Feasibility of efficient smartphone-based threshold and loudness assessments in typical home settings

Chen Xu,[1] Lena Schell-Majoor,[1] and Birger Kollmeier[1]

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

Ambient noise is a critical factor affecting the precision of mobile hearing tests conducted in home environments. Monitoring noise levels during out-of-booth measurements provides essential information about the suitability of the setting for accurate audiometric testing. When ambient noise is controlled, results are expected to be comparable to in-booth measurements. This study remotely conducted air-conduction pure-tone audiometry and adaptive categorical loudness scaling (ACALOS) tests at 0.25, 1, and 4 kHz using a smartphone, while an integrated microphone and a dosimeter app were used to quantify ambient noise levels. Additionally, a reinforced ACALOS (rACALOS) method was proposed to integrate threshold measurement into the ACALOS procedure. The rACALOS method not only improves the accuracy of threshold estimation but also increases efficiency by combining two independent procedures into a single, streamlined process. As a result, ambient noise levels were mostly below the maximum permissible level. Hearing tests conducted via smartphone demonstrated moderate-to-excellent reliability, with intraclass correlation coefficients (ICCs) exceeding 0.75, and strong validity, with biases of less than 1 dB. In simulations, the rACALOS method reduced the bias towards pre-assumed thresholds, and in behavioral experiments, it showed a stronger correlation with pure-tone audiometric thresholds than the baseline method. Overall, this study demonstrates that

administering pure-tone audiometry and ACALOS tests at home is feasible, valid, efficient, and reliable when ambient noise is sufficiently low.

Keywords: remote audiology; ambient noise; validity and reliability; categorical loudness scaling

Corresponding author: Chen Xu

Contact:

chen.xu@uni-oldenburg.de

Department of Medical Physics and Acoustics, Faculty VI

Carl von Ossietzky Universität Oldenburg, 26111, Oldenburg, Germany

ACKNOWLEDGMENTS

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

DECLARATION OF CONFLICTING INTERESTS

The authors declare that there is no conflict of interest.

INTRODUCTION

Despite the benefits of easy access and early diagnosis, a significant concern with mobile hearing tests is the lack of what Zhao et al. (2022) refer to as 'auditory hygiene'. In laboratory settings, optimal auditory hygiene is ensured through the use of soundproof booths, calibrated equipment, attentive participants, and supervision by trained personnel. In contrast, mobile audiometric tests conducted in home environments typically lack these controlled conditions, which may compromise the accuracy of the results. Thus, it is important to investigate the impact of this reduced auditory hygiene on the reliability of mobile hearing assessments.

Previous studies have demonstrated that conducting hearing tests outside of sound-treated booths can be feasible under certain conditions. Behar et al. (2021) reviewed audiometric assessments performed without booths and highlighted several viable solutions, such as testing in quiet environments with sound-attenuating headphones, using insert earphones or over-the-ear earmuffs, and employing active noise reduction earmuffs (Maclennan-Smith et al., 2013; Swanepoel et al., 2015; Brennan-Jones et al., 2016; Clark et al., 2017). Furthermore, recent research (e.g., Margolis et al., 2022; Meinke & Martin, 2023) has proposed standards for defining the maximum permissible ambient noise levels (MPANLs) for audiometric test rooms, based on the use of specific earphones (e.g., insert, supra-aural, circumaural). If ambient noise does not exceed the MPANL for a given earphone type, the environment is generally considered suitable for accurate audiometric testing.

In addition to the test environment, the choice of hearing assessment is another key consideration. Almufarrij et al. (2022) reviewed 187 web- and app-based tools for remote hearing tests, finding that pure-tone audiometry and speech-in-noise tests dominate the landscape, representing 49% and 22% of all tools, respectively. However, to our knowledge, only a few

studies (e.g., Kopun et al., 2022) have explored the remote application of categorical loudness scaling (CLS), a supra-threshold test widely used in clinical audiology for diagnostics and hearing device fitting. While Kopun et al. (2022) demonstrated the preliminary feasibility of conducting CLS remotely, three major limitations emerged: (1) the equipment used for remote testing was a laptop rather than a smartphone, (2) only five participants (N = 5) were involved in the validation study, and (3) the reliability of CLS data collected in remote settings was suboptimal and requires improvement. To address these limitations, we extended the work of Kopun et al. (2022) by increasing the sample of young adults with normal hearing, optimizing the original CLS method for use with smartphones, and by integrating an audiogram measurement procedure into the CLS procedure.

As reported in Almufarrij et al. (2022), only 12% of hearing assessment tools have undergone validation and evaluation through peer-reviewed publications, highlighting that the validity and test-retest reliability of most tools available in app stores remain unknown. Consequently, the methods for quantifying validity and reliability of audiometric tests in home environments should be clearly defined, and results on both validity and test-retest reliability must be reported. Specifically, Bland-Altman plots are often used to validate audiometric tests, such as the matrix sentence test via smart speaker (Ooster et al., 2020) or categorical loudness scaling (CLS) (Fultz et al., 2020). For test-retest reliability, intraclass correlation coefficients (ICC) are typically used to assess agreement between repeated measures (Koo & Li, 2016). Specifically for CLS, Rasetshwane et al. (2015) and Kopun et al. (2022) introduced within-run variability and across-run bias as additional measures for assessing reliability in a home environment. In the present study, we incorporate not only basic metrics such as correlation coefficient (R), bias, and root-mean-squared-error (RMSE), but also advanced statistical

47 measures from previous studies to comprehensively report the validity and test-retest reliability
48 of smartphone-based audiometric tests.

49 The adaptive CLS procedure (ACALOS, Brand & Hohmann, 2002; ISO 16832, 2006) often
50 inaccurately estimates the audiometric threshold, as indicated by a correlation coefficient of less
51 than 0.5 between the ‘true’ audiometric and estimated thresholds, reflecting a weak correlation.
52 Please note that the thresholds estimated by CLS (hereafter referred to as 'CLS thresholds') are
53 defined as the level corresponding to 2.5 categorical units (CU) on the loudness growth function,
54 as outlined by Oetting et al. (2014). Oetting et al. (2014) further demonstrated that the threshold
55 predicted by the ACALOS method did not coincide with the ‘true’ audiometric threshold. This
56 discrepancy may be at least partially attributed to the use of different stimuli—narrow-band
57 noise in ACALOS versus pulsed tones in audiometry—and distinct psychophysical paradigms,
58 namely, categorical magnitude estimation in ACALOS versus target sound detection in
59 audiometry. To reduce this discrepancy, our study introduces a reinforced ACALOS (rACALOS)
60 method, which integrates a more accurate threshold estimation process within the ACALOS
61 procedure. This rACALOS approach allows participants to perform both threshold and ACALOS
62 measurements in a single procedure rather than separate tests, thereby increasing efficiency.
63 Additionally, the rACALOS method enhances reliability at low SPLs near the hearing threshold
64 by incorporating additional trials with the aim to provide a more accurate estimate of the ‘true’
65 hearing threshold which is usually directly assessed in pure-tone audiometry.

66 To accurately estimate the 'true' hearing threshold as a reference, it is essential to account for
67 as many influencing factors as possible. In our previous work, we investigated the impact of
68 experimenter supervision on pure-tone audiometry and adaptive categorical loudness scaling
69 (ACALOS) outcomes using a smartphone-based application in a sound-attenuated booth with

both normal-hearing (NH) and hearing-impaired (HI) listeners (Xu et al., 2024b). Our findings indicated that experimenter supervision had no significant effect (Xu et al., 2024b). Additionally, to address potential distractions for listeners, we proposed and simulated a model-free adaptive procedure for robust and efficient threshold estimation—the graded response bracketing (GRaBr) approach (Xu et al., 2024a). The present study aims to further validate GRaBr by comparing its performance with established baseline methods in human participants.

Taken together, the primary objectives of this study are: 1) to experimentally evaluate the performance of the novel, efficient GRaBr and rACALOS methods in human participants; 2) to assess the validity and test-retest reliability of the smartphone-based application for pure-tone audiometry and ACALOS in a home environment with some degree of background noise, given the absence of a sound booth.

METHODS

Participants

Fifteen young adults with normal hearing (aged 20 to 35 years; 4 males, 11 females) participated in this study. All participants were members of working groups or students at the University of Oldenburg, recruited primarily through verbal announcements. The three authors did not participate in the study. All participants self-reported no hearing issues and were presumed to have normal hearing (NH). Two inclusion criteria were applied: (i) an air-conduction pure-tone average (PTA-4) at 0.5, 1, 2, and 4 kHz in the better ear had to be less than or equal to 20 dB HL, and (ii) symmetric hearing, defined as a threshold difference of no more than 20 dB between ears at any test frequency. All 15 participants met these criteria. Some listeners (N = 5) received compensation of €12 per hour for their participation, while others took

part as part of their work duties. The study was approved by the Research Ethics Committee of the University of Oldenburg (Drs. EK/2023/004).

Equipment, Procedure, and Environment

Prior to the start of remote testing, a test kit was assembled (see supplemental materials), which included a smartphone (OnePlus, Android), a USB-C charger, and HD650 circumaural headphones (Sennheiser, Wedemark, Hanover, Germany). The smartphone and headphones were pre-calibrated using a Brüel & Kjær (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, with a target calibration level of 80 dB SPL. Upon handing over the test kit, participants received a brief oral explanation of the remote experiments, and consent forms were signed before they began. Participants could initiate testing at home by connecting to the internet via WLAN and accessing the provided website. For data security, a VPN connection was established using the 'GlobalProtect' app when accessing the site. The workflow of the web-based application for remote testing was described in Xu et al. (2024b). A Raspberry Pi 3 Model B (Raspberry Pi Foundation, UK), a Linux-based microcontroller, served as the server hosting the measurement site. All behavioral data were stored on an SD card within the Raspberry Pi, located at the University of Oldenburg.

The tele-health model, following the definition in Robler et al. (2022), was a self-testing model, requiring participants to complete all remote measurements within one week and return the test kit. The home environments were primarily located in rural regions of northwestern Germany, including cities such as Oldenburg, Cloppenburg, Jever, and Bad Zwischenahn.

Noise Level Measurement

The smartphone app "Decibel X" (SkyPaw Co., Ltd) was used to measure ambient noise levels and is freely available for download on the Google Play store. The app was configured with an A-weighted frequency filter and a slow time weighting of 500 ms. Real-time, average, and maximum environmental noise levels were displayed on the smartphone screen, but no sound files were recorded during the measurement. A digital sound level meter (Voltcraft SL-100), with an accuracy of ±2 dB at 1 kHz and compliant with the EN 60651 Class 3 standard, was used to calibrate the smartphone's integrated microphone. The smartphone app's parameters, including the A-weighted filter and slow time weighting were set as closely as possible to match the digital sound level meter. The app was then calibrated with a linear gain adjustment of 13.7 dB. Please note that the same smartphone and headphones were provided to all test participants, ensuring a consistent gain across measurements. Calibration stimuli consisted of narrowband noise signals fixed at 80 dB SPL.

At the start of each measurement session, the participants were required to document the current ambient noise level (see supplementary materials for remote measurement guidelines). A total of 24 sessions were conducted, consisting of 4 listening tests (SIUD, GRaBr, ACALOS, and rACALOS; see details below) across 3 test frequencies (0.25, 1, and 4 kHz) and 2 runs (test and retest), presented in randomized order. Participants were allowed to take short breaks between sessions. No specific instructions were provided regarding how to hold the smartphone during ambient noise measurement. Although participants were encouraged (but not required) to complete all sessions in the morning or evening, they were strongly advised to monitor the real-time noise level using the "Decibel X" app throughout each session. If the real-time noise level exceeded 45 dB(A), participants were instructed to pause testing until the noise level fell below

this threshold. A limit of 45 dB(A) was chosen based on Kopun et al. (2022), who demonstrated that remote CLS results are comparable to in-lab CLS measurements when ambient noise is kept below 50 dB(A). Additionally, the time and location of each remote session were recorded.

Listening Tests

Pure-tone audiometry

Two adaptive methods, the single-interval up-down (SIUD) procedure and the graded response bracketing (GRaBr) approach, were used to measure air-conduction pure-tone hearing thresholds (Lecluyse et al., 2009; Xu et al., 2024a). Xu et al. (2024a) conducted computer simulations demonstrating that GRaBr significantly outperformed the established SIUD method in terms of robustness against both long- and short-term inattention, as well as efficiency. In this study, the self-administered listening tests conducted at home present an ideal scenario for using an inattention-aware method like GRaBr, as participants are no longer supervised by an experimenter and are therefore supposed to be more susceptible to distractions.

In both procedures, listeners were presented with two tones, one tone, or silence, and were required to indicate how many tones they heard. The sound level was adjusted adaptively based on the participants' responses: the task became more challenging following correct answers and easier after incorrect responses. The primary distinction between SIUD and GRaBr lies in the level difference between the two tones presented in most trials: fixed at 10 dB for SIUD, but variable for GRaBr. To ensure a fair comparison between the two methods, key parameters, such as the minimum number of trials, number of reversals, and starting level, were matched as closely as possible. Both procedures commenced with a cue tone set at 60 dB HL with a random bias of less than 5 dB and terminated after a minimum of 14 reversals and 10 trials. For both methods, the first four reversals in each track were discarded.

Each pure tone lasted 0.2 s, with cosine ramps of 0.02 s and a 0.3 s interval between tones. Test frequencies of 0.25, 1, and 4 kHz were used for the stimuli. In SIUD, the correct response rates were fitted to an S-shaped logistic psychometric function, and the level at the 50% correct response point (L_{50}) was estimated as the hearing threshold. For GRaBr, responses from the upper and lower tracks were fitted to two independent psychometric functions, and the hearing threshold was calculated as the mean level at the 50% correct response point of both functions (i.e., $0.5*(L_{50,upper} + L_{50,lower})$). To assess test-retest reliability, both methods (SIUD and GRaBr) were repeated, with the test and retest referred to as Run 1 and Run 2, respectively. No specific time interval was recommended between the test and retest; participants were simply instructed to complete both runs within one week.

Adaptive categorical loudness scaling

The adaptive categorical loudness scaling (ACALOS) method was used to assess the loudness growth function (Brand & Hohmann, 2002; ISO 16832, 2006). In the ACALOS task, participants rated the loudness of stimuli on an 11-point scale with descriptors ranging from ’very soft”, “soft”, “medium”, “loud”, and “very loud” with 4 unnamed intermediate categories in between, plus the two limiting categories “not heard”, and “too loud”. The stimulus levels, ranging from -10 to 105 dB, were presented in a pseudo-random order following an initial estimation of the user-specific dynamic range (Phase I, see Fig. 1), which was updated to obtain a more representative placement of test level in Phase II, encompassing 26 trials. At the end of the procedure, a loudness growth function was modeled by fitting two linear segments and a transition region using a Bezier fit, following the BTUX fitting method (Oetting et al., 2014).

However, applying ACALOS without modifications in a mobile setting for remote testing may pose challenges. Fluctuating ambient noise in home environments could affect loudness

judgments at low sound pressure levels (SPL). Furthermore, as a supra-threshold measure of loudness perception, ACALOS often fails to provide reliable categorical loudness estimates near the hearing threshold (Oetting et al., 2014). Oetting et al. (2014) reported that the mean intra-subject standard deviation of loudness levels close to the threshold was notably high (around 10 dB), yielding significant variability in the hearing threshold estimation from loudness judgments near the threshold.

To address the limitations of ACALOS near the hearing threshold, a modified method, reinforced adaptive categorical loudness scaling (rACALOS), was introduced to improve the accuracy of hearing threshold level (HTL) estimation. An example run is shown in Fig. 1. The rACALOS followed the same adaptive rules as ACALOS during Phases I and II (see above) but presented additional stimuli near the hearing threshold to better estimate HTL. The starting level of Phase III was set at the minimum level reached in Phases I and II, plus 5 dB. In this phase, a one-up-one-down adaptive rule was applied: the stimulus level increased by 5 dB if participants responded with "not heard" and decreased by 5 dB if they selected other loudness categories (e.g., "very soft," "medium"). Phase III consisted of 10 trials.

The stimuli used were one-third-octave-band low-noise noises centered at 0.25, 1, and 4 kHz (Kohlrausch et al., 1997). Each noise stimulus had a duration of 1 second with 0.05-second rise and fall ramps. To assess reliability, participants repeated both ACALOS and rACALOS measurements at all frequencies for both test and retest conditions.

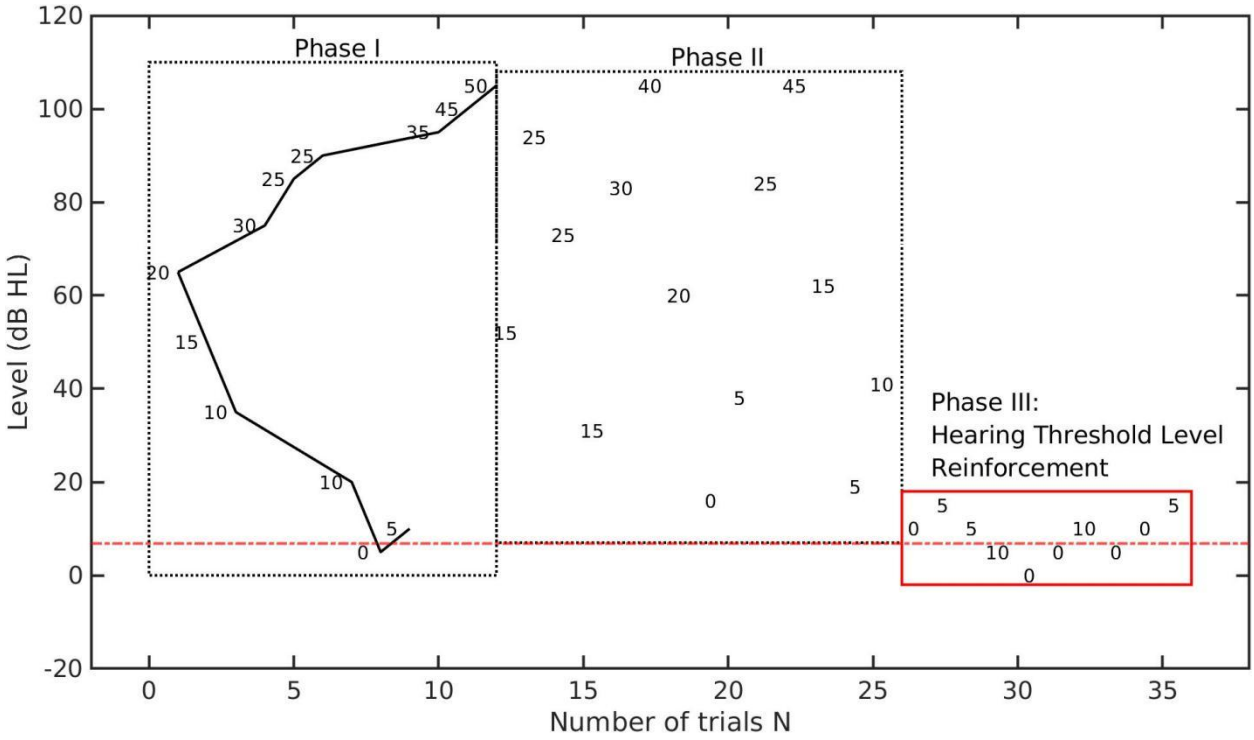


Fig. 1. An example track of the reinforced adaptive categorical loudness scaling (rCALOS), where the level (in dB HL) is plotted as a function of the number of trials N. The listener's response (in categorical units (CU)) is annotated with numbers between 0 ('not heard') and 50 ('too loud'). Left dotted rectangle region: Phase I ('dynamic range estimation'); Middle dotted rectangle region: Phase II ('presenting and re-estimation'); Right solid red rectangle region: Phase III ('hearing threshold level reinforcement'); Red dash-dotted line: target threshold. In Phase III, the step size is set to 5 dB, and the number of trials is set to 10.

Accuracy of HTL estimation for the rACALOS procedure

Computer simulations

Monte-Carlo simulations were conducted to compare the baseline ACALOS and rACALOS in terms of accuracy in estimating the hearing threshold level (HTL). The statistical behavior of the virtual listener was based on the models described by Brand et al. (2000) and Oetting et al. (2014), assuming a normal distribution. The mean response of the virtual listener was modeled

using a three-parameter loudness function consisting of two linear segments with slopes m_{low} and m_{high} , and a smoothed transition region between 15 and 35 categorical units (CU). A standard deviation of 4 CU, derived from empirical data in Brand et al. (2000), was employed. The simulated loudness judgment was drawn from a normal distribution defined by this mean (loudness function) and the standard deviation (4 CU) for a given presentation level L.

The simulated loudness responses were constrained to the range of 0 to 50 CU and rounded to the nearest 5 CU. The target loudness function parameters were set to 84.1 dB HL for L_{cu}, 0.3 for m_{low}, and 1.0 for m_{high}. Phase III of the rACALOS procedure varied the number of trials (N) between 10 and 30 in increments of 10, with step sizes of 2 and 5 dB. The Monte-Carlo simulations were executed 1000 times in total. All simulations were implemented in MATLAB R2021a (The MathWorks, Inc., Natick, MA) and Octave 5.2.0.

Behavioral experiments

In this study, we conducted behavioral experiments using a repeated-measures design, where 15 participants completed both pure-tone audiometry and ACALOS tests. We compared the estimated HTL from the ACALOS and rACALOS methods to the 'true' HTL measured by pure-tone audiometry (i.e., GRaBr and SIUD). To assess the relationship between pure-tone and ACALOS thresholds, various statistical methods were employed, i.e., correlation coefficients (R), root mean square error (RMSE), and bias, along with scatter plots to evaluate the performance of the different ACALOS methods.

Statistics

To evaluate the validity of GRaBr and rACALOS relative to standard audiometric and CLS procedures conducted in a soundproof booth, we utilized Bland-Altman plots following the approach of Fultz et al. (2020) and Giavarina (2015). Additionally, test-retest reliability for both

audiometric procedures was assessed using intraclass correlation coefficients (ICCs) as per Buhl et al. (2022). Reliability levels were categorized as poor (ICC < 0.5), moderate (ICC ≥ 0.5), good (ICC ≥ 0.75), and excellent (ICC ≥ 0.9). Following Kopun et al. (2022), we further applied mean interquartile range (MIQR) and mean signed difference (MSD) metrics to evaluate the reliability of both ACALOS procedures, with lower values indicating greater reliability. Detailed statistical methods for validity and reliability assessment are provided in Supplementary Materials S1.

RESULTS

Noise Level Measurements

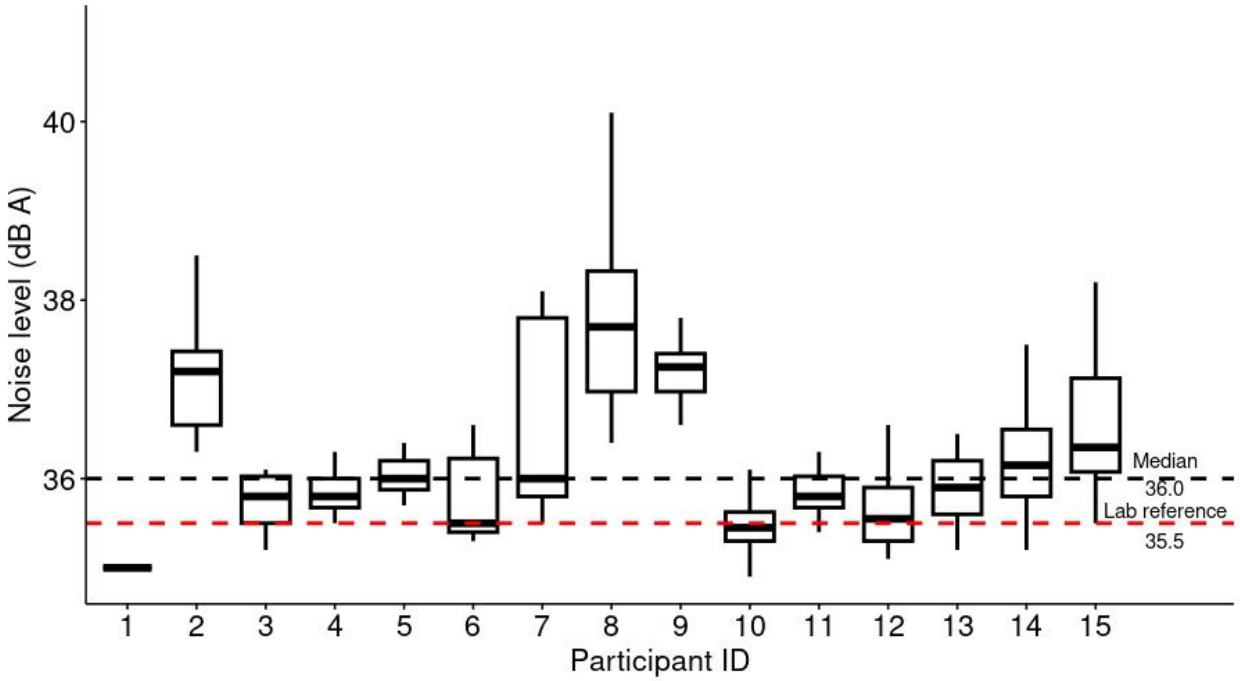
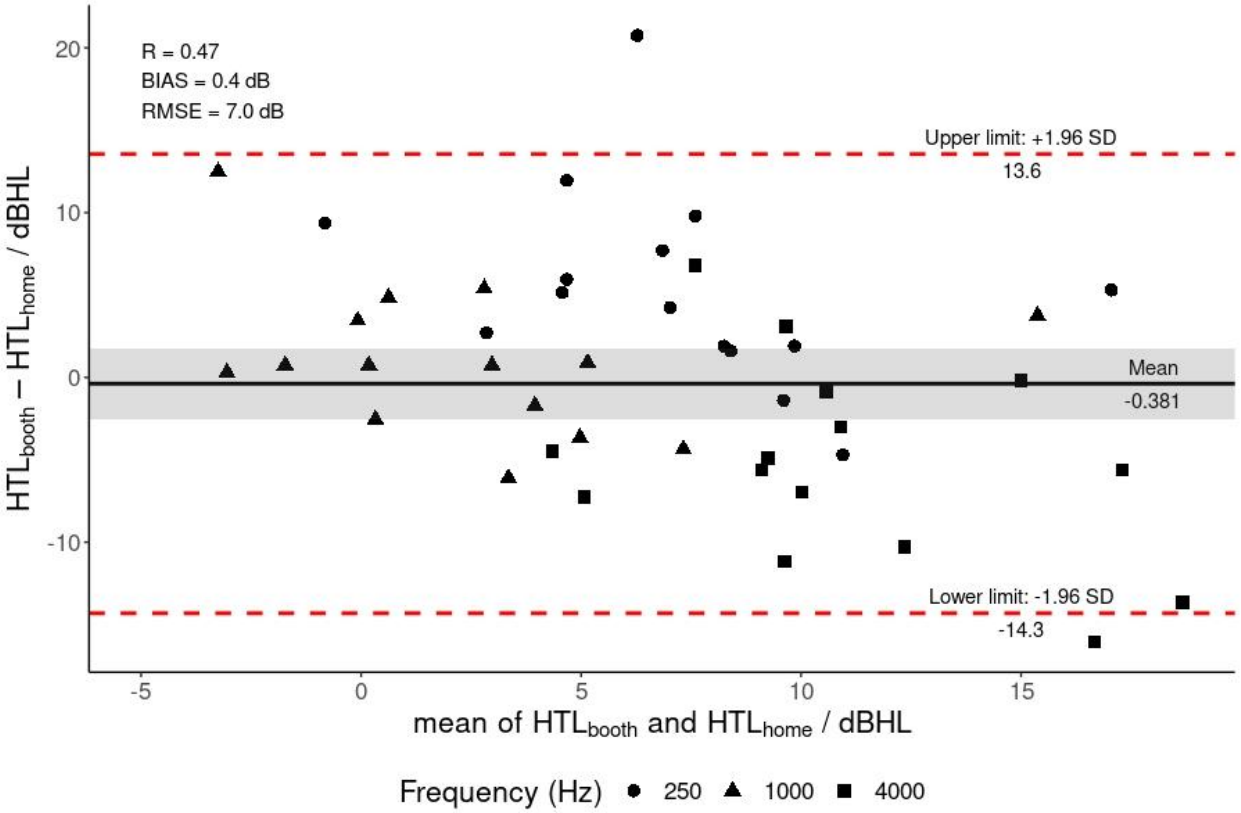


Fig. 2. Ambient noise level (in dB A) measurement across participants (N = 15). Medians, 25th and 75th percentiles, and interquartile ranges (IQR) are visualized in the box-plot while the end of the whiskers denotes the minimum and maximum, indicating the 5th and 95th percentiles respectively. Red dashed line: lab reference (i.e., ambient noise level measured within a booth). Black dashed line: median value across subjects.

Fig. 2 presents a box plot of the ambient noise levels recorded by each participant (N = 15), who documented the noise level a total of 24 times, corresponding to 24 measurement sessions at home within a week. Notably, the noise levels for all participants remained below the recommended upper limit of 45 dB A. The median noise level across subjects was 36.0 dB, which was approximately 0.5 dB higher than the reference noise level measured inside the sound-attenuated booth. Overall, the sound levels in participants' homes were considerably low and comparable to those measured within the booth, indicating a suitable test environment. A few participants (e.g., No. 2 and No. 8) lived near a train station, resulting in slightly elevated noise levels compared to others. Additionally, one participant (No. 1) misinterpreted the task and consistently rounded the recorded noise level to an integer, leading to uniform values across sessions.

263 **Validation Experiment**

264 ***GRaBr***



265

266 Fig. 3. Bland-Altman plot of hearing threshold levels (HTL) in dB HL of frequencies at 0.25, 1,
267 and 4 kHz (represented with circle, triangle, and rectangle, respectively) measured inside the
268 booth (i.e., HTL_{booth}) using the standard audiometry and at home (i.e., HTL_{home}) using the GRaBr
269 procedure. Red dashed lines: 95% level of agreement; Black solid line: bias between the two
270 measurement environments; Grey shaded rectangle area: 95% confidence interval of the bias.
271 The correlation coefficient (R), bias (BIAS), and root mean squared error (RMSE) are provided
272 in the top-left corner.

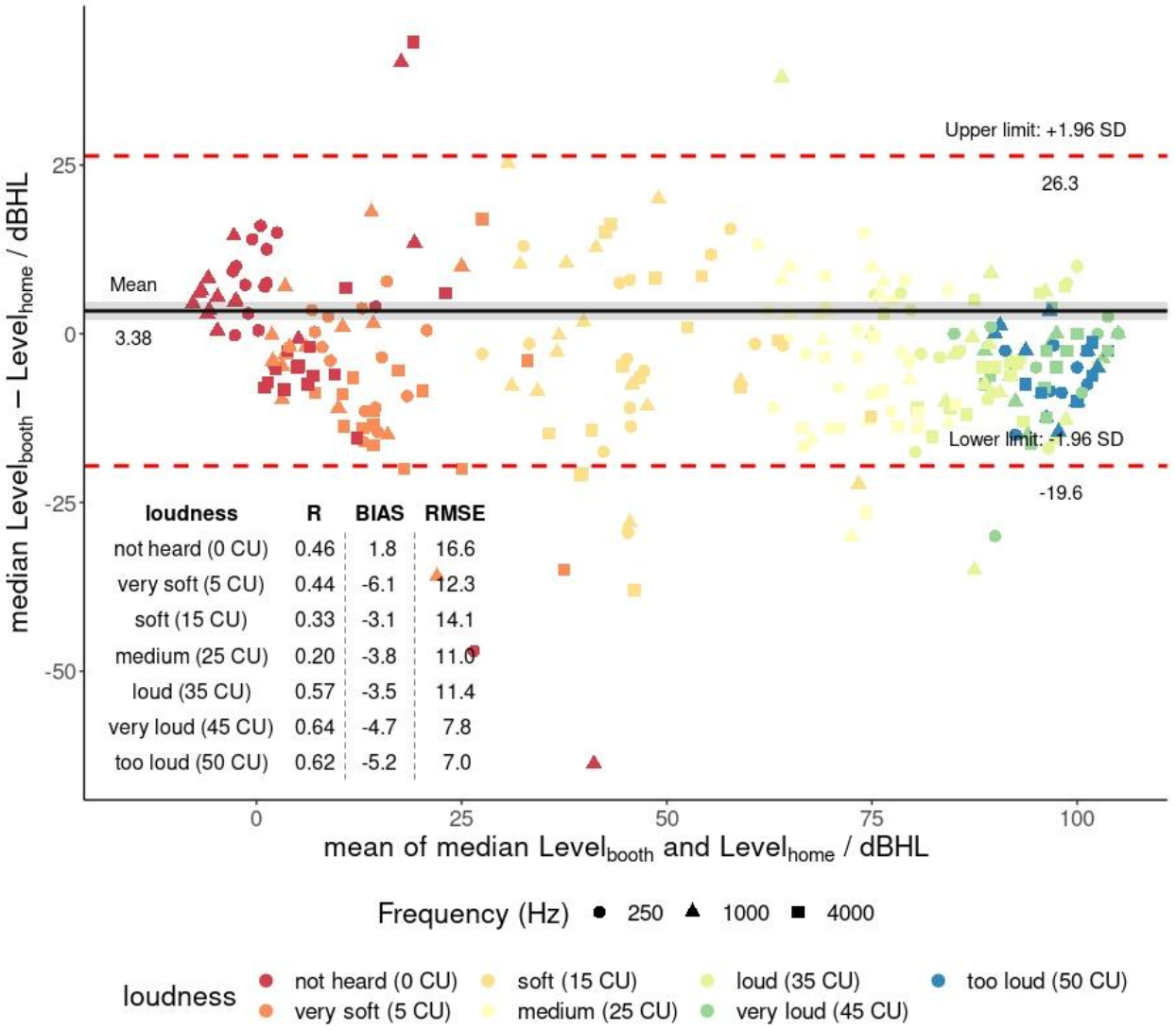
273 Fig. 3 compares pure-tone audiometry results obtained in the booth using the standard
274 audiometry versus testing at home using the GRaBr procedure at frequencies of 0.25, 1, and 4

kHz. Most data points fell within the 95% level of agreement, indicating that the at-home and in-booth measurements did not differ systematically. Furthermore, the 95% confidence interval of the bias (depicted by the shaded region) encompassed the line of equality, suggesting no significant bias between the two testing environments. Although the correlation between HTL_{booth} and HTL_{home} was moderate, both the bias and root mean squared error (RMSE) were relatively small. Overall, the comprehensive statistical analyses indicated good agreement between results from both environments, supporting the validity of the smartphone-based remote method for pure-tone audiometry as an alternative to standard assessments conducted in the booth, provided that ambient noise levels remain low.

A two-way repeated measures ANOVA was conducted to evaluate the effects of frequency (0.25, 1, and 4 kHz) and test environment (booth versus home) on hearing thresholds. As anticipated, there was no significant main effect of the test environment ($p = 0.77$); however, the main effect of frequency was significant ($p < 0.05$). Despite the lack of a significant effect from the test environment, post-hoc tests comparing HTLs between the home and booth settings indicated that thresholds measured in the booth did not significantly differ from those measured at home at 1 kHz, while a significant difference was observed at 0.25 and 4 kHz ($p < 0.05$).

Validation results for the SIUD procedure in a home environment, compared to a standard audiometer, are presented in Figure S1. The SIUD method showed a bias of 0.6 dB, indicating good validity. Additionally, the SIUD procedure differed significantly from GRaBr in measured thresholds ($p < 0.05$). Overall, the validity of both adaptive procedures was comparable, suggesting that both are suitable for remote measurements in home settings.

296 ***rACALOS***



297

298 Fig. 4. Bland-Altman plot of median levels assigned to each CU (denoted with different colors)
299 for three frequencies (represented with different shapes) for comparing two test environments,
300 i.e., inside the booth using a standard CLS procedure and at home using the rACALOS
301 procedure for each participant. A comprehensive set of statistical measures containing R, Bias,
302 and RMSE of each CU is provided in the embedded table located at the bottom-left corner. See
303 Fig. 3 for an explanation of the Bland-Altman plot and Supplementary Materials S1 for its
304 statistical implication.

Fig. 4 presents the Bland-Altman plot comparing the median levels of each categorical unit (CU) measured inside the booth using a standard CLS approach and at home using the rACALOS approach at frequencies of 0.25, 1, and 4 kHz. The 95% levels of agreement (LOA) for the upper and lower limits were 26.3 dB and -19.6 dB, respectively. Only a small number of points fell outside the 95% LOA, indicating that the rACALOS measurements in the booth did not systematically differ from those obtained remotely. The overall bias between the two environments across all participants was notably small at 3.38 dB, suggesting that the rACALOS approach demonstrates good validity compared to the standard CLS approach.

The R values for categorical units (CUs) of 35 or higher ranged from 0.57 to 0.62, indicating a moderate positive correlation. In contrast, CUs of 25 or lower exhibited an R value below 0.45, suggesting a weak correlation. The biases were generally below 5 dB, and as CU decreased, the RMSE tended to increase. This phenomenon may be attributed to the relatively high variability in individual hearing thresholds, resulting in a steeper loudness perception slope at lower levels. Consequently, this leads to reduced validity at low categorical unit (CU) levels. However, it is important to note that the slightly elevated background noise levels in the home environment did not systematically affect this variability, as both positive and negative deviations were observed between threshold levels estimated at home and those measured in the booth.

To examine the effects of three within-subject factors—test environment (booth/home), frequency (0.25/1/4 kHz), and CU (ranging from 0 to 50 CU in 5 CU increments)—on median levels corresponding to each CU, a three-way repeated measures ANOVA was conducted. As expected, the test environment showed no significant main effect, while both frequency and CU exhibited significant main effects ($p < 0.05$). A post-hoc t-test analyzed the effect of the test environment across all frequencies and CUs, revealing no significant differences in most of the

33 groups of comparison (i.e., 3 levels of frequency * 11 levels of CU), except for three groups (measurement at 4 kHz with 5, 25, and 45 CU).

The results of the validation experiment comparing the original ACALOS procedure with the standard CLS procedure are shown in Fig. S2 of the supplementary material, indicating good validity comparable to that of the rACALOS procedure discussed above. Furthermore, ACALOS differed significantly from the rACALOS approach ($p < 0.05$), primarily reflecting the higher sampling and weighting of the loudness data at low levels by rACALOS.

Test-Retest Reliability Experiment

SIUD and GRaBr

The GRaBr procedure showed test-retest intraclass correlation coefficient (ICC) values exceeding 0.75 ($p < 0.05$), indicating good reliability across all three frequencies, whereas the SIUD procedure yielded ICC values ranging from 0.59 to 0.77 ($p < 0.05$), reflecting moderate test-retest reliability. This difference was significant ($p < 0.05$), i.e., GRaBr demonstrated significantly higher test-retest reliability than SIUD based on these metrics. Further details on reliability statistics can be found in Supplementary Document S2 and Table S1.

A significant main effect of frequency was observed ($p < 0.05$). Moreover, pairwise t-tests were performed to assess reliability by comparing the two runs for both adaptive procedures across all three frequencies, showing no significant differences between runs in most cases, except for GRaBr at 1 kHz ($p < 0.05$).

ACALOS and rACALOS

The reliability of the ACALOS and rACALOS procedures was assessed using across-run bias (quantified by mean signed difference, MSD) and within-run variability (measured by mean interquartile range, MIQR). Both adaptive procedures demonstrated an MSD of less than 5 dB at

all frequencies, indicating a small across-run bias. Most MIQR values did not exceed 10 dB for either procedure at the three frequencies, although they were typically larger than 10 dB at 5, 10, and 15 CU, reflecting a consistent within-run variability. Overall, these metrics suggested that both ACALOS and rACALOS exhibited strong reliability. Please refer to Supplementary Material S3 and Table S2 for detailed information on the reliability comparison of the ACALOS and rACALOS procedures.

A repeated measures ANOVA revealed a significant main effect of the procedure, indicating a statistically significant difference between ACALOS and rACALOS ($p < 0.05$). Since the rACALOS and ACALOS procedures are identical in Phases I and II, this difference is likely attributable to the additional trials included in Phase III of the rACALOS procedure (see Fig. 1).

No significant effect was found for frequency, and as expected, the two runs (test and retest measurements) did not differ. A subsequent post-hoc t-test compared median levels of the ACALOS and rACALOS procedures between runs 1 and 2 across three frequencies and 11 categories, indicating that median levels for run 1 did not significantly differ from those for run 2 in most cases (31 out of 33 groups of comparison = 3 levels of frequency * 11 levels of CU), except for two groups (measurements at 0.25 kHz for 25 and 40 CUs).

Accuracy of HTL Estimation for the rACALOS procedure

Computer simulations

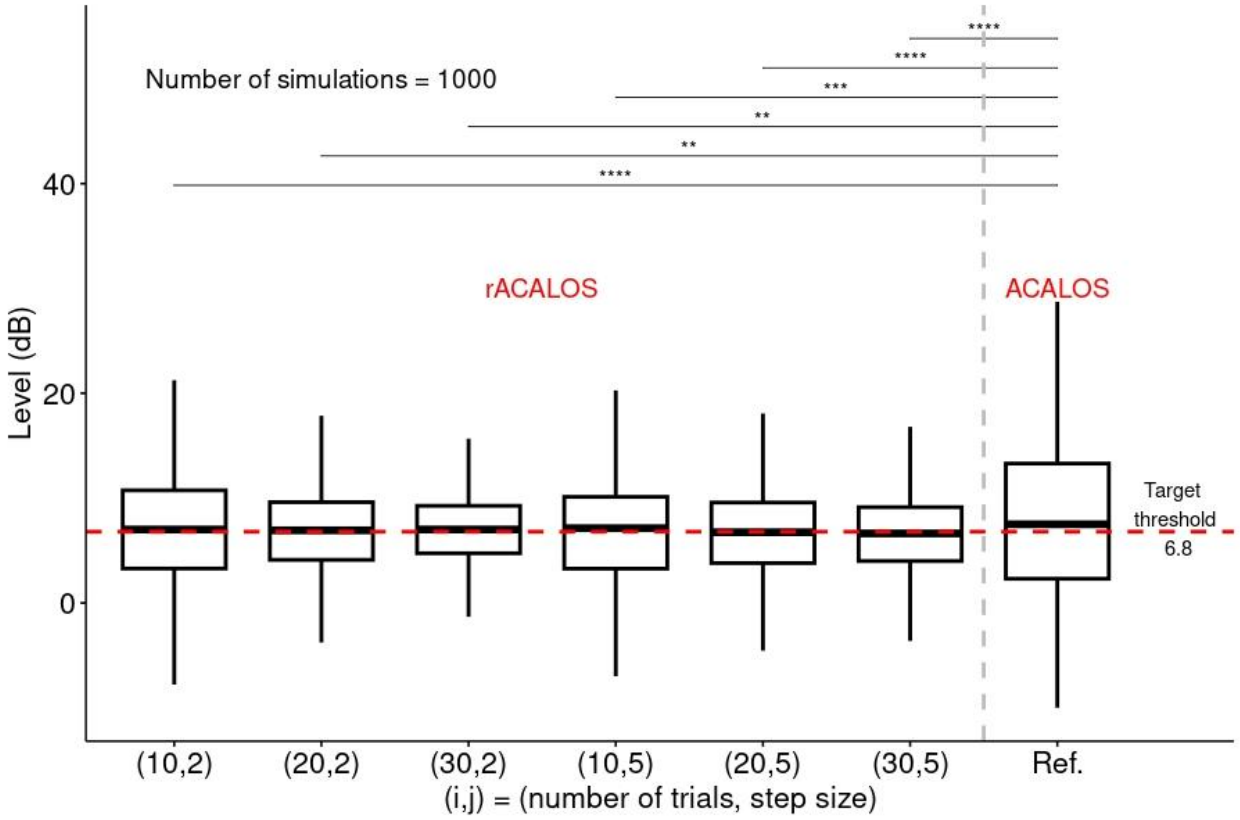
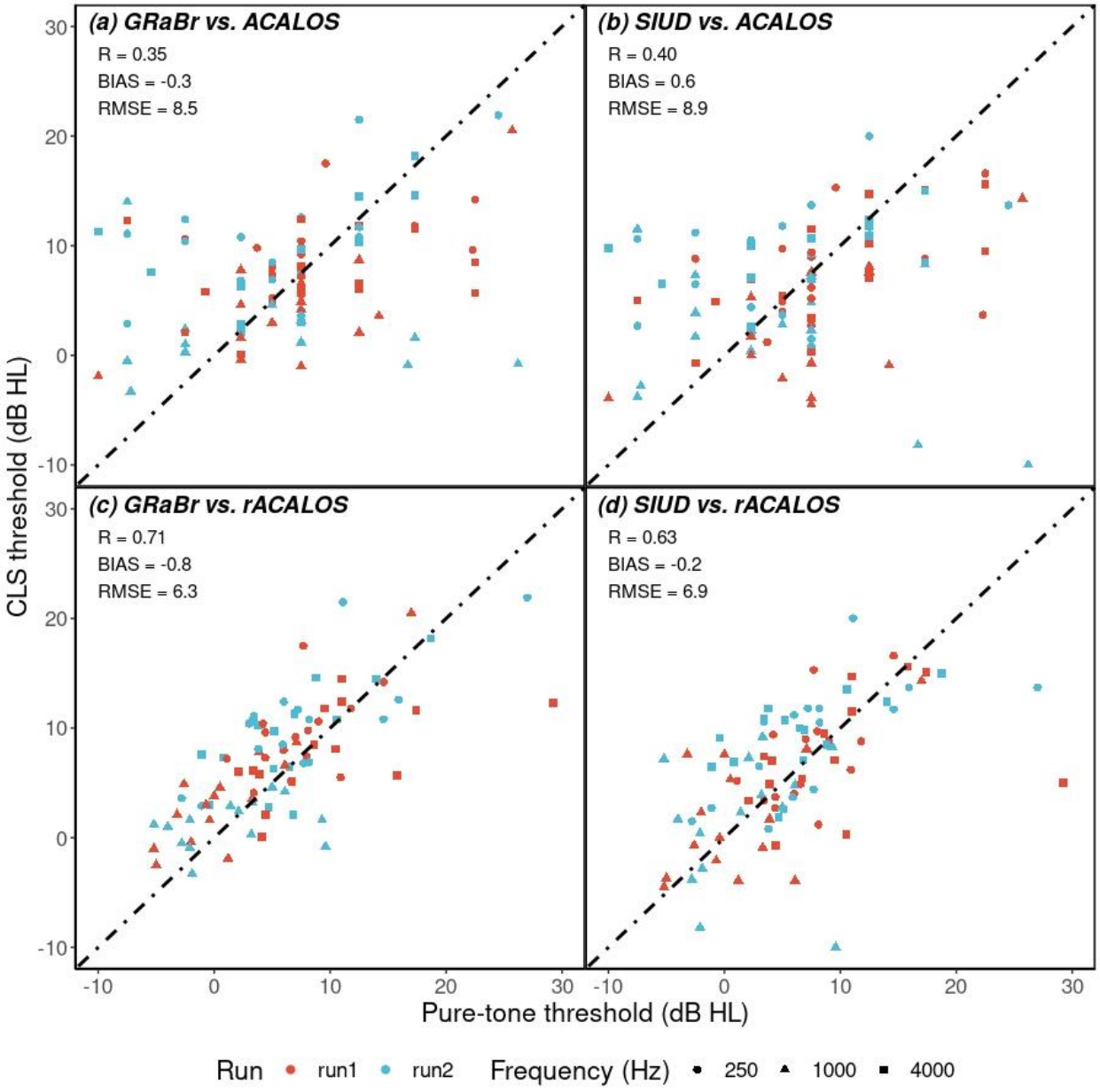


Fig. 5. Estimated hearing thresholds in dB (i.e., level of the loudness growth function corresponding to 2.5 CU) of rACALOS variants (to the left of the vertical dashed line) using reinforcement at the hearing threshold level obtained with Monte-Carlo simulations (N = 1000 runs) in comparison to the baseline ACALOS (reference group "Ref."). The parameter combinations (i,j) are displayed where i and j denote the number of trials and step size in the reinforcement phase. Red horizontal dashed line: target ('true') threshold. cf. Fig. 2 for an explanation of the box-and-whiskers plot. The statistical outcome of the pair-wise comparison against the reference group is visualized. The level of significance for p values is labeled with stars above the lines.

Computer simulations (N = 1000 runs) of thresholds estimated from the ACALOS and rACALOS methods under various parameter combinations are presented in Fig. 5. The medians from the rACALOS method were closer to the target threshold compared to ACALOS, and the interquartile ranges (IQRs) for rACALOS were significantly smaller than those for ACALOS, as indicated by F-tests ($p < 0.05$). This indicates that rACALOS provides a more accurate estimation of the hearing threshold level (HTL) than the original method. Additionally, increasing the number of trials resulted in a decrease in IQR, suggesting that the precision of both methods can be enhanced by increasing the number of trials even though more measurement time is required. Furthermore, methods utilizing a smaller step size exhibited significantly narrower IQRs compared to those with a larger step size, as suggested by F-tests ($p < 0.05$).

A two-way ANOVA was conducted to evaluate the effects of the number of trials (10, 20, and 30) and step size (2 and 5 dB) on the simulated thresholds. The analysis indicated that both factors significantly impacted the simulated thresholds ($p < 0.05$). Subsequently, a pair-wise t-test was performed to compare the simulated hearing thresholds of ACALOS (set as the reference) and rACALOS, with p-values adjusted using the Bonferroni method. The results revealed a significant difference in simulated thresholds between ACALOS and rACALOS across all parameter sets ($p < 0.05$) After carefully balancing high accuracy and relatively fast convergence, a step size of 5 dB was selected, and the number of trials was set to 10 for the remainder of this study.

399 ***Behavioral experiments***



400

401 Fig. 6. Scatter plot for comparison between pure-tone (abscissa) and CLS (ordinate) thresholds in
402 dB HL of N = 15 individual listeners. Frequency is labeled with different shapes while the run is
403 denoted with different colors (run1: red, run2: blue). A set of statistical metrics (R, Bias, and
404 RMSE) are reported in the top-left corner. For rACALOS, 10 additional trials with a step size of
405 5 dB were used.

Pure-tone audiometric thresholds are plotted against CLS thresholds for two runs and three frequencies in Fig. 6. Compared to ACALOS, the majority of rACALOS points were consistently and closely clustered around the diagonal line, indicating that thresholds estimated by the rACALOS method aligned more closely with pure-tone thresholds than those from baseline ACALOS and, hence, provide improved accuracy in threshold estimation. Quantitatively, R values increased by 36% for GRaBr and 23% for SIUD when ACALOS was reinforced near the hearing threshold level. Additionally, RMSE values for the rACALOS method decreased by approximately 2 dB compared to the baseline, while biases remained unchanged. Overall, the reinforcement of baseline ACALOS positively influenced cross-correlation and reduced error.

The highest correlation coefficient and lowest RMSE were observed between GRaBr thresholds and rACALOS, followed by SIUD and rACALOS. In contrast, the unmodified ACALOS procedure showed lower correlation coefficients and higher RMSEs for both threshold estimation methods, indicating the superior performance of rACALOS, as confirmed by t-tests ($p < 0.05$).

DISCUSSION

Noise Level Measurements

The median ambient noise level across participants' homes was 36.0 dB A, which is generally comparable to the reference noise level in a soundproof booth. As expected, the measurement results from the home environment aligned well with those obtained inside the booth. Additionally, our findings comply with the American National Standards Institute (ANSI) S3.1–1999 (R2018) standard for maximum permissible ambient noise levels (MPANL) for supra-aural and insert earphones with covered ears, although they exceed the MPANL

recommendation for uncovered ears, as established for audiogram measurements. Furthermore, our measured noise levels did not surpass the updated MPANL, which was extended by Margolis et al. (2022) for three types of circumaural earphones. Overall, these results demonstrate why our listening tests conducted in a home environment can achieve accuracy comparable to those performed inside a booth.

Our measured ambient noise levels are lower than those reported in most earlier studies (e.g., 40 dB A by Storey et al. (2014), 46 dB A in a non-outpatient clinic by Brennan-Jones et al. (2016), and between 33.7 and 46.3 dB SPL in a 'natural' environment by Swanepoel et al. (2015)) that aimed to control ambient noise during audiometric tests. However, our levels are higher than those in a few studies, such as 34.6 dB A in a non-sound-treated clinical room by Serpanos et al. (2022) and 35 dB A in exam rooms by Bean et al. (2022). It is likely that our participants conducted the smartphone-based listening tests at home in rural areas during the morning or evening, whereas other studies typically test in clinical settings located in urban areas during the daytime, which tend to be noisier. Consequently, our overall measurement environments contained less ambient noise.

In addition to meeting the MPANL for pure-tone audiometry, our study adheres to the MPANL of 50 dB A specified for the ACALOS test outside a sound-treated booth, as suggested by Kopun et al. (2022). Therefore, we expect that our measured ACALOS results in a home environment will be comparable to those obtained inside a booth (see the discussion of the validation study for ACALOS below).

Pure-Tone Audiometry

Pure-tone audiometry conducted outside the booth on a smartphone in a quiet environment is generally valid and reliable when compared to in-booth measurements. While SIUD

452 demonstrates moderate reliability, GRaBr shows good reliability (the ICC values are greater than
453 0.75 ($p < 0.05$)) for remote smartphone-based assessments, making GRaBr the significantly more
454 reliable option ($p < 0.05$), as expected from the simulations reported by Xu et al. (2024a). Our
455 findings align with recent studies examining the validity of boothless pure-tone audiometry
456 (Maclennan-Smith et al., 2013; Storey et al., 2014; Swanepoel et al., 2015; Brennan-Jones et al.,
457 2016; Serpanos et al., 2022). The bias between in-booth and at-home measurements is 0.4 dB,
458 which falls within the empirical ranges reported by Maclennan-Smith et al. (2013) (-0.6 to 1.1
459 dB) and Swanepoel et al. (2015) (-2.0 to 1.5 dB). However, the correlation coefficient R (0.47)
460 in our study is notably lower than that reported by Maclennan-Smith et al. (2013), where R
461 exceeded 0.92 for both ears at frequencies between 0.25 and 8 kHz. This discrepancy may be
462 attributed to the much smaller range of thresholds across our participants: our study included 15
463 young adults with normal hearing, whereas Maclennan-Smith et al. (2013) had a larger sample of
464 147 elderly participants with hearing impairments, 59% of whom exhibited a pure-tone average
465 (PTA) greater than 25 dB. As Swanepoel et al. (2010) noted, hearing-impaired listeners typically
466 show higher correlation coefficients than those with normal hearing due to reduced sensitivity
467 and lesser impact from ambient noise. However, our test sample with young, normal hearing
468 listeners puts a higher demand on the quietness of the acoustic environment and the reliability of
469 the test procedure.

470 The test-retest reliability aligns well with findings from previous studies, such as those by
471 Swanepoel et al. (2015) and Hazan et al. (2022). The bias ($N = 11$) between test and retest
472 measurements was 1.8, 0.0, and 1.4 dB at 0.25, 1, and 4 kHz, respectively, consistent with the
473 findings of Swanepoel et al. (2015), where the bias also remained below 2 dB. The correlation
474 coefficient R at 1 kHz aligns with Hazan et al. (2022), although it is smaller at 4 kHz. Hazan et al.

(2022) suggested that test-retest performance improves with poorer hearing; since our study focused on young normal-hearing (NH) listeners with better hearing abilities, it is plausible that this contributed to the lower R-value observed at 4 kHz. Additionally, while Hazan et al. (2022) automatically rejected hearing thresholds when the ambient noise level at certain frequencies exceeded the stimulus level, we did not filter out such outliers.

The threshold offset between GRaBr and SIUD was approximately 1 dB, with GRaBr demonstrating a smaller standard deviation of thresholds. This trend mirrors findings from a simulation study, suggesting that the theoretical framework established by Xu et al. (2024a) accurately predicts outcomes in behavioral experiments. Since GRaBr presents more trials near the threshold level compared to SIUD, it is reasonable to conclude that the uncertainty, as indicated by the standard deviation, is significantly lower for GRaBr than for SIUD ($p < 0.05$). This confirms the preference for GRaBr over SIUD for smartphone usage, attributed to its superior performance as highlighted in the simulation study.

Adaptive Categorical Loudness Scaling

Remote adaptive categorical loudness scaling (ACALOS) and its reinforced version (rACALOS) conducted at home demonstrated strong validity and test-retest reliability. Our findings align with the validation study by Kopun et al. (2022) and reliability studies by Rasetshwane et al. (2015), Fultz et al. (2020), and Kopun et al. (2022). The systematic bias of 3.4 dB between in-booth and at-home measurements in our study is notably lower than the 5.4 dB reported by Kopun et al. (2022), suggesting improved accuracy in our results. One possible explanation could be the difference in environmental noise, as the average ambient noise level reported by Kopun et al. (2022) was approximately 10 dB higher than in our study, likely contributing to the larger bias in their measurements. Furthermore, differences in methodology

may also explain the discrepancy; while Kopun et al. (2022) applied the standard ISO 3682 method, we employed an optimized procedure based on Oetting et al. (2014), which may have enhanced the precision of our measurements.

Both ACALOS methods demonstrated high test-retest reliability, quantified by mean IQR (within-run variability) and MSD (across-run bias). At 1 kHz, the mean IQR as a function of CU for both ACALOS methods was generally consistent with the data from Rasetshwane et al. (2015) and Kopun et al. (2022). Specifically, the mean IQR at 5 CU for rACALOS closely matched that of Kopun et al. (2022) and was smaller than that reported by Rasetshwane et al. (2015), suggesting good stability near the hearing threshold. Additionally, at 4 kHz, the mean IQR at 5 CU for rACALOS was smaller than in both empirical studies, likely due to the reinforcement at the HTL. Overall, rACALOS exhibited the least variability at the threshold level compared to baseline ACALOS, as well as the results reported in these two studies, indicating its superior performance in reducing the variability at the threshold.

Regarding across-run bias at 1 and 4 kHz, similar to the findings of Rasetshwane et al. (2015), the mean signed differences (MSD) of both ACALOS methods in our study were approximately 2-3 dB smaller than those reported by Kopun et al. (2022). This can be attributed to our stricter requirements for the acoustic conditions, including a lower maximum permissible ambient noise level, which likely reduced ambient noise interference and resulted in smaller across-run bias. While the ACALOS method showed a smaller MSD at 4 kHz, it had a larger MSD at 1 kHz compared to rACALOS. Fultz et al. (2020) evaluated the reliability of four different CLS methods—(1) fixed-level procedure (FL), (2) slope-adaptive procedure (SA), (3) maximum expected information-median (MEI-Med), and (4) maximum expected information-maximum likelihood (MEI-ML). The bias in Fultz et al.'s study across these methods at both frequencies

was larger than ours. A potential reason for this discrepancy could be the inherent limitations of the newly developed CLS methods, as Fultz et al. (2020) noted that the adaptive track of the MEI method was suboptimal due to listener variability represented in the multi-category psychometric function. With the addition of more trials, particularly those near the threshold, our method is expected to yield less variability in threshold estimates compared to other approaches, thereby reducing bias.

Accuracy of HTL Estimation

Computer simulations indicate that rACALOS provides more precise estimates of hearing thresholds compared to the baseline ACALOS, largely due to the increased number of stimuli presented near the threshold level (see Fig. 1). One limitation of the original ACALOS is its potential failure to provide a low variability of the estimated hearing threshold level (HTL), as highlighted by Oetting et al. (2014), most likely due to evenly distributing the fit error across the whole dynamic range. This is mitigated in rACALOS by reinforcing responses in the HTL region. Additionally, increasing the number of trials (N) and using a smaller step size can reduce error and enhance measurement accuracy, although this comes at the cost of reduced efficiency (e.g., Kollmeier et al., 1988). These findings align with earlier studies, such as Lecluyse et al. (2009), which support the trade-off between precision and efficiency.

Table 1 presents a comparison between our current study and several state-of-the-art works (Fultz et al., 2020; Trevino et al., 2016; Sanchez-Lopez et al., 2021) by evaluating the cross-correlation between CLS and pure-tone thresholds. Multiple CLS methods, including FL, MEL-Med, MEL-ML, SA, ACALOS, and rACALOS, were used to estimate thresholds, which were then compared with pure-tone thresholds measured using various audiometric methods such as a clinical audiometer, SIUD, and GRaBr. In the studies by Fultz et al. (2020) and Trevino et al.

(2016), R values ranged from 0.21 to 0.26 for all four CLS methods, indicating a relatively weak cross-correlation. Additionally, the RMSEs and biases in these studies were notably large, suggesting that CLS thresholds did not align well with pure-tone thresholds. In contrast, Sanchez-Lopez et al. (2021) applied a baseline ACALOS method using the same audiometric procedure as Fultz et al. (2020), and while the R-value did not significantly improve, both RMSE and bias were notably reduced. In our study, we employed SIUD and GRaBr to measure pure-tone thresholds, yielding a stronger cross-correlation and smaller bias, although the RMSE was slightly larger or comparable to that reported by Sanchez-Lopez et al. (2021).

Considering all the studies, the rACALOS method consistently produces thresholds closest to pure-tone thresholds, outperforming other ACALOS methods. However, it is important to note that rACALOS requires more measurement time due to the increased number of trials focused on converging near the HTL. Additionally, using precise audiometry methods such as SIUD and GRaBr may yield stronger correlations with CLS thresholds, despite the fact that many studies still regard pure-tone thresholds obtained via clinical audiometers as the ‘gold standard’. It is also crucial to recognize that this comparison is based on a small sample of young NH listeners, and the conclusions may differ if HI listeners are included or if a larger participant pool is studied. This consideration is particularly relevant for potential discrepancies between the narrowband noise thresholds estimated by the CLS methods used here and the pulsed pure-tone thresholds assessed via audiograms. While threshold differences in our study sample of young NH listeners were minimal, variations in stimulus characteristics—such as spectral extent and modulation spectrum—may yield threshold differences in naïve listeners with hearing impairments. Nonetheless, these differences are expected to be minimal, as the low-noise, third-

566 octave-band noise utilized here is effectively equivalent to a frequency-modulated sinusoid with
567 minor envelope fluctuations and an instantaneous frequency confined well within a critical band.

568 Table 1. Comparison including ours and several state-of-the-art studies between various
569 pure-tone audiometry methods and CLS methods in terms of threshold level employing a set of
570 statistical measures (R, RMSE, and Bias). N = number of participants. The largest R, the
571 smallest RMSE, and bias between different combinations of audiometric and CLS methods are
572 highlighted in bold.

	Audiometric method	CLS method	N	R	RMSE	Bias
Fultz et al. 2020; Trevino et al. 2016	Audiometer	FL	17	0.21	12.2	-6.9
		MEL-Med		0.26	25.3	-18.0
		MEL-ML		0.26	15.5	-10.6
		SA		0.21	15.7	-8.4
Sanchez-Lopez et al. 2021	Audiometer	ACALOS	11	0.24	7.1	-2.3
current	SIUD	ACALOS	15	0.44	9.4	1.5
	GRaBr			0.38	9.0	1.0
	SIUD	rACALOS		0.59	7.8	0.5
	GRaBr			**0.71**	**6.9**	**0.04**

573

574 **Advantages of rACALOS**

575 **Increased time efficiency**: The rACALOS procedure combines two listening tests—pure-
576 tone audiometry and ACALOS—into a single, integrated protocol. This approach significantly
577 reduces the measurement time required for participants by eliminating the need for separate tests.

578 **Improved HTL accuracy**: Compared to the original ACALOS, rACALOS includes
579 additional trials near the hearing threshold level (HTL), enhancing the precision of HTL

estimation (see Table 1 for details). These modifications enable the seamless integration of audiometric measurement into the ACALOS framework.

Consistent user interface and no additional training requirements: The rACALOS procedure uses the same interface as ACALOS, so participants familiarized with ACALOS require no extra training to complete the new protocol.

Limitations and Outlook

In this study, we conducted smartphone-based listening tests outside of a sound booth, preceded by ambient noise level measurements. Given that most tests occurred in rather quiet acoustical conditions (i.e., little environmental noise pollution), the testing environment generally exhibited a low background noise level. However, many individuals live in urban regions with significant vehicle or industrial noise, where real-world environments are typically much noisier. Testing in such noisy conditions warrants further investigation. Potential solutions, such as circumaural muffs or noise-canceling earphones (NCE), could prove effective. For instance, Saliba et al. (2017) evaluated mobile-based audiometry under 50 dB A background noise, using passive and active noise cancellation by placing circumaural muffs over insert headphones, successfully reducing noise. Similarly, Clark et al. (2017) tested NCE (BoseQuietComfort 15) in a patient consultation room and found that NCE sufficiently attenuated ambient noise below the ANSI standards.

A key concern for out-of-booth audiometric tests is distraction. As noted by Margolis et al. (2022), background noise not only causes direct masking but also acts as a source of distraction. Their study demonstrated that increasing background noise levels led to elevated hearing thresholds and higher subjective ratings of distraction. Xu et al. (2024a) further supported these findings, characterizing distraction from internal noise (e.g., background noise) as long-term

inattention. They also proposed and simulated short-term inattention—where listeners are distracted by external events—during mobile hearing tests, though this has yet to be validated with human participants.

Another limitation of this study is the use of an integrated microphone for noise measurement. Studies like Kopun et al. (2022) recommend using an external microphone, such as the MicW iBoundary, which provides higher accuracy in capturing frequency characteristics and calibration precision compared to the internal microphone used here. Enhanced calibration of smartphone microphones could be achieved with an external reference sound, such as a whistle tone produced by a standard empty beer bottle (Scharf et al., 2024). However, achieving more accurate calibration and a detailed assessment of ambient noise spectra is beyond the scope of this proof-of-concept study, which involved a limited sample size. Future research will expand the sample size and include participants with sensorineural hearing loss for comparison.

Finally, Shen et al. (2018) and Kursun et al. (2023) introduced a quick categorical loudness scaling (qCLS) procedure based on a Bayesian adaptive method, which can estimate equal loudness contours within just 5 minutes. Given its efficiency and accuracy, incorporating qCLS into future smartphone-based loudness tests is worth considering. However, it remains uncertain whether qCLS can estimate hearing thresholds as precisely as the rACALOS developed in this study, highlighting the need for further research to evaluate its threshold accuracy in comparison.

CONCLUSION

This proof-of-concept study demonstrates that smartphone-based hearing tests—specifically pure-tone audiometry and categorical loudness scaling—can be effectively conducted remotely in participants' homes, provided that background noise levels are sufficiently low (e.g., below the MPANLs standard). The key findings from our experiments can be summarized as follows:

Validation Experiment: Our results indicate that air-conduction pure-tone audiometry and categorical loudness scaling yield equivalent outcomes in two test environments (i.e., at home and inside a sound-attenuated booth) at frequencies of 0.25, 1, and 4 kHz, suggesting satisfactory validity.

Test-Retest Reliability Experiment: Despite background noise levels reaching up to 45 dB A in a home environment, both audiometric tests exhibited moderate-to-good test-retest reliability, with the reliability at 1 kHz being higher than at the other two frequencies.

Performance of GRaBr: GRaBr demonstrated greater reliability than SIUD across all three frequencies, evidenced by a higher (intraclass) correlation and a lower RMSE value. Consequently, GRaBr is preferred for mobile audiometry outside of the booth due to its enhanced reliability.

Performance of rACALOS: Both computer simulations and human experiments confirm that thresholds estimated by rACALOS are closer to those measured using standard audiometric procedures compared to baseline ACALOS, indicating that the rACALOS method improves HTL estimation. In real-world environments, this reinforcement strategy may be particularly beneficial, as low SPL test stimuli are more susceptible to interference from background noise. In addition, the rACALOS method can integrate threshold measurement with the ACALOS test, resulting in greater efficiency compared to conducting the two tests separately. Therefore, the rACALOS approach holds promise for efficient remote assessments using mobile devices in the future.

GLOSSARY

Abbreviation	Meaning
ACALOS	adaptive categorical loudness scaling

ANOVA	analysis of variance
B&K	Brüel&Kjaer
BTUX	fitting method for loudness function in ACALOS
CLS	categorical loudness scaling
CU	categorical units
FL	fixed-level procedure
GRaBr	graded response bracketing
HI	hearing impaired
HTL	hearing threshold level (at 2.5 CU on the loudness function)
ICC	intraclass cross-correlation
IQR	interquartile ranges
LOA	level of agreement
MEL-Med	maximum expected information-maximum likelihood
MEL-ML	maximum expected information-median
MIQR	mean interquartile range
MPANLs	maximum permissible ambient noise levels
MSD	mean signed difference
NCE	noise reduction earphones
NH	normal hearing
PTA	pure-tone average
qCLS	quick categorical loudness scaling
rACALOS	reinforced adaptive categorical loudness scaling
RMSE	root mean squared error
SA	slope-adaptive procedure
SIUD	single interval up and down
SPL	sound pressure level

REFERENCES (BIBLIOGRAPHIC)

650 Akeroyd, M. A., Arlinger, S., Bentler, R. A., Boothroyd, A., Dillier, N., Dreschler, W. A., ... &
651 Kollmeier, B. (2015). International Collegium of Rehabilitative Audiology (ICRA)
652 recommendations for the construction of multilingual speech tests: ICRA Working Group
653 on Multilingual Speech Tests. International journal of audiology, 54(sup2), 17-22.
654 Almufarrij, I., Dillon, H., Dawes, P., Moore, D. R., Yeung, W., Charalambous, A. P., ... &
655 Munro, K. J. (2022). Web-and app-based tools for remote hearing assessment: a scoping
656 review. International Journal of Audiology, 1-14.
657 American National Standards Institute. Maximum Permissible Ambient Noise Levels for
658 Audiometric Test Rooms. (ANSI S3.1–R2018).New York, NY: American National
659 Standards Institute; 2018
660 Bean, B. N., Roberts, R. A., Picou, E. M., Angley, G. P., & Edwards, A. J. (2022). Automated
661 audiometry in quiet and simulated exam room noise for listeners with normal hearing and
662 impaired hearing. Journal of the American Academy of Audiology, 33(01), 006-013.
663 Behar, A. (2021). Audiometric tests without booths. International Journal of Environmental
664 Research and Public Health, 18(6), 3073.
665 Bianco, R., Mills, G., de Kerangal, M., Rosen, S., & Chait, M. (2021). Reward enhances online
666 participants' engagement with a demanding auditory task. Trends in Hearing, 25,
667 23312165211025941.
668 Brand, T., & Hohmann, V. (2002). An adaptive procedure for categorical loudness scaling. The
669 Journal of the Acoustical Society of America, 112(4), 1597-1604.
670 Brand, T., 2000. Analysis and Optimization of Psychophysical Procedures in Audi-ology.
671 Universität Oldenburg, Germany. PhD thesis.

672 Brennan-Jones, C. G., Eikelboom, R. H., Swanepoel, D. W., Friedland, P. L., & Atlas, M. D.
673 (2016). Clinical validation of automated audiometry with continuous noise-monitoring in
674 a clinically heterogeneous population outside a sound-treated environment. International
675 journal of audiology, 55(9), 507-513.
676 Buhl, M., Akin, G., Saak, S., Eysholdt, U., Radeloff, A., Kollmeier, B., & Hildebrandt, A. (2022).
677 Expert validation of prediction models for a clinical decision-support system in audiology.
678 Frontiers in Neurology, 13, 960012.
679 Clark, J. G., Brady, M., Earl, B. R., Scheifele, P. M., Snyder, L., & Clark, S. D. (2017). Use of
680 noise cancellation earphones in out-of-booth audiometric evaluations. International
681 Journal of Audiology, 56(12), 989-996.
682 Fultz, S. E., Neely, S. T., Kopun, J. G., & Rasetshwane, D. M. (2020). Maximum expected
683 information approach for improving efficiency of categorical loudness scaling. Frontiers
684 in Psychology, 11, 578352.
685 Giavarina, D. (2015). Understanding bland altman analysis. Biochemia medica, 25(2), 141-151.
686 Hazan, A., Luberadzka, J., Rivilla, J., Snik, A., Albers, B., Méndez, N., ... & Kinsbergen, J.
687 (2022). Home-Based Audiometry With a Smartphone App: Reliable Results?. American
688 Journal of Audiology, 31(3S), 914-922.
689 ISO 16832, 2006. Acousticsd Loudness Scaling by Means of Categories. Standard of the
690 International Organization for Standardization, Geneva, Switzerland.
691 Kohlrausch, A., Fassel, R., Van Der Heijden, M., Kortekaas, R., Van De Par, S., Oxenham, A. J.,
692 & Püschel, D. (1997). Detection of tones in low-noise noise: Further evidence for the role
693 of envelope fluctuations. Acta Acustica united with Acustica, 83(4), 659-669.

694 Kollmeier, B., Gilkey, R. H., & Sieben, U. K. (1988). Adaptive staircase techniques in
695 psychoacoustics: A comparison of human data and a mathematical model. The Journal of
696 the Acoustical Society of America, 83(5), 1852-1862.
697 Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation
698 coefficients for reliability research. Journal of chiropractic medicine, 15(2), 155-163.
699 Kopun, J. G., Turner, M., Harris, S. E., Kamerer, A. M., Neely, S. T., & Rasetshwane, D. M.
700 (2022). Evaluation of Remote Categorical Loudness Scaling. American journal of
701 audiology, 31(1), 45-56.
702 Kursun, Bertan & Petersen, Erik & Shen, Yi. (2023). Exploring Self-directed Hearing-aid Fitting
703 with No Booth And No Audiogram. 10.13140/RG.2.2.19575.19360.
704 Lecluyse, W., & Meddis, R. (2009). A simple single-interval adaptive procedure for estimating
705 thresholds in normal and impaired listeners. The Journal of the Acoustical Society of
706 America, 126(5), 2570-2579.
707 Maclennan-Smith, F., Swanepoel, D. W., & Hall III, J. W. (2013). Validity of diagnostic pure-
708 tone audiometry without a sound-treated environment in older adults. International
709 journal of audiology, 52(2), 66-73.
710 Margolis, R. H., Saly, G. L., & Wilson, R. H. (2022). Ambient Noise Monitoring during Pure-
711 Tone Audiometry. Journal of the American Academy of Audiology, 33(01), 045-056.
712 Meinke, D. K., & Martin, W. H. (2023). Boothless audiometry: Ambient noise considerations.
713 The Journal of the Acoustical Society of America, 153(1), 26-39.
714 Min, S. H., & Zhou, J. (2021). Smplot: An R package for easy and elegant data visualization.
715 Frontiers in Genetics, 12, 2582.

716 Oetting, D., Brand, T., & Ewert, S. D. (2014). Optimized loudness-function estimation for
717 categorical loudness scaling data. Hearing Research, 316, 16-27.
718 Ooster, J., Krueger, M., Bach, J. H., Wagener, K. C., Kollmeier, B., & Meyer, B. T. (2020).
719 Speech audiometry at home: automated listening tests via smart speakers with normal-
720 hearing and hearing-impaired listeners. Trends in Hearing, 24, 2331216520970011.
721 Peng, Z. E., Waz, S., Buss, E., Shen, Y., Richards, V., Bharadwaj, H., ... & Venezia, J. H. (2022).
722 Remote testing for psychological and physiological acoustics. The Journal of the
723 Acoustical Society of America, 151(5), 3116-3128.
724 Rasetshwane, D. M., Trevino, A. C., Gombert, J. N., Liebig-Trehearn, L., Kopun, J. G., Jesteadt,
725 W., ... & Gorga, M. P. (2015). Categorical loudness scaling and equal-loudness contours
726 in listeners with normal hearing and hearing loss. The Journal of the Acoustical Society
727 of America, 137(4), 1899-1913.
728 Revelle, W. (2018). psych: Procedures for psychological, psychometric, and personality research.
729 Robler, S. K., Coco, L., & Krumm, M. (2022). Telehealth solutions for assessing auditory
730 outcomes related to noise and ototoxic exposures in clinic and research. The Journal of
731 the Acoustical Society of America, 152(3), 1737-1754.
732 Saliba, J., Al-Reefi, M., Carriere, J. S., Verma, N., Provencal, C., & Rappaport, J. M. (2017).
733 Accuracy of mobile-based audiometry in the evaluation of hearing loss in quiet and noisy
734 environments. Otolaryngology–Head and Neck Surgery, 156(4), 706-711.
735 Sanchez-Lopez, R., Nielsen, S. G., El-Haj-Ali, M., Bianchi, F., Fereczkowski, M., Cañete, O.
736 M., ... & Santurette, S. (2021). Auditory tests for characterizing hearing deficits in
737 listeners with various hearing abilities: The BEAR test battery. Frontiers in neuroscience,
738 15, 724007.

739 Scharf, M. K., Huber, R., Schulte, M., & Kollmeier, B. (2024). Microphone calibration
740 estimation for mobile audiological tests with resonating bottles. International Journal of
741 Audiology, 1-7.
742 Serpanos, Y. C., Hobbs, M., Nunez, K., Gambino, L., & Butler, J. (2022). Adapting audiology
743 procedures during the pandemic: Validity and efficacy of testing outside a sound booth.
744 American Journal of Audiology, 31(1), 91-100.
745 Shen, Y., Zhang, C., & Zhang, Z. (2018). Feasibility of interleaved Bayesian adaptive procedures
746 in estimating the equal-loudness contour. The Journal of the Acoustical Society of
747 America, 144(4), 2363-2374.
748 Storey, K. K., Muñoz, K., Nelson, L., Larsen, J., & White, K. (2014). Ambient noise impact on
749 accuracy of automated hearing assessment. International Journal of Audiology, 53(10),
750 730-736.
751 Swanepoel, D. W., Matthysen, C., Eikelboom, R. H., Clark, J. L., & Hall III, J. W. (2015). Pure-
752 tone audiometry outside a sound booth using earphone attentuation, integrated noise
753 monitoring, and automation. International Journal of Audiology, 54(11), 777-785.
754 Swanepoel, D. W., Mngemane, S., Molemong, S., Mkwanazi, H., & Tutshini, S. (2010). Hearing
755 assessment—reliability, accuracy, and efficiency of automated audiometry. Telemedicine
756 and e-Health, 16(5), 557-563.
757 Trevino, A. C., Jesteadt, W., & Neely, S. T. (2016). Development of a multi-category
758 psychometric function to model categorical loudness measurements. The Journal of the
759 Acoustical Society of America, 140(4), 2571-2583.
760 Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., ... & Yutani,
761 H. (2019). Welcome to the Tidyverse. Journal of open source software, 4(43), 1686.

762 Xu, C., Hülsmeier, D., Buhl, M., & Kollmeier, B. (2024a). How Does Inattention Influence the
763 Robustness and Efficiency of Adaptive Procedures in the Context of Psychoacoustic
764 Assessments via Smartphone? Manuscript accepted by Trends in Hearing.
765 Xu, C., Schell-Majoor, L., & Kollmeier, B. (2024b). Development and verification of non-
766 supervised smartphone-based methods for assessing pure-tone thresholds and loudness
767 perception. Manuscript accepted by the International Journal of Audiology.
768 Zhao, S., Brown, C. A., Holt, L. L., & Dick, F. (2022). Robust and Efficient Online Auditory
769 Psychophysics. Trends in hearing, 26, 23312165221118792.