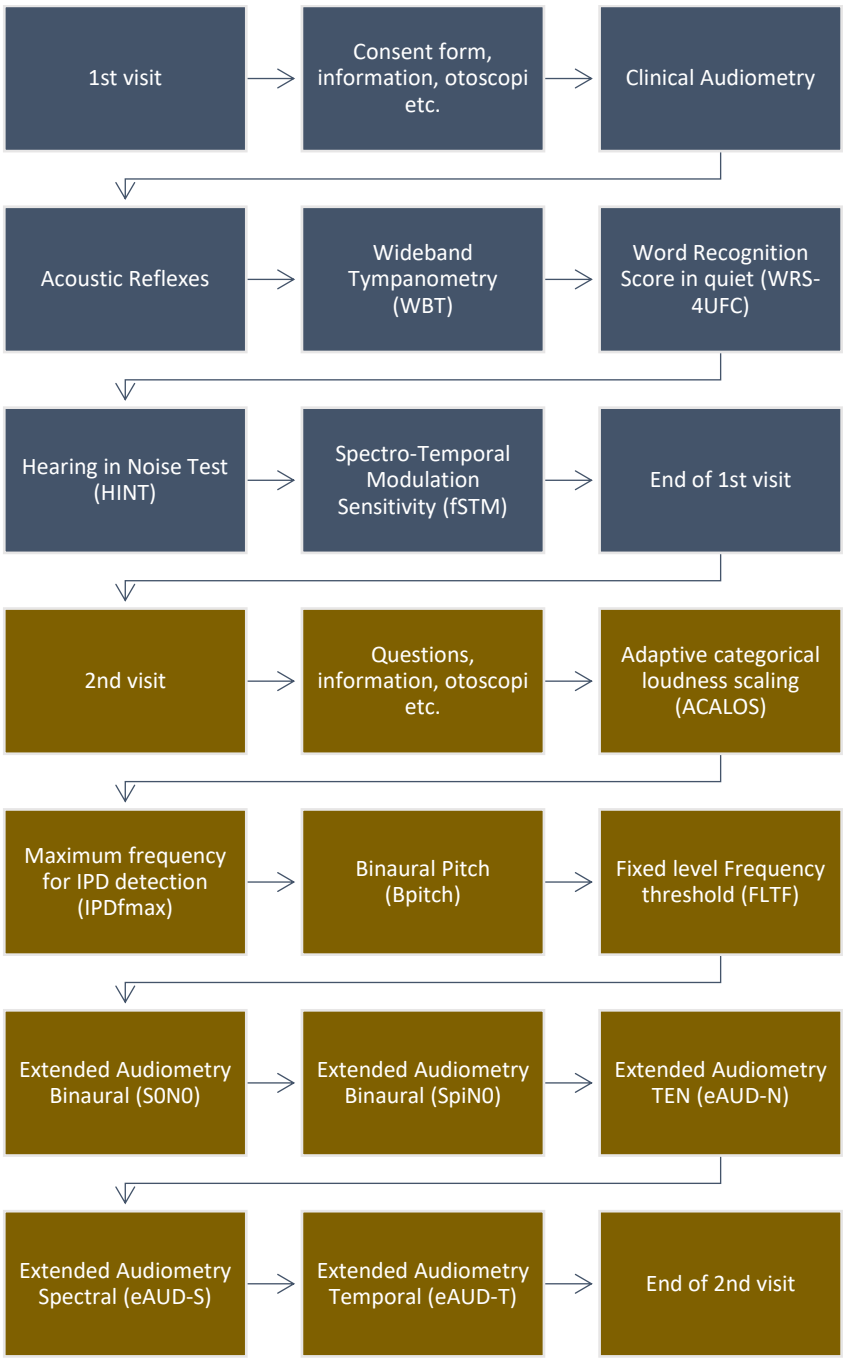

Figure 1: Diagram of the order of the tests in the two visits.

The same order was kept for the 1st and 2nd visits and for all the listeners as depicted in Figure 1. If some measurements could not be completed during a visit, they would be measured in a later visit. Systematical training was only used for the IPDfmax test. Instructions with a little training was done systematically for fSTM, Bpitch and FLTF test. Each measurement, except binaural tests, were first measured on the right ear unless the participant said that left ear is the better ear.

Audiometry and audibility

Pure-tone Audiometry:

The pure-tone audiometry is still the "gold standard" in audiology, not only for fitting hearing aids but also for diagnostics. Overall, no alternative measure has provided enough evidence that could support the substitution or modification of this test. In the BEAR project, the standard (ISO 8253-1, 2010) was followed. However, it seems that the average at low and high frequencies or even the slope of the audiometric curve can provide more consistent information for classification purposes (Moore, 2016; Vlaming et al., 2011). The time estimation for a complete pure-tone audiometry is ~20 minutes.

Condition	Frequencies (kHz)	Ears	Outcome measures	Duration
Air-condution	0.125, 0.25, 0.5, 1, 2, 4, 8 Optional: 0.75, 1,5, 3 and 6	Left and Right	AUD_LF AUD_HF	8-12 min
Bone-conduction	0.25, 0.5, 1, 2, 4	Left and Right	Air-Bone GAP	6-10 min

Fixed-level Frequency threshold (eAUD-HF)

The task consists of a tone detection presented at 80 dB SPL. In the current implementation, the target is a warble tone, which is particularly useful to avoid

standing waves in the ear canal. Furthermore, the procedure used here is a yes/no task using a single-interval adjustment-matrix (SIAM) as described in Kaernbach (1990). In each trial, the target can be present or not. If the target is detected the frequency is increased according to the step-size; if it is not detected the frequency is decreased. However, if the stimulus is not presented (*catch trial*) but the listener provides a positive response, the frequency is decreased compared to the previous trial. Thus, the bias and criterion are controlled during the experiment which yields in a response pattern that is considered less arbitrary than the Békesy method.

Parameter	Values	Comments
Procedure	SIAM (Kaernbach, 1990)	
Conditions	Single condition	
Ears	Left and Right	
Stimuli	Warble tones in quiet	
Stimulus level	80 dB SPL	
Tracking variable	Frequency in logarithmic scale	
Starting frequency and range	Starting frequency: 8 kHz Range: 2 – 20 kHz	
Step size	1/2, 1/5 and 1/10 octave	
Reversals	2 discarded, 4 measurements	
Repetitions	2	
Duration	5 minutes	This includes the explanation of the task.
Outcome measure	FLFT	Fixed-level frequency threshold. Maximum detected frequency at 80 dB SPL

Instructions: *"You will hear tone with high pitch. The tone will be played at different pitch each time. If you can hear the tone press "YES", if there's no tone, press/select the "NO" option on the screen. Sometimes, you can be in doubt, press "no" if you are not sure whether you heard the tone".*

Speech perception tests

Word recognition scores in quiet (WRS-4UFC):

In the BEAR test battery, a 4-unforced-choice paradigm has been introduced. After the word is presented, four alternatives were shown on the screen, as well as a question mark. The four words have been carefully chosen previously. The target is placed randomly in one of the four buttons, together with the 3 words with the lowest Levenshtein distance (Sanders & Chin, 2009) that are also part of the Dantale I corpus.

Parameter	Values	Comments
Procedure	Constant stimuli	
Conditions	PTA + 40, 30, 20, and 10 dB PTA: pure-tone audiometric thresholds average (0,5 – 2 kHz)	The PTA is calculated by the software.
Ears	Left and Right	
Corpus	Dantale I	
Lists	25 monosyllabic words	
Duration	12 minutes (both ears)	This includes the explanation of the task.
Outcome measure	maxDS SRTQ	Maximum discrimination score Speech reception threshold Roll-over index

Instructions: *"You will hear a word. Four similar words will appear on the screen plus a question mark (?). After some time, it will get more difficult. Find as many as you can hear, if you can't hear anything, just press the question mark (?)."*

Hearing in noise test (HINT)

In the BEAR test battery, Danish HINT was used as in (Nielsen & Dau, 2011). Additionally, a list presented at a fixed signal-to-noise ratio of 4 dB SNR was scored for

the entire 20-sentences list and presented as a sentence recognition score. The outcome measures of this test were the speech reception threshold and the sentence recognition score at 4 dB SNR.

Parameter	Values	Comments
Procedure	1) 1up-1down 2) Sentence score at a fixed SNR = +4dB SNR	
Conditions	1) SRT (50%) in speech 2) Fixed 4 dB SNR	
Noise type	Speech-shape stationary noise (daHINT noise)	
Ears	Left and Right	
Speech Corpus	HINT (CLUE)	
Lists	20 sentences	
Noise Level	PTA + 30 dB PTA: pure-tone audiometric thresholds average (0,5 – 2 kHz)	Adjusted manually by the examiner
Duration	12 minutes (both ears)	This includes the explanation of the task.
Outcome measure	SRT_N SS_4dB	Speech reception threshold in noise Sentence Score for a fixed +4 dB SNR

Instructions: *"You will hear a sentence in noise, and your task is to repeat back the sentence. It is getting more difficult. Repeat as much as you can hear, if you can't hear anything, just say pass."*

Binaural processing abilities

Maximum frequency for IPD detection (IPD_{fmax})

Interaural phase difference (IPD) detection abilities have been connected to the sensitivity to temporal fine structure (Brian C J Moore, 2007). The Maximum frequency for IPD detection when the signal in both ears has an IPD = 180º for determining has been successfully measured in hearing-impaired listeners (Santurette & Dau, 2012;

Neher et al. 2011) showing a reduced sensitivity in some of the cases that were not correlated to the loss of audibility. In the BEAR test battery, the stimulus duration and procedure are identical to the method proposed by Füllgrabe, Harland, Sęk, & Moore (2017), as this procedure has been found reliable and without training effects in older listeners. However, the step-size considered here differs slightly by reducing the step size first in steps of 2/3-octaves, then 1/3 octave and finally half of a 1/3-octave. These modifications should not affect the results in terms of accuracy.

Parameter	Values	Comments
Procedure	1up-2down (~70% psychometric function) 2 AFC	
Conditions	Single condition in a binaural	
Stimuli	Pure-tone with inverted phase in the contralateral ear.	
Level	35 dB sensation level (SL)	Based on the pure-tone audiometric threshold. The thresholds of intermediate frequencies were interpolated.
Presentation of the stimuli	Sequence ABAB and AAAA presented in random order. The listener has to identified the interval containing ABAB.	A: diotic pure-tone B: IPD=180°
Duration	Total sequence duration = 1.9s Pause between 2 intervales = 0.5s	A or B = 0.4s Pause = 0.1s
Tracking variable	Frequency in logarithmic scale	log(f)
Step size	Decreasing step-size 2/3, 1/3 and 1/6 octave	
Reversals	6	
Repetitions	2	
Time	7 minutes	Including training using constant stimuli and ILDs = 6 dB instead of IPDs.
Outcome measure	IPD_FMAX	Maximum frequency for detecting an interaural phase difference of 180°

Instructions: *"You will hear 2 sequences of sounds. Within each sequence there are four sounds presented. You may perceive that some of the sounds in one of the sequences are moving between ears or in your head. Your task is to select the interval which you hear*

the sound moving (from the screen options). It is important to wait until you have heard both intervals before choosing the correct sequence".

Extended binaural audiometry in noise (eAUD-B)

Masking level differences consist of two measurements; 1) the masked thresholds for detecting a pure tone in one-octave band noise presented diotically, 2) masked thresholds with the same noise but in antiphase in one of the ears (dichotic) (Brown & Musiek, 2013). As a result, a masking release of about 15 dB is expected in a healthy ear (Durlach, 1963). This measurement has been connected to Temporal fine structure (TFS) sensitivity (Strelcyk & Dau, 2009) and binaural pitch perception (Santurette & Dau, 2012) and it seems to be a promising test for characterizing the binaural performance.

Although new audiometer models have recently included masking level differences (MLD) following the procedure proposed in Brown & Musiek (2013), in the BEAR test battery, this test has been included as a part of the extended audiometry (eAUD).

The test is a simple tone detection task in threshold equalizing noise (TEN) in two conditions:

1) S_0N_0: Noise and tone have the same phase in both ears.
2) $S_{pi}N_{0:}$ The tone is played in anti-phase in both ears.

The advantage of measuring MLD in similar conditions as the eAUD is that the binaural and 2 monoaural measures can be also compared.

Parameter	Values	Comments
Procedure	SIAM (Kaernbach, 1990)	
Conditions	Diotic condition (S0N0) Dichotic condition (SpiN0)	

Parameter	Values	Comments
Stimuli	The tone in TEN noise	
Frequencies	500	
Noise Level	70 dB HL	
Tracking variable	Level of the tone	
Step size	Decreasing step size 10, 5, 2 dB	
Reversals	2 discarded, 4 measurement	
Repetitions	2	
Time	6-7 minutes	
Outcome measures	BMR	Binaural masking level difference.

Binaural Pitch

Binaural pitch is a test that was previously used in Santurette & Dau (2012) as a pitch contour detection and identification task. The task consists of the detection of a melody embedded in noise. Each run consists of a set of 10 diotic and 10 dichotic melodies allocated randomly along with a sound file of 2 minutes length. While the diotic melody can be detected monoaurally, the dichotic melody can be only perceived if the binaural processing abilities are intact. This is because the tones that form the melody are indeed generated by adding phase-difference patterns to the noise presented in the two ears, which creates a pitch percept (Cramer & Huggins, 1958). The listener is asked to press the button each time he or she can hear the pitch contour. Then, the noise starts and a *training* pitch contour is played diotically at a higher level. Subsequently, the diotic and dichotic melodies are presented. Finally, a score is obtained for the diotic and dichotic conditions. Figure 2 shows the user interface of the binaural pitch test.

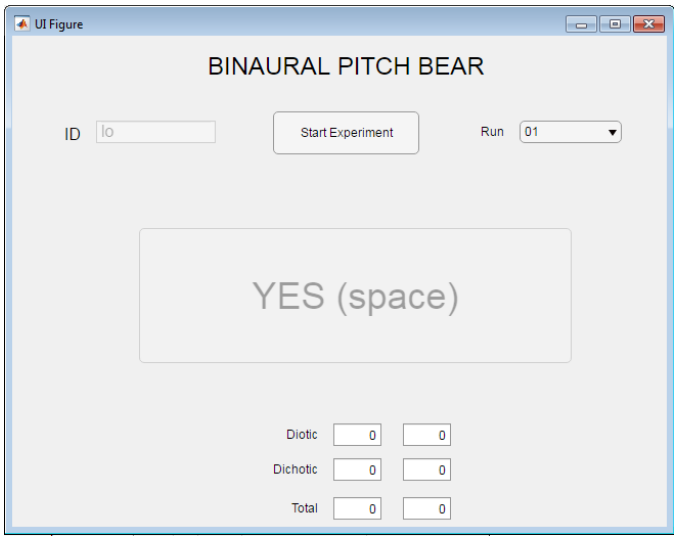

Figure 2: The user interface of the Binaural Pitch test.

Parameter	Values	Comments
Task	Pitch contour detection	
Stimuli	Diotic and Dichotic pitch contours embed in a noise	
Presentation Level	70 dB SPL	
Number of presentations	10 Diotic 10 Dichotic	
Repetitions	2	BP_20 refers to the two repetitions
Time	5 minutes	
Outcome measures	BP_20	Detection score of the dichotic stimuli

Instruction: You will hear a continuous noise. Within that noise, a melody will be played (consisting of three tones). Every time you hear the melody, you have to click on the "YES".

Loudness perception

Adaptive categorical loudness scaling (ACALOS)

The assessment of loudness perception is a matter of interest to the audiology community. ACALOS is a standardized procedure (ISO 16832, 2006) for measuring loudness, which provides information about the growth of loudness and the most comfortable levels. In previous studies, its relations to auditory thresholds (Al-Salim et al. 2010), basilar membrane compression (Jürgens, Kollmeier, Brand, & Ewert, 2011) and fitting of dynamic compression in HAs (Oetting, Hohmann, Appell, Kollmeier, & Ewert, 2016) have been investigated.

The method consists of the categorical scaling of a 1/3-octave noise presented at a certain level. In each presentation, the listener is asked to give a category between *"not heard"* and "*extremely loud"*. Shows the user interface where the categories are on a 13-point scale (see Figure 3). The presentation level of the next stimulus is calculated based on the previous trials (Brand & Hohmann, 2002). In the BEAR test battery, ACALOS was measured monoaurally in each ear. Figure 3 shows the user interface of ACALOS.

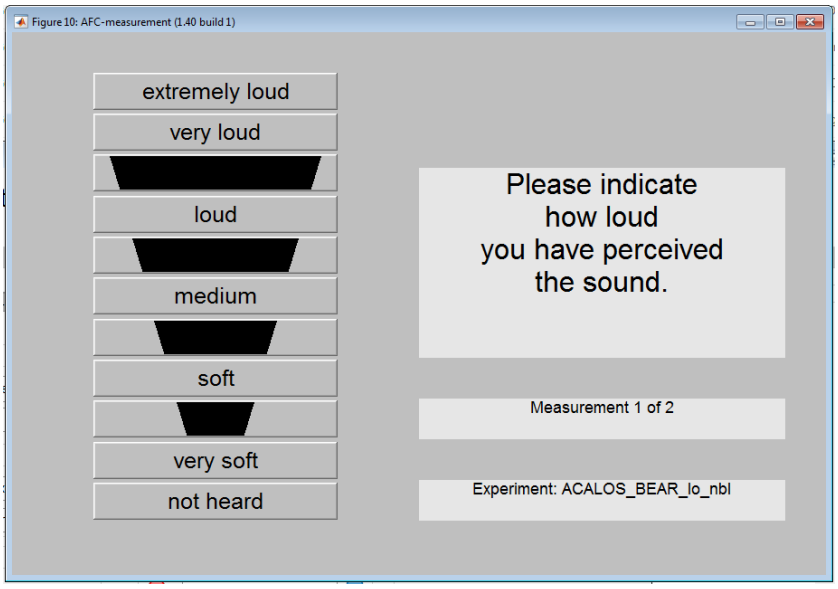

Figure 3: User interface of the ACALOS test.

Parameter	Values	Comments
Procedure	ACALOS (Brand & Hohmann, 2002)	
Ears	Monoaurally, Left and Right	
Stimuli	1/3-octave noises centered at 250 – 500 -1000 – 2000 – 4000 -6000 Hz	
Level	Adaptive level from -10 to 105 dB HL	
Repetitions	1	
Time	20 minutes	
Outcome measure	HTL_LF HTL_HF MCL_LF MCL_HF UCL_LF UCL_HF DynR_L DynR_R Locut_LF Locut_HF Slope_LF Slope_HF m_high_LF m_high_HF OHC_LF OHC_HF	HTL: Hearing thresholds estimation. (1 CU). MCL: Most Comfortable level (25 CU). UCL: Uncomfortable level (50 CU). DynR: Dynamic Range. Locut: ACALOS output parameter, the level where the linear parts intersect. Slope: m_low output parameter. The slope of the lower linear part.

Parameter	Values	Comments
		M_high: output parameter, the slope of the higher linear part. OHC: Outer hair cell loss estimation from the ACALOS results.

Instruction: *"You will hear a sound, and you have to rate how loud in volume you perceived the sound. There will be sounds, which will change in volume and frequency. You have to choose on the screen, scale from not heard to extremely loud, how you perceived the sound. You do the rating by choosing one of the buttons on the screen".*

"You can choose any option (including the bars between the statements)"

"There is one measurement for each ear. The test takes some time, and if you feel like it is taking long, don't worries it should take around 10 min per ear. You can take a short break in-between".

Spectro-Temporal Resolution

Extended audiometry in noise (eAUD)

The assessment of temporal and spectral resolution is based on the difference between the detection in noise and the detection when the temporal or spectral characteristics of the noise make the detection much easier and a release from masking is observed. The masked thresholds were performed with the level of the masker at 70 dB HL and it consists of 3 conditions as sketched in Figure 4:

eAUD-N (Noise): Threshold equalized noise (TEN).

eAUD-S (Spectral): Noise is off-frequency.

eAUD-T (Temporal): Noise is temporally modulated.

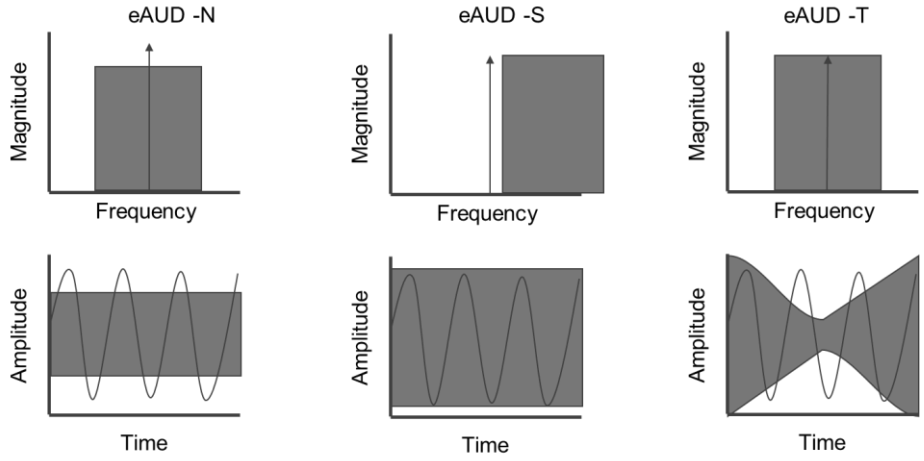

Figure 4: Sketch of the conditions of the extended audiometry (eAUD).

The eAUD-S, in combination with the TEN HL test (equivalent to eAUD-N), can provide an estimate of frequency selectivity. Noise is a 3-octave band TEN played at 70 dB HL in both cases. However, eAUD-S uses simultaneous masking but off-frequency, where the lower cut-off frequency in normalized frequency is 1.10 (Figure 4). Therefore, a masking release is expected if the auditory filters are sharply tuned (normal-hearing listeners). This release of masking can be related to the tip-to-tail distance presented in PTCs.

The eAUD-T, together with the eAUD-N, provides an estimate of the temporal resolution in line with the F-T test and the concept of a temporal resolution factor introduced by Zwicker & Schorn (1982). Here, the same 1-octave-wide TEN noise is temporally modulated with a modulation frequency of 4 Hz (Figure 4). This *unmasks* the target and provides a masking release because of the listening in the dips advantage.

Parameter	Values	Comments
Procedure	SIAM (KAERNBACH, 1990)	
Conditions	Noise Temporally modulated noise (f_m=4Hz) Shifted noise (f_l=1.3f_c)	f_m: modulation frequency f_l: lower cut-off frequency f_c: center frequency of the noise. This is equal to the frequency of the tone.

Parameter	Values	Comments
Ears	Left and Right ear independently	
Stimuli	Warble tone in TEN noise	
Frequencies	500 and 2000 Hz	
Noise Level	70 dB HL	
Tracking variable	Level of the tone	
Step size	Decreasing step size 10, 5, 2 dB	
Reversals	2 discarded, 4 measurement	
Repetitions	2	
Outcome measures	TiN_LF TMR_LF SMR_LF TiN_HF TMR_HF SMR_HF	TiN: Tone in noise in dB HL TMR: Temporal masking release SMR: Spectral masking release
Time	25 minutes	

Instructions: *"You will hear a noise and a tone. If you can hear the tone press "YES", if there is only noise, press/select the "NO" option on the screen. Sometimes it will be difficult to say so you can be in doubt, press "no" if you are not sure whether you heard the tone".*

Spectro-temporal modulation sensitivity

Fast spectro-temporal modulation sensitivity (fSTM)

In the BEAR project, a fast spectro-temporal sensitivity test (fSTM) was suggested. A pilot study not shown here investigated different alternatives of a fSTM test, including the one using SIAM. The fast STM sensitivity measurement consists of a YES/NO task in a constant stimuli procedure with catch trials. The stimulus presented is a sequence of 4 noises following an ABAB pattern. While A segments are unmodulated noises, B segments are spectro-temporally modulated. The catch trial consists of an ABAB sequence where the modulation is well below the threshold obtained in NH in previous studies.

Parameter	Values	Comments
Procedure	sSTM (screening based the score obtained on 10 presentations at -3 dB) fSTM: SIAM (KAERNBACH, 1990)	
Conditions	3-octave noise carrier centered at 800 Hz. fm=4Hz, Ω=2c/o 1-octave noise carrier at 4kHz fm=4Hz, Ω=4c/o	The low-frequency stimulus is similar to the one in Bernstein et al. (2016).
Stimuli	Sequence ABAB where A is unmodulated noise and B is modulated.	
Duration	Total Interval duration: 2s	Each A or B are 0.5 s long with a ramp in and out of 5 ms between them.
Tracking variable	Modulation depth in logarithmic scale 20log(m)	
Steps	5, 2, and 1dB	
Reversals	2 discarded, 4 measurement	
Repetitions	2	
Outcome measures	Estimation of the 80% percent of the psychometric function (dB)	
Time	Screening test: 1.5 minutes Test: 10-15 min	

Parameter	Values	Comments
	sSTM_8 sSTM_4k fSTM_8 fSTM_4k	sSTM: screening STM test. Sensitivity (d') for – 3 dB condition fSTM: Spectro-temporal modulation detection threshold

Instructions: "*You will hear a sequence of four sounds. You have to listen if there is a difference between the sounds or if the sounds are all the same. This difference can be very small. If you hear any difference within the sequence press/select "YES", if they all sound the same, press/select the "NO" option on the screen.* For example;

Sh, Sh, Sh, Sh = no difference

Sh, SS, Sh, SS = difference"

B Clinical feasibility of the test battery

Introduction

Intraclass correlation (ICC) is a way of measuring the reliability of a measurement method (Goldsmith & Stratford, 1997). A higher ICC value (ranges between 0 - 1) indicates a higher correlation between the test and retest measurement, which is an indication of higher reliability. It is important to state that there is no standard value for acceptable reliability, and a low ICC could not only reflect the low degree of measurement agreement but also relate to the lack of variability among the sampled subjects. Koo and Li (2016) suggest that an ICC value of less than 0.5 is an indication of poor reliability. Values varying between 0.5 and 0.75 indicated a mediate reliability, and an ICC value of above 0.75 indicates good reliability. An ICC value of above 0.9 indicated excellent reliability of the measurement method. Both the Pearson's r and the ICC can give misleading results, as they are very sensitive to the spread of data between subjects. Therefore Downham et. al. (2005) suggest to also investigate if there is a systematic bias of the data and a measurement error. In this study, the test-retest reliability of the test battery has been explored by using the intraclass correlation (Koo & Li, 2016), Pearson's correlation, systematic change in means (Downham et al., 2005) and standard error of measurements (SEM; Goldsmith & Stratford, 1997).

Methods

Participants

Test-retest measurements were performed in seven HI and three NH people for all tests of the test battery. The average PTA of the HI listeners was 31 dB HL, and all with bilateral HL. The retest session (second set of two visits) was measured 4 months after the first visit. The same

type of equipment was used in both sessions, but four HI listeners were measured at a different location for the first visit. There were two different examiners for the first test session. The same person examined all the subject in the second session.

Measures of reliability

Interclass cross-correlation calculation has different forms based on different assumptions. By looking at the flowchart presented in the paper by Koo and Li (2016), a two-way mixed effect model was chosen for the analysis. The absolute agreement was chosen, as Koo and Li (2016) state that the test-retest reliability study would be meaningless if there was no agreement between repeated measures. Depending on the test, either a single measurement or the mean of k measurements was used as the type of measurement. The equation below shows the formula for the two-way mixed model, with absolute agreement

$$ICC = \frac{MS_R - MS_E}{MS_R + (k-1)MS_E + \frac{k}{n(MS_C - MS_E)}} \quad (1)$$

where MS_R = mean square for rows; MS_W = mean square for residual sources of variance; MS_E = mean square for error; MS_C = mean square for columns; n = number of subjects and k = number of raters/measurements.

Pearson's r is another way of investigating the reliability and gives very similar results to the ICC. While the ICC is looking at the distance of the point from a straight line that is going through the origin, Pearson's r is looking at the distance from any kind of linear line (Koo & Li, 2016).

For investigating if there is a systematic bias of the data and a measurement error, both the mean difference in results between the two sessions and the 95% confidence interval is calculated for all tests. If the mean difference ($\bar{d}$) had a negative value, this indicates that the results from the first session tend to be larger than the second one. If the confidence interval is including zero, it can be concluded that there is no systematic bias between the two sessions.

The standard error of measurement (SEM) is also calculated. SEM is a way of calculating measurement error (Goldsmith & Stratford, 1997) and is a way to compare different measurement methods. Because it is in the same units as the original measure, also the SEM is calculated in percentage to compare different measurement method with different units.

$$SEM = \sigma_T\sqrt{(1 - ICC)} \quad (2)$$

where σ_T is the total sample standard deviation, and ICC is the ICC value shown in equation 2.

$$SEM\ \% = \frac{SEM}{\bar{m}} \times 100$$

where $\bar{m}$ is the mean of all measurements.

Results and discussion

Table I: The ICC and Pearson's r values for all tests of the test battery

Test	*Condition*	*ICC*	*R*	*Systematic change*	*SEM (%)*
WRS	*10dB*	*ICC = 0.591, p = 0.001*	*0.59*	*d = 0.04, CI = [-0.04,0.11]*	*0.13 (23.05)*
	20dB	*ICC = 0.291, p = 0.096*	*0.28*	*d = -0.007, CI = [-0.06,0.04]*	*0.078 (9.84)*

Test	Condition	ICC	R	Systematic change	SEM (%)
	30dB	*ICC = 0.251, p = 0.128*	*0.25*	*d = -0.005, CI = [-0.02,0.01]*	*0.03 (3.16)*
	40dB	*ICC = 0.475, p = 0.011*	*0.48*	*d = -0.009, CI = [-0.03, 0.01]*	*0.04 (4.29)*
HINT	***SRT***	***ICC = 0.611, p = 0.001***	*0.60*	*d = 0.17, CI = [-0.46,0.79]*	*1.02 (211.54)*
	SS+4dB SNR	*ICC = 0.574, p = 0.002*	*0.58*	*d = -2.96, CI = [-7.73,1.82]*	*7.94 (9.56)*
STM	*LF*	***ICC = 0.916, p = 0.00***	*0.85*	*d = -0.09 CI = [-0.84, 0.67]*	*0.93 (12.26)*
	HF	*ICC = 0.548, p = 0.003*	*0.59*	*d = 0.51, CI = [-0.27, 1.29]*	*1.31 (37.4)*
ACALOS	***HTL***	***ICC = 0.946, p = 0.000***	*0.95*	*d = -0.13, CI = [-1.27, 1.00]*	*4.59 (17.53)*
	MCL	*ICC = 0.678, p = 0.000*	*0.68*	*d = 0.53, CI = [-1.10, 2.16]*	*6.59 (7.86)*
	Slope	***ICC = 0.821, p = 0.000***	*0.82*	*d = -0.002, CI = [-0.02, 0.02]*	*0.07 (15.51)*
Binaural Pitch	***Dichotic***	***ICC = 0.987**, p = 0.000*	*0.99*	*d = -2, CI = [-5.54, 1.54]*	*3.99 (4.91)*
	Total	*ICC = 0.983, p =0.000*	*0.99*	*d = -0.5, CI = [-2.61, 1.61]*	*2.27 (2.52)*
Frequency tracking procedures	*IPD_{fmax}*	*ICC = 0.950, p = 0.000*	*0.96*	*d = -15.84, CI = [-66.44, 34.75]*	*65.39 (6.37)*
	eAUD-HF (FLFT)	*ICC = 0.890, p = 0.000*	*0.89*	*d = 212.71, CI = [-89.7, 515.1]*	*495.3 ()*
eAUD-B	*Bo*	*ICC = 0.327, p = 0.101*	*0.41*	*d = 1.99, CI = [0.24, 3.74]*	*2.28 (3.24)*
	Bp	*ICC = 0.673, p = 0.007*	*0.70*	*d = 1.10, CI = [-1.62, 3.83]*	*3.1 (5.48)*
	BMR	***ICC = 0.783, p = 0.002***	*0.77*	*d = 0.89, CI = [-1.08, 2.86]*	*2.25 (16.2)*
eAUD-N	*LF*	*ICC = 0.325, p = 0.05*	*0.40*	*d = 1.27, CI = [0.09, 2.44]*	*2.02 (2.87)*
	HF	*ICC = 0.551, p = 0.005*	*0.54*	*d = 0.29, CI = [-0.99, 1.56]*	*2.11 (2.89)*
eAUD-S	***S: LF***	***ICC = 0.851**, p = 0.00*	*0.85*	*d = -0.36, CI = [-1.45, 0.73]*	*1.78 (3.34)*
	S: HF	*ICC = 0.954, p = 0.000*	*0.95*	*d = 0.51, CI = [-0.66, 1.69]*	*1.92 (4.08)*
	SMR:LF	*ICC = 0.651, p = 0.004*	*0.68*	*d = 1.48, CI = [0.04, 2.91]*	*2.47 (14.24)*
	SMR: HF	***ICC = 0.858, p = 0.000***	*0.85*	*d = -0.32, CI = [-2.09, 1.46]*	*2.85 (11.19)*

Test	Condition	ICC	R	Systematic change	SEM (%)
eAUD-T	T: LF	ICC = 0.665, p = 0.002	0.77	d = 1.35, CI = [0.49, 2.21]	1.64 (2.59)
	T: HF	ICC = 0.875, p = 0.000	0.89	d = -0.96, CI = [-2.00, 0.09]	1.78 (2.88)
	TMR: LF	ICC = 0.192, p = 0.205	0.19	d = -0.06, CI = [-1.40, 1.27]	2.17 (30.24)
	TMR: HF	ICC = 0.668, p = 0.003	0.71	d = 1.13, CI = [-0.38, 2.63]	2.54 (23.96)

The test-retest reliability of the test battery has been investigated, looking at the ICC, Pearson's R, systematic changes in the data and the SEM. Some of the tests, such as IPD, Binaural Pitch and FLFT showed a good to excellent test-retest reliability with all ICC values above 0.89. There was also no indication of a systematic bias and the SEM showed low values that were below 7% of the total mean for each test. The ACALOS outcome measures also showed good reliability with ICC ranging from 0.67 to 0.95 (ICC_{HTL} = 0.95, ICC_{MCL} = 0.67, ICC_{Slope} = 0.82). There was no indication of a systematic change in the data, and the SEM values for both, the HTL and MCL, varied around 5 dB, which was the same as the uncertainty in the one expected in pure-tone audiometry.

For the WRS test, lower ICC values were found, indicating poorer reliability. This could be a result of the participants having the alternatives visually in front of them, and even though they could not hear the word, they chose the word closest to it. The mean difference between test and retest for the 30 dB and 40 dB conditions is 4%, which is only one difference incorrect words (1/25). A mediate reliability was shown for both outcomes of the HINT measurements, with ICC varying between 0.57 and 0.61. One reason for this mediate reliability could be the choice of lists and lack of randomization in the first session. As mentioned in the main document, there was a small list effect between the two ears that can also play a role here. There was no indication of

any systematic changes in the data, and the SEM values were relatively small, which indicates good reliability.

The STM measurements showed an excellent reliability for the LF condition (ICC = 0.91) and a mediate reliability for the HF condition (ICC = 0.548). In addition, the SEM values showed better reliability for the LF condition. A reason for this could be that many subjects could not simply detect any modulation for the HF condition and therefore answered randomly. A poor to mediate reliability is shown for each condition of the binaural extended audiometry (eAUD-B; ICC_{S0N0} = 0.327, ICC_{SpiN0} = 0.673, ICC_{BMR} = 0.783). Diotic condition (S_0N_0) showed the lowest ICC value and the lowest spread of the data. However, there was also a shift in the data, with higher values for the first session. For the $S_{pi}N_0$ condition and the BMR, there was no shift in the data. The S_0N_0 was used for calculating the BMR, and reliability can be questionable.

The TiN condition of the extended audiometry (eAUD-N) showed a poor to mediate reliability (ICC_{LF} = 0.325, ICC_{HF} = 0.551). The results of the LF condition also showed a shift towards higher values for the first session. The TiN part of the extended audiometry was used for calculating both the SMR and the TMR which is crucial to understand the following results. The temporal condition showed mediate reliability, and a systematic change showing higher values for the first session. The spectral condition of the extended audiometry (eAUD-S) showed results that are promising for its implementation in the clinics with a good to excellent reliability for both conditions (ICC_{LF} = 0.851, ICC_{HF} = 0.954), however, the reliability of the spectral masking release was lower. Moreover, all conditions of the extended audiometry show somehow the same standard error of measurements (SEM), around 2 dB, which is the same as the minimum step size.

Time efficiency of the test battery

The examiners kept track of the time used by each of the participants in completing the test battery. In the case of some events, for example additional repetitions needed, annotated the events cautiously for later investigation. Regarding the test procedure in Appendix A, additional repetitions of the threshold estimations were need if: 1) a repetition was considered as an outlier if a given threshold was greater than three scaled median absolute deviations of the three repetitions; or 2) the responses of the listeners during the tracking procedure were inconsistent or reached the maximum or minimum possible values. In that case, the measurement was considered an invalid or "missing" data point.

The results of the timing are shown in Figure 5, probability and mean number of extra repetitions per listener are shown in Table 2.

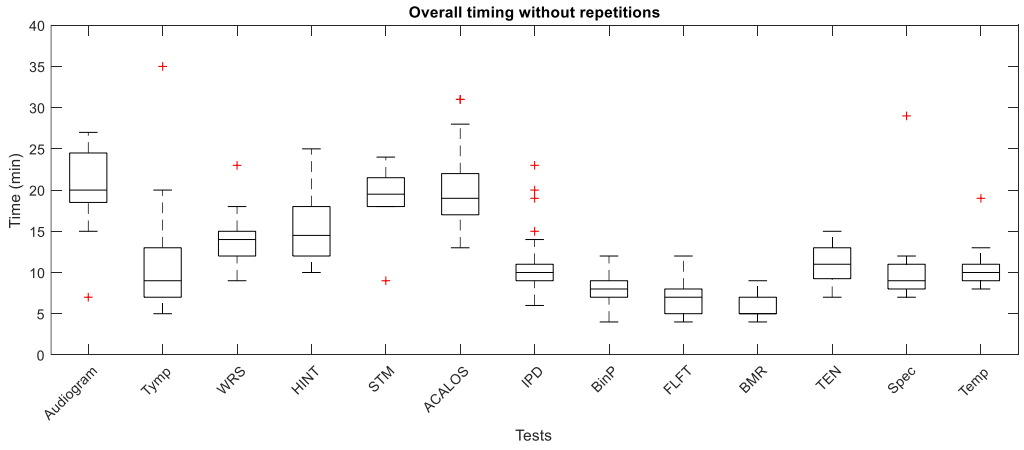

Figure 5: The overall time of the different tests in the test battery.

Table 2: Table with the probability of needing repetitions, and the probability of having missing values (unreliable data). The rows with no values in the probability of missing values column is due to the test not producing this missing values. The mean number of extra repetitions is only averaging over the number of extra repetitions.

Test	*Probability of extra repetitions (%)*	*Probability of missing values (%)*	*Total probability of having to repeat (%)*	*Mean number of extra repetitions*
HINT	*1.32*	-	*1.32*	*4 (only one subject)*
STM	*42.86*	*90.79*	*88.16*	*4.32*
IPD	*10.77*	*10.97*	*20.55*	*1.87*
Binaural Pitch	*8.11*	-	*8.11*	*0.167*
FLFT	*5.63*	*4.05*	*9.46*	*1.85*
S_0N_0	*42.59*	*27.03*	*58.11*	*2*
$S_{pi}N_0$	*20.59*	*9.11*	*27.03*	*1.85*
eAUD-N	*66.67*	*46.58*	*82.19*	*3*
eAUD-S	*48.57*	*52.70*	*75.68*	*3.07*
eAUD-T	*53.85*	*46.58*	*75.34*	*3.27*

Tests such as WRS, HINT, ACALOS, Binaural Pitch, and FLFT showed a low number of total repetitions needed. This can be partly explained by the procedure used. Looking at the timing of these tests, ACALOS and HINT had a larger spread, meaning that the estimated test-time was more subject-dependent than in the case of other tests.

The timing data of the STM tests includes also the STM screening in the beginning of the test. The probability of having to repeat the measurement was 42.86 %. Looking at the missing values, 90.79% of the subjects had missing values giving 88.16 % probability of having to repeat at least one condition. Many listeners had substantial difficulties for detecting the STM at HF condition what was unexpected and the examiners performed several extra repetitions before giving up. For the last 30 listeners, if the threshold was “missing” after more than 2 extra repetitions, the result was consider “missing”. Furthermore, STM was measured using the SIAM

tracking procedure which has catch trials and a conservative criterion for giving a certain threshold as valid. If the listener missed many catch trials or if the responses are done too quickly, there is a high risk of having a no valid (missing) estimate. For many listeners that was the case for the HF condition but not the LF condition. The number of repetitions needed was largely spread for the STM test, ranging from 1 - 14. The average number of extra repetitions were around four. This supports the idea of modifying the tracking procedure of this test for further investigations.

The IPD_{fmax} showed a low variation of the timing data with some outlier. This can be due to extra systematic training provided before the test. The probability of repetition was in total 20.55% meaning that one in five subjects had to repeat either as a result of the two repetitions not being close enough or invalid threshold estimates (missing values).

The binaural part of the extended audiometry are a relatively short measurement with a total time around 5 min. Looking at the probability of repetition, both parts of the binaural extended audiometry shows a large probability of repetitions. The S_0N_0 part showed 58% total probability of repetitions, meaning that more than half the subjects needed to repeat. The duration of the eAUD-N, eAUD-S and eAUD-T, without any extra repetitions, was around 10 minutes each. However, looking at the probability of needing repetitions, these were more than 75% for all three subtests. This suggests that three out of four subjects needed to repeat the at least one of the conditions, as a result of either missing value or the two repetitions being too far from each other. It can also be seen that the mean extra repetitions was three.

Final remarks

Overall, the tests with evidence of good to excellent reliability will be considered for a shorter clinical version of the test battery that will be implemented in the Danish public hospitals. The SIAM tracking procedure was chosen to resemble the procedure of a pure-tone audiometry. However, several additional repetitions were needed in all the test using SIAM to provide reliable results. Therefore, different tracking procedures have to be considered for the clinical version of the test battery.

C Author Contributions

RSL implemented the test battery, carried out pilot experiments, analyzed the data, interpreted the results and wrote the manuscript. SS, FB, MF, TD and RSL conceptualized the research study, discussed the inclusion criteria for the tests and set the requirements in terms of the number of participants and the expected hearing loss variability. Data acquisition was conducted by SGN and ME with the support of MW and OMC under the responsibility of TN, SS and TD. TN and ME designed the clinical protocol, the inclusion criteria of the participants and the recruitment process. SS, TD, FB, MF and TN supervised the work in different stages of the study. SGN conducted the test-retest reliability study, analysed the data and wrote a report with the results assisted by OMC and RSL. All authors critically reviewed and significantly contributed to the manuscript and approved the final version for publication.

References

Al-Salim, S. C., Kopun, J. G., Neely, S. T., Jesteadt, W., Stiegemann, B., & Gorga, M. P. (2010). Reliability of Categorical Loudness Scaling and Its Relation to Threshold. *Ear and Hearing*, *31*(4), 567–578. https://doi.org/10.1097/AUD.0b013e3181da4d15

Brand, T., & Hohmann, V. (2002). An adaptive procedure for categorical loudness scaling. *The Journal of the Acoustical Society of America*, *112*(4), 1597–1604. https://doi.org/10.1121/1.1502902

Brown, M., & Musiek, F. (2013). Pathways: The Fundamentals of Masking Level Differences for Assessing Auditory Function. *The Hearing Journal*, *66*(1), 16.

https://doi.org/10.1097/01.HJ.0000425772.41884.1d

Cramer, E. M., & Huggins, W. H. (1958). Creation of Pitch through Binaural Interaction. *The Journal of the Acoustical Society of America*, *30*(5), 413–417. https://doi.org/10.1121/1.1909628

Downham, D. Y., Holmbäck, A. M., & Lexell, J. (2005). Reliability of measurements in medical research and clinical practice. In *Studies in Multidisciplinarity Vol 3* (pp. 147–163). https://doi.org/10.1016/S1571-0831(06)80013-4

Durlach, N. I. (1963). Equalization and Cancellation Theory of Binaural Masking-Level Differences. *The Journal of the Acoustical Society of America*, *35*(8), 1206–1218. https://doi.org/10.1121/1.1918675

Ewert, S. (2013). AFC - A modular framework for running psychoacoustic experiments and computational perception models. *Proceedings of the International Conference on Acoustics AIA-DAGA 2013*, 1326–1329. Retrieved from www.aforcedchoice.com

Füllgrabe, C., Harland, A. J., Sęk, A. P., & Moore, B. C. J. (2017). Development of a method for determining binaural sensitivity to temporal fine structure. *International Journal of Audiology*, *56*(12), 926–935. https://doi.org/10.1080/14992027.2017.1366078

Goldsmith, C. H., & Stratford, P. W. (1997). Use of the Standard Error as a Reliability Index of Interest : An Applied Example Using Elbow Flexor Strength Data. *Physical Therapy*, *77*(7), 745–750.

IEC 60318-1. (2009). Electroacoustics - Simulators of human head and ear - Part 1: Ear

simulator for the calibration of supraaural earphones. *International Standard*. Retrieved from https://webstore.iec.ch/publication/1443

ISO 16832. (2006). *Acoustics - Loudness scaling by means of categories* (Vol. 2006, pp. 1–12). Vol. 2006, pp. 1–12. Retrieved from http://www.iso.org/iso/catalogue_detail.htm?csnumber=32442

ISO 8253-1. (2010). Acoustics - Audiometric test methods - Part 1: Pure-tone air and bone conduction audiometry. *International Organization for Standardization*. Retrieved from http://www.iso.org/iso/catalogue_detail.htm?csnumber=43601

Jürgens, T., Kollmeier, B., Brand, T., & Ewert, S. D. (2011). Assessment of auditory nonlinearity for listeners with different hearing losses using temporal masking and categorical loudness scaling. *Hearing Research*, *280*(1–2), 177–191. https://doi.org/10.1016/j.heares.2011.05.016

Kaernbach, C. (1990). A single-interval adjustment-matrix (SIAM) procedure for unbiased adaptive testing. *The Journal of the Acoustical Society of America*, *88*(6), 2645–2655. https://doi.org/10.1121/1.399985

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*. https://doi.org/10.1016/j.jcm.2016.02.012

Moore, B. C. J. (2016). A review of the perceptual effects of hearing loss for frequencies above 3 kHz. *Http://Dx.Doi.Org/10.1080/14992027.2016.1204565*, *2027*(July). https://doi.org/10.1080/14992027.2016.1204565

Moore, B.C.J., & Sek, A. (2016). Preferred Compression Speed for Speech and Music and Its Relationship to Sensitivity to Temporal Fine Structure. *Trends in Hearing*, *20*(0), 1–15. https://doi.org/10.1177/2331216516640486

Moore, Brian C J. (2007). Cochlear Hearing Loss. In *Cochlear Hearing Loss: Physiological, Psychological and Technical Issues*. https://doi.org/10.1002/9780470987889

Neher, T., Laugesen, S., Søgaard Jensen, N., & Kragelund, L. (2011). Can basic auditory and cognitive measures predict hearing-impaired listeners' localization and spatial speech recognition abilities? *The Journal of the Acoustical Society of America*, *130*(3), 1542–1558. https://doi.org/10.1121/1.3608122

Nielsen, J., & Dau, T. (2011). The Danish hearing in noise test. *International Journal of Audiology*, *50*(3), 202–208. https://doi.org/10.3109/14992027.2010.524254

Oetting, D., Hohmann, V., Appell, J.-E., Kollmeier, B., & Ewert, S. D. (2016). Spectral and binaural loudness summation for hearing-impaired listeners. *Hearing Research*, *335*, 179–192. https://doi.org/10.1016/j.heares.2016.03.010

Rosowski, J. J., Stenfelt, S., & Lilly, D. (2013). An Overview of Wideband Immittance Measurements Techniques and Terminology. *Ear and Hearing*, *34*, 9s-16s. https://doi.org/10.1097/AUD.0b013e31829d5a14

Sanders, N. C., & Chin, S. B. (2009). Phonological Distance Measures*. *Journal of Quantitative Linguistics*, *16*(1), 96–114. https://doi.org/10.1080/09296170802514138

Santurette, S., & Dau, T. (2012). Relating binaural pitch perception to the individual listener's

auditory profile. *The Journal of the Acoustical Society of America*, *131*(4), 2968–2986. https://doi.org/10.1121/1.3689554

Strelcyk, O., & Dau, T. (2009). Relations between frequency selectivity, temporal fine-structure processing, and speech reception in impaired hearing. *The Journal of the Acoustical Society of America*, *125*, 3328–3345. https://doi.org/10.1121/1.3097469

Vlaming, M. S. M. G., Kollmeier, B., Dreschler, W. A., Martin, R., Wouters, J., Grover, B., … Mohammadh, T. (2011). Hearcom: Hearing in the communication society. *Acta Acustica United with Acustica*, *97*(2), 175–192. https://doi.org/10.3813/AAA.918397

Wolff, A., Houmøller, S. S., Hougaard, D., Gaihede, M., Hammershøi, D., & Schmidt, J. (2020). Health-related Quality of Life in Hearing Impaired Danish Adults before and after Hearing Aid Rehabilitation. *International Journal of Audiology*.

Zwicker, E., & Schorn, K. (1982). Temporal resolution in hard-of-hearing patients. *Audiology : Official Organ of the International Society of Audiology*, *21*(6), 474–492. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7181741