Development and verification of non-supervised smartphone-based methods for assessing pure-tone thresholds and loudness perception

Chen Xu[a]*, Lena Schell-Majoor[a] and Birger Kollmeier[a]

[a]*Medizinische Physik and Cluster of Excellence Hearing4all, Universität Oldenburg, D-26111 Oldenburg, Germany*

Contact: Chen Xu

chen.xu@uni-oldenburg.de

Department of Medical Physics and Acoustics, Faculty VI

University of Oldenburg, 26111, Oldenburg, Germany

Development and verification of non-supervised smartphone-based methods for assessing pure-tone thresholds and loudness perception

Objective: The benefit of using smartphones for hearing tests in a non-supervised, rapid, and contactless way has drawn a lot of interest, especially if supra-threshold measures are assessed that go beyond audiogram-based measures alone. It is unclear, nevertheless, how well these measures compare to more supervised and regulated manual audiometric assessments. The aim of this study is to validate such smartphone-based methods against standardized laboratory assessments.

Design: Pure-tone audiometry and categorical loudness scaling (CLS) were used. Three conditions with varying degrees of supervision were created and compared. In order to assess binaural and spectral loudness summation, both narrowband monaural and broadband binaural noise have been examined as CLS test stimuli.

Study sample: N = 21 individuals with normal hearing and N = 16 participants with mild-to-moderate hearing loss.

Results: The tests conducted here did not show any distinctions between smartphone-based and laboratory-based methods.

Conclusions: Non-supervised listening tests via smartphone may serve as a valid, reliable, and cost-effective approach, e.g., for pure-tone audiometry, CLS, and the evaluation of binaural and spectral loudness summation. In addition, the supra-threshold tests can be constructed to be invariant against missing calibration and external noise which makes them more robust for smartphone usage than audiogram measures.

Keywords: remote audiology; categorical loudness scaling; pure-tone audiometry; self-supervision; mobile health

Introduction

Although the clinical routine audiometry tests (e.g., tone audiometry and speech audiometry) are highly valid and reliable to evaluate hearing ability, their practical drawbacks in terms of time consumption and personal intensiveness are not negligible (Colsman et al., 2020). Hence, employing a smartphone to conduct non-supervised listening tests - at least for simple routine cases where no medical supervision is required - might be a cost-effective alternative and attracts considerable interest. The current study aims at validating this approach by comparing non-supervised threshold and supra-threshold tests to classical laboratory-based audiometric assessments in a controlled way.

Previously, many studies have demonstrated that smartphone-based non-supervised methods are plausible and applicable to measure air-conduction pure-tone audiometry. Swanepoel et al. (2014) and Yousuf Hussein et al. (2016) developed and calibrated the hearScreen^{TM} app, examined in 15 normal hearing adults and 162 children, and compared it with clinical audiometry. Their results revealed that the smartphone-based pure-tone audiometry measurement was comparable to clinical audiometry. Later, Abu-Ghanem et al. (2016) evaluated the smartphone application ‘uHear’ for a questionnaire and a pure-tone audiometry test in the 26 participants aged 65 years and older and reported that there was an agreement between the app and audiometer assessment for most of the test participants in all frequencies. The app yielded a sensitivity of 100% and a specificity of 80% compared with clinical audiometry. More recently, Hazan et al. (2022) designed an experiment to test the reliability of a smartphone app ‘DuoTone’ on 1641 participants from a cloud database. Their results suggested that the test-retest reliability of the app did not differ from the standard audiometry performed in the clinics.

However, the validity and reliability of the smartphone tone audiometry apps are limited due to inherent limitations of the procedures employed. To the authors' knowledge, nearly all of the current smartphone apps use either a modified Hughson-Westlake (Hughson et al., 1944) procedure or a not revealed procedure in order to circumvent any patent issues. The modified Hughson-Westlake procedure, most often employed in clinical audiometry, is widely adopted by clinicians due to its simple administration, little patient training, and easy implementation. Thus, most smartphone apps are directly adapted from clinical methods in order to be comparable with a clinical hearing test. However, according to the findings by Lecluyse and Meddis (2009) and Xu et al. (2023), the modified Hughson-Westlake procedure might be inaccurate and overestimate the true threshold if administered in a self-paced, unsupervised way due to occasional inattentiveness of the listeners. Following the recommendations of Lecluyse and Meddis (2009) and Xu et al. (2023), the present study therefore adopts the non-clinical adaptive procedure (e.g., the single interval up and down procedure SIUD, proposed by Lecluyse & Meddis, 2009) to assess air-conduction pure-tone audiometry on a smartphone and compares the acquired results with the laboratory-based measurements.

Another limitation of smartphone apps to measure the individual audiogram is their dependence on an absolute calibration of the earphones employed which can not be warranted, e.g., for Android devices. This problem is mostly circumvented by using supra-threshold tests (e.g., digit-in-noise (DIN) test, Smits et al., 2004) that assess a certain relative quantity of stimulus components (e.g., signal-to-noise ratio at threshold) and are largely independent of the absolute presentation level. Hence, supra-threshold auditory measures (e.g., speech-in-noise (SIN) tests) have attracted much attention in the last years for hearing screening via smartphone.

In clinical audiology, such supra-threshold tests are used to specify individual functional deficits. Of these, the assessment of loudness growth with increasing level or stimulus bandwidth is of clinical interest, e.g., determining the recruitment phenomenon and for fitting hearing devices (Kollmeier & Hohmann, 1995; Oetting et al., 2016; Koppun et al., 2022). Individual loudness perception is commonly measured employing the categorical loudness scaling (CLS) technique and quantified with a monotonic loudness growth function (Brand & Hohmann, 2002; Oetting et al., 2014). The task of the CLS requires participants to select the descriptors from an 11-point scale, e.g., 'too loud', 'medium', 'soft', etc., based on their loudness perception. The CLS is a supra-threshold listening test that has been included in the 'auditory profile' (i.e., a comprehensive and well-specified set of audiological test procedures described in Van Esch et al., 2013) and has also recently been proposed for usage in machine-learning-supported auditory profiles by Saak et al. (2022).

The standardized adaptive procedure to perform CLS measurement (i.e., Adaptive Categorical Loudness Scaling, **ACALOS**) was introduced by Brand and Hohmann (2002) and standardized in ISO 16832 (2006). CLS has a broad application in clinical audiology, not only as a diagnostic tool but also to fit hearing aids or cochlea implants. For diagnostic purposes, an increase in loudness growth with stimulus level – clinically termed as recruitment phenomenon and assumed to be due to dysfunctional outer hair cells (Hallpike & Hood, 1959; Buus & Florentine, 2002) – can well be characterized by CLS (e.g., Kollmeier & Hohmann, 1995, Launer, 1995, and Rasetshwane et al., 2015). Jürgens et al. (2011) proposed to estimate the hearing loss attributable to outer hair cells (OHC) by applying CLS and concluded that CLS could be a measure of auditory nonlinearity. Further diagnostical applications of CLS were described, e.g., by Shiraki et al. (2022) as a means to better characterize patients with

certain patterns in Bekesy audiometry and by Erinc et al. (2022) and Hébert et al. (2013) as a means to better characterize patients with tinnitus and hyperacusis.

With respect to using CLS as a tool for hearing device fitting, many studies have demonstrated that individualized loudness compensation for narrowband signals can lead to a better-individualized treatment with hearing devices (see Kollmeier & Hohmann, 1995, Kollmeier & Kießling, 2018, Oetting et al., 2018, and Fereczkowski et al., 2023 for hearing aids and Müller-Deile et al., 2021 for cochlea implants). Despite its theoretical advantage to characterize supra-threshold functional hearing deficits and of the compensation by an appropriately, individually fitted hearing device, the usage of CLS for clinical purposes has been limited due to several reasons:

a) Time constraints in clinical settings that interfere with the usage of more sophisticated methods beyond the minimum set of clinical routine procedures. However, self-paced, smartphone-based procedures might take over that do not impose such a time-consuming burden on the professional audiologists.

b) Previous forms of CLS have been discredited by an influential paper by Elberling (1999) arguing that the uncertainty in hearing aid gain setting will not be reduced by CLS. However, their claim was based on the questionable assumption of a perfectly-known individual threshold. More refined measuring and evaluation techniques in CLS (e.g., Brand & Hohmann, 2002, and Oetting et al., 2014, 2016) demonstrate a low correlation between scaling slope estimate and individual threshold, thus demonstrating the importance of the individually obtained loudness growth function for hearing loss compensation.

c) Recent insights into the individually strongly varying loudness summation across frequency and across ears as demonstrated by Oetting et al. (2016) who reported that using narrowband gain compensation, levels of HI listeners to reach ‘medium loud’

were lower than for NH listeners when broadband signals were presented. Participants with the same hearing thresholds perceived loudness substantially differently for binaural broadband signals. Thus Oetting et al. (2016) recommended that the broadband and binaural loudness scaling should be included for hearing-aid fitting. To further investigate the potential consequences of the spectral and binaural loudness summation, Van Beurden et al. (2018) extended the study of Oetting et al. (2016) by recruiting more test participants with a broader range of hearing loss. Spectral loudness summation of HI listeners was detected to be greater than of NH listeners for both monaurally and binaurally presented signals. The effect of hearing loss did not significantly influence the binaural loudness summation. In agreement with Oetting et al. (2016), Van Beurden et al. (2018) found large individual variations in HI listeners for binaural broadband signals. In this study, we, therefore, follow the recommendations by Oetting et al. (2016) and Van Beurden et al. (2018, 2021) to employ not only narrowband signals presented unilaterally, but also broadband signals presented bilaterally for both NH and HI listeners.

Even though CLS is an applicable and useful measurement for clinical diagnostics and assessment of hearing loss compensation as introduced above (e.g., Rasetshwane et al., 2015; Fultz et al., 2020), it is not yet accessible for a smartphone or any other mobile device. There is only one study published so far that introduced a remote CLS measurement on a laptop and compared it with the laboratory setting (Kopun et al., 2022). However, they did not examine the test persons via smartphone and did not consider HI participants. Furthermore, Kopun et al. (2022) only measured 5 participants for the validation study. One possible obstacle against self-controlled CLS measurement in an unrestricted environment is the influence of background noise (which might cause a bias at low stimulus levels that might be confused with a

126 recruitment phenomenon) or any inattention effect of the participant (as simulated in Xu
127 et al., 2023). Hence, in this paper, one of our objectives is to examine the plausibility
128 and validity of the smartphone-based app for CLS measurement under different degrees
129 of control in experimental settings.

130 Taken together, the following research questions should be answered by our
131 study by performing three sub-experiments (i.e., Exp 1: pure-tone audiometry; Exp 2:
132 adaptive categorical loudness scaling; Exp 3: binaural and spectral loudness summation)
133 that all employ normal-hearing and hearing-impaired listeners and compare laboratory
134 situations with self-steered, smartphone-based setups:

135 - Are the results of the smartphone-based pure-tone audiometry and categorical
136 loudness scaling quantitatively comparable to a laboratory-based assessment when using
137 various statistical measures (e.g., correlation coefficient R, root mean square error, etc.)?

138 - Which factors (e.g., the way of supervision, the degree of hearing loss, and test
139 frequency) might influence the differences between smartphone-based and laboratory-
140 based measurements?

141 - Is the smartphone test able to detect individual differences in binaural and
142 spectral loudness summation in a similar way as laboratory-based measures?

Materials and methods

Pure-tone audiometry

Pure-tone audiometry was assessed via a single-interval-up-and-down (SIUD) procedure, introduced by Lecluyse and Meddis (2009). Listeners were presented with a probe tone and a cue tone which had a 10 dB higher sound level than the probe tone and had a 20% chance to be muted, and required to indicate how many tones they have heard. The smartphone user interface of SIUD is provided in the left bottom corner of Fig. 1a. If listeners answered correctly, the task became harder by decreasing the sound level of the following trial. In the end, the track converged at the level of the listener's hearing threshold. The behavioral data were fitted to a logistics psychometric function and the level (L_{50}) corresponding to 50% of the psychometric function was estimated as the hearing threshold.

The stimuli were pure tones consisting of 0.2 s duration for each tone, 20 ms cosine ramps, and 0.2 s for a break between two tones at 0.25, 1, and 4 kHz frequencies for both ears. The starting level for the probe tone was set up at 50 dB with a random offset between 0 and 5 dB. There was a fixed level difference of 10 dB between the probe and cue tone. If listeners reported that the first stimulus was not heard, the initial level increased until it was audible. The initial step size was chosen as 10 dB and reduced to 2 dB after the first reversal. The track terminated when at least 14 reversals and 10 trials were reached. The trials before the 4th reversal were excluded from the threshold calculation.

Adaptive categorical loudness scaling

Adaptive categorical loudness scaling (ACALOS), described by Brand and Hohmann

(2002) and ISO 16832 (2006), was applied to measure an individual's loudness growth function. There were in total 11 categorical scale values distributed on the 50-point categorical units (CU) scale according to Heller (1985), i.e., the verbal values 'very soft' (5 CU), 'soft' (15 CU), 'medium' (25 CU), 'loud' (35 CU) and 'very loud' (45 CU), four intermediate categories without verbal labels, and the two limiting categories 'not heard' (0 CU) and 'too loud' (50 CU) in ACALOS. Listeners needed to rate the stimuli based on their individual loudness perception given the 11 categories. The user interface for the smartphone of ACALOS is shown in the left bottom corner of Fig. 1b. ACALOS mainly comprised two phases, i.e., 'dynamic range estimation' and 'presenting and re-estimation'. During dynamic range estimation, the procedure started at 65 dB and presented upward and downward stimuli in an interleaved manner to obtain a rough estimate of the dynamic range between 0 CU and 50 CU. The individual loudness function was then fine-tuned in the second phase by presenting stimuli at 5 levels estimated from the first phase corresponding to the categorical loudness of 5, 15, 25, 35, and 45 CU in a randomized order.

The 'BTUX' method, which was introduced by Oetting et al. (2014), was employed to fit a loudness growth function. The descriptive parameters (i.e., hearing threshold level (HTL) corresponding to 2.5 CU, median loudness level (MLL at 25 CU), and uncomfortable loudness level (UCL at 50 CU), respectively) were derived from the fitted loudness growth function (Oetting et al., 2014). Furthermore, the most comfortable loudness (MCL) was estimated as the sound level at 20 CU of the growth function (Van Esch et al., 2013). Finally, the dynamic range (DR) was calculated as the difference between UCL and HTL.

The narrowband stimuli were one-third-octave-band low-noise noises (Kohlrausch et al., 1997) centered at 0.25, 1, and 4 kHz (later referred to as LNN250,

LNN1000, and LNN4000, respectively). The broadband stimulus was uniformly exciting noise (UEN17) with equal energy in each of the 17 critical frequency bands, defined in Zwicker (1961). All stimuli (i.e., three narrowband and one broadband stimuli) were presented monaurally for both ears. In addition, LNN1000 and UEN17 were played bilaterally. The duration for all signals was 1 s with 50 ms rise and fall ramps.

Smartphone application design

Fig. 1 shows the overview of the employed smartphone application. The web-app was developed based on the flask (version 1.1.2) framework in python (Python Software Foundation, version 3.10.6) while the database was based on SQLite3 (version 3.37.2). Both frameworks are open source.

The workflow to conduct a non-supervised listening test on a smartphone was as follows: the listener first registered an account and signed up to the dashboard. There are some general instructions, e.g., study background, user consent, and test environments, displayed in text format on the dashboard. Then the listener needed to indicate which measurement to perform by clicking on the appropriate button. Subsequently, some specific guidelines for the chosen listening test were shown. The listener started the measurement by clicking on the 'start' button and the stimuli were automatically presented to the listener. The listener considered and later responded to questions, i.e., 'How loud was the sound?' or 'How many tones have you heard?' for CLS and pure-tone audiometry assessments, respectively. The response data were returned to the server via WLAN and stored in the cloud database. Based on the incoming response, the server prepared the adjusted stimulus (here, the adjustment mainly refers to the sound levels for both listening tests) and played it back to the listener. The listener was redirected to the dashboard when the listening test was

completed. No data were stored locally on the smartphones but, instead, were primarily stored on the server.

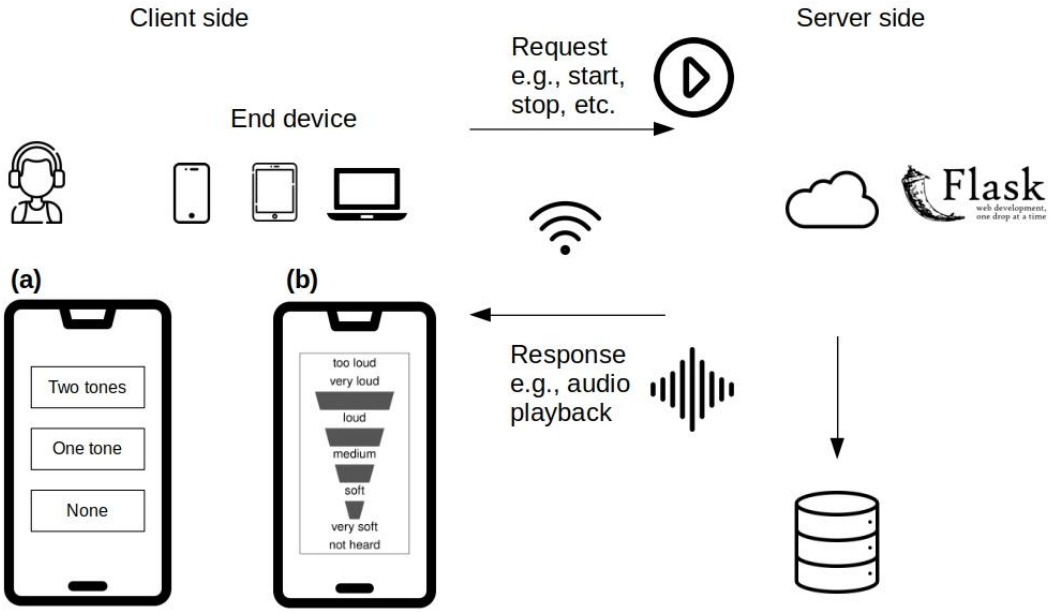


Fig. 1. Overview of the smartphone application. On the client side, the user interface of two assessments (i.e., (a) pure-tone audiometry and (b) categorical loudness scaling) is shown. On the server side, the web application framework 'FLASK' is available for processing requests from a listener. The measurement data was not stored locally but in the cloud database.

Subject groups

21 normal hearing (NH, aged between 20 and 35 years; 7 males, 14 females) and 16 hearing impaired listeners (HI, aged between 67 and 88 years; 11 males, 5 females) participated in the study. The participants in the NH group are mainly members of the working group and students of the university. The HI listeners were recruited via the database of Hörzentrum Oldenburg gGmbH. The mild-to-moderately impaired listeners with sensorineural hearing loss exhibited pure-tone averages (PTA) varying between 26.3 and 42.5 dB while NH listeners yielded thresholds at or below 15 dB for all frequencies between 250 Hz and 4 kHz. The differences in PTA between the left and

right ears of HI listeners did not exceed 10 dB, indicating that the hearing loss of all HI listeners was symmetric. All participants did not have any previous experience with smartphone hearing tests. The listeners received an expenditure compensation of 12 euros per hour for their participation in the study. The research ethics committee of the University of Oldenburg approved the proposal (Drs. EK/2022/011) for this study.

Table 1. Summarized statistics (i.e., average and standard deviation) on pure-tone thresholds (in dB HL) of the group of hearing-impaired listeners (N =16) for both ears with frequency varying from 0.125 to 8 kHz (11 frequencies) measured by the clinical audiogram (IEC 60645-1, 2002).

		Frequency (Hz)											
		125	250	500	750	1000	1500	2000	3000	4000	6000	8000	PTA[a]
Left	Ave.	12.5	12.1	18.9	22.9	24.6	31.4	38.9	49.6	53.9	59.6	67.1	34.1
	SD	7.3	7.8	7.9	8.9	7.5	7.9	8.4	9.5	10.0	8.0	7.3	4.9
Right	Ave.	15.0	16.1	21.4	23.6	24.3	30.7	35.3	45.0	50.0	56.4	66.1	32.8
	SD	9.0	10.6	9.3	9.1	8.1	9.2	9.1	10.7	10.0	9.1	9.2	5.5

[a] PTA (i.e., pure-tone average) denotes the average thresholds of 0.5, 1, 2, and 4 kHz

Table 1. provides the means and standard deviations of the clinical audiogram (IEC 60645-1, 2002) as a function of 11 frequencies, together with pure tone average (PTA), for both ears of HI listeners measured by an audiologist with HDA200 headphones. The PTAs for better ears of HI listeners were 31.8 (± 5.3) while the mean PTA difference was less than 2 dB.

Test conditions

Table 2. reports the difference in the experimental design of the three conditions. Overall, the experimental design is a repeated measures design. The main difference among the different conditions was the degree of supervision. Condition I was a fully-

supervised, manual measurement as reference. Condition III was a non-supervised assessment. Condition II was semi-supervised, i.e., the test examiner was available on request for questions while the experiment ran automatically under the control of the same adaptive procedure as for condition III. Specifically, the examiner did not have access to the log data in condition II and only answered questions.

Table 2. Experimental design for the three conditions employed that differed in the degree of supervision. Condition I implements a fully-supervised, manual laboratory measurement while condition III implements a non-supervised, automatic smartphone assessment, and condition II ("in-between condition") represents a semi-supervised condition using a self-controlled data acquisition on a laboratory setup. All three experiments had the same acoustic environment (i.e., a sound-treated listening booth).

	Supervision	Automation	Sound	Apparatus	Headphone	Calibration	Environment
Condition I (Reference)	Fully	Manual	Focusrite Scarlett 2i2	HP ENVY x360 Laptop	Sennheiser HDA 200	Yes	Sound-treated booth
Condition II	Semi[a]	Automated					
Condition III	Non		Built-in	OnePlus Android			

[a] Test supervisor available on request

Furthermore, the sound card for conditions I & II was Scarlett 2i2, while the built-in sound card of the smartphone was employed in condition III. HP ENVY x360 laptop was used for conditions I & II while the Android smartphone (OnePlus Nord N10 5G 128 GB, google chrome downloaded) was used for condition III. The same calibrated smartphone was provided to all participants.

Finally, in all three conditions the same HDA200 headphone was employed in a sound-attenuated booth. All conditions were calibrated employing a B&K artificial ear

4153, a B&K 0.5-inch microphone 4134, a B&K microphone pre-amplifier 2669, and a B&K measuring amplifier 2610. The target level for calibration was 80 dB SPL.

Data analysis

Psychophysical parameters

As already mentioned before, L_{50} (i.e., the half-way point of the psychometric function) in the pure-tone audiometry experiment was estimated as the hearing threshold, described in Eq. 1:

$$p(L) = 1/(1 + e^{-4s(L-L_{50})}) \quad (1)$$

where p(L) is the probability of correct responses, L defines the sound level, and s denotes the slope of the half-way point of the function. Moreover, the signed difference between condition II or III, respectively and I is defined as $L_{50,II/III} - L_{50,I}$, where $L_{50,II/III}$ denotes the hearing threshold measured in condition II or III, respectively, while $L_{50,I}$ is the hearing threshold measured in condition I. In addition, the absolute value of the difference is described as $|L_{50,II/III} - L_{50,I}|$. Finally, the root mean square error (RMSE) and the Pearson correlation coefficient (R) of conditions II and III against I are calculated.

For the categorical loudness scaling experiment, loudness functions as defined in Brand and Hohmann (2002) and Oetting et al., (2014) were employed (cf. Eq. 2), which consist of two linear parts and one transition region using a Bezier fit:

$$F(L) = \begin{cases} 25CU + m_{low}(L - L_{cut}) \text{ for } L \leq L_{15} \\ bez(L, L_{cut}, L_{15}, L_{35}) \text{ for } L_{15} < L < L_{35} \\ 25CU + m_{high}(L - L_{cut}) \text{ for } L \geq L_{35} \end{cases} \quad (2)$$

where m_{low} and m_{high} denote the slope value of the low and high linear part, L_{cut} is the intersection level of the two linear parts, L_{15} and L_{35} are the levels of the 'soft'

293 and 'loud' category respectively, and bez is a quadratic smoothing function between L_{15}
294 and L_{35}. The Pearson correlation coefficient (R), root mean square error (RMSE), and
295 bias of levels for each category (in total 11 categories) are calculated. For binaural
296 loudness summation, the level difference for equal loudness (LDEL) is calculated as:

$$\mathrm{LDEL} = L_b - L_l \quad (3)$$

297 where L_b and L_l are defined as the level for binaural and monaural presentation
298 of the **left ear** at the same category unit (i.e., equal loudness) respectively. The LDEL of
299 the **left ear** for spectral loudness summation is described as:

$$\mathrm{LDEL} = L_{LNN} - L_{UEN17} \quad (4)$$

300 where L_{LNN} and L_{UEN17} denote the level for low-noise narrowband noise and
301 UEN17 broadband noise at the same category unit respectively. All algorithms for
302 experimental data fitting were developed in MATLAB R2021a (The MathWorks, Inc.,
303 Natick, MA).

304 *Statistical analysis*

305 A mixed-design ANOVA was applied using degree of hearing loss (two levels: NH/HI)
306 as a between-subject factor, condition (three levels: I/II/III), and frequency (three levels:
307 0.25, 1, and 4 kHz) as within-subject factors. Furthermore, a post-hoc analysis among
308 conditions using a pair-wise t-test was carried out, where the p value was corrected with
309 'Bonferroni'. In the post-hoc analysis, condition I was set up as a reference group. If p
310 value < 0.05 (*), 0.01 (**), 0.001 (***), and 0.0001 (****), the statistical test is
311 considered as being significant, highly significant, very highly significant, and
312 extremely significant, respectively, while if p value >= 0.05 (ns), the result is not
313 significant, implying that there is no difference between two conditions. The
314 'Tidyverse' package (Wickham et al., 2019) developed in the software environment 'R'

(R Foundation for Statistical Computing) was employed for the statistical analysis of the mixed-design ANOVA and the post-hoc analysis.

Results

Experiment I: Pure-Tone Audiometry

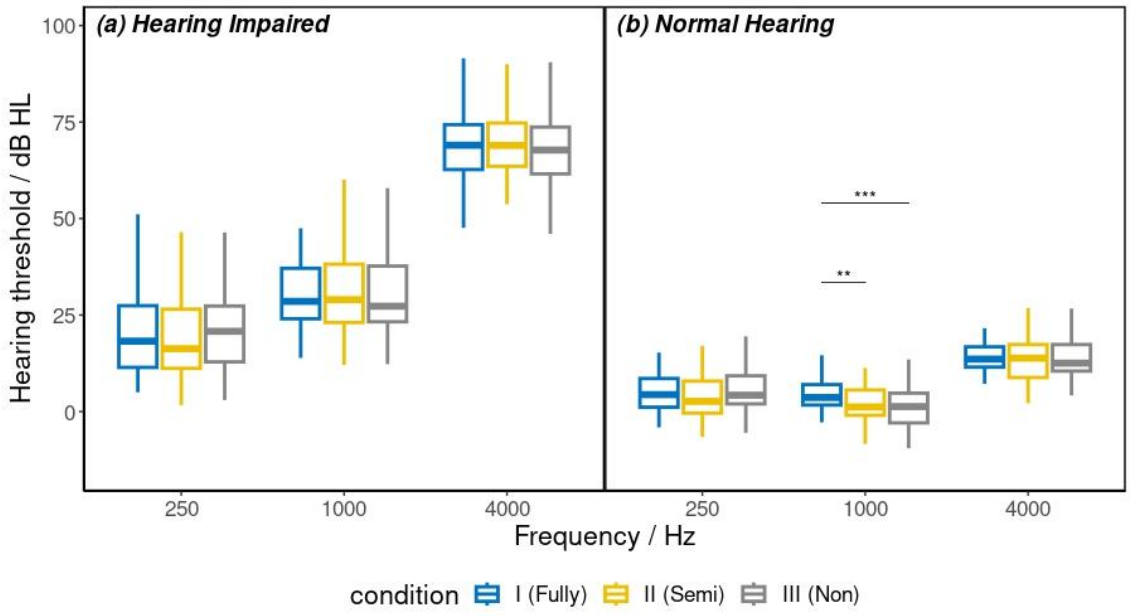


Fig. 2. Hearing threshold L_{50} (in dB HL) grouped by three conditions (I: Fully-, II: Semi-, III: Non-supervised) for (a) hearing impaired (HI) and (b) normal hearing (NH) listeners as a function of three frequencies (0.25, 1, and 4 kHz). Condition I was set up as a reference. The medians, 25%, 75% percentiles, and interquartile ranges (IQR) are given in the respective bar-and-whiskers plot. The ends of the whiskers describe values within 1.5*IQR of the 25% and 75% percentiles. In case of statistically significant differences, the level of significance is labeled with stars above the lines.

Fig. 2 compares hearing thresholds for HI and NH participants at 0.25, 1, and 4 kHz frequencies among the three conditions with decreasing amount of supervision. In general, median thresholds of conditions II and III were in line with those of condition I for all groups and frequencies. As expected, median thresholds of HI were higher than

NH for all three frequencies. Furthermore, median thresholds of HI listeners at 4 kHz were the highest, followed by 1 kHz and 0.25 kHz.

A three-way mixed-design ANOVA was performed to analyze the effect of the degree of hearing loss (NH/HI), frequency (0.25, 1, and 4 kHz), and condition (I/II/III) on hearing threshold L_{50}, revealing that there was a significant difference in hearing thresholds for the degree of hearing loss ($p < 0.05$) and frequency ($p < 0.05$) while no significant difference for condition ($p = 0.22$) was detected. The post-hoc analysis compared hearing thresholds of conditions II and III against I, indicating that conditions II and III did not significantly differ from condition I for all three frequencies within both listener groups except for the NH group at 1 kHz.

Statistical values, i.e., the signed difference, its absolute value, RMSE, R, and p value significance level of the thresholds L_{50} of conditions II and III against the reference condition I for two listener groups and three frequencies, are summarized in Table 3. Comparing the thresholds between conditions II and I (upper half of Table 3), mean signed differences were less than 2 dB in most cases, while mean absolute differences were around 3 dB. All RMSE values were smaller than 5 dB. The R values of HI listeners were higher than 0.9, suggesting a strong positive correlation while the R values of NH listeners were higher than 0.65, indicating a moderately positive correlation. Regarding the comparison of L_{50} between conditions III and I (bottom half of Table 3), mean signed differences were less than 1 dB except for NH listeners at 1 kHz. Similar to the comparison between conditions II and I, the mean absolute differences were around 3 dB, RMSE values were less than 5 dB and there was a strong correlation in the NH group while a moderately positive correlation was found in HI listeners.

Table 3. Quantitative comparison of the measured thresholds L_{50} between conditions II and I, and III and I in terms of signed difference, absolute difference, RMSE, R, and p value significance[a] level.

	Subject	Frequency	Mean signed difference (SD)	Mean absolute difference	RMSE	R	P value significance
Cond.II-Cond.I	HI	250	-0.7 ± 3.36	2.7 ± 2.07	3.4	0.96	ns
		1000	0.1 ± 4.28	3.3 ± 2.69	4.2	0.94	ns
		4000	0.2 ± 2.81	2.3 ± 1.58	2.8	0.97	ns
	NH	250	-1.9 ± 4.28	3.8 ± 2.69	4.6	0.69	ns
		1000	-2.4 ± 4.23	3.8 ± 3.07	4.8	0.78	**
		4000	-0.6 ± 4.99	3.8 ± 3.21	5.0	0.66	ns
Cond.III-Cond.I	HI	250	0.8 ± 4.01	3.1 ± 2.60	4.0	0.94	ns
		1000	0.3 ± 5.71	3.4 ± 4.58	5.6	0.88	ns
		4000	-0.5 ± 4.46	3.5 ± 2.72	4.4	0.92	ns
	NH	250	0.1 ± 4.78	3.7 ± 2.99	4.7	0.60	ns
		1000	-3.9 ± 4.00	4.4 ± 3.31	5.5	0.80	***
		4000	-0.4 ± 3.83	3.0 ± 2.39	3.8	0.76	ns

[a] ns: not significant; *p<0.05; **p<0.01; ***p<0.001; ****p<0.000

Experiment II: Adaptive Categorical Loudness Scaling

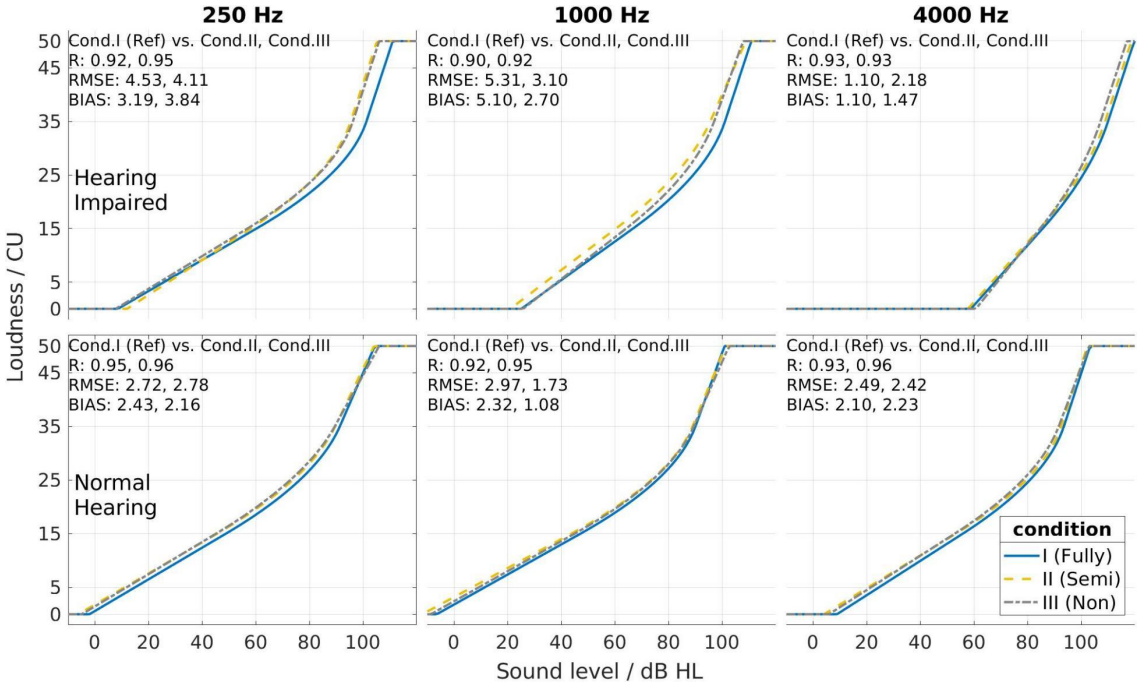


Fig. 3. Average loudness growth function (i.e., loudness in CU as a function of sound level in dB HL) of the three experimental conditions employed (condition I = fully-supervised; II = semi-supervised; III = non-supervised) for HI (upper row) and NH (bottom row) listeners at 0.25 kHz (left column), 1 kHz (middle column), and 4 kHz (right column). The Pearson correlation coefficients (R), root mean square errors (RMSE), and biases between two conditions II and III against I (reference) of levels for each category units are provided in the upper left corner of each sub-figure.

Fig. 3 plots the average loudness function of three conditions for HI and NH listeners at 0.25, 1, and 4 kHz frequencies. For all frequencies and listener groups, the average loudness functions of conditions II and III were consistent with condition I. The average loudness functions of HI listeners generally showed steeper growth than NH listeners, especially at 4 kHz, which could be explained by the 'loudness recruitment', as mentioned above. HI listeners exhibited a significant increase in the slope of the loudness function with an increase in frequency which was not observed in NH listeners.

Quantitatively speaking, the Rs of conditions II/III against I were higher than 0.9 for both NH and HI listeners at all three frequencies, indicating a rather high correlation of average loudness functions between conditions II and I, and between conditions III and I. HI listeners exhibited RMSE values less than 5 dB for most of the cases except for the comparison between conditions I and II at 1 kHz. NH listeners even produced a less than 3 dB RMSE value for all cases. Similarly, the bias for HI listeners was less than 4 dB and for NH listeners less than 3 dB with one exception occurring for HI listeners between conditions I and II at 1 kHz. Overall, the statistical measures suggested that the loudness function of conditions II and III showed a great agreement with condition I.

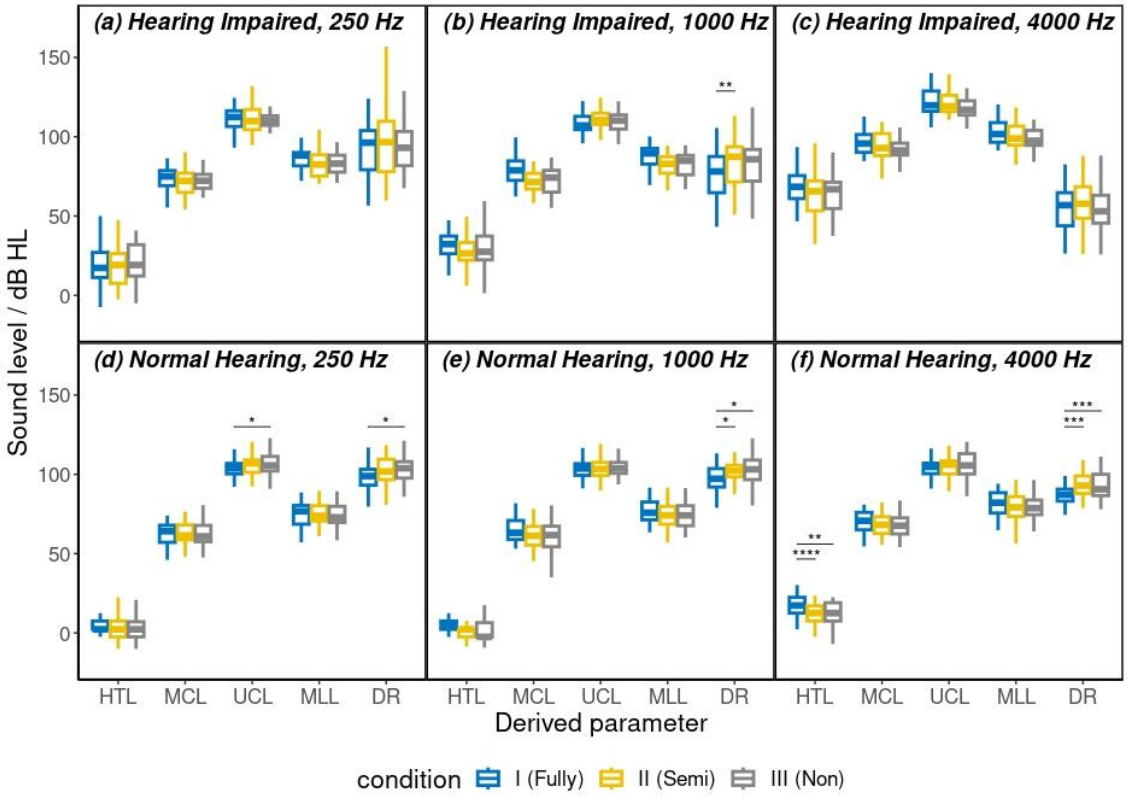


Fig. 4. Five descriptive and intuitive parameters (in dB HL) derived from the loudness function of three conditions for HI and NH listeners at 0.25, 1, and 4 kHz frequencies. HTL: hearing threshold level (2.5 CU); MCL: most comfortable loudness level (20 CU); UCL: uncomfortable loudness level (50 CU); MLL: median loudness level (25 CU); DR: dynamic range (UCL-HTL). See Fig. 2 for an explanation of the bar-and-whiskers plot.

Five descriptive parameters (i.e., HTL, MCL, UCL, MLL, and DR) of three conditions for HI and NH listeners at 0.25, 1, and 4 kHz are shown in Fig. 4. The median descriptive parameters for all three frequencies and both listener groups in conditions II and III were close to the condition I. Moreover, the median levels of the five descriptive parameters did not change with an increase in frequency for NH listeners. As expected, the median levels of HTL increased while DR decreased with an increase in frequency for HI listeners. The IQRs of HTL and DR were larger for HI listeners compared to NH listeners.

The effect of hearing impairment (NH/HI), frequency (0.25, 1, and 4 kHz), and condition (I/II/III) on five descriptive parameters (HTL, MCL, UCL, MLL, and DR) was assessed via the five different mixed-design ANOVA tests. The results revealed that there was a significant main effect of the degree of hearing loss and frequency on all five descriptive parameters ($p < 0.05$). Moreover, the factor condition was not significant on UCL ($p = 0.12$) while was significant on the other four descriptive parameters ($p < 0.05$).

A pair-wise t-test was performed to assess whether there was a significant difference in levels between conditions II and I, and III and I, respectively. For HI listeners, all five descriptive parameters of conditions II and III did not significantly differ from the condition I at all frequencies except for DR at 1 kHz. Furthermore, for NH listeners at 0.25 kHz, there was a significant difference in UCL and DR between conditions III and I ($p < 0.05$). At 1 kHz, the differences in DR between conditions II and I, and conditions III and I were significant. At 4 kHz, the differences across conditions of HTL and DR were significant.

As the t-test and the mixed-designed ANOVA test typically assume the 'homogeneity of variance' (i.e., all groups have the same variance), our data might violate the assumption (i.e., NH listeners have a smaller variance than HI listeners, as shown in Fig. 4), and thus the validity of the statistical tests might be affected. This would lead to falsely rejecting the null hypothesis (i.e., the factor condition is not supposed to be significant but reported to be significant).

Taken together, while for most cases the five parameters did not differ between the reference condition I and the less supervised conditions II and III, respectively, statistically significant differences only existed in a few groups, suggesting that these significant differences might not be systematic differences but rather random differences. In addition, the magnitudes of the significant differences in the NH and HI groups were overall less than 5 dB, indicating that the differences might not be clinically relevant. As we always measured condition I first, the sequence or training effect might explain such a difference.

Experiment III: Binaural and Spectral Loudness Summation

Binaural loudness summation

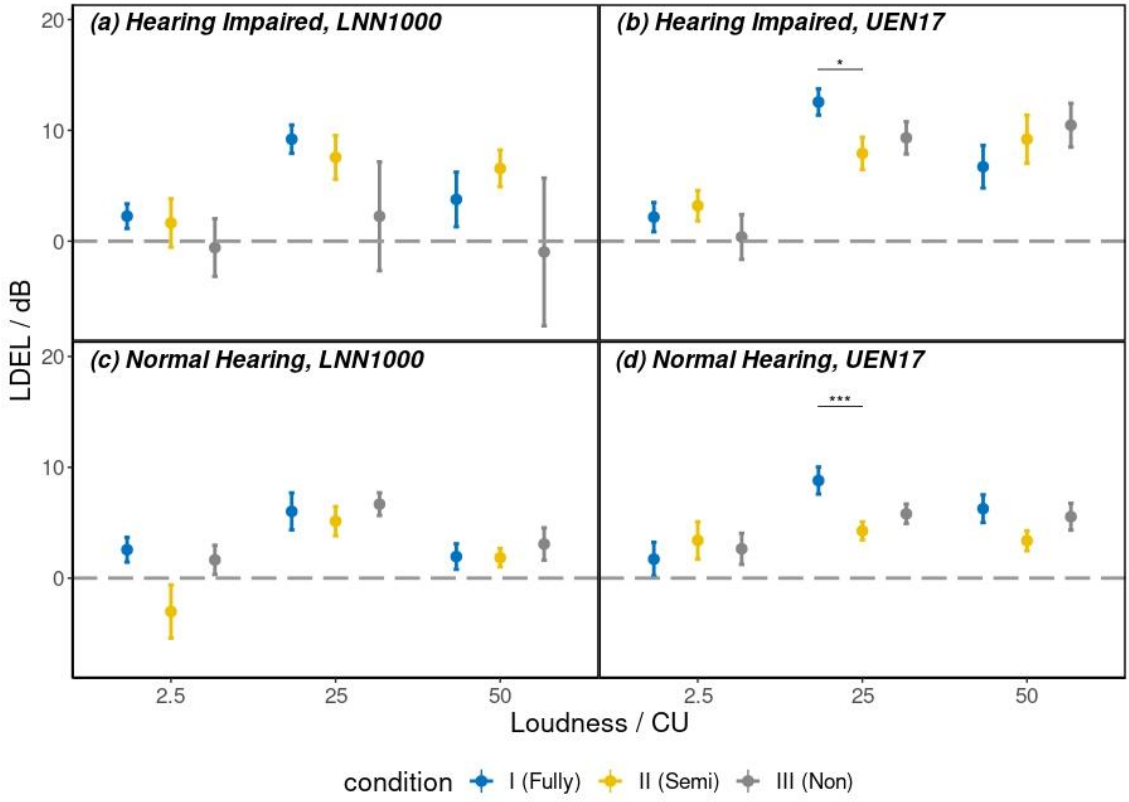


Fig. 5. Mean and standard deviation (denoted by whiskers) of level difference for equal loudness (LDEL, in dB) between binaural and monaural (left ear) presentation for equal loudness at 2.5, 25, and 50 CU using narrowband noise (LNN1000) and UEN17 broadband noise, respectively, for HI (upper row) and NH (bottom row) listeners. Conditions I, II, and III are differentiated with three colors. Grey dashed line: 0 dB. LNN1000: one-third-octave-band centered at 1 kHz low-noise noise; UEN17: uniformly exciting noise at 17 critical bands.

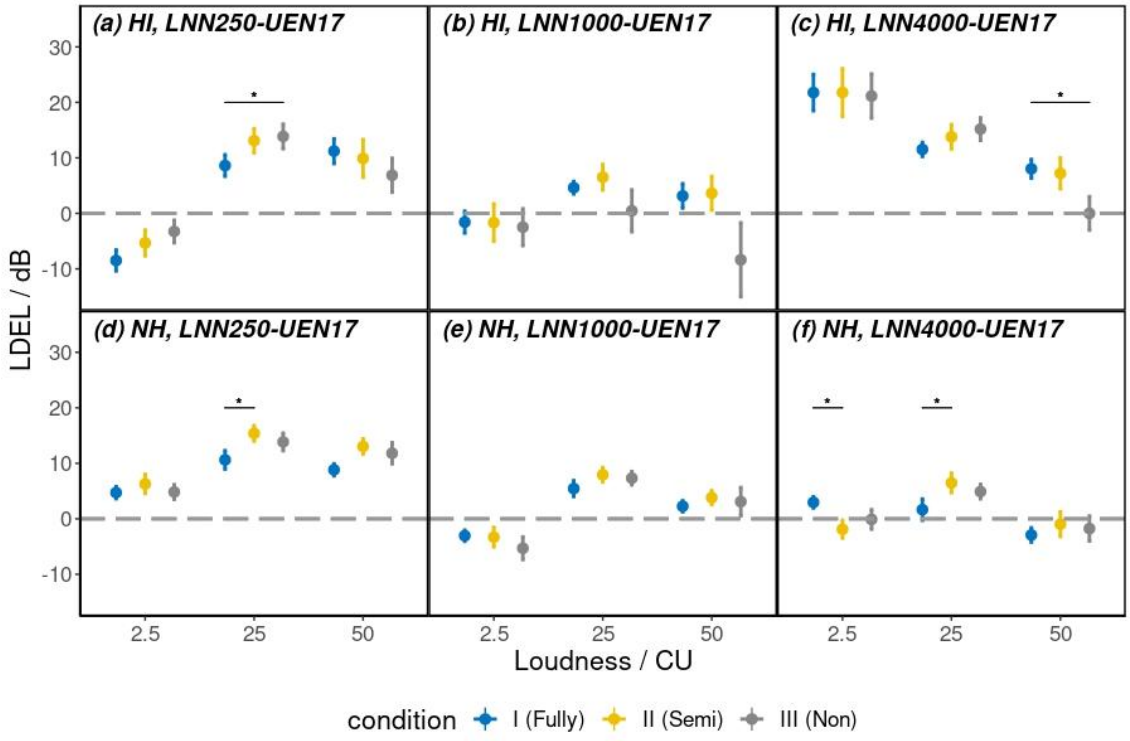
Mean and standard deviation of the level differences for equal loudness (LDELs) as a function of loudness in CU of HI and NH participants for LNN1000 and UEN17 among three conditions are shown in Fig. 5. In most cases, the mean LDELs of conditions III and II were in agreement with those of condition I. It is notable that the standard deviation of LDEL of the condition III for LNN1000 at 25 and 50 CU for HI listeners

was considerably larger than conditions II and I. Binaural loudness summation was signaled by mean LDELs significantly larger than 0, which was observed in most groups. Exceptions were observed for the HI listener at 2.5 and 50 CU of the condition III and NH listener at 2.5 CU of the condition II stimulated by LNN1000. Generally, the LDELs of 25 CU were the highest except for HI listeners of conditions II and III stimulated by UEN17.

A four-way mixed-design ANOVA (NH/HI as a between-subject factor, condition (I/II/III), frequency (0.25, 1, and 4 kHz), and loudness (2.5, 25, and 50 CU) as within-subject factors) was conducted to assess the effect on LDEL. There was a significant main effect on the degree of hearing loss ($p < 0.05$), frequency ($p < 0.05$), and loudness CU ($p < 0.05$). The main effect of the condition was, however, not significant ($p = 0.4$).

Despite the insignificant main effect of the condition, the post-hoc analysis employing a pairwise t-test with 'Bonferroni' adjustment was carried out on LDEL, where condition I was the reference. In general, the LDEL of conditions II and III did not differ from condition I. However, a significant difference occurred in some pairs, i.e., between conditions I and II at 25 CU for both NH ($p < 0.05$) and HI ($p < 0.001$) stimulated by the UEN17 broadband signal. Even though these differences were statistically significant, the mean values of the differences were roughly 6 dB. Thus, similar to the results above, the significant differences in statistics might not be clinically relevant differences.

466 *Spectral loudness summation*



467

468 Fig. 6. Mean and standard deviation (denoted by whiskers) of level difference for equal
469 loudness (LDEL) between three narrowband stimuli (LNN250, left; LNN1000, middle;
470 LNN4000, right) and one broadband stimulus (UEN17) for equal loudness at 2.5, 25,
471 and 50 CU for HI (upper row) and NH (bottom row) listeners. Grey dashed line: 0 dB.
472 All signals were presented monaurally on the left ear. LNN250, 1000, 4000: one-third-
473 octave-band centered at 0.25, 1, and 4 kHz low-noise noise; UEN17: unified excitation
474 noise at 17 critical bands.

475

476 Fig. 6 shows LDEL (with error bars) of three conditions as a function of loudness in CU
477 between LNN250 and UEN17 (left), LNN1000 and UEN17 (middle), and LNN4000
478 and UEN17 (right) for HI (upper) and NH (bottom) listeners. Generally, the mean
479 difference of LDEL between conditions II and I, and between III and I was small with
480 values smaller than 10 dB. For HI listeners, the mean LDELs at 25 and 50 CU were
481 greater than 0 while lower than 0 at 2.5 CU concerning the comparison between
482 LNN250 and UEN17. However, the mean LDELs of NH listeners were larger than 0 at

three CU. Comparing the LDELs between LNN1000 and UEN17, both NH and HI listeners exhibited a negative LDEL at 2.5 CU while positive at 25 and 50 CU for three conditions with one exception of the HI listener for the condition III at 50 CU. Regarding the mean LDEL difference between LNN4000 and UEN17, NH and HI participants showed a substantial difference: the mean LDELs of HI listeners were always positive, while NH listeners were around 0.

A four-way mixed-design ANOVA was conducted to assess the effect of the degree of hearing loss (HI/NH), condition (I, II, and III), comparison (LNN250-UEN17, LNN1000-UEN17, LNN4000-UEN17), and loudness (2.5, 25, and 50 CU) on LDELs, in which the first factor was set up as a between-subject factor while the latter three as within-subject factors. The statistical outcome of ANOVA revealed that the main effect of all four factors was significant ($p < 0.05$).

A pair-wise t-test as a post-hoc analysis was performed to check whether the LDEL between conditions II/III and I was significantly different. The results suggested that in most cases, the LDEL of conditions II and III did not significantly differ from the LDEL of the condition I. For HI participants, there was only a significant difference between conditions I and III on LDEL at 25 CU in comparison pairs of LNN250-UEN17 and at 50 CU of LNN4000-UEN17 ($p < 0.05$). For NH listeners, only the difference in LDEL between conditions I and II was significant at 25 CU of LNN250-UEN17, and at 2.5 and 25 CU of LNN4000-UEN17 ($p < 0.05$).

Discussion

Performing pure-tone audiometry and categorical loudness scaling on a smartphone was demonstrated here to be feasible if the smartphone is calibrated properly, the ambient noise is under control and the adaptive procedure provides high precision. The test outcome on a smartphone appears to be valid since it is aligned with the laboratory

508 measurement in most cases. The way of supervision does not have a general impact on
509 the measurement results, i.e., the non-supervised automated tests performed here are in
510 principle equivalent to the fully-supervised manual tests.

511 The smartphone hearing tests employed here are applicable and accessible not
512 only for normal hearing participants but also for persons with a hearing loss. It is useful
513 and not difficult for HI listeners to administer the measurements themselves on a
514 smartphone if they are familiar with the procedures. On top of the commonly employed
515 unaided ACALOS measurement, i.e., narrowband signal presented unilaterally, the
516 broadband stimulus for binaural presentation is also evaluated on a smartphone and
517 does not show a large difference compared to the lab test. The usage of a variety of
518 stimuli for adaptive categorical loudness scaling might support fine-tuning for a non-
519 linear hearing aid on a smartphone in the future.

520 ***Pure-tone audiometry***

521 A number of studies have considered the difference between an app-based tone-in-quiet
522 measurement and the clinical audiogram in order to validate the respective app on the
523 mobile device. They either do the comparison in a clinical sound-insulated environment
524 for both cases (e.g., Swanepoel et al., 2010; Colsman et al., 2020; Hazan et al., 2022) -
525 which is supposed to yield no difference due to the acoustic presentation mode – or in a
526 “quiet everyday environment” for the app (e.g., Kam et al., 2012; Abu-Ghanem et al.,
527 2016) where any observed difference may be due to acoustical reasons (i.e., low-
528 frequency noise components that can hardly be suppressed by ear-level devices), due to
529 procedural differences (e.g., distraction due to attention-demanding occurrences in daily
530 life, see Xu et al., 2023), or due to device calibrations. Typically, those studies that
531 perform the validation under similar clinical, acoustically controlled, and distraction-
532 sparse conditions as in our study agree with our study by reporting only a very small

mean (signed) difference (e.g., within 5 dB as revealed in Thai-Van et al., 2022) across conditions. In addition, in a relatively noisy environment, fewer differences are expected for hearing-impaired listeners since their audiometric results would only be affected by higher ambient noise levels than normal-hearing listeners. As pointed out by Swanepoel et al. (2010), as HI listeners typically have a reduced hearing sensitivity, the apparent awareness of the internal noise level in NH listeners is largely eliminated.

Adaptive categorical loudness scaling

To our knowledge, there is no study so far evaluating categorical loudness scaling on a smartphone. Our experimental results provide the first evidence that it is plausible and valid to perform non-supervised CLS measurement on a smartphone both for NH and HI listeners. In addition, there is only one study so far, i.e., Kopun et al. (2022), which evaluated the CLS measurement on a laptop remotely in comparison to a clinical database. This is comparable with the comparison between conditions II and I in our study on the group level. Kopun et al. (2022) reported that for NH participants (N = 5), the mean signed difference averaged across categories was 5.9 and 4.9 at 1 and 4 kHz, respectively. The mean signed difference of our study is much smaller, i.e., 2.3 and 2.1 for 1 and 4 kHz. First, the fitting of the loudness function might play a role. Kopun et al. (2022) simply calculated the median level of each category to describe the individual loudness function without fitting the data to a 2-segment linear function. Second, the outliers were not removed, leading to non-monotonic loudness growth. This contrasts to our study where we fitted the individual responses based on the method introduced in Oetting et al. (2014) to obtain an individual monotonic loudness function. Third, the test environment might make an impact. We conducted all experiments in a sound-attenuated booth to eliminate the influence of environmental noise. Kopun et al. (2022), however, did in-lab measurements at a sound-treated booth while remote laptop

558 measurements at home. Although Kopun et al. (2022) attempted to control and check
559 the noise level between runs in the remote measurements, the fluctuating environmental
560 noise might influence the loudness judgment during the run. Fourth, Kopun et al. (2022)
561 used a different calibrated headphone (i.e., Sennheiser HD 280 Pro). Lastly, the time
562 gap between conditions II and I in Kopun et al. (2022) ranged from 2 years 6 months to
563 2 years 9 months while our time gap was less than a day. Overall, these differences not
564 only in the experimental setup but also in the data processing would explain why our
565 study exhibits a higher reproducibility than the earlier study, indicated by a smaller
566 mean signed difference.

567 The descriptive parameters (i.e., HTL, MLL, UCL, and DR) of our study
568 measured with a smartphone for NH listeners match quite well with the reference values
569 reported in Oetting et al. (2016). The mean difference of the 4 parameters between
570 Oetting et al. (2016) (N = 9) and our results is less than 2 dB at 0.25 kHz while lying
571 within one standard deviation at 1 and 4 kHz. Furthermore, our measured MLLs and
572 DRs are quite consistent with the empirical values for young NH listeners (N = 11) and
573 HI listeners (N = 70) provided by Sanchez-Lopez et al. (2021). The median MCLs and
574 DRs of NH listeners reported by Sanchez-Lopez et al. (2021) were 70 and 97.5 dB HL
575 at low frequencies, and 75 and 92.5 dB HL at high frequencies while the median MCLs
576 and DRs of listeners measured by us were 73.5 and 103.5 dB HL for low frequencies,
577 and 78.7 and 90.6 dB HL for high frequencies. The difference between Sanchez-Lopez
578 et al. (2021) and our study is around 5-6 dB and relatively small. Comparing the HI
579 listeners of Sanchez-Lopez et al. (2021), most of our measured parameters for both low
580 and high frequencies stay within the 25% and 75% percentile range of Sanchez-Lopez
581 et al. (2021) except for MCLs at high frequencies. One possible reason might be
582 different high frequency measurements: we only measured 4 kHz while Sanchez-Lopez

et al. (2021) measured 2, 4, and 6 kHz and averaged the values of MCL. Another explanation could be that individual (within-subject) preference for MCLs might vary. Overall, the descriptive parameters measured by a smartphone show good consistency with the empirical values reported in the literature for both NH and HI listeners.

The three conditions differing in degree of supervision with calibrated hardware appear not to systematically influence the results of CLS in terms of both loudness growth functions and derived parameters (as shown in Fig. 3 and revealed by the mix-designed ANOVA), implying that we could let the participants test themselves on a smartphone for the CLS test, which meets our expectations. One reason to explain the results might be that the task for loudness judgment is rather intuitive and natural based on the feedback from our participants covering both NH and HI listeners. In addition, CLS is a supra-threshold measurement, which is expected to be less prone to influence by factors such as hardware and environment. Unlike some other speech-related tasks, e.g., the speech-in-noise test or listening effort test which are rather cognitively demanding, the CLS task does not involve speech comprehension, and, therefore, should be rather robust without any additional assistance from experimenters.

Binaural and Spectral Loudness Summation

Level differences for equal loudness (LDELs,) - that quantify the binaural and spectral loudness summation - mostly do not show differences between the standard in-lab and smartphone measurements. This indicates that the smartphone measurements could detect the binaural and spectral loudness summation as well as the assessment conducted in a laboratory. However, we find that the factor condition shows a significant main effect on LDEL for the spectral loudness summation and in some groups, there is a significant difference in LDEL between conditions, as revealed by the post-hoc t-tests. Despite the (unexpected) significant difference in statistics, the values

608 of the difference in LDEL are generally below 10 dB, which might not be considered to
609 be clinically significant (e.g., in Thai-Van et al., 2022, 10 dB difference is defined as a
610 criterion to determine the 'clinical equivalence').

611 A similar amount of binaural loudness summation for NH listeners can be
612 observed in our study as reported by Oetting et al. (2016), indicating that the binaural
613 LDELs for both broadband and narrowband signals are highest at 25 CU and lowest at
614 2.5 and 50 CU. Furthermore, the broadband signal exhibits higher LDELs than the
615 narrowband signal. For broadband signals, a higher individual variability at high
616 loudness could be observed for HI than for the NH listeners, which is compatible with
617 Oetting et al. (2016). Whilby et al. (2006) examined 1-kHz pure tones for HI listeners,
618 suggesting that LDELs were around 6 dB at medium loudness levels, decreased towards
619 lower levels, and exhibited high individual variability. Their findings are quite
620 comparable with our results, although we employ a different stimulus (i.e., 1 kHz one-
621 third octave noise).

622 Concerning the spectral loudness summation experiment, our results in general
623 are in line with Brand and Hohmann (2001). They reported that spectral LDELs were
624 around 25 dB for speech shaped noise at medium loudness, and decreased towards
625 lower and higher loudness for NH listeners (N = 8). We have a similar trend but smaller
626 values of LDELs. This might be explained by the applied broadband signal: in our case,
627 it is UEN17 while speech-shaped noise with different speech spectra was employed by
628 Brand and Hohmann (2001). For HI listeners (N = 8), Brand and Hohmann (2001)
629 showed that LDELs were approximately 10 dB and decreased with lower loudness,
630 which is in line with our results.

631 Loudness scaling and loudness matching appear to be the two main tools to
632 assess loudness summation for practical applications. Van Beurden et al. (2021)

compared the two measurement procedures and concluded that both procedures provided valid and reliable results. Loudness scaling, on one hand, provides information on the entire loudness range. It requires a simple categorical judgment task, which is quite intuitive even for the elderly and naïve participants while loudness matching is less intuitive and needs more instructions for the listeners who have to "equalize apples and pears", i.e., are forced to judge two differently perceived stimuli as being equal in one domain which is a challenge for inexperienced persons. On the other hand, loudness scaling might be more time-consuming than loudness matching. Even though we do not systematically compare the two methods on a smartphone, we prefer to apply loudness scaling on mobile devices since the feedback from our participants indicates that it is rather straightforward and easy to measure while using an acceptable measurement time.

Limitations and outlook

One major limitation of the pure-tone audiometry in this study is that we only measured three frequencies, which mainly cover the speech range. For more refined clinical diagnostics, it might be of interest to measure in total 11 frequencies for both ears similar to the clinical audiogram. However, for a rough classification of hearing loss and given the limited additional information of additional audiogram frequencies at the cost of a higher time effort, the choice of three frequencies is a compromise.

Our current study only considers conducting the smartphone measurements in a sound-treated booth in order to eliminate any effects of the environment on the measurement outcome (e.g., distraction or background noise). It is worthwhile to consider experiments outside the booth while still ensuring the quality of the audiometric data. A possible solution could be monitoring the real-time noise level during the measurement as Kopun et al. (2022), Swanepoel et al. (2014; 2015), Maclennan-Smith et al. (2013), and Serpanos et al. (2022) did. Another approach for

out-of-booth measurement could be using noise cancellation earphones (e.g., Clark et al., 2017).

The headphone employed here is a professional audiometric headphone (Sennheiser HDA200), which appears to be expensive and not publicly accessible. Van der Aerschot et al. (2016) recommended that affordable headphones, e.g., Sennheiser HD202 could be applied for pure-tone audiometry assessment. Moreover, the cheap headphones Sennheiser HD 280 Pro circumaural headphone was utilized by Kopun et al. (2022). Pickens et al. (2018) suggested that both, the Pioneer HDJ-2000 (Pioneer, Bunkyo, Tokyo, Japan) and the Sennheiser HD280 Pro (Sennheiser, Wedemark, Hanover, Germany) headphones, could be employed for mobile pure-tone audiometry assessment. The true wireless stereo (TWS) earbuds for pure-tone audiometry introduced by Guo et al. (2021) might also be considered as a daily-accessible alternative to the audiology headphone.

In our current study, we calibrated the smartphone output accurately in order to eliminate the influence of calibration and make it comparable to the standard laboratory measurement. However, in everyday life, the smartphone is normally not calibrated. How to treat the uncalibrated mobile device and additional hardware in non-laboratory setups remains a challenge. Kisić et al. (2022), for instance, proposed that human speech might be an appropriate and stable test signal for microphone calibration while Scharf et al. (2023) considered the whistling sound of a 0.33 l beer bottle as a rough calibration signal.

Conclusions

Three different experiments were designed to validate the usage of smartphone-based, non-supervised audiometric tests by studying the influence of the degree of supervision on audiometric tests to be performed with mobile devices:

683 - Experiment I (pure-tone audiometry) indicates that the way of supervision does
684 not influence the measurement outcome. More specifically, the mean signed difference
685 and mean absolute difference between smartphone and laboratory audiometry of NH
686 and HI listeners exhibit less than 1 dB and 4 dB, respectively, in most cases.
687 - Experiment II (Adaptive CLS measurement) reveals that supervision does not
688 affect the outcome values of categorical loudness scaling (i.e., the derived loudness
689 growth functions of NH and HI listeners). The bias between smartphone and in-lab
690 loudness function is considerably small and yields 2.67 and 1.8 dB for NH and HI
691 participants, respectively. In addition, the 5 intuitive parameters (i.e., HTL, MCL, MLL,
692 UCL, and DR) of smartphone CLS do not differ from the standard CLS assessment.
693 - Experiment III (binaural and spectral loudness summation) implies that
694 binaural and spectral loudness summation can be derived by employing a smartphone in
695 a way consistent with lab experiments. The LDELs measured on a smartphone between
696 unilateral and bilateral presentation to quantify binaural loudness summation for both
697 NH and HI listeners concerning both narrowband and broadband signals are consistent
698 with those measured inside an acoustics laboratory. A similar trend is observed for the
699 spectral loudness summation. Furthermore, the individual variations of HI listeners in
700 loudness summation at loudness uncomfortable levels for binaural broadband signals
701 are considerably large. Thus, in line with Oetting et al. (2016), including a binaural
702 broadband signal for measuring the loudness perception appears to be a valid
703 prerequisite for hearing aid fitting.
704 In conclusion, both audiometric tests considered here can be used for non-
705 supervised smartphone-based hearing examination and are expected to yield very
706 similar results as being conducted in a controlled laboratory experiment.

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2177/1 - Project ID 390895286.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Reference

Abu-Ghanem, S., Handzel, O., Ness, L., Ben-Artzi-Blima, M., Fait-Ghelbendorf, K., & Himmelfarb, M. (2016). Smartphone-based audiometric test for screening hearing loss in the elderly. European archives of oto-rhino-laryngology, 273(2), 333-339.

Bisgaard, N., Vlaming, M. S., & Dahlquist, M. (2010). Standard audiograms for the IEC 60118-15 measurement procedure. Trends in amplification, 14(2), 113-120.

Brand, T., & Hohmann, V. (2001). Effect of Hearing Loss, Centre Frequency, and Bandwidth on the Shape of Loudness Functions in Categorical Loudness Scaling: Efecto de la hipoacusia, la frecuencia central y el ancho de banda, en la configuración de la funciones de sonoridad en una escala categóries de sonoridad. Audiology, 40(2), 92-103.

Brand, T., & Hohmann, V. (2002). An adaptive procedure for categorical loudness scaling. The Journal of the Acoustical Society of America, 112(4), 1597-1604.

Buus, S., & Florentine, M. (2002). Growth of loudness in listeners with cochlear hearing losses: Recruitment reconsidered. JARO: Journal of the Association for Research in Otolaryngology, 3(2), 120.

Clark, J. G., Brady, M., Earl, B. R., Scheifele, P. M., Snyder, L., & Clark, S. D. (2017). Use of noise cancellation earphones in out-of-booth audiometric evaluations. International Journal of Audiology, 56(12), 989-996.

Colsman, A., Supp, G. G., Neumann, J., & Schneider, T. R. (2020). Evaluation of accuracy and reliability of a mobile screening audiometer in normal hearing adults. Frontiers in psychology, 11, 744.

736 Elberling, C. (1999). Loudness scaling revisited. Journal of the American Academy of
737 Audiology, 10(05), 248-260.
738 Erinc, M., & Derinsu, U. (2022). Behavioural and Electrophysiological Evaluation of
739 Loudness Growth in Clinically Normal Hearing Tinnitus Patients with and
740 without Hyperacusis. Audiology and Neurotology, 27(6), 469-477.
741 Fereczkowski, M., & Neher, T. (2023). Predicting Aided Outcome With Aided Word
742 Recognition Scores Measured With Linear Amplification at Above-
743 conversational Levels. Ear and Hearing, 44(1), 155-166.
744 Fultz, S. E., Neely, S. T., Kopun, J. G., & Rasetshwane, D. M.(2020). Maximum
745 expected information approach for improv-ing efficiency of categorical loudness
746 scaling.Frontiers in Psy-chology, 11,32–63.
747 https://doi.org/10.3389/fpsyg.2020.578352
748 Guo, Z., Yu, G., Zhou, H., Wang, X., Lu, Y., & Meng, Q. (2021). Utilizing True
749 Wireless Stereo Earbuds in Automated Pure-Tone Audiometry. Trends in
750 Hearing, 25, 23312165211057367.
751 Hallpike, C. S., & Hood, J. D. (1959). Observations upon the neurological mechanism
752 of the loudness recruitment phenomenon. Acta Oto-laryngologica, 50(3-6), 472-
753 486.
754 Hazan, A., Luberadzka, J., Rivilla, J., Snik, A., Albers, B., Méndez, N., ... &
755 Kinsbergen, J. (2022). Home-Based Audiometry With a Smartphone App:
756 Reliable Results?. American Journal of Audiology, 31(3S), 914-922.
757 Heller, O. (1985). Hörfeldaudiometrie mit dem Verfahren der Kategorienunterteilung
758 (KU). Psychologische Beitrage.
759 Hughson, W., Westlake, H. et al. (1944). “Manual for program outline for rehabilitation
760 of aural casualties both military and civilian,” Trans Am Acad Ophthalmol
761 Otolaryngol 48(Suppl), 1–15.
762 Hébert, S., Fournier, P., & Noreña, A. (2013). The auditory sensitivity is increased in
763 tinnitus ears. Journal of Neuroscience, 33(6), 2356-2364.
764 IEC 60645-1, 2002. Electroacoustics - Audiometric Equipment - Part 1: Equipmentfor
765 Pure-tone Audiometry. Standard of the International
766 ElectrotechnicalCommission, Geneva, Switzerland.
767 ISO 16832, 2006. AcousticsdLoudness Scaling by Means of Categories. Standard ofthe
768 International Organization for Standardization, Geneva, Switzerland.

769 Jürgens, T., Kollmeier, B., Brand, T., & Ewert, S. D. (2011). Assessment of auditory
770 nonlinearity for listeners with different hearing losses using temporal masking
771 and categorical loudness scaling. Hearing Research, 280(1-2), 177-191.
772 Kam, A. C. S., Sung, J. K. K., Lee, T., Wong, T. K. C., & van Hasselt, A. (2012).
773 Clinical evaluation of a computerized self-administered hearing test.
774 International Journal of audiology, 51(8), 606-610.
775 Kisić, D., Horvat, M., Jambrošić, K., & Franček, P. (2022). The Potential of Speech as
776 the Calibration Sound for Level Calibration of Non-Laboratory Listening Test
777 Setups. Applied Sciences, 12(14), 7202.
778 Kohlrausch, A., Fassel, R., Van Der Heijden, M., Kortekaas, R., Van De Par, S.,
779 Oxenham, A. J., & Püschel, D. (1997). Detection of tones in low-noise noise:
780 Further evidence for the role of envelope fluctuations. Acta Acustica united with
781 Acustica, 83(4), 659-669.
782 Kollmeier, B., & Hohmann, V. (1995). Loudness estimation and compensation for
783 impaired listeners employing a categorical scale. Advances in hearing research,
784 441-453.
785 Kollmeier, B., & Kiessling, J. (2018). Functionality of hearing aids: State-of-the-art and
786 future model-based solutions. International journal of audiology, 57(sup3), S3-
787 S28.
788 Kopun, J. G., Turner, M., Harris, S. E., Kamerer, A. M., Neely, S. T., & Rasetshwane,
789 D. M. (2022). Evaluation of Remote Categorical Loudness Scaling. American
790 journal of audiology, 31(1), 45-56.
791 Lecluyse, W., & Meddis, R. (2009). A simple single-interval adaptive procedure for
792 estimating thresholds in normal and impaired listeners. The Journal of the
793 Acoustical Society of America, 126(5), 2570-2579.
794 Maclennan-Smith, F., Swanepoel, D. W., & Hall III, J. W. (2013). Validity of
795 diagnostic pure-tone audiometry without a sound-treated environment in older
796 adults. International journal of audiology, 52(2), 66-73.
797 Oetting, D., Brand, T., & Ewert, S. D. (2014). Optimized loudness-function estimation
798 for categorical loudness scaling data. Hearing Research, 316, 16-27.
799 Oetting, D., Hohmann, V., Appell, J. E., Kollmeier, B., & Ewert, S. D. (2016). Spectral
800 and binaural loudness summation for hearing-impaired listeners. Hearing
801 Research, 335, 179-192.

Oetting, D., Hohmann, V., Appell, J. E., Kollmeier, B., & Ewert, S. D. (2018). Restoring perceived loudness for listeners with hearing loss. Ear and hearing, 39(4), 664-678.

Pickens, A. W., Robertson, L. D., Smith, M. L., Zheng, Q., & Song, S. (2018). Headphone evaluation for app-based automated mobile hearing screening. International Archives of Otorhinolaryngology, 22(04), 358-363.

Rasetshwane, D. M., Trevino, A. C., Gombert, J. N., Liebig-Trehearn, L., Kopun, J. G., Jesteadt, W., ... & Gorga, M. P. (2015). Categorical loudness scaling and equal-loudness contours in listeners with normal hearing and hearing loss. The Journal of the Acoustical Society of America, 137(4), 1899-1913.

Saak, S., Huelsmeier, D., Kollmeier, B., & Buhl, M. (2022). A flexible data-driven audiological patient stratification method for deriving auditory profiles. Frontiers in Neurology, 13, 959582.

Sanchez-Lopez, R., Nielsen, S. G., El-Haj-Ali, M., Bianchi, F., Fereczkowski, M., Cañete, O. M., ... & Santurette, S. (2021). Auditory tests for characterizing hearing deficits in listeners with various hearing abilities: The BEAR test battery. Frontiers in neuroscience, 15.

Scharf, M. K., Schulte, M., Huber, R., & Kollmeier, B. Microphone Calibration Estimation for Smartphones with Resonating Beer Bottles.

Seluakumaran, K., & Shaharudin, M. N. (2021). Calibration and initial validation of a low-cost computer-based screening audiometer coupled to consumer insert phone-earmuff combination for boothless audiometry. International journal of audiology, 1-9.

Serpanos, Y. C., Hobbs, M., Nunez, K., Gambino, L., & Butler, J. (2022). Adapting Audiology Procedures During the Pandemic: Validity and Efficacy of Testing Outside a Sound Booth. American Journal of Audiology, 31(1), 91-100.

Shiraki, S., Sato, T., Ikeda, R., Suzuki, J., Honkura, Y., Sakamoto, S., ... & Kawase, T. (2022). Loudness functions for patients with functional hearing loss. International Journal of Audiology, 61(1), 59-65.

Smits, C., Kapteyn, T. S., & Houtgast, T. (2004). Development and validation of an automatic speech-in-noise screening test by telephone. International journal of audiology, 43(1), 15-28.

Swanepoel, D. W., Matthysen, C., Eikelboom, R. H., Clark, J. L., & Hall III, J. W. (2015). Pure-tone audiometry outside a sound booth using earphone attentuation,

integrated noise monitoring, and automation. International Journal of Audiology, 54(11), 777-785.

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.

Swanepoel, D. W., Myburgh, H. C., Howe, D. M., Mahomed, F., and Eikelboom, R. H. (2014). "Smartphone hearing screening with integrated quality control and data management," International journal of audiology 53(12), 841–849.

Thai-Van, H., Joly, C. A., Idriss, S., Melki, J. B., Desmettre, M., Bonneuil, M., ... & Reynard, P. (2022). Online digital audiometry vs. conventional audiometry: a multi-centre comparative clinical study. International Journal of Audiology, 1-6.

Van Beurden, M., Boymans, M., van Geleuken, M., Oetting, D., Kollmeier, B., & Dreschler, W. A. (2018). Potential consequences of spectral and binaural loudness summation for bilateral hearing aid fitting. Trends in Hearing, 22, 2331216518805690.

Van Beurden, M., Boymans, M., van Geleuken, M., Oetting, D., Kollmeier, B., & Dreschler, W. A. (2021). Uni-and bilateral spectral loudness summation and binaural loudness summation with loudness matching and categorical loudness scaling. International Journal of Audiology, 60(5), 350-358.

Van der Aerschot, M., Swanepoel, D. W., Mahomed-Asmail, F., Myburgh, H. C., & Eikelboom, R. H. (2016). Affordable headphones for accessible screening audiometry: An evaluation of the Sennheiser HD202 II supra-aural headphone. International Journal of Audiology, 55(11), 616-622.

Van Esch, T. E., Kollmeier, B., Vormann, M., Lyzenga, J., Houtgast, T., Hällgren, M., ... & Dreschler, W. A. (2013). Evaluation of the preliminary auditory profile test battery in an international multi-centre study. International journal of audiology, 52(5), 305-321.

Whilby, S., Florentine, M., Wagner, E., & Marozeau, J. (2006). Monaural and binaural loudness of 5-and 200-ms tones in normal and impaired hearing. The Journal of the Acoustical Society of America, 119(6), 3931-3939.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., ... & Yutani, H. (2019). Welcome to the Tidyverse. Journal of open source software, 4(43), 1686.

869 Xu, C., Hülsmeier, D., Buhl, M., & Kollmeier, B. (2023, June 26). How Does
870 Inattention Influence the Robustness and Efficiency of Adaptive Procedures in
871 the Context of Psychoacoustic Assessments via Smartphone?.
872 https://doi.org/10.31234/osf.io/9ytd6
873 Yousuf Hussein, S., Wet Swanepoel, D., Biagio de Jager, L., Myburgh, H. C.,
874 Eikelboom, R. H., & Hugo, J. (2016). Smartphone hearing screening in mHealth
875 assisted community-based primary care. Journal of telemedicine and telecare,
876 22(7), 405-412.
877 Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands
878 (Frequenzgruppen). J. Acoust. Soc. Am. 33:248. doi: 10.1121/1.1908630