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 factors such as ambient noise, user attention, and correct headphone placement are controlled, the results are expected to be more 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 by the listener using a smartphone, while an integrated microphone and a dosimeter app were used to quantify ambient noise levels. Additionally, a reinforced ACALOS (rACALOS) method is proposed to integrate threshold measurement into the ACALOS procedure by presenting additional stimuli near the hearing threshold. 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 exceeding 0.75, and strong validity, with biases of less than 1 dB and upper and lower limits of agreement within ±15 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 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

Pure-tone audiometry, a threshold-based testing approach, is widely used in audiology for detecting hearing loss and guiding audiogram-based hearing aid fitting. In contrast, supra-threshold testing methods, such as categorical loudness scaling (CLS), assess loudness perception and are typically employed to identify supra-threshold deficits and enable loudness-based fitting. In addition, these supra-threshold listening tests are inherently robust to device calibration offsets, as their parameter estimation is based on level differences (Almufarrij et al., 2022), making them particularly suitable for mobile listening assessments. However, there is currently no standardized approach that integrates both threshold and CLS assessments—an integration that would be particularly advantageous for remote or mobile hearing evaluations. To address this gap, we propose the reinforced adaptive categorical loudness scaling (rACALOS) test, an extension of the adaptive CLS (ACALOS) test, designed to unify audiometric and CLS evaluations within a single procedure which can easily be adopted for smartphone-based testing.

This unified approach offers several advantages. First, it integrates both threshold and supra-threshold assessments, enabling a more comprehensive evaluation of hearing function and facilitating hearing aid fitting. Second, it improves efficiency by replacing two separate procedures with a single streamlined test. Third, rACALOS addresses limitations of ACALOS, particularly its inaccuracy and unreliability near the hearing threshold (Oetting et al., 2014). Furthermore, rACALOS is designed for remote, unsupervised testing, allowing participants to perform assessments at home, thereby enhancing accessibility and feasibility for large-scale studies. Because supra-threshold measures span a broad range of sound levels—or even rely on level differences—they are generally less sensitive to calibration errors than threshold-based measures such as the audiogram. This robustness also applies when combining threshold and

supra-threshold measures, thereby reducing the precision requirements for level calibration on mobile devices. Consequently, while scientific studies such as the present one still require proper calibration for each participant, a rough calibration with a precision of approximately 5–10 dB may be sufficient for practical applications (see Scharf et al. for a pragmatic field method).

The primary goal of this study is to evaluate the validity and test–retest reliability of smartphone-based audiometry and adaptive categorical loudness scaling (ACALOS), including the novel rACALOS method, in a home environment. This study addresses three key limitations, outlined as follows:

A) the impact of ambient noise on remote, unsupervised listening tests

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.

B) the unknown validity and reliability of remote listening tests

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, while only 12% of these tools have undergone validation and evaluation through peer-reviewed publications. 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. We are interested in both audiometric and categorical loudness scaling (CLS) tests, as they are fundamental not only for identifying hearing deficits—such as loudness recruitment, characterized by a reduced dynamic range between threshold and uncomfortable listening level—but also for the fitting of hearing devices. 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.

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). A Bland–Altman plot was used to assess the agreement between the newly introduced method and the reference method, as it provides a reliable evaluation that is independent of potential systematic bias. Unlike correlation analysis—which can indicate a strong linear relationship even when a substantial bias exists—the Bland–Altman approach reveals whether the two methods yield comparable results across the measurement range.

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

C) the need for optimization toward a unified, reinforced CLS test (rACALOS)

The adaptive CLS procedure (ACALOS, Brand & Hohmann, 2002; ISO 16832, 2006) often inaccurately estimates the audiometric threshold, as indicated by a correlation coefficient of less than 0.5 between the ‘true’ audiometric and estimated thresholds, reflecting a weak correlation.

Please note that the thresholds estimated by CLS (hereafter referred to as 'CLS thresholds') are defined as the level corresponding to 2.5 categorical units (CU) on the loudness growth function, as outlined by Oetting et al. (2014). 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). We are interested in CLS thresholds because they can help estimate audiometric thresholds, which are the outcome of audiometry. This allows the rACALOS approach to integrate pure-tone audiometry and CLS into a single process, enhancing the efficiency of remote testing.

Oetting et al. (2014) further demonstrated that the threshold predicted by the ACALOS method did not coincide with the 'true' audiometric threshold. This discrepancy may be at least partially attributed to the use of different stimuli—narrow-band noise in ACALOS versus pulsed tones in audiometry—and distinct psychophysical paradigms, namely, categorical magnitude estimation in ACALOS versus target sound detection in audiometry. To reduce this discrepancy, our study introduces a reinforced ACALOS (rACALOS) method, which integrates a more accurate threshold estimation process within the ACALOS procedure. This rACALOS approach allows participants to perform both threshold and ACALOS measurements in a single procedure rather than separate tests, thereby increasing efficiency. Additionally, the rACALOS method enhances reliability at low SPLs near the hearing threshold by incorporating additional trials with the aim to provide a more accurate estimate of the 'true' hearing threshold which is usually directly assessed in pure-tone audiometry. Although the rACALOS approach in this study was implemented in a research setting, where participants independently conducted the test at home using a calibrated smartphone, its intended application extends beyond research and may include clinical diagnostics and mobile health settings.

To accurately estimate the 'true' hearing threshold as a reference, it is essential to account for as many influencing factors as possible. In our previous work, we investigated the impact of experimenter supervision on pure-tone audiometry and adaptive categorical loudness scaling (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 GRaBr method is an adaptive approach that replaces the conventional yes–no task used in clinical threshold measurements with a two-interval paradigm, in which listeners are presented with two sounds at different levels and asked to indicate how many sounds they perceived—e.g., none, one, or two. The present study aims to further validate GRaBr by comparing its performance with established baseline methods (i.e., the single-interval up-down (SIUD), please see Lecluyse et al. (2009) for details) in human participants. Compared to the SIUD approach, GRaBr adaptively adjusts the level difference between the two sounds in each trial and has demonstrated improved efficiency and greater robustness to inattention in computer simulations (Xu et al., 2024a).

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 in comparison to simulations (please see Supplementary Materials S4 for the simulation results); 2) to assess the validity and test-retest reliability of the unsupervised smartphone-based application for pure-tone audiometry and ACALOS using a repeated measures design 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. There was only one test kit in use, and the setup was calibrated once at the beginning of data collection. Upon handing over the test kit, participants received a brief oral explanation of the remote experiments, and consent forms were signed before they began. Please note that the oral instructions consist of three parts: test environment

check, internet connection check, and steps for conducting the remote experiments. The details are provided below. 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.

Listeners were asked to complete remote testing at home on two separate days within a week—one for the test measurement and the other for the retest measurement. Each day's testing lasted approximately half an hour, resulting in a total testing time of one hour per participant. The home environments were primarily located in rural regions of northwestern Germany and 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 calibration procedure of the app-based sound level meter (SLM) in conjunction with the internal microphone was conducted as follows: First, the digital SLM, which had been calibrated

in our acoustics laboratory, was placed alongside the app-based SLM. Second, narrowband noise signals at 80 dB SPL were presented via a laptop to both the digital and app-based SLMs. Third, the real-time noise level of the app-based SLM was adjusted. In our case, a gain correction of 13.7 dB was required to match the calibrated digital SLM.

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. Please note that the same smartphone and headphones were provided to all test participants, ensuring a consistent gain across measurements.

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 based on the pre-experiment oral instructions. 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. In each trial of both procedures, two tones were presented: a probe tone and a cue tone at a higher level. The cue tone had a 20% chance of being muted, a condition known as a catch trial in psychology. The primary distinction between SIUD and GRaBr lies in the level difference between the probe and cue tone presented in most trials: fixed at 10 dB for SIUD, but variable for GRaBr (i.e., the level difference in the GRaBr procedure was initially set to 10 dB, then reduced to 5 dB after the first instance in which listeners reported hearing only one tone, and further decreased to 2 dB after the second such instance). The SIUD procedure began with an initial step size of 10 dB. Once the first 'one' tone response was recorded, the sound level was adjusted to the midpoint of the previous two levels. Thereafter, a 2 dB step size was applied (see Fig. 3A in Xu et al. (2024a) for an example run). The GRaBr

procedure started with an initial step size of 8 dB. After the first reversal, the step size was reduced to 6 dB, then halved to 3 dB after the second reversal. Finally, after the third reversal, the step size was set to 1 dB (see Fig. 3B in Xu et al. (2024a) for an example run). To ensure a fair comparison between the two methods, key parameters, such as the minimum number of trials, minimum number of reversals, and starting level, were matched as closely as possible. The initial sound level of the cue tone for both procedures was drawn from a discrete uniform distribution between 55 dB and 65 dB, with 1 dB increments (11 possible values). Both procedures 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 in a randomized order 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 (i.e., the tracks of the cue tones and the probe tones) 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. Please refer to Xu et al. (2024a) for details on the adaptive procedures.

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). Each frequency was assessed in a separate run, with evaluations conducted at 0.25, 1, and 4 kHz.

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). 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. In contrast, the mean intra-subject standard deviation at the level corresponding to 50 categorical units on the loudness function (i.e., the estimated uncomfortable loudness level) was below 5 dB, indicating that this measure is more reliable than the threshold estimate.

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.

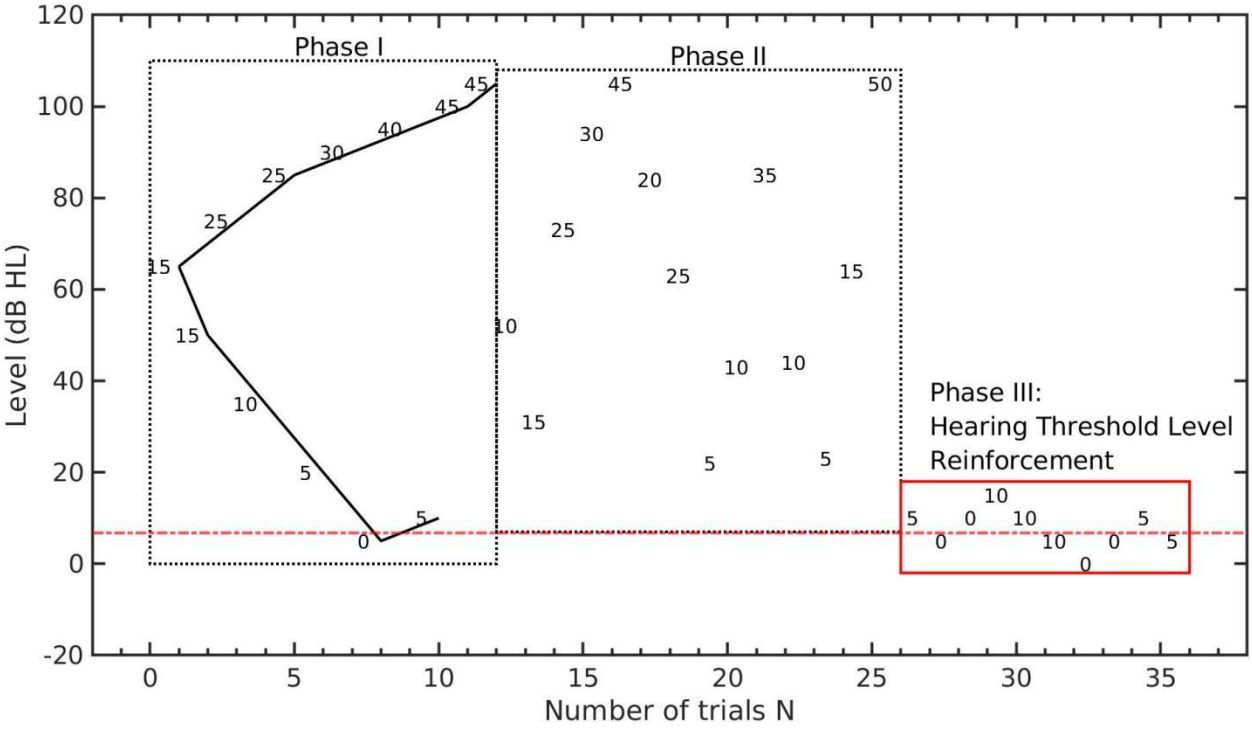

Fig. 1. An example track of the reinforced adaptive categorical loudness scaling (rACALOS), where the level (in dB HL) is plotted as a function of the number of trials N. The listener's response (in categorical units (CU)) is annotated with numbers between 0 ('not heard') and 50 ('too loud'). Left dotted 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.

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. In-booth audiometry and ACALOS testing were performed in a soundproof booth and supervised by a researcher. We compared the estimated HTL from the ACALOS and rACALOS methods to the 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.

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). Pure-tone thresholds obtained using the clinical method in a soundproof booth by a researcher were treated as the gold standard. 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.

Overall, this study involved three main sets of comparisons. First, to evaluate validity, we compared the results of unsupervised audiometry (GRaBr and SIUD) and loudness scaling (rACALOS and ACALOS) conducted at home against the corresponding in-booth standard measurements administered by a researcher. Second, to assess test-retest reliability, we compared results from two home-based runs (Run 1 vs. Run 2) for all four tests: SIUD, GRaBr, ACALOS, and rACALOS. Third, to examine HTL estimation accuracy, we compared the hearing thresholds derived from the ACALOS and rACALOS methods with those obtained via GRaBr and SIUD.

RESULTS

Noise Level Measurements

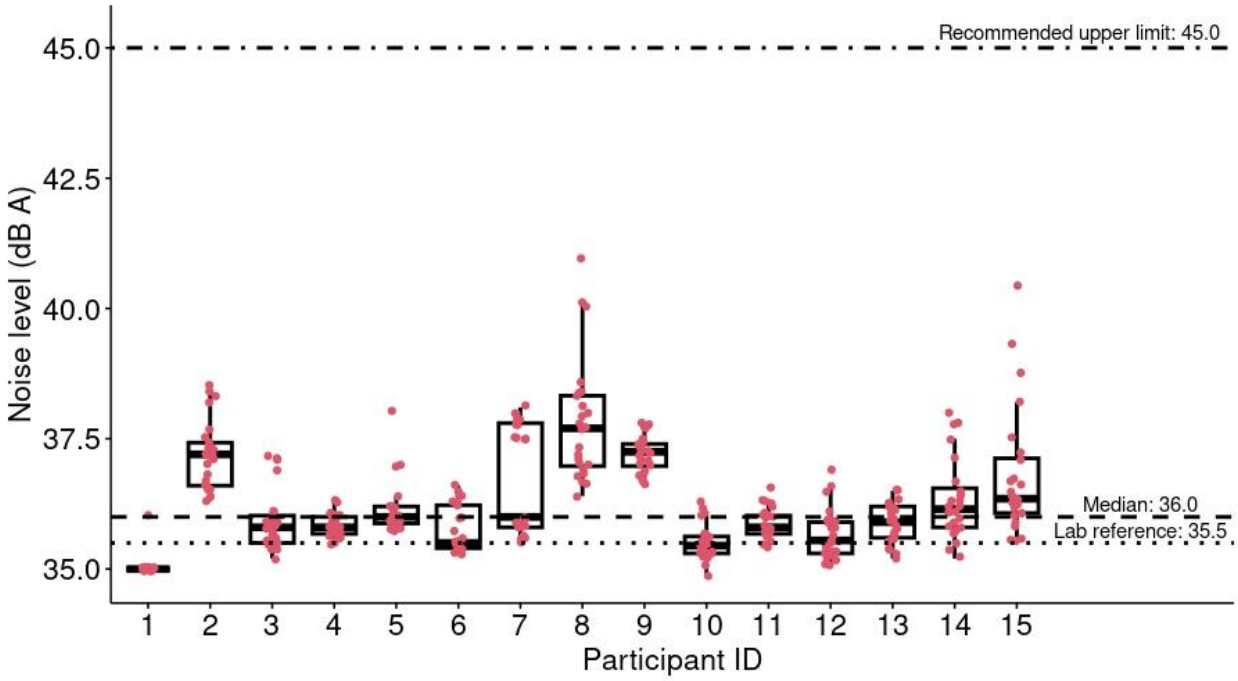

Fig. 2. Ambient noise level (in dB A) measurement across participants (N = 15). Medians, 25th and 75th percentiles, and interquartile ranges (IQR) are visualized in the box-plot while the end of the whiskers denotes the minimum and maximum, indicating the 5th and 95th percentiles

respectively. Individual data points are displayed in the jitter plot. Dotted line: lab reference (i.e., ambient noise level measured within a booth). Dashed line: median value across subjects. Dot-dashed line: recommended upper limit.

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. Please note that the instantaneous ambient noise level was recorded at the start of each measurement session. 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. The interquartile range (IQR) across participants was 1.2 dB. 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.

Test-Retest Reliability

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 ($F(2, 28) = 15.46$, $p < 0.05$). 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 ($t(14) = 2.85$, $p < 0.05$).

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 ($p < 0.05$). Compared to the ACALOS procedure, the rACALOS procedure yielded lower mean sound levels (by 1.9 dB) across participants and CU. 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 ($F(2, 28) = 2.51$, $p = 0.10$), and as expected, the two runs (test and retest measurements) did not differ ($F(1, 14) = 1.97$, $p = 0.18$). 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 25CU ($t(29) = 2.08$) and 40 CU ($t(29) = 2.21$)).

Figure 3 illustrates the test-retest reliability of the rACALOS procedure in a home environment using a Bland-Altman plot to compare runs 1 and 2. The overall bias between the two runs was 0.74 dB, indicating minimal deviation and suggesting high reliability of the rACALOS procedure even in a home setting. The 95% limits of agreement were found to be located at -20.3 dB and 21.8 dB, respectively. The correlation coefficient (R) for categorical units (CUs) of 0, 5, and 25 was below 0.6, whereas for all other CUs, R exceeded 0.7. The bias for each CU remained below 2 dB. The root mean squared error (RMSE) for CUs 0, 5, and 15 exceeded 10 dB, while for all other CUs, it was below 10 dB. Overall, RMSE values tended to decrease as CU increased.

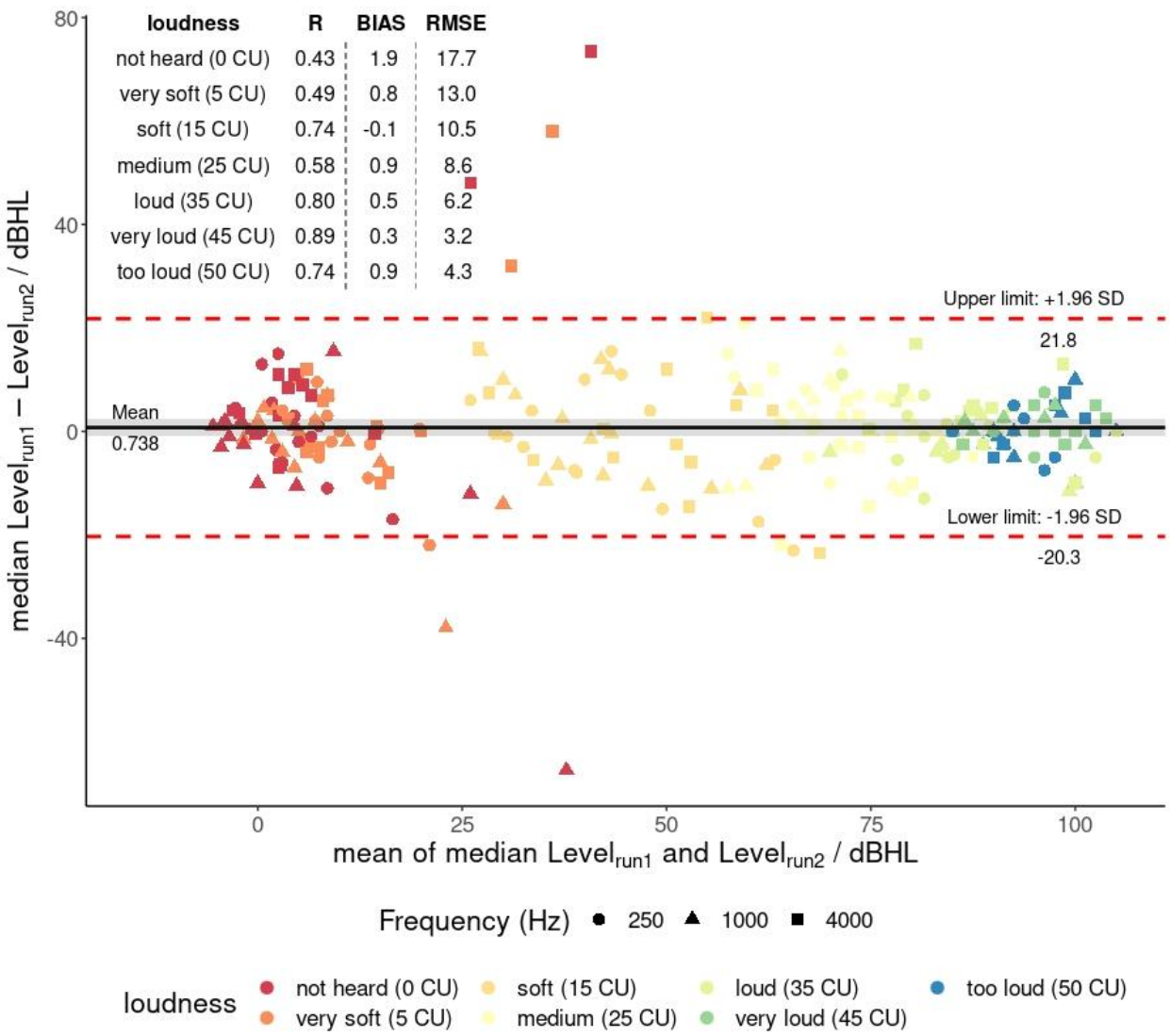

Fig. 3. Bland-Altman plot assessing the reliability of **the rACALOS procedure** by comparing run 1 and run 2 in a home environment. Each point represents the median level of a categorical unit (CU) for each participant. CUs are distinguished by color, while frequencies are represented by different shapes. The correlation coefficient (R), bias (BIAS), and root mean squared error (RMSE) for each CU are provided in the top-left corner of the figure. Red dashed lines: 95% limits of agreement; Black solid line: bias between the test and retest measurements; Grey shaded rectangle area: 95% confidence interval of the bias.

Validation Experiment

GRaBr

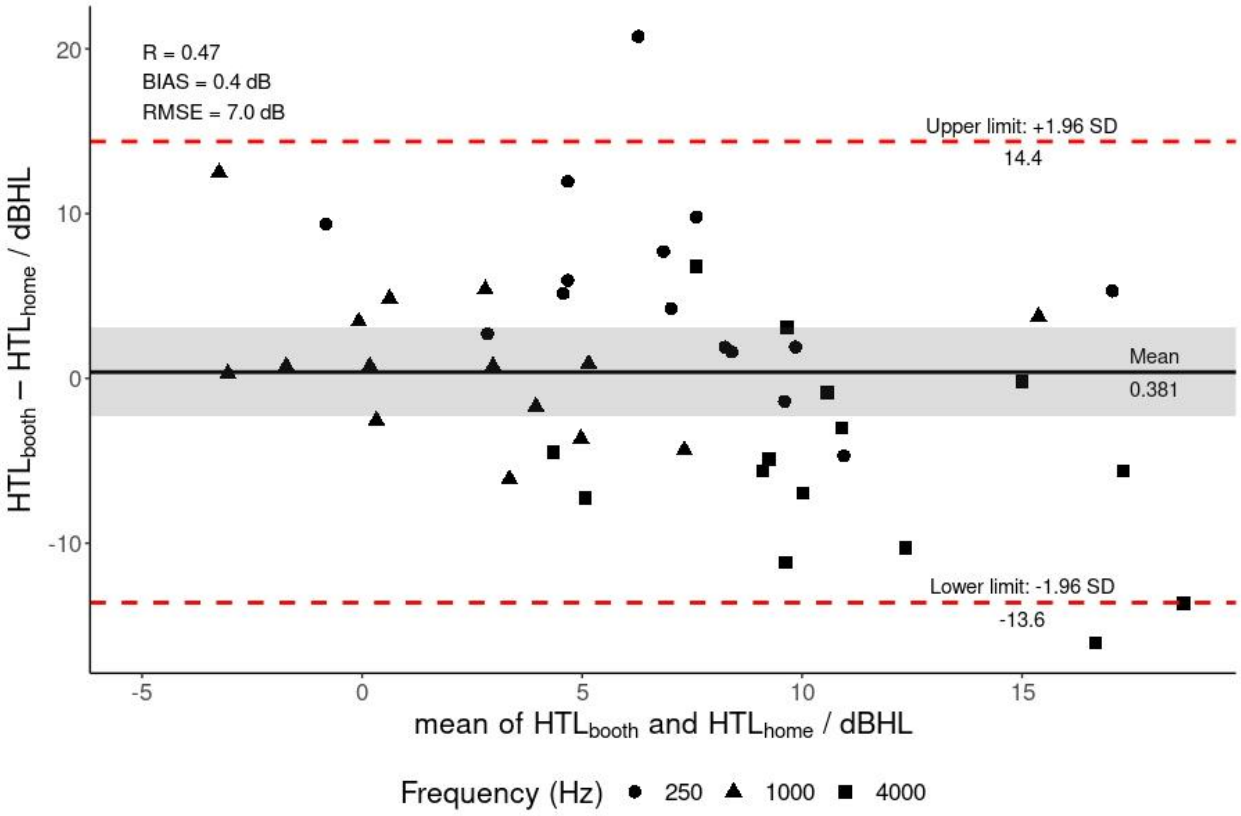

Fig. 4. Bland-Altman plot of hearing threshold levels (HTL) in dB HL of frequencies at 0.25, 1, and 4 kHz (represented with circle, triangle, and rectangle, respectively) measured inside the booth (i.e., HTL_{booth}) using the standard audiometry and at home (i.e., HTL_{home}) using **the GRaBr procedure**. The correlation coefficient (R), bias (BIAS), and root mean squared error (RMSE) are provided in the top-left corner. Please refer to Fig. 3 for an explanation of the Bland-Altman plot and Supplementary Materials S1 for its statistical implication.

Fig. 4 compares pure-tone audiometry results obtained in the booth using the standard audiometry versus testing at home using the GRaBr procedure at frequencies of 0.25, 1, and 4 kHz. The differences between the two measurement conditions did not significantly deviate from

a normal distribution, as confirmed by the Shapiro–Wilk test ($p = 0.92$). The average difference (systematic bias) between hearing thresholds measured in the booth and at home was 0.4 dB, which is well below the commonly accepted clinical margin of 5 dB (Thai-Van et al., 2023), indicating clinically acceptable agreement in terms of bias. The correlation between standard clinical audiometry and home-based audiometry using the GRaBr procedure was moderate, likely reflecting the limited variability in hearing thresholds among the normal-hearing participants. The root mean squared error (RMSE) was 7 dB, consistent with values reported in previous validation studies of app-based audiometry (e.g., Wiseman et al., 2023), suggesting comparable accuracy. However, the 95% limits of agreement (−13.6 to 14.4 dB) slightly exceeded the typical clinical threshold variability of ±10 dB (Fultz et al., 2020), indicating that the combined variability arising from both methodological differences (home vs. booth) and procedural differences (GRaBr vs. standard clinical audiometry) is somewhat greater than what is typically expected in clinical settings. Furthermore, the 95% confidence interval of the bias (depicted by the shaded region) encompassed the line of equality, suggesting no bias between the two testing environments. 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 ($F(1, 14) = 0.09$, $p = 0.77$); however, the main effect of frequency was significant ($F(2, 28) = 20.80$, $p < 0.05$). The interaction between frequency and test environment was statistically significant, $F(2, 28) = 36.01$,

$p < 0.05$. 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 ($t(14) = -2.89$ for 0.25 kHz, $t(14) = 2.93$ for 4 kHz, $p < 0.05$).

Validation results for the SIUD procedure in a home environment, compared to a standard audiometer, are presented in Fig. S1. The SIUD method showed a bias of 0.6 dB, indicating good validity. Furthermore, the RMSE of the SIUD procedure in a home environment, compared to standard audiometry in a laboratory setting, was 6.8 dB. Additionally, the SIUD procedure differed significantly from GRaBr in measured thresholds ($p < 0.05$). The mean hearing threshold difference between the GRaBr and SIUD procedures across participants was approximately 0.75 dB, with the GRaBr procedure yielding higher average thresholds. Overall, the validity of both adaptive procedures was comparable, suggesting that both are suitable for remote measurements in home settings.

rACALOS

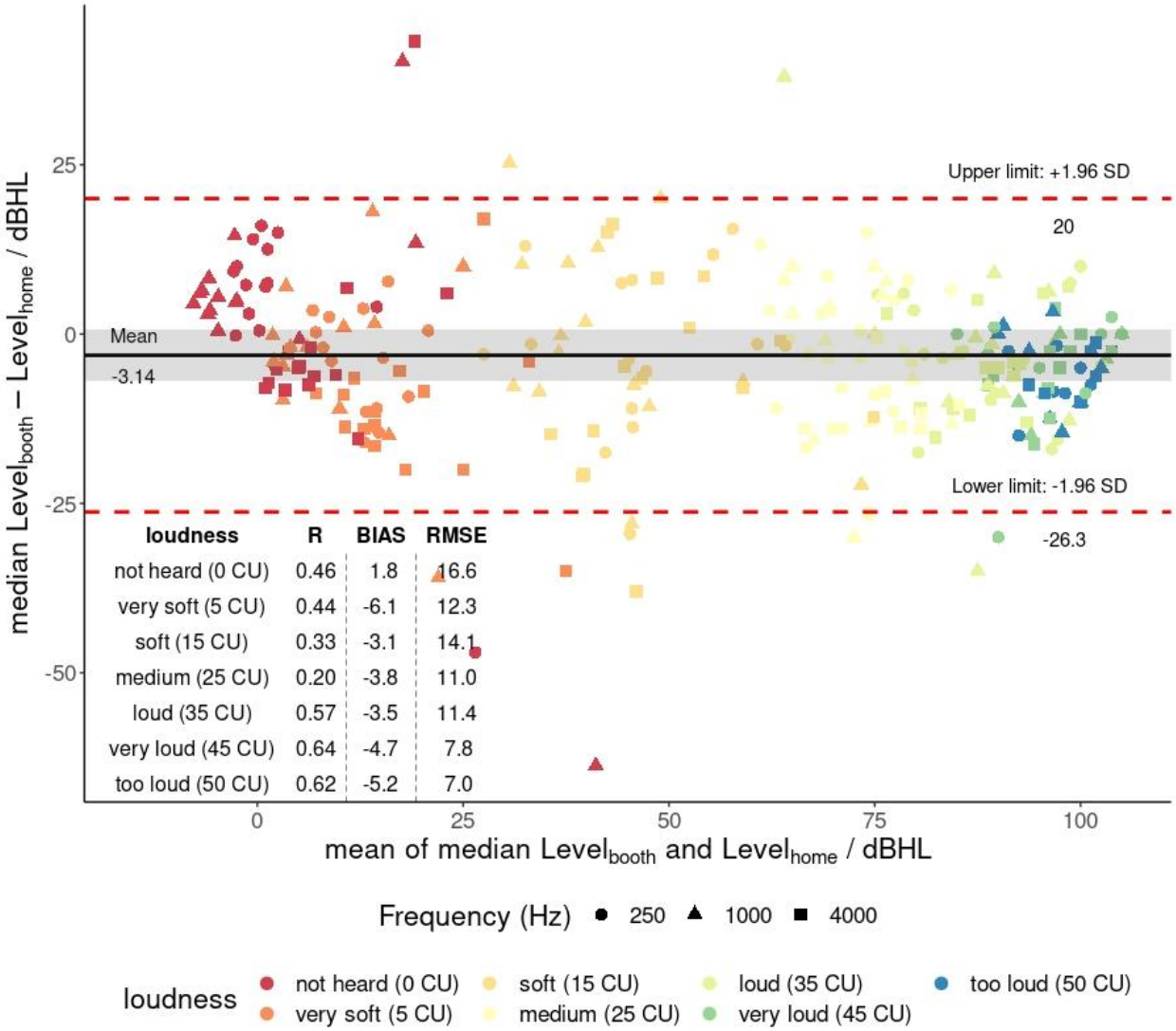

Fig. 5. Bland-Altman plot of median levels assigned to each CU (denoted with different colors) for three frequencies (represented with different shapes) for comparing two test environments, i.e., inside the booth using a standard CLS procedure and at home using **the rACALOS procedure** for each participant. A comprehensive set of statistical measures containing R, Bias, and RMSE of each CU is provided in the embedded table located at the bottom-left corner. See Fig. 3 for an explanation of the Bland-Altman plot and Supplementary Materials S1 for its statistical implication.

Fig. 5 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% upper and lower limits of agreement (LOA) were 26.3 dB and -19.6 dB, respectively. The Shapiro-Wilk test indicated that the differences between the two measurements followed a normal distribution ($p = 0.22$). 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. 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 reduced agreement between both measurements 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.

Compared to Fig. 3 (reliability of rACALOS at home), the range of the 95% limits of agreement was broader, and the overall bias was larger. Additionally, the R values for each CU were smaller, while the bias and RMSE values were generally larger, indicating that the validity experiment from Fig. 5 shows less consistency than the reliability experiment from Fig.3 This outcome was expected, as measurements taken in two different locations typically yield less consistent results than repeated measurements at the same location.

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 ($F(2, 28) = 5.33$ for frequency, $F(3, 38) = 353.30$ for CU, $p < 0.05$). 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 CU ($t(14) = 3.03$), 25 CU ($t(14) = 3.78$), and 45 CU ($t(14) = 2.33$)).

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.

Accuracy of HTL Estimation for the rACALOS procedure

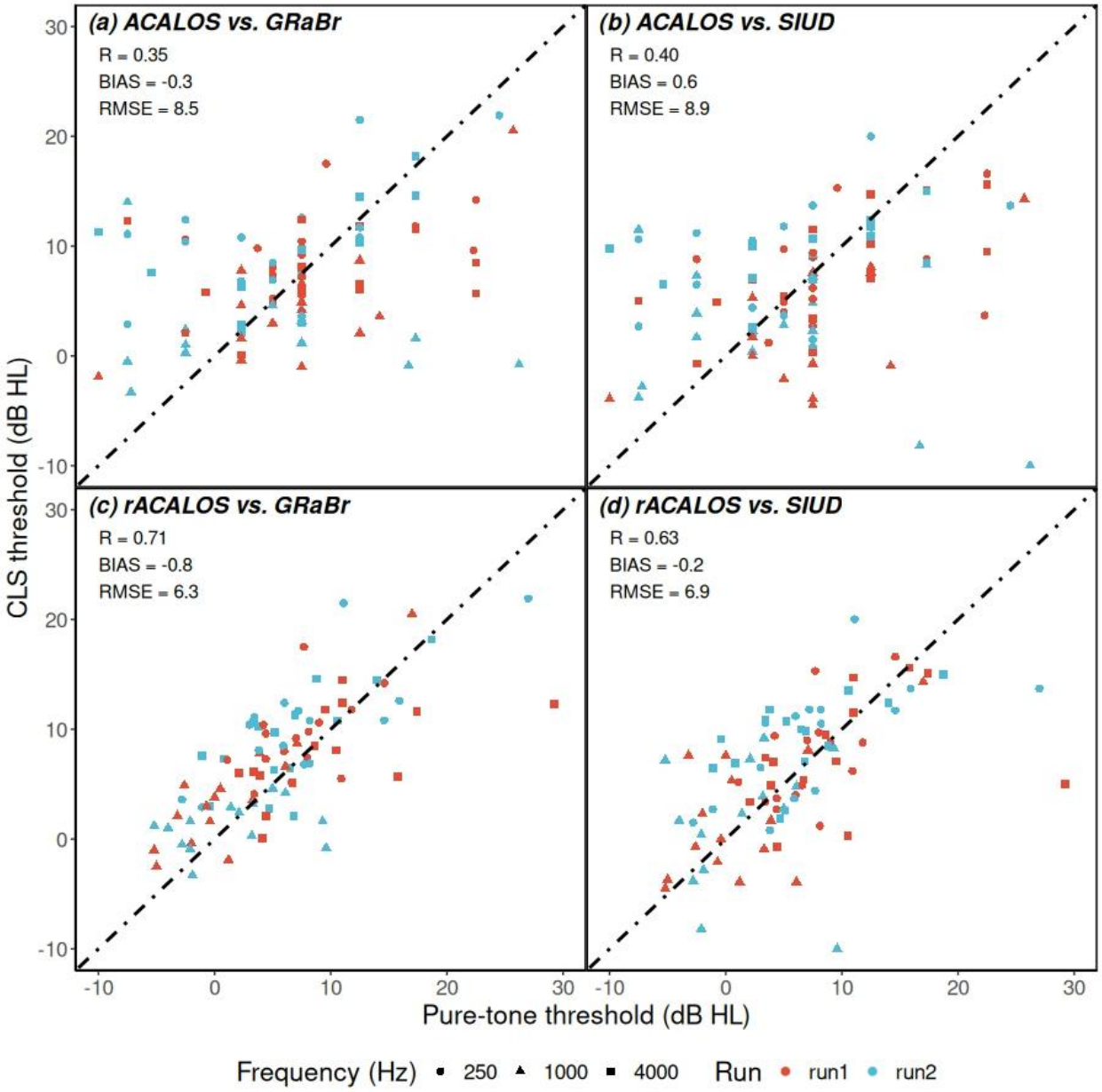

Fig. 6. Scatter plot for comparison between pure-tone (abscissa) and CLS (ordinate) thresholds in dB HL of N = 15 individual listeners. Frequency is labeled with different shapes while the run is denoted with different colors (run1: red, run2: blue). A set of statistical metrics (R, Bias, and RMSE) are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

This study systematically investigated the impact of ambient noise on smartphone-based listening tests. The results demonstrate the feasibility of conducting unsupervised audiometry and ACALOS tests outside a sound-treated booth, provided that ambient noise levels are maintained within acceptable limits. We also evaluated the validity and reliability of two novel adaptive methods—GRaBr for audiometry and rACALOS for categorical loudness scaling—implemented on smartphones. Both GRaBr and rACALOS demonstrated high validity compared to standard methods and exhibited moderate-to-excellent test-retest reliability.

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. This may be because the ambient noise levels at home and in the test booth overlapped to some extent, despite the booth generally being expected to have lower background noise levels than typical home environments. As a consequence, 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 demonstrates moderate reliability, GRaBr shows good reliability (the ICC values are greater than 0.75 ($p < 0.05$)) for remote smartphone-based assessments, making GRaBr the significantly more reliable option ($p < 0.05$), as expected from the simulations reported by Xu et al. (2024a). Our findings align with recent studies examining the validity of boothless pure-tone audiometry (Maclennan-Smith et al., 2013; Storey et al., 2014; Swanepoel et al., 2015; Brennan-Jones et al., 2016; Serpanos et al., 2022). The bias between in-booth and at-home measurements is 0.4 dB, which falls within the empirical ranges reported by Maclennan-Smith et al. (2013) (-0.6 to 1.1 dB) and Swanepoel et al. (2015) (-2.0 to 1.5 dB). However, the correlation coefficient R (0.47) in our study is notably lower than that reported by Maclennan-Smith et al. (2013), where R exceeded 0.92 for both ears at frequencies between 0.25 and 8 kHz. This discrepancy may be attributed to the much smaller range of thresholds across our participants: our study included 15 young adults with normal hearing, whereas Maclennan-Smith et al. (2013) had a larger sample of 147 elderly participants with hearing impairments, 59% of whom exhibited a pure-tone average

(PTA) greater than 25 dB. As Swanepoel et al. (2010) noted, hearing-impaired listeners typically show higher correlation coefficients than those with normal hearing due to reduced sensitivity and lesser impact from ambient noise. However, our test sample with young, normal hearing listeners puts a higher demand on the quietness of the acoustic environment and the reliability of the test procedure.

The biases observed for the SIUD and GRaBr procedures were −0.6 dB and 0.4 dB, respectively, both well within the commonly accepted clinical margin of ±5 dB and thus considered clinically acceptable. The 95% limits of agreement (LOA) were -14.1 to 13.0 dB for SIUD and -13.6 to 14.4 dB for GRaBr, slightly exceeding the typical clinical threshold of ±10 dB. While such variability may be tolerable for preliminary screening or population-level assessments, it may limit the applicability of unsupervised home-based testing for precise individual diagnostics. The relatively wide limits of agreement are expected, as variability likely accumulates due to differences in both the procedures (SIUD or GRaBr versus clinical audiometry) and the testing environments (booth versus home). It could also be attributed to factors such as individual differences in task engagement, ambient environmental conditions, or variability in smartphone acoustics. Future research should explore strategies to reduce this variability, including enhanced user instructions, noise monitoring, or calibration procedures.

The test-retest reliability aligns well with findings from previous studies, such as those by Swanepoel et al. (2015) and Hazan et al. (2022). The bias (N = 11) between test and retest measurements was 1.8, 0.0, and 1.4 dB at 0.25, 1, and 4 kHz, respectively, consistent with the findings of Swanepoel et al. (2015), where the bias also remained below 2 dB. The correlation 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 (i.e., 7.7 dB) closely matched that of Kopun et al. (2022) and was about 7 dB 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 (i.e., 6.7 dB) was smaller than in both empirical studies (i.e., about 10 dB for Kopun et al. (2022) and 15 dB for Rasetshwane et al. (2015)), 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

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. This finding is consistent with

computer simulations (see supplementary materials) that examined threshold variability across different parameter combinations of rACALOS compared to the original ACALOS procedure. 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, while pure-tone thresholds obtained using clinical audiometers are still widely regarded as the 'gold standard', more precise and criterion-free methods such as SIUD and GRaBr may produce stronger correlations with CLS thresholds. 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-octave-band noise utilized here is effectively equivalent to a frequency-modulated sinusoid with minor envelope fluctuations and an instantaneous frequency confined well within a critical band.

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

	Audiometric method	CLS method	N	R (spearman)	RMSE (dB)	Bias (dB)
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**

Advantages of rACALOS

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

Improved HTL accuracy: Compared to the original ACALOS, rACALOS includes 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.

690 **Limitations and Outlook**

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

703 A key concern for out-of-booth audiometric tests is distraction. As noted by Margolis et al.
704 (2022), background noise not only causes direct masking but also acts as a source of distraction.
705 Their study demonstrated that increasing background noise levels led to elevated hearing
706 thresholds and higher subjective ratings of distraction. Xu et al. (2024a) further supported these
707 findings, characterizing distraction from internal noise (e.g., background noise) as long-term
708 inattention. They also proposed and simulated short-term inattention—where listeners are
709 distracted by external events—during mobile hearing tests, though this has yet to be validated
710 with human participants.

711 Another limitation of this study is the use of an integrated microphone for noise measurement.
712 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.

In the future, we plan to expand our study by increasing the number of normal-hearing (NH) participants and incorporating hearing-impaired (HI) participants. Compared to NH listeners, we expect the validity of pure-tone audiometry and ACALOS tests in HI listeners to be comparable, as supported by previous studies (e.g., Hazan et al., 2022; Bean et al., 2022; Xu et al., 2024b), suggesting that hearing loss does not significantly affect test validity. Regarding reliability, HI listeners are expected to show similar or sometimes even higher test-retest reliability in audiometry than NH listeners (Hazan et al., 2022), likely due to their elevated hearing thresholds, which reduce the impact of ambient noise. Similarly, in the ACALOS procedure, HI listeners should exhibit similar or even greater reliability at lower levels of the loudness growth function, as they are less affected by background noise. This may have a limited impact on the accuracy of the loudness growth slope fitted to the data, which is generally increased in hearing-impaired listeners exhibiting recruitment. However, since the slope estimate is derived from loudness judgments across multiple supra-threshold levels, it is considered a reliable and robust measure, even when obtained through self-administered smartphone assessments.

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: Human experiments confirms that thresholds estimated by rACALOS are closer to those measured using audiometric procedures (i.e., SIUD and GRaBr) 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	limit 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)

Akeroyd, M. A., Arlinger, S., Bentler, R. A., Boothroyd, A., Dillier, N., Dreschler, W. A., ... & Kollmeier, B. (2015). International Collegium of Rehabilitative Audiology (ICRA) recommendations for the construction of multilingual speech tests: ICRA Working Group on Multilingual Speech Tests. International journal of audiology, 54(sup2), 17-22.

Almufarrij, I., Dillon, H., Dawes, P., Moore, D. R., Yeung, W., Charalambous, A. P., ... & Munro, K. J. (2022). Web-and app-based tools for remote hearing assessment: a scoping review. International Journal of Audiology, 1-14.

American National Standards Institute. Maximum Permissible Ambient Noise Levels for Audiometric Test Rooms. (ANSI S3.1–R2018).New York, NY: American National Standards Institute; 2018

Bean, B. N., Roberts, R. A., Picou, E. M., Angley, G. P., & Edwards, A. J. (2022). Automated audiometry in quiet and simulated exam room noise for listeners with normal hearing and impaired hearing. Journal of the American Academy of Audiology, 33(01), 006-013.

Behar, A. (2021). Audiometric tests without booths. International Journal of Environmental Research and Public Health, 18(6), 3073.

Bianco, R., Mills, G., de Kerangal, M., Rosen, S., & Chait, M. (2021). Reward enhances online participants' engagement with a demanding auditory task. Trends in Hearing, 25, 23312165211025941.

Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. Statistical methods in medical research, 8(2), 135-160.

Bland, J. M., & Altman, D. G. (2007). Agreement between methods of measurement with multiple observations per individual. Journal of biopharmaceutical statistics, 17(4), 571-582.

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

Brand, T., 2000. Analysis and Optimization of Psychophysical Procedures in Audi-ology. Universität Oldenburg, Germany. PhD thesis.

Brennan-Jones, C. G., Eikelboom, R. H., Swanepoel, D. W., Friedland, P. L., & Atlas, M. D. (2016). Clinical validation of automated audiometry with continuous noise-monitoring in a clinically heterogeneous population outside a sound-treated environment. International journal of audiology, 55(9), 507-513.

Buhl, M., Akin, G., Saak, S., Eysholdt, U., Radeloff, A., Kollmeier, B., & Hildebrandt, A. (2022). Expert validation of prediction models for a clinical decision-support system in audiology. Frontiers in Neurology, 13, 960012.

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.

Fultz, S. E., Neely, S. T., Kopun, J. G., & Rasetshwane, D. M. (2020). Maximum expected information approach for improving efficiency of categorical loudness scaling. Frontiers in Psychology, 11, 578352.

Giavarina, D. (2015). Understanding bland altman analysis. Biochemia medica, 25(2), 141-151.

Hazan, A., Luberadzka, J., Rivilla, J., Snik, A., Albers, B., Méndez, N., ... & Kinsbergen, J. (2022). Home-Based Audiometry With a Smartphone App: Reliable Results?. American Journal of Audiology, 31(3S), 914-922.

ISO 16832, 2006. Acousticsd Loudness Scaling by Means of Categories. Standard of the International Organization for Standardization, Geneva, Switzerland.

Kohlrausch, A., Fassel, R., Van Der Heijden, M., Kortekaas, R., Van De Par, S., Oxenham, A. J., & Püschel, D. (1997). Detection of tones in low-noise noise: Further evidence for the role of envelope fluctuations. Acta Acustica united with Acustica, 83(4), 659-669.

Kollmeier, B., Gilkey, R. H., & Sieben, U. K. (1988). Adaptive staircase techniques in psychoacoustics: A comparison of human data and a mathematical model. The Journal of the Acoustical Society of America, 83(5), 1852-1862.

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

Kopun, J. G., Turner, M., Harris, S. E., Kamerer, A. M., Neely, S. T., & Rasetshwane, D. M. (2022). Evaluation of Remote Categorical Loudness Scaling. American journal of audiology, 31(1), 45-56.

Kursun, Bertan & Petersen, Erik & Shen, Yi. (2023). Exploring Self-directed Hearing-aid Fitting with No Booth And No Audiogram. 10.13140/RG.2.2.19575.19360.

Lecluyse, W., & Meddis, R. (2009). A simple single-interval adaptive procedure for estimating thresholds in normal and impaired listeners. The Journal of the Acoustical Society of America, 126(5), 2570-2579.

Maclennan-Smith, F., Swanepoel, D. W., & Hall III, J. W. (2013). Validity of diagnostic pure-tone audiometry without a sound-treated environment in older adults. International journal of audiology, 52(2), 66-73.

Margolis, R. H., Saly, G. L., & Wilson, R. H. (2022). Ambient Noise Monitoring during Pure-Tone Audiometry. Journal of the American Academy of Audiology, 33(01), 045-056.

Meinke, D. K., & Martin, W. H. (2023). Boothless audiometry: Ambient noise considerations. The Journal of the Acoustical Society of America, 153(1), 26-39.

Min, S. H., & Zhou, J. (2021). Smplot: An R package for easy and elegant data visualization. Frontiers in Genetics, 12, 2582.

Oetting, D., Brand, T., & Ewert, S. D. (2014). Optimized loudness-function estimation for categorical loudness scaling data. Hearing Research, 316, 16-27.

Ooster, J., Krueger, M., Bach, J. H., Wagener, K. C., Kollmeier, B., & Meyer, B. T. (2020). Speech audiometry at home: automated listening tests via smart speakers with normal-hearing and hearing-impaired listeners. Trends in Hearing, 24, 2331216520970011.

Peng, Z. E., Waz, S., Buss, E., Shen, Y., Richards, V., Bharadwaj, H., ... & Venezia, J. H. (2022). Remote testing for psychological and physiological acoustics. The Journal of the Acoustical Society of America, 151(5), 3116-3128.

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.

Revelle, W. (2018). psych: Procedures for psychological, psychometric, and personality research.

Robler, S. K., Coco, L., & Krumm, M. (2022). Telehealth solutions for assessing auditory outcomes related to noise and ototoxic exposures in clinic and research. The Journal of the Acoustical Society of America, 152(3), 1737-1754.

Saliba, J., Al-Reefi, M., Carriere, J. S., Verma, N., Provencal, C., & Rappaport, J. M. (2017). Accuracy of mobile-based audiometry in the evaluation of hearing loss in quiet and noisy environments. Otolaryngology–Head and Neck Surgery, 156(4), 706-711.

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, 724007.

Scharf, M. K., Huber, R., Schulte, M., & Kollmeier, B. (2024). Microphone calibration estimation for mobile audiological tests with resonating bottles. International Journal of Audiology, 1-7. DOI: 10.1080/14992027.2024.2395416

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.

Shen, Y., Zhang, C., & Zhang, Z. (2018). Feasibility of interleaved Bayesian adaptive procedures in estimating the equal-loudness contour. The Journal of the Acoustical Society of America, 144(4), 2363-2374.

Storey, K. K., Muñoz, K., Nelson, L., Larsen, J., & White, K. (2014). Ambient noise impact on accuracy of automated hearing assessment. International Journal of Audiology, 53(10), 730-736.

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.

Thai-Van, H., Joly, C. A., Idriss, S., Melki, J. B., Desmettre, M., Bonneuil, M., ... & Reynard, P. (2023). Online digital audiometry vs. conventional audiometry: a multi-centre comparative clinical study. International Journal of Audiology, 62(4), 362-367.

Trevino, A. C., Jesteadt, W., & Neely, S. T. (2016). Development of a multi-category psychometric function to model categorical loudness measurements. The Journal of the Acoustical Society of America, 140(4), 2571-2583.

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.

Wiseman, K., Slotkin, J., Spratford, M., Haggerty, A., Heusinkvelt, M., Weintraub, S., ... & McCreery, R. (2023). Validation of a tablet-based assessment of auditory sensitivity for researchers. Behavior research methods, 55(6), 2838-2852.

Xu, C., Hülsmeier, D., Buhl, M., & Kollmeier, B. (2024a). How Does Inattention Influence the Robustness and Efficiency of Adaptive Procedures in the Context of Psychoacoustic Assessments via Smartphone?. Trends in Hearing, 28, 23312165241288051.

Xu, C., Schell-Majoor, L., & Kollmeier, B. (2024b). Development and verification of non-supervised smartphone-based methods for assessing pure-tone thresholds and loudness perception. International Journal of Audiology, 1-11.

Zhao, S., Brown, C. A., Holt, L. L., & Dick, F. (2022). Robust and Efficient Online Auditory Psychophysics. Trends in hearing, 26, 23312165221118792.