FÖRDERPREIS **2025**

PLATZ 3

Dr. Chen Xu

Crucial Elements of a Virtual Hearing Clinic on Mobile Devices: Psychophysics, Diagnostic Parameter Estimation, and Validation

Von der Fakultät VI für Medizin und Gesundheits-wissenschaften der Carl von Ossietzky Universität Oldenburg

Dissertation

EUHA

Europäische Union der Hörakustiker e.V.

Crucial Elements of a Virtual Hearing Clinic on Mobile Devices: Psychophysics, Diagnostic Parameter Estimation, and Validation

Von der Fakultät VI für Medizin und Gesundheitswissenschaften der Carl von Ossietzky Universität Oldenburg zur Erlangung des Grades und Titels eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

eingereichte Dissertation

von Herrn Chen Xu

geboren am 01.06.1995 in Jiangsu

Erstbetreuer: Prof. Dr. Dr. Birger Kollmeier

Eingereicht: 11.02.2025

Carl von Ossietzky
Universität
Oldenburg

Crucial Elements of a Virtual Hearing Clinic on Mobile Devices: Psychophysics, Diagnostic Parameter Estimation, and Validation

Dissertation

Submitted:	11.02.2025
By:	Chen Xu born on 01.06.1995 in Jiangsu, China
Matriculation number:	6401027
Supervisor:	Prof. Dr. Dr. Birger Kollmeier

Carl von Ossietzky Universität Oldenburg
Department of Medical Physics and Acoustics, Faculty VI
26111 Oldenburg
Phone: +49 (0) 441 798 5497
Homepage: https://uol.de/mediphysik

Abstract

Motivation:

Smartphone-based listening tests have expanded the reach of traditional laboratory-based assessments, offering convenient, low-threshold access to hearing evaluations and early diagnostics of hearing loss. Despite these advantages, such tests often suffer from a lack of controlled environments, absence of test supervisors, uncalibrated devices, and inattentive participants, resulting in potential inaccuracies and unreliable outcomes.

Objective:

The primary objectives are threefold: first, to develop and validate smartphone-based listening tests, including air-conduction pure-tone audiometry and categorical loudness scaling (CLS); second, to analyze the impact of factors such as inattention, supervision, and ambient noise on test performance; and third, to optimize adaptive procedures for mobile device implementation.

Design:

In the first sub-study, seven adaptive procedures were co-simulated in combination with two categories of inattentive listeners. The simulated listeners were parameterized with three levels of inattention and varying false alarm rates. The robustness of the adaptive procedures was quantified using bias and root-mean-square error (RMSE), while efficiency was measured through the rate of convergence and normalized efficiency. In addition, the graded response bracketing (GRaBr), a model-free and efficient adaptive procedure resistant to inattention, was introduced for audiogram measurement.

In the second sub-study, smartphone-based pure-tone audiometry and adaptive categorical loudness scaling (ACALOS) were compared to their traditional counterparts. The influence of supervision was systematically evaluated. Additionally, smartphone-based assessments of binaural and spectral loudness summation were validated.

In the third sub-study, the test-retest reliability and validity were investigated using smartphone-based pure-tone audiometry and ACALOS conducted outside a sound booth, in a home environment with controlled ambient noise. Additionally, a reinforced

adaptive categorical loudness scaling (rACALOS) method was introduced to integrate threshold measurement into the CLS procedure.

Study sample:

Numerical experiments utilizing Monte-Carlo simulations were conducted on 1,000 virtual listeners in Sub-study 1. Additionally, empirical experiments were performed on 21 participants with normal hearing and 16 participants with mild-to-moderate hearing loss in Sub-study 2. Finally, in Sub-study 3, 15 young adults with normal hearing were recruited.

Results:

Inattention significantly affected the robustness and efficiency of adaptive procedures in smartphone-based listening tests. However, when ambient noise was controlled and mobile devices were calibrated, the results of smartphone-based tests were comparable to those of laboratory-based tests. Human supervision did not affect the accuracy of the listening tests. Notably, the graded response bracketing (GRaBr) method and the reinforced adaptive categorical loudness scaling (rACALOS) outperformed baseline methods with regard to time efficiency, accuracy, and robustness against inattention in measuring audiograms and loudness growth functions.

Conclusions:

This thesis provides critical prerequisites for smartphone-based listening tests to be performed accurately and reliably without supervision, making them a cost-effective alternative to traditional clinical routine tests.

Zusammenfassung

Motivation:

Smartphone-basierte Hörtests haben den Umfang herkömmlicher, laborbasierter Hörtests erheblich erweitert und bieten den Teilnehmern einfachen Zugang zur Bewertung ihres Hörvermögens und somit frühen Diagnosen von Hörverlust. Trotz dieser Vorteile leiden solche Tests oft unkontrollierten Umgebungen, fehlenden Testaufsichten, unkalibrierten Geräten und unaufmerksamen Teilnehmern, was zu Ungenauigkeiten und unzuverlässigen Ergebnissen führen kann.

Ziel:

Es gibt drei primäre Ziele: Erstens, die Entwicklung und Validierung von smartphone-basierten Hörtests, einschließlich Luftleitungston-Audiometrie und Kategorischer Lautheitsskalierung (CLS); zweitens, die Analyse der Auswirkungen von Faktoren wie Unaufmerksamkeit, Aufsicht und Umgebungsgeräuschen auf die Testergebnisse; und drittens, die Optimierung adaptiver Verfahren für die Implementierung auf mobilen Geräten.

Design:

In der ersten Teilstudie wurden sieben adaptive Verfahren in Kombination mit zwei Kategorien unaufmerksamer Hörer ko-simuliert. Die simulierten Hörer wurden mit drei Stufen der Unaufmerksamkeit und variierenden Falschalarmraten parametrisiert. Die Robustheit der adaptiven Verfahren wurde anhand von Bias und Root-Mean-Square-Error (RMSE) quantifiziert, während die Effizienz durch die Konvergenzrate und die normalisierte Effizienz gemessen wurde. Zudem wurde Graded Response Bracketing (GRaBr), ein modellfreies und effizientes adaptives Verfahren, das gegenüber Unaufmerksamkeit robust ist, für die Messung von Audiogrammen eingeführt.

In der zweiten Teilstudie wurden smartphone-basierte Tonaudiometrie und Adaptive Categorical Loudness Scaling (ACALOS) mit ihren traditionellen Gegenstücken verglichen. Der Einfluss von Aufsicht wurde systematisch untersucht. Darüber hinaus wurden smartphone-basierte Messungen der binauralen und spektralen

Lautheitssummation validiert.

In der dritten Teilstudie wurden die Test-Retest-Reliabilität und Validität von smartphone-basierter Reintonaudiometrie und ACALOS außerhalb einer Schallkabine, in einer häuslichen Umgebung mit kontrolliertem Umgebungsgeräusch, untersucht. Zusätzlich wurde die Methode Reinforced Adaptive Categorical Loudness Scaling (rACALOS) eingeführt, um die Schwellenwertmessung in das CLS-Verfahren zu integrieren.

Studienprobe:

Numerische Experimente unter Verwendung von Monte-Carlo-Simulationen wurden in Teilstudie 1 mit 1.000 virtuellen Probanden durchgeführt. Darüber hinaus wurden empirische Experimente mit 21 normalhörenden Teilnehmern und 16 Teilnehmern mit leichtem bis mittlerem Hörverlust in Teilstudie 2 durchgeführt. In Teilstudie 3 wurden schließlich 15 junge Erwachsene mit normalem Hörvermögen rekrutiert.

Ergebnisse:

Unaufmerksamkeit hatte einen erheblichen Einfluss auf die Robustheit und Effizienz adaptiver Verfahren in smartphone-basierten Hörtests. Wenn jedoch das Umgebungsrauschen kontrolliert und die mobilen Geräte kalibriert wurden, waren die Ergebnisse der smartphone-basierten Tests mit denen laborbasierter Tests vergleichbar. Die Anwesenheit einer Aufsichtsperson hatte keinen Einfluss auf die Genauigkeit der Hörtests. Besonders hervorzuheben ist, dass die Methoden Graded Response Bracketing (GRaBr) und Reinforced Adaptive Categorical Loudness Scaling (rACALOS) die herkömmlichen Verfahren in Bezug auf Zeiteffizienz, Genauigkeit und Robustheit gegenüber Unaufmerksamkeit bei der Messung von Audiogrammen und Lautheitswachstumsfunktionen übertrafen.

Fazit:

Smartphone-basierte Hörtests können genau und zuverlässig ohne Aufsicht durchgeführt werden und sind eine kosteneffiziente Alternative zu herkömmlichen klinischen Routinetests.

List of publications

Peer-reviewed journals

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

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

3. **Xu, C.**, Schell-Majoor, L., & Kollmeier, B. (2024). Feasibility of efficient smartphone-based threshold and loudness assessments in typical home settings. medRxiv, 2024-11.

Conferences

1. **Xu, C.**, Schell-Majoor, L., & Kollmeier, B. (2024a). Predict standard audiogram from a loudness scaling test employing unsupervised, supervised, and explainable machine learning techniques. In Proc. ”Fortschritte der Akustik - DAGA’24”, Hannover, Germany.

2. **Xu, C.**, Schell-Majoor, L., & Kollmeier, B. (2024b). Towards a robust and optimum prediction of audiometric profiles from non-audiometric features. In Audiological Research Cores in Europe (ARCHES), Leuven, Belgium.

3. **Xu, C.**, Schell-Majoor, L., & Kollmeier, B. (2023a). Development and verification of self-supervised smartphone-based methods for assessing pure-tone audiometry and loudness growth function. In 16th European Federation Audiology Societies Congress, Sibenik, Croatia.

4. **Xu, C.**, Schell-Majoor, L., & Kollmeier, B. (2023b). Smartphone-based hearing tests for a Virtual Hearing Clinic: Influence of ambient noise on the absolute

threshold and loudness scaling at home. In Virtual Conference on Computational Audiology VCCA June 29-30, online.

5. **Xu, C.**, Hülsmeier, D., Buhl, M., & Kollmeier, B. (2022). How Robust and Efficient Are Different Adaptive Hearing Threshold Procedures for Use With Mobile Devices. In Audiological Research Cores in Europe (ARCHES), Amsterdam, The Netherlands.

Open access datasets

1. **Xu, C.**, Hülsmeier, D., Buhl, M., & Kollmeier, B. (2023). Simulation data of the inattention paper [Data set]. Zenodo. https://doi.org/10.5281/zenodo.8077505

Abbreviations

ACALOS adaptive categorical loudness scaling

AFC alternative forced choice

ANOVA analysis of variance

APTA automated pure-tone audiometry

B&K Brüel&Kjaer

BTUX fitting method for loudness function in ACALOS

CLS categorical loudness scaling

CU categorical units

FC fully-concentrated

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

MC moderately-concentrated

MEL-Med maximum expected information-maximum likelihood

MEL-ML maximum expected information-median

MIQR mean interquartile range

MLP	maximum likelihood procedure
MPANLs	maximum permissible ambient noise levels
MSD	mean signed difference
NCE	noise reduction earphones
NC	non-concentrated
NE	normalized efficiency
NH	normal hearing
nIFC	n-interval forced-choice
PF	psychometric function
PTA	pure-tone average
qCLS	quick categorical loudness scaling
rACALOS	reinforced adaptive categorical loudness scaling
RMSE	root mean square error
SA	slope-adaptive procedure
SIAM	single interval adjustment matrix
SIUD	single interval up and down
SPL	sound pressure level
UML	updated maximum likelihood

Table of Contents

1 General introduction

Hearing loss currently affects over 1.5 billion people worldwide (Kushalnagar, 2019). According to the World Health Organization (WHO), 430 million individuals, including 34 million children, experience significant hearing problems. WHO projects that by 2050, 2.5 billion people will have hearing loss, with at least 700 million requiring assistive devices. The rapid advancement of mobile technology offers potential for low-cost assistive solutions, which may help raise awareness of hearing loss and encourage the adoption of hearing devices.

Recently, mobile devices such as smartphones, tablets, and wearables have become widely available to the majority of the global population. Utilizing these devices to conduct hearing research (e.g., detecting hearing loss, providing remote hearing aid fitting, monitoring everyday hearing status, evaluating hearing aid fitting in daily lives) has attracted considerable interest (Kollmeier et al., 2023; Mok et al., 2023; Almufarrij et al., 2022). Generally, these mobile devices offer easy access to hearing tests. Additionally, smartphone-based hearing assessments can provide preliminary diagnostics at an early stage, potentially motivating the usage of hearing devices for hearing loss intervention in case of a negative test result. These tests can be completed quickly and independently by participants, and the results from mobile devices can be accurate and comparable to those collected in clinics or laboratories (Swanepoel et al., 2010; 2014; 2015; Almufarrij et al., 2022; Xu et al., 2024b). Overall, substantial evidence suggests that applying mobile devices for hearing examinations is beneficial (Guo et al., 2021; Almufarrij et al., 2022; Zhao et al., 2022; Kollmeier et al., 2023).

Despite these benefits, there are concerns about directly deploying hearing tests developed in laboratories onto mobile devices. Laboratory auditory experiments typically assume that participants are fully concentrated and under the complete supervision of experimenters (i.e., participants are well-trained and fully understand the content of the experiments). These experiments are usually performed in sound-attenuated booths to eliminate environmental noise, ensuring high 'auditory hygiene' (Zhao et al., 2022) or the availability of unlimited experimental/environmental and cognitive resources (Kollmeier et al., 2023). In contrast, smartphone-based auditory measurements are highly uncontrollable. Participants may easily loose attention and

often need to complete the hearing tests on their own. Moreover, ambient noise is commonly present in out-of-booth measurements (e.g., at participants' homes or in clinics). Additionally, mobile devices and headphones are often uncalibrated, which can result in test stimuli not being precisely presented as intended. Consequently, the measurement accuracy may decrease compared to controlled in-lab measurements (Peng et al., 2020; 2022). Hence, a systematic investigation is required to determine whether and how the constraints of environmental and cognitive resources during smartphone-based listening tests affect the validity, test-retest reliability, and efficiency of the measurements.

In this thesis, several major factors were therefore investigated that might impact the accuracy of smartphone-based hearing tests (i.e., **inattention** as a cognitive factor, **supervision** as a psychological factor, **ambient noise** as environmental factor). These factors are comprehensively and systematically studied from different aspects. An additional aim is to validate the diagnostic modules for the browser-based app Virtual Hearing Clinic (VHC) and to report the test-retest reliability. Finally, the efficiency of smartphone-based listening tests will be compared.

- Influence of inattention

Challenges: In smartphone-based listening tests, participants may be more prone to distractions compared to standard in-laboratory measurements. Potential sources of distraction include incoming messages, emails, and background noise. Additionally, certain participant groups, such as children or individuals with neurological conditions (e.g., stroke), may be particularly susceptible to inattention. This lack of focus can adversely affect measurement outcomes, leading to reduced accuracy, reliability, and efficiency. Therefore, it is essential to account for the potential impact of inattention when designing and conducting smartphone-based auditory assessments.

Limitations: Although inattention is a critical factor influencing the outcomes of smartphone-based listening tests, many well-established adaptive tracking methods do not account for it, as they generally assume that participants are fully attentive and provide consistent responses. Consequently, directly applying these traditional adaptive approaches to assess listening abilities on mobile devices may be inappropriate. To

address this limitation, adaptive procedures should be optimized and refined to mitigate the effects of inattention. Specifically, methods that are robust against inattention and efficient should be prioritized for smartphone-based auditory assessments.

Backgrounds:

A. **Inattention model:** In psychophysics, participants' behavior is typically modeled using an S-shaped logistic psychometric function (Brand & Kollmeier, 2002). This function describes the relationship between physical stimuli (e.g., sound level) and participant responses (e.g., if a certain signal has been perceived or not). A logistic psychometric function has four free parameters: L_{50}, s, p_{min}, and p_{max}. These parameters represent the sound level at 50% yes response (hearing threshold level), the slope, and the lower and upper asymptotes, respectively. Green (1995) introduced the concept of the inattentive observer by incorporating a lapse rate parameter (p_{max}) to account for attentional variability. In Green's model, a listener with $p_{max} = 1$ is considered fully attentive and perfectly concentrated, whereas a listener with $p_{max} < 1$ is deemed inattentive. This inattentive listener model has been widely recognized and adopted in subsequent studies (e.g., Rinderknecht et al., 2018; Manning et al., 2018) to evaluate the robustness of adaptive auditory procedures. Rinderknecht et al. (2018) later characterized this modeling approach as 'sustained inattention'.
B. **Adaptive procedure:** Previously, the clinical approach based on the Hughson-Westlake procedure (Hughson et al., 1944) was regarded as the 'golden' standard for determining the audiogram. Alternative procedures employing maximum likelihood estimation or Bayesian principles have also demonstrated high efficiency and reliability in threshold measurement (e.g., Green, 1993; Shen & Richards, 2012; Watson, 2017). Moreover, Kaernbach (1990) proposed the single-interval adjustment-matrix (SIAM) procedure, an adaptive and efficient method for measuring thresholds using a simple yes-no task. Additionally, a model-free, single-interval up-and-down approach introduced by Lecluyse and Meddis (2009) has shown accuracy in estimating thresholds for both normal-hearing and hearing-impaired listeners. However, as mentioned above, inattention remains a persistent challenge in mobile auditory testing environments. The optimal adaptive tracking method to mitigate the impact of inattention in such contexts is yet to be identified.

Therefore, the objective of this study is to determine the most suitable adaptive procedure for addressing inattention during testing.

C. **Simulations:** Monte Carlo simulations are commonly used to compare different adaptive procedures across various simulated inattentive listeners, as demonstrated in previous studies (e.g., Shen & Richards, 2012). These simulations randomly generate a set of events, such as simulated adaptive tracks, and estimate parameters associated with these events, such as the root-mean-square error (RMSE) between estimated and target thresholds. The RMSE provides insight into the robustness of adaptive procedures, while efficiency can be assessed through the rate of convergence—quantified as the standard error of the estimated thresholds as a function of the number of trials (see Kollmeier et al., 1988; Brand & Kollmeier, 2002, for details). When the parameters of the simulated listeners (e.g., thresholds, slopes) align with those of actual participants, the simulated results can be expected to approximate experimental outcomes.

- Influence of supervision

Challenges: Proper supervision is essential for both psychoacoustic experiments conducted in laboratory settings and conventional clinical audiometric tests. In speech-in-noise tests, for instance, experimenters must train participants to ensure they understand the procedures and monitor their performance. Data may be discarded if participants perform suboptimally (Leek et al., 2000). Previous studies have shown significant differences in test performance between naive (untrained) and experienced (trained) listeners (Gu & Green, 1994).

In clinical settings, audiometric tests, such as audiograms, require the expertise of audiologists or otolaryngologists. These specialists provide critical support by guiding patients, demonstrating the use of medical devices (e.g., audiometers), and ensuring the quality of measurement data. Unreliable or invalid responses are often discarded, necessitating retesting. The role of these professionals is indispensable for obtaining accurate and reliable results in both research and clinical contexts.

Smartphone-based hearing assessments are often conducted by participants independently, a condition referred to as "non-supervision." However, supervision is a critical factor in ensuring the accuracy and reliability of measurement procedures. The

absence of supervision in such assessments may negatively affect test performance, yet this issue has received limited attention. Most previous studies validating smartphone-based hearing tests (e.g., Swanepoel et al., 2010) have not explicitly examined the role of supervision. Thus, investigating the effect of supervision on the performance of smartphone-based listening assessments is essential.

Limitations: Two significant limitations have been identified. First, most current smartphone-based applications predominantly adopt clinical adaptive procedures to assess audiograms. However, as noted by Lecluyse and Meddis (2009), clinical adaptive procedures may yield inaccurate threshold estimations due to factors such as inattention as introduced before. This highlights the need for non-clinical, self-paced adaptive procedures that offer greater precision.

Second, earlier studies have primarily focused on the feasibility of smartphone-based assessments for audiograms and speech-in-noise or speech-in-quiet tests. In contrast, few studies have investigated the validity of categorical loudness scaling (CLS) tests on mobile devices, despite their critical role in diagnosing hearing disorders such as tinnitus and hyperacusis (e.g., Erinc et al., 2022; Hébert et al., 2013) and in the fitting of hearing devices (Kollmeier & Hohmann, 1995; Kollmeier & Kießling, 2018; Oetting et al., 2018). Additionally, studies exploring binaural and spectral loudness summation using smartphones remain scarce, even though these processes are vital for optimizing hearing aid fittings (Oetting et al., 2018). Consequently, the development and validation of smartphone-based tests for both CLS and binaural and spectral loudness summation are essential. One primary objective of this thesis therefore is to validate smartphone-based audiometric and CLS tests. This creates a research gap regarding whether supervised in-lab hearing tests yield similar results when applied to mobile devices, which usually lack supervision.

Backgrounds:

A. **Human supervision:** Previous studies (e.g., Swanepoel et al., 2010; Colsman et al., 2020) have investigated and compared two levels of supervision in listening tests: fully supervised and non-supervised conditions. In fully supervised conditions, typical of traditional listening assessments, an audiologist is present to administer the tests for participants. In contrast, non-supervised conditions lack audiologist

involvement, requiring participants to conduct all aspects of the experiments independently. However, to the best of our knowledge, the intermediate semi-supervised condition—where participants perform the tests independently but have access to a supervisor for questions, without the supervisor accessing the log data—has received little attention in the literature.

B. **Smartphone-based pure-tone audiometry:** Several studies have demonstrated the validity and reliability of smartphone-based audiometry compared to standard audiometry. For instance, prior research (e.g., Swanepoel et al., 2014; Yousuf Hussein et al., 2016; Van Tonder et al., 2017) has shown that smartphone-based audiometric results are consistent with standard audiograms, with mean threshold differences typically less than 5 dB across the standard 11 audiometric frequencies (Thai-Van et al., 2023). Furthermore, test-retest reliability is high, as most mean differences between repeated measurements on smartphones are also below 5 dB (Hazan et al., 2022). Please note that the stimuli employed in these studies are normally pure tones with a fixed duration of 1 s at frequencies ranging from 0.125 kHz to 8 kHz. These findings are robust across participants with normal hearing and those with hearing impairments, as well as across various age groups. Additionally, smartphone-based audiometry is notably time-efficient, typically requiring less than 10 minutes to assess both ears (Swanepoel et al., 2014). Given the use of calibrated devices, including headphones and smartphones, these studies underscore that smartphone-based audiometry aligns closely with standard audiometric practices.

C. **Smartphone-based CLS test:** The Categorical Loudness Scaling (CLS) test is a supra-threshold auditory assessment used to evaluate loudness perception. Participants rate the loudness of sounds on an 11-point scale, which includes labeled categories such as "very soft," "soft," "medium," "loud," and "very loud," along with four unnamed intermediate levels and two boundary categories: "not heard" and "too loud." The Adaptive Categorical Loudness Scaling (ACALOS) method, introduced by Brand and Hohmann (2002), was standardized in ISO 16832 (2006). To date, no study has conducted the CLS test using smartphones. The only existing study evaluating CLS in a remote setting was conducted on laptops (Kopun et al., 2022). Regarding validity, their findings showed no significant difference

between remote laptop-based CLS tests and standard laboratory-based CLS tests in a sample of five normal-hearing participants, indicating good validity for the laptop-based approach. However, across-run biases in the remote setting were larger compared to the laboratory setting in a group of 21 adult participants, suggesting that the reliability of the laptop-based CLS test is lower and could benefit from future improvements. The stimuli used in Kopun et al. (2022) were pure tones that were 1,000 ms in duration with 20-ms rise/fall times at 1 and 4 kHz. ACALOS typically employs low-noise narrowband noises as measurement stimuli due to their natural modulation properties, lack of influence on the fine structure of the absolute threshold, and other advantageous characteristics.

- Influence of ambient noise

Challenges:

In psychoacoustics laboratories, sound-proof booths eliminate environmental noise, ensuring accurate measurements. However, smartphone-based hearing tests, often conducted at home, lack such facilities, making ambient noise a potential factor affecting test accuracy. As highlighted by Margolis et al. (2022), ambient noise can cause direct masking and distraction, negatively impacting results. Therefore, controlling ambient noise is crucial for out-of-booth measurements.

Limitations:

There are three main limitations to consider:

First, while ambient noise is a critical factor influencing the results of smartphone-based listening tests, few studies have addressed or measured its impact during smartphone-based ACALOS assessments conducted outside sound-proof booths. In contrast, several studies on smartphone-based audiograms have controlled for ambient noise (e.g., Swanepoel et al., 2015; Storey et al., 2014; Brennan-Jones et al., 2016). Additionally, most studies measure environmental noise in non-clinical or ‘natural’ settings, yet typical home environments remain underexplored.

Second, as noted by Almufarrij et al. (2022), only 12% of mobile hearing assessment applications available in app stores are validated through peer-reviewed

publications, leaving the validity and test-retest reliability of most apps unverified. Therefore, it is essential to examine the validity and reliability of the developed app in this study and compare its performance with findings from previous research.

Third, the ACALOS method for assessing loudness growth functions requires adaptation for mobile testing. Ambient noise may interfere more significantly with loudness perception at the hearing threshold level (HTL) compared to supra-threshold levels (e.g., loudness discomfort levels). Furthermore, the HTL estimated by ACALOS shows low consistency with audiometric thresholds obtained via traditional audiograms, with a correlation coefficient of only about 0.25 (Kinkel, 2007). Consequently, updating the original ACALOS method is necessary to enhance its suitability for mobile testing, rather than applying it directly.

Backgrounds:

A. **Ambient noise monitoring**: It is generally feasible to perform pure-tone audiometry outside a sound-proof booth, provided the ambient noise level remains below recommended thresholds, such as the maximum permissible ambient noise levels (MPANLs). If the noise level does not exceed these thresholds, the testing environment is deemed suitable, and the audiometric results are expected to be as accurate as standard in-lab measurements. However, no specific standards currently exist for ACALOS measurements. A study by Kopun et al. (2022) suggests that environments with ambient noise levels below 50 dB(A) may be appropriate for remote ACALOS assessments. Ambient noise can be measured in mobile testing settings using smartphone applications (e.g., Decibel X by SkyPaw Co., Ltd) along with integrated or external smartphone microphones. Mobile devices must be precisely calibrated to accurately measure ambient noise. For certain well-known smartphone models, particularly iOS devices, calibration may not be necessary due to minimal hardware variation across devices of the same type. In contrast, for Android smartphones, accurate measurements require calibration, as the hardware variations are larger. This necessitates either device-specific calibration or the inclusion of estimated calibration factors. Therefore, if noise levels are controlled, the results of both listening tests are expected to align closely with standard assessments.

B. **Metrics for quantifying validity and reliability**: The validity of smartphone-based audiogram and ACALOS tests can be evaluated against standard measurements using Bland-Altman plots, as demonstrated in previous studies (Fultz et al., 2020; Giavarina, 2015). The reliability of smartphone-based audiograms can be assessed through test-retest comparisons using intraclass correlation coefficients (ICCs), following the criteria outlined by Buhl et al. (2022): poor (ICC < 0.5), moderate (ICC ≥ 0.5), good (ICC ≥ 0.75), and excellent (ICC ≥ 0.9). For ACALOS, reliability can be measured using the mean interquartile range (MIQR) and mean signed difference (MSD), as proposed by Kopun et al. (2022). Lower values of MIQR and MSD indicate higher reliability for the approach.

Taken together, based on the review of the available literature, as well as on the limitations outlined, the first objective of this thesis is to explore the validity, test-retest reliability, and efficiency of smartphone-based listening tests, given the unclear performance of these tests on mobile devices. Second, since the selection of appropriate listening tests for mobile applications remains uncertain, the second objective is to identify the optimal and minimal set of such tests. Furthermore, three key factors—namely inattention, supervision, and ambient noise—will be examined in the main body of this thesis. The detailed outline of this thesis is presented below.

- Outline of the thesis

Chapter 1 introduces the general background of the study, emphasizing the potential benefits of modern, advanced mobile device technology in hearing loss prevention, diagnostics, and rehabilitation. These benefits include improved accessibility to listening tests and early diagnostics for hearing disorders. However, several challenges remain: first, the validity, reliability, and efficiency of various smartphone-based listening tests are not yet well understood; second, the criteria for selecting appropriate listening tests for mobile platforms are unclear. Chapter 1 also presents an integrated literature review, detailing previous attempts to address these challenges and identifying the limitations of earlier studies. Based on this review, three critical factors—inattention, supervision, and ambient noise—are identified and investigated in the subsequent chapters (Chapters 2-4), which form the main body of the thesis.

Chapter 2 explores the impact of inattention on the performance of adaptive procedures used in smartphone-based audiogram measurements. Monte Carlo simulations are employed to evaluate the robustness and efficiency of various adaptive procedures, both model-based and model-free, under conditions simulating inattentive listener behavior. The chapter concludes by recommending an optimal adaptive procedure for smartphone-based audiograms, with detailed justifications for its selection.

Chapter 3 investigates the role of supervision in listening tests, comparing three supervision modes. Both smartphone-based audiogram and adaptive categorical scaling (ACALOS) tests are developed and validated against standard laboratory-based tests with both normal-hearing and hearing-impaired participants across three different frequencies. The feasibility of smartphone-based binaural and spectral loudness summation is also examined.

Chapter 4 focuses on the influence of ambient noise on listening tests conducted using mobile devices. Experiments are designed to monitor and control ambient noise in home environments, extending the scope of Chapter 3's experiments outside sound-proof booths. This chapter reports the validity and reliability of smartphone-based audiogram and ACALOS tests under noisy conditions. Additionally, it evaluates the novel adaptive procedure GRaBr, introduced in Chapter 2, for audiogram measurements and the reinforced ACALOS (rACALOS) procedure, proposed in the current chapter, for CLS measurements.

Chapter 5 provides a comprehensive discussion of the findings. It addresses the two primary research questions posed in Chapter 1: (1) the validity, test-retest reliability, and efficiency of smartphone-based listening tests, and (2) the selection of appropriate listening tests for mobile platforms. The chapter also explores potential applications of smartphone-based listening tests, including auditory profile determination and the establishment of a national hearing health cohort. Finally, the limitations of the study and future research directions are discussed.

In summary, Chapters 1 and 5 provide the general introduction and discussion, respectively, while Chapters 2-4 investigate the key factors (inattention, supervision, and ambient noise) affecting smartphone-based listening tests. Chapter 2 focuses on the efficiency of adaptive procedures for smartphone-based audiograms. Chapter 3

evaluates the validity of both smartphone-based audiogram and ACALOS tests in sound-proof environments, while Chapter 4 examines their validity and reliability in home environments. Furthermore, Chapter 4 compares the optimized adaptive procedures proposed in Chapters 2 and 4, utilizing human participants. Even though these chapters address the key prerequisites for developing a Virtual Hearing Clinic for widespread application on smartphones and demonstrate its initial implementation on real systems, significant progress is still required to achieve the ultimate goal of this thesis: developing a system of easily accessible, reliable, and valid hearing tests as part of the broader objective, Hearing4All.

2 Influence of inattention[1]

Abstract

Inattention plays a critical role in the accuracy of threshold measurements, e.g., when using mobile devices. To describe the influence of distraction, long- and short-term inattention models based on either a stationary or a non-stationary psychometric function were developed and used to generate three simulated listeners: fully-, moderately-, and non-concentrated listeners. Six established adaptive procedures were assessed via Monte-Carlo simulations in combination with the inattention models and compared with a newly proposed method: the graded response bracketing procedure (GRaBr). Robustness was examined by bias and root mean square error between the 'true' and estimated thresholds while efficiency was evaluated using rates of convergence and a normalized efficiency index. The findings show that inattention has a detrimental impact on adaptive procedure performance—especially for the short-term inattentive listener—and that several model-based procedures relying on a consistent response behavior of the listener are prone to errors owing to inattention. The model-free procedure GRaBr, on the other hand, is considerably robust and efficient in spite of the (assumed) inattention. As a result, adaptive techniques with desired properties (i.e., high robustness and efficiency) as revealed in our simulations—such as GRaBr—appear to be advantageous for mobile devices or in laboratory tests with untrained subjects.

Keywords: inattention model; mobile listening test; model-free adaptive procedure; Monte-Carlo simulations

[1] This section is a formatted reprint of

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

2.1 Introduction

Measuring sensory thresholds is one of the fundamental topics in psychophysics and central for hearing assessment, e.g., in hearing screening, in characterizing auditory functions, or in rehabilitative audiology. There are many methods established to obtain threshold measurements efficiently, including various adaptive procedures that steer the stimulus level according to the previous responses of the participant (Treutwein, 1995; Leek, 2001). Psychophysical procedures that are used to calculate sensory thresholds typically rely on participants to be attentive so that they can produce consistent responses. However, Green (1995), observed that participants can be inattentive and produce responses that are unrelated to the stimulus (Wichmann & Hill, 2001), and modeled sustained inattention by adjusting the lapse rate of the psychometric function. Green's (1955) stationary inattention model (herein referred to as "long-term inattention") has been widely adopted by many other studies to evaluate the robustness of the adaptive procedures against inattention (e.g., Shen & Richards, 2012; Rinderknecht et al., 2018; Manning et al., 2018). However, with the recent advent of remote, self-driven, and even smartphone-based hearing testing, a completely different setting of threshold measurements comes into play (e.g., Bisitz & Silzle, 2011; Ooster et al., 2020; Luengen et al., 2021). Such occasional inattention—termed 'short-term inattention'—requires a different, non-stationary attention model, where the individual state of attention is randomly drawn to subsequently determine the respective response probability. This differs from sustained inattention, which is modeled using a fixed, stationary probability.

The study aims to unravel how this type of assumed short-term inattention influences the result of the various hearing threshold measurement procedures in contrast to long-term inattention behavior known from the literature (Green, 1995; Rinderknecht et al., 2018). The second aim is to quantify, normalize, and eventually optimize the robustness and efficiency of the adaptive procedures to be used for smartphone measurements in the future. This is a prerequisite for our research question: Do adaptive procedures differ in their robustness against both types of inattention and how do these differences affect their efficiency?

To measure auditory thresholds efficiently, Kaernbach (1990) proposed the single interval adjustment matrix (SIAM) approach based on a simple yes-no task for testing. SIAM was validated by Shepherd et al. (2011) using auditory stimuli for its ability to measure absolute threshold fast, reliably, and accurately for human participants. The SIAM procedure utilizes the outcome of the signal detection matrix (i.e., hit, miss, false alarm, and correct rejection) to adjust the sound level in an adaptive manner. Green (1990; 1993; 1995) and Gu and Green (1994) introduced a single interval adaptive approach employing maximum likelihood procedure (MLP). The MLP procedure consists of two steps: maximum likelihood estimation and stimulus selection. In the maximum likelihood estimation, different psychometric functions are proposed as hypotheses. Then the likelihood of each hypothesis is calculated and the function with the highest likelihood is selected to obtain the level of the next trial from the inverse function at the p-target, i.e., the threshold level that corresponds to the (target) probability p at the estimated psychometric function (e.g., 50% for a yes-no task, and 75% for the two-alternative forced choice (2AFC) task, see Grassi & Soranzo, 2009, respectively, as typical examples from a certain range of values). The MLP method appears to be rather efficient and was validated with human subjects by Amitay et al. (2006) and Leek et al. (2000). However, Green (1995) found that the MLP yielded a poor estimate of thresholds if participants were inattentive. This "unforgiving" property of model-based or parametric methods results from the fact that the whole track history influences the respective next level placement. This motivated the introduction of hybrid methods (e.g., Hall, 1981) where an adaptive, non-parametric level placement procedure with a shorter memory is combined with a maximum likelihood (ML) method for the final threshold estimate. More recently, Shen and Richards (2012) optimized the original MLP method and designed an updated maximum likelihood (UML) procedure, aiming at improving the low accuracy of threshold estimates resulting from lapses in attention. In the UML procedure, the stimulus selection process takes the interim estimate of lapse rate into account.

The adaptive method parameter estimation by sequential testing (PEST) is among the first non-parametric adaptive psychophysical testing methods (Taylor & Creelman, 1967; Gescheider, 2013). The PEST method compares the respective correct response rate with the target probability and determines the level of the subsequent stimulus by interpolation using a series of diminishing step sizes. The same long-term memory

problem as with the MLP methods exists for Bayesian adaptive procedures that build upon the PEST method such as best PEST (Pentland, 1980), QUEST (Watson & Pelli, 1983), and the state-of-the-art QUEST+ approach (Watson, 2017). QUEST+ utilizes the minimum entropy principle to select the respective next stimulus level and maximum likelihood theory to estimate the final values of the parameters. Specifically, QUEST+ searches for the most informative stimulus by minimizing the entropy of the posterior probability density. When taking the interim estimate of lapse rate during the stimulus selection process into consideration, these methods including UML are not as "unforgiving" as those discussed before and, hence, achieve accurate and efficient threshold estimation inside the lab (Watson, 2017).

One problem of the single-interval Yes/No procedure (e.g., MLP in Gu & Green, 1994)—when used in combination with the adaptive rules discussed so far—is the need to control (or at least to detect) the individual detection criterion as described by signal detection theory (Green & Swets, 1966). This is usually done by inserting sham trials (also referred to as 'catch trials' in psychology), i.e., trials that do not contain a signal to estimate the false alarm rate concurrently with the correct detection rate. Alternatively, n-interval forced-choice (nIFC) methods are used where only one randomly selected interval contains the target signal and the other (n-1) intervals the reference. However, the necessity of these additional blank intervals (as well as sham trials) increases the measurement time and thus reduces the efficiency of the procedure for estimating thresholds. Furthermore, naïve subjects might get frustrated if they do not perceive the intended signal frequently, as pointed out by Lecluyse and Meddis (2009), and tend to loose the cue for a stable detection. This calls for a minimum of sham trials or blank intervals and for providing suprathreshold stimulus levels not too rarely during a track.

To accommodate both requests, Lecluyse and Meddis (2009) and Meddis and Lecluyse (2011) recommend the single interval up and down (SIUD) procedure for tone detection. The SIUD involves presenting two tones—the probe tone and an additional cue tone with a fixed level increase of 10 dB—while participants indicate how many tones they have heard (0, 1, or 2 tones). The responses are used to track the threshold of the probe tone and to detect false alarms recorded in (rare) sham trials where the cue tone is absent, leading to an abortion of the track. Although human experiments suggest that the SIUD procedure is accurate and efficient for threshold measurement, there is no

systematic assessment of the influence of inattention on the robustness and efficiency of this procedure. Also, the presentation of the cue tone requires a significant amount of measurement time. Additionally, as the level of the cue tone exceeds the probe tone level by a fixed amount of 10 dB, this difference might be appropriate for the initial portion of the adaptive track, but too large for the final portion to be helpful and informative for the determination of the threshold as the cue tones are always audible in the final portion of the adaptive track.

In the SIUD procedure, the information about the audibility of the cue tone with the (much) higher level is discarded in regular, non-sham trials. This has no negative effect on the outcome of the threshold estimation process since the cue tone audibility information relates to the saturation region of the psychometric function at approx. 100% which does not decrease the uncertainty about the threshold level. However, discarding the information to be gained from the cue tones in most trials (i.e., nearly 50% of the stimuli presented) could result in a poorer efficiency of the procedure in terms of a decrease in measurement uncertainty per unit of measurement time spent.

We therefore suggest a smaller, adaptively adjusted difference between the probe and cue tone in order to "bracket" the threshold and to exploit the detectability of the cue tone by the tracking procedure as well. This is expected to increase the efficiency of the procedure as more cue tones are presented near the threshold level. Hence, based on the SIUD procedure proposed by Lecluyse and Meddis (2009), we suggest the Graded Response Bracketing procedure (GRaBr), and compare the GRaBr procedure with the procedures discussed so far.

Although adaptive, response-criterion-compensating procedures (i.e., SIAM, MLP, UML, QUEST+, SIUD, and GRaBr) are advantageous in laboratory measurements and efficient for achieving a certain level of accuracy, sometimes they are not assigned importance in clinical practice for pure-tone threshold estimation (Lecluyse & Meddis, 2009). Instead, practitioners in audiology or otolaryngology make a compromise between speed and simplicity vs. accuracy. They primarily use manual methods for pure-tone threshold estimation such as the Hughson–Westlake procedure because of its simple administration, little patient training, and easy implementation (Hughson et al., 1944). Bisitz and Silzle (2011) reported a self-administered threshold measurement

approach, i.e., the APTA procedure, which is based on the Hughson-Westlake method, whose modified versions are widely applied in clinical audiogram measurements (Hughson et al., 1944; Guo et al., 2021). It is an ascending method (i.e., the procedure typically starts with an inaudible sound level and gradually increases the sound level until the participants indicate they can hear the tone) to assess the listener's hearing threshold, where the listener's task is to indicate whether a tone is heard or not.

Taken together, a number of well-motivated, established procedures for adaptively testing thresholds exist for laboratory use, but a consistent comparison with respect to their efficiency and robustness against long- and short-term inattention is still missing. This, however, is an important prerequisite for selecting the most appropriate procedure for mobile testing. To address this gap, we performed Monte Carlo simulations for the adaptive procedures listed above, i.e., APTA, GRaBr, MLP, QUEST+, SIAM, SIUD, and UML, while systematically varying the two modes of inattention. The performance of all adaptive procedures was evaluated in terms of the robustness of the (simulated) observable against inattention, as well as their normalized efficiency which accounts for the "effective" time used to derive a threshold estimate including any sham trials or aborted tracks.

2.2 Methods

2.2.1 Inattention model

The behavior of a virtual listener is typically modeled with a stationary psychometric function, i.e., the relationship between stimulus intensity (e.g., sound level) and a test subject's response (e.g., the proportion of 'yes' responses). The psychometric function was formulated as a four-parameter transformed logistic function by Brand and Kollmeier (2002) and Green et al. (1966):

$$p(L, \varphi) = p_{min} + (p_{max} - p_{min})/(1 + e^{-4s(L-L_{50})}) \quad (2.1)$$

where p is the probability of 'yes' responses, L defines the sound level, and φ describes the parameter vector. p_{min} indicates the lower boundary of the function (also referred to as the false alarm rate for a yes/no paradigm). p_{max} denotes the upper asymptote of the function (the miss rate is calculated by 1-p_{max}). s is the slope of the

function at the half-way point. L_{50} describes the threshold at the half-way point between the minimum and maximum of the psychometric function. The default parameter vector φ was set as (p_{min} = 0, p_{max} = 1, s = 0.125, L_{50} = 15), where the target threshold was 15 dB.

As illustrated in Fig. 2.1(A), we define the inattentive listener distracted by internal noise as a "long-term inattentive listener" exhibiting sustained inattention. This characterization assumes a stationary probabilistic process that persists throughout the entire measurement track. Conversely, in Fig. 2.1(B), we describe the inattentive listener as a "short-term inattentive listener" with occasional lapses in attention, where we assume a non-stationary behavior characterized by sporadic bursts of inattention during the measurement track. Note that—averaged across a whole measurement track—the average effect of the short-term inattention model would be reflected in a deformation of the average psychometric function (i.e., an increase of the lower asymptote and decrease of the upper asymptote for a Yes/No task) which resembles the psychometric function already employed by the long-term inattentive listener model. However, the interaction between the adaptive procedure and a non-stationary psychometric function (modeled here as a nested random process) would not adequately be covered by the long-term inattentive listener model which uses the same psychometric function in all (simulated) trials. Moreover, the long-term inattention model assumes that the likelihood of a 'yes' response under inattention p('yes'|inattention) = p_{min} whereas the short-term inattention model allows p('yes'|inattention) to be specified as an arbitrary probability. For participants using smartphones to perform listening tests, it is likely that inattentiveness causes to respond with 'yes' (p('yes'|inattention) = 1) or 'no' (p('yes'|inattention) = 0) with equal probability. Therefore, in this scenario, p('yes'|inattention) may differ from the value of p_{min}, provided by the long-term inattention model, and instead be closer to 0.5.

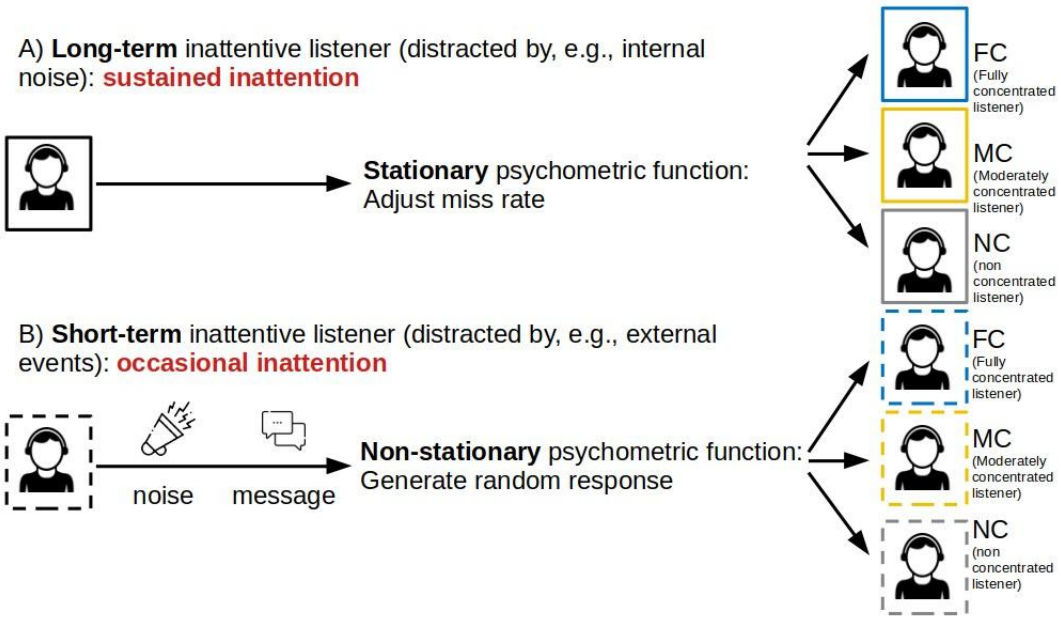

Fig. 2.1. Inattentive listener simulation. A) long-term inattentive listener (solid boxes), B) short-term inattentive listener (dashed boxes). FC: Fully concentrated listener (blue), MC: Moderately concentrated listener (yellow), NC: Non concentrated listener (grey). The parameter p_{max} of the long-term FC, MC, and NC listeners is set to 1, 0.95, and 0.9 respectively. The short-term FC, MC, and NC listeners respond randomly in 0%, 10%, and 20% of trials (corresponding to the parameter p_{inatt}), respectively.

The psychometric functions (PF) generated for both types of inattentive listeners are illustrated in Fig. 2.2 (A) for the long-term and (B) for the short-term inattentive listener. The variations in the rate of inattention for both types of inattentive listeners are termed non-, moderately-, and fully-concentrated listeners (abbreviated as NC, MC, and FC listeners, respectively). The FC listener as a reference group is identical for both types of inattentive listeners. In case of long-term inattention, following Green (1995), we vary the inattention by setting the upper asymptote p_{max} of the regular PF from 0.9 to 1.0 with a spacing of 0.05 (i.e., 0.9, 0.95, and 1.0). For short-term inattention, a regular PF is assumed in most trials, while in up to 20% of all trials (i.e., p_{inatt} = 0, 0.1 or 0.2), a random response behavior is foreseen, i.e., a constant PF with p('yes'|inattention) = 0.5. This value of p('yes'|inattention) = 0.5 is chosen as the most likely value if no information about the listener is available. It also illustrates the potential impact of scenarios where p('yes'|inattention) is higher than any of the p_{min} values implemented in the current study.

For both types of inattention, we also vary the lower asymptote p_{min} between 0 and 0.1 with a 0.05 step size (i.e., 0, 0.05, and 0.1) to account for different false alarm rates (in case of a yes/no paradigm), as shown in Fig. 2.2. A total of 18 simulated listeners are established: 3 levels of inattention (FC, MC, and NC listeners) * 3 levels of false alarm rate (0, 0.05, and 0.1) * 2 types of inattention (long-term and short-term). Note that a symmetric shape of the PF (i.e., p_{max} = 1- p_{min}) is assumed in 6 of these simulated listeners whereas the more general case of an asymmetric PF is assumed for 12 of them.

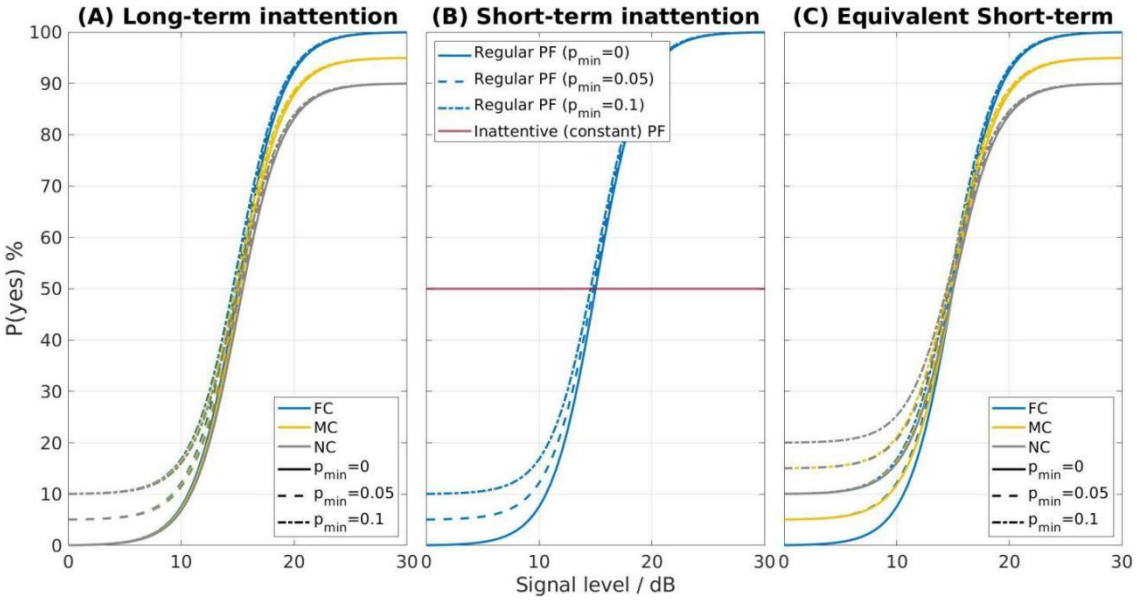

Fig. 2.2. Observer model for (A) long-term inattention (B) short-term inattention (C) equivalent short-term inattention. In the long-term inattention model, p_{max} is set to 1, 0.95, and 0.9 for the FC, MC, and NC listener. In the short-term inattention model, a constant psychometric function (PF), i.e., p('yes'|inattention) = 0.5 is employed for up to 20% of all trials (p_{inatt} = 0, 0.1 or 0.2), denoted as the FC, MC, and NC listener, respectively, while a regular logistic PF is applied in the remaining trials. This results in the "equivalent expected PF" for short-term inattention in (C), i.e., the expectation value of the PF from the nested random process of the short-term inattention model. In addition, three levels of false alarm rate p_{min} (0, 0.05, and 0.1) are included for both long- and short-term inattention models.

The short-term inattention observer model switches randomly between a regular and a constant psychometric function. Mathematically, a nested process with two states using a conditional function is employed. In the i_{th} trial, a random decision is first made

to determine the state (i.e., a state representing decision-making based on the sensory input if i belongs to the set Q with P(i ∈ Q) = 1- p_{inatt} and a state for randomly guessing if i belongs to the complementary set $\overline{Q}$ with P(i ∈ $\overline{Q}$) = p_{inatt}). Subsequently, a random decision is made if the respective response is 'yes' or 'no', which is controlled by the respective conditional psychometric function given as:

$$p(L, \varphi, i)_{short} = \begin{cases} p_{min} + \dfrac{1 - p_{min}}{1 + e^{-4s(L - L_{50})}} & \text{if } i \in Q \\ 0.5 \text{ if } i \in \overline{Q} \end{cases} \tag{2.2}$$

The long-term equivalent of the PF from (2.2)—loosely denoted as "average" PF across a whole track—is given as the expectation value from this nested random process. We denote this expectation value of the PF effectively resulting across a whole track as the "equivalent expected PF" for short-term inattention, which is given in Fig. 2.2 (C). Note that it does not approach unity for the upper asymptote, but rather the value 1- $p_{inatt}/2$. Likewise, the lower asymptote does not approach p_{min}, but rather is increased by the value $p_{inatt}/2$:

$$\begin{aligned} p_{max,short} &= 1 - p_{inatt}/2 \\ p_{min,short} &= p_{min} + p_{inatt}/2 \end{aligned} \tag{2.3}$$

Hence, the long-term behavior of the PF for the short-term inattentive listener—expressed by the equivalent expected PF—is very similar to the psychometric function employed for the long-term inattention model (if $p_{max,short}$ is replaced by p_{max} and if $p_{min,short}$ is replaced by p_{min}). However, the trial-by-trial behavior differs considerably between the long- and short-term inattention models. The single-trial PF shape of the long-term inattention model is trial-independent as it does not vary throughout a track whereas the PF shape of the short-term inattention model varies from trial to trial as the listener randomly switches between two different states..

When comparing Fig. 2.2 (A) and (C), the upper asymptote $p_{max,short}$ of the equivalent expected PF for short-term inattention is in line with p_{max} of the long-term inattentive listener due to the respective choice of the parameter p_{inatt} for the FC, MC and NC listener. However, there is a discrepancy in the lower asymptote between p_{min}

and $p_{min,short}$. Therefore, the equivalent expected PF of the short-term inattention model can not be made equal to the long-term inattention model at the same level of inattention and false alarm rate. An equivalence can only be made if the short-term inattention model is compared to the long-term inattention model at the same level of inattention but at different levels of false alarm rate, e.g., the short-term MC listener with $p_{min} = 0$ should be compared with the long-term MC listener with $p_{min} = 0.05$.

Additionally, when the constant PF in the short-term inattention model equals the false alarm rate, the short-term inattention model would be equivalent to the long-term inattention model. This assumes that participants' decision-making during inattentive trials still follows the same sensory process as in attentive trials, i.e., all trials could be described by the well-established pure sensory process outlined in Green (1995). However, in this study, we assume during inattention a uniform distribution of responses with a constant PF (i.e., p('yes'|inattention) = 0.5) which does not necessarily equal the false alarm rate. Note that this uniform distribution in inattentive trials follows from the maximum uncertainty or minimum entropy principle and assumes that participants generally make judgments independently of stimulus intensity in the inattentive trials. Thus, the whole process is rather described as a nested model of two independent processes.

2.2.2 Adaptive procedures

The properties of seven employed adaptive procedures that will be compared in this paper are provided in Table 2.1 and exemplary tracks are visualized in Fig. 2.3. The target threshold was fixed at 15 dB. MLP, QUEST+, SIAM, and UML are model-based or parametric procedures whereas the other procedures are model-free (Audiffren & Bresciani, 2022). Typically, the parameter space (i.e., ranges of parameters L_{50}, s, p_{min}, p_{max}, as well as procedure-specific parameters) together with the stimulus space (i.e., sound level) are required to be specified beforehand for those model-based procedures. MLP, QUEST+, and UML mainly employ Bayes' rule for stimulus placement and will herein be referred to as Bayesian procedures. Only APTA is a variant of the clinical method. Most adaptive procedures (e.g., MLP, QUEST+, SIAM, UML, and APTA) utilize a yes/no task, whereas the SIUD and GRaBr utilize a variant of the standard yes/no task (i.e., counting how many tones are detected, with the three response options:

none, one, and two tones). 20% catch trials are implemented in SIUD and GRaBr while the other adaptive procedures contain no catch trials. Two intervals are presented in SIUD and GRaBr whereas the other adaptive procedures have only one interval.

The six established adaptive procedures (i.e., SIUD, APTA, QUEST+, MLP, UML, and SIAM) followed as closely as possible the respective protocols introduced by Lecluyse and Meddis (2009), Bisitz and Silzle (2011), Watson (2017), Green (1993), Shen and Richards (2012), and Kaernbach (1990). The starting level was set using the strategy described in Lecluyse and Meddis (2009) for all the procedures to ensure a fair comparison. The starting level of all methods except for APTA and QUEST+ followed a discrete uniform distribution ranging between 35 dB and 45 dB with a step size of 1 dB (11 values). The APTA procedure began at -10 dB SPL (inaudible level) while QUEST+ determined the starting level based on its own rule by averaging over the upper and lower limit of the stimulus range. Thresholds were estimated for each procedure based on the median of all sound levels (indicated with a green line in Fig. 2.3) but the trials before the first reversal were discarded for the SIAM, SIUD, and GRaBr procedures. Please note that APTA followed a specific rule to estimate the threshold (i.e., the sound level of the last 'yes' response was determined as the threshold). All procedures were run until the N = 50th trial.

Table 2.1. Summarized characteristics of the employed adaptive procedures.

Procedure	Model-based[a]	Bayesian	Clinical	Proportion of catch trails	Number of intervals on each trial	Parameter space	Stimulus space	Literature
SIUD	-	-	-	20%	2	N/A	N/A	Lecluyse & Meddis (2009)
GRaBr	-	-	-	20%	2	N/A	N/A	present
APTA	-	-	×	-	1	N/A	N/A	Bisitz & Silzle (2011)
QUEST+	×	×	-	-	1	L_{50} = [-10 50] s = 0.125 p_{min} = 0 p_{max} = 1	[-10, 50]	Watson (2017)

MLP	×	×	-	-	1	L_{50} = [-10 50] s = 0.125 p_{min} = [0 0.1] p_{max} = 1	[-10, 50]	Green (1993)
UML	×	×	-	-	1	L_{50} = [-10 50] s = 0.125 p_{min} = 0 p_{max} = 1	[-10, 50]	Shen & Richards (2012)
SIAM	×	-	-	-	1	t = 0.75	N/A	Kaernbach (1990)

a The model-based (i.e., parametric) procedures hypothesize the parameters of the psychometric function while the model-free do not.

SIUD: Lecluyse and Meddis (2009) and Meddis and Lecluyse (2011) introduced an adaptive absolute threshold measurement based on a simple variant of the yes/no task: normally two tones were presented in a trial and the listeners were required to count how many tones they heard (0,1, or 2, respectively). One of the tones (denoted as cue tone) had a fixed 10 dB higher level than the probe tone and had a 20% chance to be muted. Catch trials for adaptive procedures were first introduced by Gu and Green (1994) to observe the false alarm rate p_{min} of the psychometric function (i.e. proportion of yes responses to catch trials). The false alarms are used to calculate the 'catch out rate' which is defined as the number of false alarms divided by the total number of catch trials. Since Leek et al. (2000) validated that the catch out rates were small (around 5%), Lecluyse and Meddis (2009) suggested that instead of observing the catch out rate, the threshold task should restart when a false alarm occurs (this is referred to as an 'abortion incident'). Thus, following Lecluyse and Meddis (2009), one track restarted in our study when a false alarm happened. The cue tone levels are visualized with the dashed line in Fig. 2.3 (A). The step size was set up to 10 dB at the beginning. The sound level was set to the middle point of the previous two levels after the first 'one' tone response was recorded. Afterwards, a 2 dB step size was used.

GRaBr: The level interval between the two tones of SIUD was always fixed at 10 dB which may be an inefficient use of measurement time, especially at the end of an adaptive track (as the cue tones are always audible) where a bracketing of the target level within a smaller level interval appears feasible. Therefore, we modified the

original SIUD procedure by making the level difference changeable, as shown in Fig. 2.3 (B). If the response was 2 or 0, the levels of the two tones decreased or increased with a certain step size, respectively, and the level difference remained unchanged. However, if the response indicated 1 tone (indicating that the threshold was bracketed by both presentation levels), a reduced level interval was applied to the two tones, and both tones were concurrently adjusted in level with a given step size.

As mentioned above, the starting level of the probe tone was drawn from a discrete uniform distribution ranging between 35 dB and 45 dB with a spacing of 1 dB (11 values), while the starting level of the cue tone was 10 dB higher than the probe tone. The level difference between the cue and probe tone was halved to 5 dB when participants first reported that they only heard one tone, and then the level was reduced to 2 dB after the second time they reported that they only heard one tone. The initial step size of the GRaBr procedure for the cue tone was 8 dB, reduced to 6 dB after the first reversal, halved to 3 dB after the second reversal, and eventually to 1 dB after the 3rd reversal, which follows the recommendations by Lecluyse and Meddis (2009).

APTA: Fig. 2.3 (C) depicts one run of the APTA procedure. The level kept increasing until the first ’yes’ response was detected. The level was reset to -10 dB and kept increasing until the second ’yes’ response was given. Afterward, the level was chosen to be 5 dB lower than the level at the second ’yes’ response. A run terminated if at least 7 ‘yes’ responses were detected and the maximum level deviation of the last two ’yes’ responses was less than 3 dB. The threshold was chosen as the sound level of the last ‘yes’ response, plotted with a green line in Fig. 2.3 (C).

QUEST+: QUEST+ is a generalization of the original QUEST procedure for threshold measurement, shown in Fig. 2.3 (D) (Watson & Pelli, 1983; Watson, 2017). In this study, the yes-no task was used for the QUEST+ procedure. The parameter space and stimulus space to initialize QUEST+ were reported in Table 2.1. L_{50} was set up as a free parameter to be estimated in the range between -10 and 50 dB.

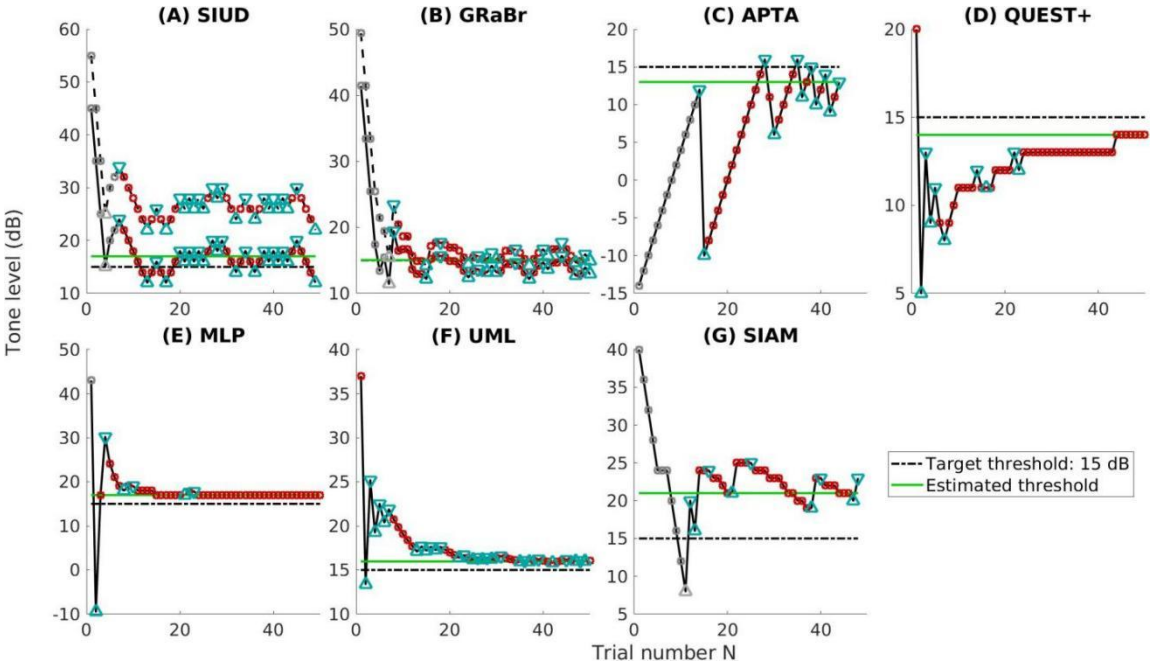

Fig. 2.3. Examples for an adaptive run, i.e., tone level as a function of the number of trials for the seven procedures considered here. A: SIUD=Single Interval Up and Down (Lecluyse & Meddis, 2009), B: GRaBr=Graded Response Bracketing, C: APTA=Automated Pure Tone Audiometry (Bisitz & Silzle, 2011), D: QUEST+ (Watson, 2017), E: MLP=Maximum Likelihood Procedure (Green, 1993), F: UML=Updated Maximum Likelihood (Shen & Richards, 2012), G: SIAM=Single Interval Adjustment Matrix (Kaernbach, 1990). Green line: threshold estimate. Triangles: trials at a reversal point. Grey circles: discarded trials for threshold estimation. Red circle: active trials. Solid line: main track for threshold estimation. Dashed line: level track of the cue tone trials. Dot-dashed line: target threshold 15 dB.

MLP: Green (1990; 1993) designed a procedure based on maximum likelihood estimation (MLP) to measure the hearing threshold in a yes-no task, shown in Fig. 2.3 (E). Following the suggestion of Grassi and Soranzo (2009), the optimal p-target (also corresponding to the sweet point in Brand and Kollmeier (2002) and Green (1995)) was adopted to be 0.6310 in the current study and p_{min} (also referred to as the false-alarm proportion in Green (1995)) varied from 0 to 0.1, with a step size of 0.05 (3 values). The range of hypothetical midpoints (i.e., L_{50}) of the psychometric function was defined as -10 dB to 50 dB (choice of stimulus space as required for the procedure, cf. Table 2.1).

UML: As is shown in Fig. 2.3 (F), the sixth adaptive procedure is UML, which is an extension of the MLP method (Shen & Richards, 2012). A 2-down 1-up sweet point selection rule was employed. The sweet point is defined as the point at the most informative level of the underlying psychometric function (Lecluyse & Meddis, 2009). Presenting the stimulus at the sweet point would minimize the variability in the threshold estimate (Shen & Richards, 2012). By the end of each track, the mean of the posterior parameter distribution (See Shen and Richards (2012) for details) was applied to estimate the L_{50}. The configurations of UML are reported in Table 2.1.

SIAM: Kaernbach (1990) described the unbiased adaptive SIAM procedure which is based on a yes-no task to measure the tone detection threshold, demonstrated in Fig. 2.3 (G). For each presentation, there was a 50% chance of containing a tone, and the participants were required to answer whether they could hear a tone or not. 75% correct tone detection (denoted as 't' in Table 2.1) was used in the SIAM procedure. The initial step size of the SIAM procedure was 4 dB, halved after the 2nd reversal, and eventually reduced to 1 dB onward after the 3rd reversal.

2.2.3 Computer simulations

All algorithms and simulations were developed in MATLAB R2021a (The MathWorks, Inc., Natick, MA) and Octave 5.2.0. The Matlab implementations of the four model-based procedures, i.e., QUEST+, MLP, UML, and SIAM were provided by Jones et al. (2018), Grassi and Soranzo (2009), Shen et al. (2015), and Schädler et al. (2020), respectively. We implemented the SIUD procedure while Hörzentrum Oldenburg gGmbH provided the Matlab toolbox for the APTA procedure.

A total of 1000 Monte-Carlo simulations were utilized as a numerical method to randomly produce a number of events (here: simulated adaptive tracks) and estimate the underlying parameters of these events (i.e., the average outcome of a given procedure, its standard deviation, and convergence rate). 1000 Monte-Carlo runs were simulated for each simulated listener and each false alarm rate. Monte-Carlo simulations have already been applied to compare different adaptive procedures in many earlier studies (e.g., Hall, 1981; Herbert et al., 2022).

2.2.4 Evaluation

Robustness

Bias: To assess the robustness of adaptive procedures, the bias (also known as the signed difference) between the threshold estimates $\widehat{L_{50}}$ and the true hearing threshold L_{50} in the k_{th} simulation is calculated in Eq. 2.4:

$$\text{Bias(k)} = \widehat{L_{50,k}} - L_{50} \tag{2.4}$$

The true threshold L_{50} is determined by the level at the center of the range for the psychometric function (Lecluyse & Meddis, 2009). A positive bias indicates that the true threshold is overestimated while a negative bias implies an underestimation of the true threshold.

Root mean square error: We calculated the root-mean-square error (RMSE) to examine the robustness of different adaptive procedures under different conditions, using the following formula:

$$\text{RMSE} = \sqrt{\sum_{k=1}^{N} \frac{(\widehat{L_{50,k}} - L_{50})^2}{N}} \tag{2.5}$$

N is the number of simulations. RMSE is always greater than or equal to zero, and a larger RMSE indicates worse performance of the procedure. To estimate the mean and standard deviation of RMSE and conduct the t and ANOVA tests on RMSE, we performed bootstrapping by drawing samples on the estimated threshold 1000 times (i.e., 1000 bootstrap replicates) with replacement, where each sample contained N = 10000 data points

Efficiency

Normalized efficiency: Taylor and Creelman (1967) proposed the sweat factor as an efficiency index. The sweat factor was widely adopted to compare adaptive procedures that had different rules and number of trials (Amitay et al., 2006; Leek, 2001; Saberi & Green, 1997; Treutwein, 1995; Audiffren & Bresciani, 2022). The empirical sweat factor SF_{emp} is defined as the product of the number of trials (both catch trials and

intervals are not included in determining the number of trials N) and the variance of the threshold estimate derived from these trials, expressed by the formula:

$$SF_{emp} = N\sigma_{L_{50}}^2 \quad (2.6)$$

N denotes the number of trials and $\sigma_{L_{50}}$ the standard deviation of the threshold estimate (how $\sigma_{L_{50}}$ is calculated is given in Eq. 2.8). We normalized the SF_{emp} by introducing the normalized efficiency (NE):

$$NE = \frac{1 - p_{aborted}}{\tau N\sigma_{L_{50}}^2} \quad (2.7)$$

$p_{aborted}$ is defined as the percent of the tracks that were aborted during the Monte-Carlo simulations. Only those methods that employ catch trials (i.e., SIUD and GRaBr) were characterized by a non-zero $p_{aborted}$. For all other methods $p_{aborted}$ was set to 0. Furthermore, the time consumption index τ indicates the time approximately consumed for each trial which is set to 1.5 for the one-interval procedures (i.e., SIAM, APTA, QUEST+, MLP, and UML) while 2.5 is assumed for the two-tone procedures (i.e., SIUD and GRaBr). Here, we assume that the duration of one tone is 0.5s while the response time is 1s. Hence, a one-interval procedure would require 1.5s in total for the time consumption index τ. In addition, the pause between two tones is 0.5s. Therefore, two-tone procedures would need 2*0.5s (duration of one tone) + 0.5s (pause) + 1s (response time) = 2.5s. This reflects the fact that the response interval is usually much smaller than the stimulus presentation time (including pauses before stimulus onset).

Rate of convergence: Following earlier studies (Kollmeier et al., 1988; Brand & Kollmeier, 2002; Shen & Richards, 2014), we plotted the standard deviation of threshold estimates $\sigma_{L_{50}}$ to compare the rate of convergence, where the standard deviation at the i_{th} trial $\sigma_{L_{50}}(i)$ is given by:

$$\sigma_{L_{50}}(i) = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}(\widehat{L_{50,k,i}} - \mu)^2} \quad (2.8)$$

where μ is the mean of the threshold estimates and N is the number of simulations. The 'Tidyverse' package (Wickham et al., 2019) developed in R (R Foundation for Statistical Computing) was employed for the statistical analysis of the ANOVA and the post-hoc t-test. Four different four-factor ANOVAs were conducted to examine the effect of i) type of inattention (two levels: long-term/short-term) ii) degree of inattention (fully-/moderately-/non-concentrated) iii) level of false alarm rate (i.e., 0, 0.05, and 0.1) iv) adaptive procedures on four dependent variables (i.e., bias, root mean square error, normalized efficiency, and standard deviation of threshold estimates).

2.3 Results

2.3.1 Robustness

Bias

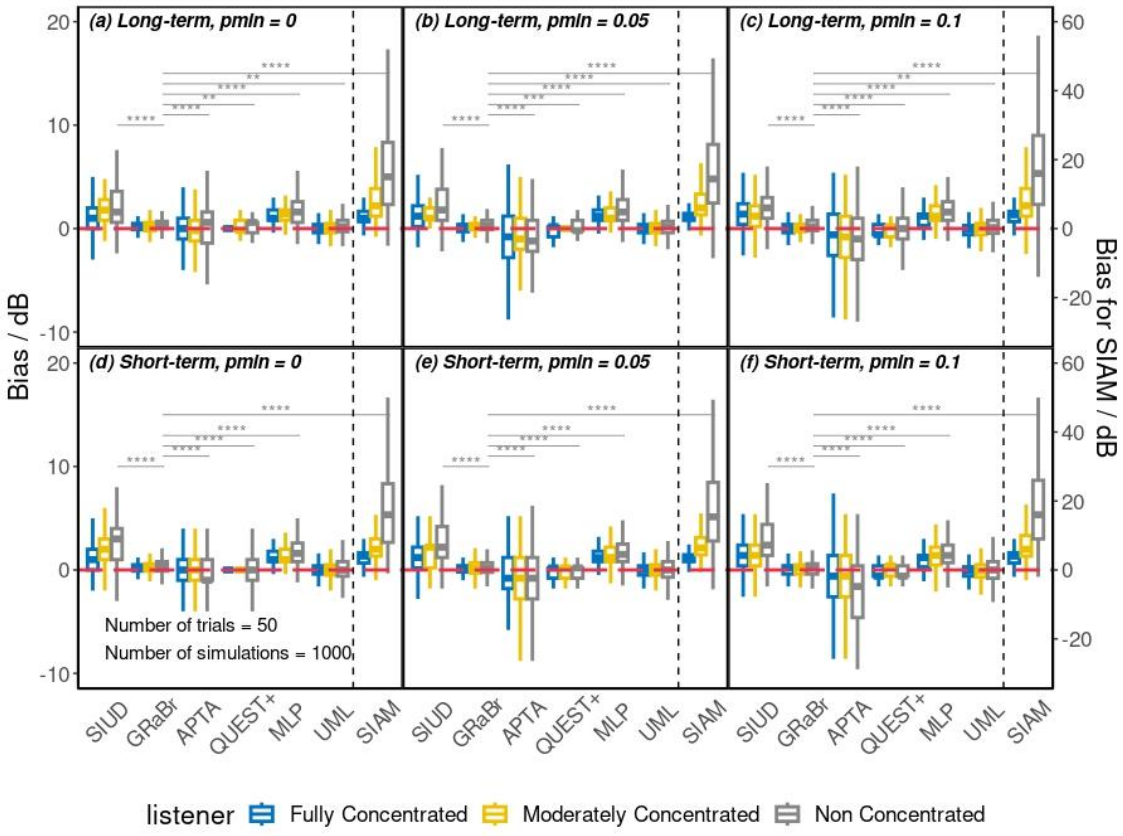

Fig. 2.4. Bias of the threshold estimates $\widehat{L_{50}}$ grouped by three simulated listeners (fully-, moderately-, and non-concentrated listener) across seven adaptive methods (SIUD, GRaBr, APTA, QUEST+, MLP, UML, and SIAM) for the long- and short-term

inattentive listener with three different false alarm rates p_{min}. Note that the bias of SIAM is plotted with a different scale (given at the right side of the figure) since a scaling factor of 1/3 had to be applied to display the data into the same plot as other procedures. See Fig. 2.3 for an explanation of the abbreviations of the adaptive procedures. Threshold estimation was compared after 50 trials to allow for a fair comparison. Dashed reference line: 0 dB. Median, 25th and 75th percentiles, and interquartile ranges (IQR) are represented in bar-and-whisker plots. The ends of the whiskers describe values within 1.5*IQR of the 25th and 75th percentiles. The statistical outcome of the pair-wise comparison against GRaBr for the NC listener is visualized via grey solid lines. The level of significance for p values is labeled with stars above the lines. Only comparisons that are statistically significant are depicted.

Fig. 2.4 shows the bias of the threshold estimates $\widehat{L_{50}}$ for seven adaptive procedures grouped by three simulated listeners for both the long- and short-term inattentive listener with three levels of false alarm rates (0, 0.05, and 0.1). The upper and bottom rows in Fig. 2.4 depict the results of the long- and short-term inattentive listeners, respectively. First, the performance in terms of bias of the long-term inattentive listener was roughly comparable to the short-term inattentive listener as the miss rate of the long-term inattentive listener was aligned with the short-term inattentive listener despite some mismatches in false alarm rate, as shown in Fig. 2.2. Second, as expected, the FC listener was the least biased while the NC listener was the most biased among the three inattentive listeners. As the level of inattention increased (i.e., from the FC to the NC listener), in most cases, the bias increased. GRaBr appeared to be less influenced by the level of inattention than the other procedures whereas SIAM was more influenced by the level of inattention. Third, as the false alarm rate increased, the median bias for most adaptive approaches increased. In general, the median biases of GRaBr, QUEST+, and UML were relatively close to 0 while the median biases of the other adaptive procedures substantially deviated from 0. Moreover, a negative median bias (i.e., an underestimation) was likely to occur if the false alarm rate increased from 0 to 0.1, especially for APTA. Overall, it is evident that an increase in the level of inattention and false alarm rate results in a larger bias and, therefore, worse performance for all adaptive procedures.

The main effect of all four factors (i.e., type and degree of inattention, false alarm rate, and adaptive procedure) was statistically significant, revealed by a four-way ANOVA test ($p < 0.05$). Subsequently, a pair-wise t-test with Bonferroni correction was carried out to compare the bias of adaptive procedures for the long- and short-term inattentive listeners over three simulated listeners with three different false alarm rates. Overall, most adaptive procedures significantly differed from each other in terms of bias ($p < 0.05$). However, there was one exception: for the long-term MC listener and the short-term NC listener, GRaBr did not differ from UML for all false alarm rates. The complete statistical comparisons are provided in the supplementary document (See Tables 2.S1, 2.S2, and 2.S11).

Root-mean-square error (RMSE)

Comparisons across seven adaptive procedures in terms of RMSE are reported in Table 2.2. It is apparent that GRaBr produced the smallest mean RMSE among all procedures in all conditions whereas APTA and SIAM had relatively large mean RMSE values. Similar to the results reported above, RMSE increased in case the level of inattention or the level of false alarm rate increased. Generally, the RMSE values of the long-term inattentive listener were comparable to the short-term inattentive listener.

All four factors (i.e., type of inattention, level of inattention, level of false alarm rate, and adaptive procedure) were significant on RMSE, revealed by a four-factor ANOVA test ($p < 0.05$). The pairwise t-tests revealed that the RMSEs from most adaptive procedures significantly differed from each other ($p < 0.05$). There was no significant difference between SIUD and APTA for the MC listener if p_{min} was 0.05. APTA did not differ from UML for the short-term MC listener if p_{min} was 0. All other pairs were tested to be significantly different (see Tables 2.S3, 2.S4, and 2.S12 of the supplementary materials for the complete statistical results).

Table 2.2. Mean and standard deviation (mean ± SD) of the root mean square error (RMSE) for the seven adaptive procedures. The smaller the RMSE, the more robust the procedure is. Refer to Fig. 2.3 for an explanation of the abbreviations for the adaptive procedure. The smallest mean RMSE value of each simulated listener (rows) is emphasized in bold.

		p_{min}	SIUD	GRaBr	APTA	QUEST+	MLP	UML	SIAM
Long -term	FC	0	1.7±0.1	**0.5±0.0**	2.1±0.2	0.7±0.0	1.4±0.1	0.6±0.0	3.8±0.2
		0.05	1.8±0.2	**0.6±0.0**	2.7±0.3	0.7±0.1	1.4±0.1	0.7±0.1	3.9±0.2
		0.1	1.8±0.1	**0.6±0.0**	3.9±0.5	1.0±0.2	1.4±0.1	0.9±0.2	4.0±0.2
	MC	0	2.5±0.3	**0.6±0.0**	2.2±0.2	0.7±0.1	4.1±1.0	1.0±0.4	11.5±1.2
		0.05	2.7±0.6	**0.6±0.0**	2.7±0.3	0.8±0.1	4.2±1.0	1.3±0.5	11.0±1.3
		0.1	2.8±0.6	**0.6±0.1**	3.8±0.5	1.0±0.2	4.3±0.9	1.5±0.5	12.2±1.3
	NC	0	4.4±0.8	**0.7±0.1**	2.3±0.2	1.0±0.2	7.1±1.2	2.4±0.8	21.0±1.5
		0.05	4.4±0.7	**0.7±0.1**	2.9±0.3	1.1±0.3	7.3±1.1	3.0±0.7	21.3±1.6
		0.1	4.3±0.8	**0.7±0.1**	4.0±0.6	1.4±0.3	6.8±1.0	2.9±0.7	22.0±1.5
Short -term	FC	0	1.8±0.2	**0.5±0.0**	2.1±0.2	0.7±0.1	1.3±0.1	0.6±0.0	4.0±0.2
		0.05	1.8±0.1	**0.5±0.0**	2.7±0.3	0.7±0.1	1.4±0.1	0.7±0.1	3.9±0.2
		0.1	1.7±0.1	**0.6±0.0**	3.7±0.5	1.0±0.2	1.5±0.2	1.0±0.3	4.1±0.2
	MC	0	2.9±0.6	**0.7±0.1**	2.2±0.2	0.8±0.1	4.3±0.9	0.9±0.3	9.5±0.9
		0.05	3.1±0.6	**0.7±0.1**	3.1±0.4	0.9±0.1	4.4±0.9	1.0±0.2	9.7±1.0
		0.1	2.9±0.5	**0.7±0.1**	5.0±0.9	1.3±0.3	3.8±0.9	1.6±0.4	9.7±0.9
	NC	0	4.8±1.1	**0.8±0.1**	2.7±0.4	1.1±0.3	7.4±1.1	2.7±0.7	20.7±1.3
		0.05	5.1±1.0	**0.9±0.1**	4.7±0.9	1.5±0.3	7.3±1.2	2.9±0.7	20.7±1.3
		0.1	5.9±1.1	**0.9±0.2**	9.1±1.4	2.0±0.4	7.1±1.2	3.1±0.6	21.1±1.3

2.3.2 Efficiency

Normalized efficiency

Fig. 2.5 shows the results of the comparison across seven adaptive methods in terms of the normalized efficiency. Overall, adaptive procedures, e.g., GRaBr, QUEST+, and UML exhibited a high normalized efficiency while SIUD, SIAM, and APTA produced a relatively low normalized efficiency. Moreover, the FC listener was the most efficient among the three simulated listeners. Additionally, the lower the false alarm rate, the more efficient the adaptive procedure was.

The main effect of four factors, i.e., type of inattention, degree of inattention, false alarm rate, and adaptive procedure on the normalized efficiency was accessed via a four-way ANOVA test. All four factors produced a significant main effect ($p < 0.05$). The influence of adaptive procedure on the normalized efficiency over long- and short-term inattentive listeners for FC, MC, and NC simulated listeners with three levels of false alarm rate was investigated employing a pair-wise t test. The results (see supplementary material in Tables 2.S5, 2.S6, and 2.S13 for the complete statistical outcome) revealed that there was a significant difference in the normalized efficiency between adaptive procedures for all simulated listeners, all inattentiveness types, and all false alarm rates ($p < 0.05$). Compared with the original SIUD procedure, the normalized efficiencies of the GRaBr procedure were significantly higher ($p < 0.05$), indicating that the modification of the GRaBr procedure with respect to the original SIUD procedure shows a positive effect. Finally, UML exhibited a significantly higher normalized efficiency than the baseline MLP procedure ($p < 0.05$).

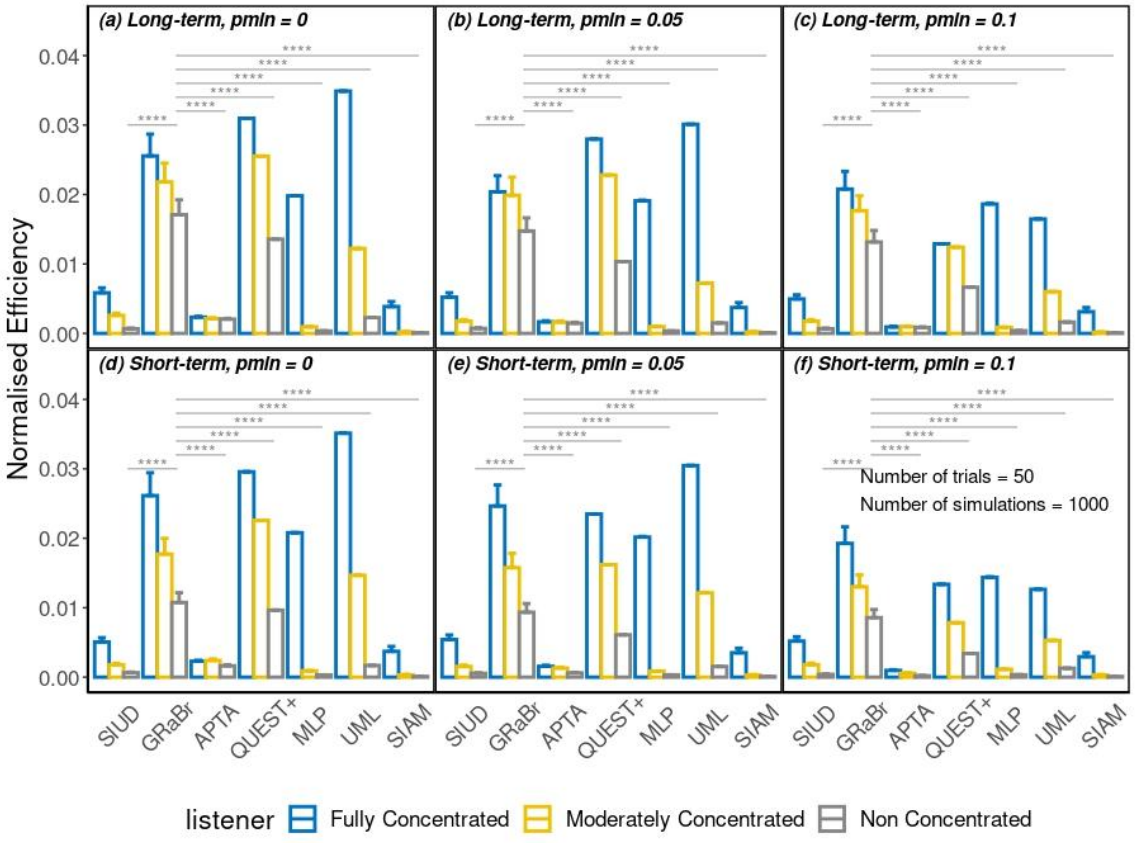

Fig. 2.5. Mean and SD of the normalized efficiency grouped by three levels of inattention across seven adaptive methods for two types of inattentive listeners with

three levels of false alarm rate. Only statistical results of the pair-wise comparison against GRaBr for the NC listener (grey solid lines) are plotted. The higher the normalized efficiency, the more efficient the procedure is. Only comparisons that are statistically significant are depicted by grey lines and respective level of significance. See Fig. 2.3 for an explanation of the abbreviations. Some error bars are so small so that they are nearly invisible.

Rate of convergence

The standard deviation of threshold estimates $\sigma_{L_{50}}$, plotted as a function of the number of trials for seven adaptive procedures, is shown in Fig. 2.6. The results of the long- and short-term inattentive listener are presented in Fig. 2.6A and 2.6B, respectively. The standard deviations monotonically decreased as the number of trials increased for most adaptive procedures. Therefore, most adaptive procedures converged. On the contrary, no clear convergence was observed for SIAM if the listener was not fully concentrated. GRaBr and QUEST+ exhibited considerably lower standard deviations for a given number of trials in most conditions. GRaBr produced lower standard deviations than the baseline SIUD procedure while UML yielded lower standard deviations than the original MLP procedure. Increasing the level of inattention or false alarm rate generally led to an elevated standard deviation. In other words, an adaptive procedure converged more quickly for the FC listener with a lower false alarm rate than the NC listener with a higher false alarm rate.

An ANOVA test implied that the main effect of all four factors (i.e., type and degree of inattention, false alarm rate, and adaptive procedure) on the average standard deviation across trials was significant ($p < 0.05$). The effect of adaptive procedures on for the long- and short-term inattentive listener over three simulated listeners with three false alarm rates was assessed via pair-wise t-test. As expected, most of the adaptive procedures did not differ from each other for the FC listener, indicating that all adaptive procedures were similarly efficient for the ideal listener. However, most adaptive procedures significantly differed from each other for the MC and NC listeners with several exceptions: GRaBr did not differ from UML and QUEST+, and there was no significant difference between GRaBr and APTA for the NC listener in case p_{min} was 0.

See the supplementary material in Tables 2.S7, 2.S8, and 2.S14 for the complete statistical results.

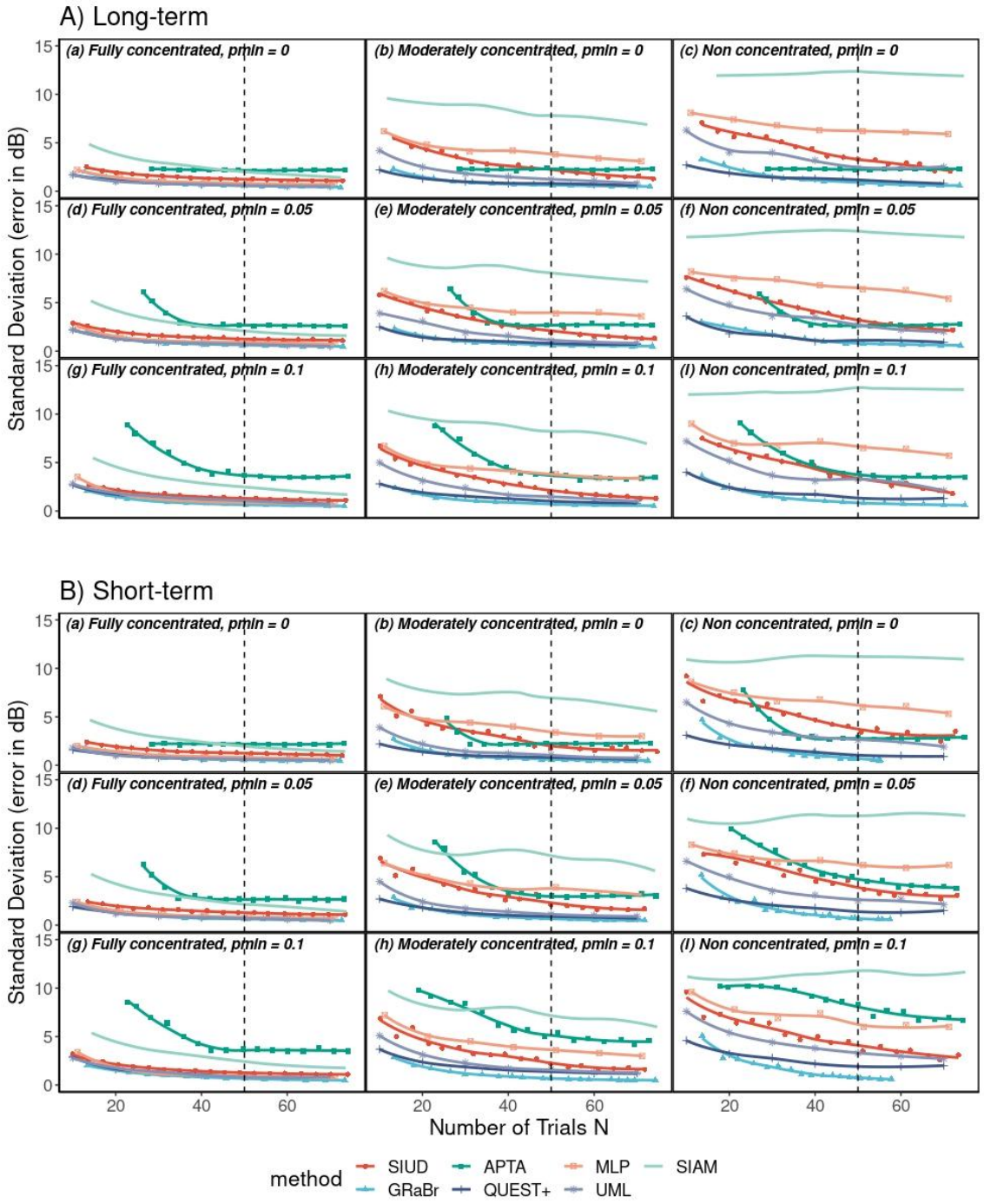

Fig. 2.6. Standard deviation of threshold estimates, $\widehat{L_{50}}$, as a function of the number of trials for seven adaptive procedures. Vertical dashed line: number of trials employed in Fig. 2.4, Fig. 2.5, and Table 2.2. See Fig. 2.3 for an explanation of the abbreviations for adaptive procedures.

2.4 Discussion

We evaluated seven adaptive procedures (three model-free and four model-based), in terms of robustness against inattention and efficiency, by Monte Carlo simulations. Two inattention models were employed for this purpose, termed long- and short-term inattention, where the level of inattention and the false alarm rate varied. Some of the well-established procedures (i.e., APTA representing the standard clinical procedure, SIUD, SIAM, and MLP) exhibited surprisingly little robustness against inattention and a considerable drop in efficiency even with moderate levels of long- and short-term inattention. The proposed procedure GRaBr, on the other hand, was rather accurate and robust for threshold measurements, revealed by a small bias and RMSE of threshold estimates against the "true" threshold. GRaBr also yielded considerable efficiency for all levels of inattention, indicated by the normalized efficiency index and rate of convergence. GRaBr outperformed the baseline SIUD while its performance was comparable with the state-of-the-art model-based procedure, i.e., QUEST+.

2.4.1 Adaptive procedures

APTA: For a false alarm rate of $p_{min} = 0$, the median estimated thresholds of the APTA procedure are close to the "true" threshold for the long- and short-term FC, MC, and NC listeners, thus the APTA procedure appears to be an unbiased estimation procedure for all types of inattentive observers with a low false alarm rate. These simulation results are consistent with human experiments conducted by Swanepoel et al. (2010) who used automatic audiometry using smartphones and compared the results with manual audiometry. There was no significant difference between automatic and manual audiometry, indicating that the APTA procedure was unbiased and robust. Moreover, Guo et al. (2021) evaluated the accuracy of the automatic audiometry application using easily accessible true wireless stereo earbuds. The verification experiment suggested that the APTA procedure was accurate enough for the threshold measurement. However, if p_{min} is larger than or equal to 0.05, the threshold estimate is no longer accurate. Therefore, experimenters should choose APTA carefully when measuring audiograms and ensure that the participants make confident decisions as much as they can.

If both the false alarm rate and the miss rate are low, the standard deviation of the APTA procedure is overall comparable to the other adaptive procedures, thus yielding APTA to be comparatively efficient. This is in line with the findings of Swanepoel et al. (2010) who examined the efficiency of the APTA procedure under the restriction of a limited measurement time (i.e., an average of 7.2-7.7 minutes for both ears of normal hearing subjects across 7 frequencies which corresponds to approx. 24 trials per adaptive track). However, APTA might have a convergence problem for participants who have a large false alarm rate and miss rate: in some groups, e.g., the short-term NC listener with a 0.1 false alarm rate, the shallow slope of the decrease in standard deviation across the number of trials (cf. Fig. 2.6) indicates that the precision increases less than expected from the $1/\sqrt{N}$ - law for independent estimations in each trial. This lack of convergence might be due to the accumulation of inconsistent trials across the whole measurement track that all add up to losing the correct target level orientation. As a consequence, it is advisable to supervise the participants to maintain a low level of false alarm rate and miss rate when performing the APTA procedure. Otherwise, the adaptive track might not be efficient.

QUEST+: Generally, the QUEST+ procedure could estimate thresholds both accurately and efficiently even for different degrees of inattention and false alarm rates, which is in line with previous studies, e.g., Audiffren and Bresciani (2022). Our present study mainly adjusts the miss rate of the logistic psychometric function to model long-term inattention. It is not surprising that QUEST+ handles such an inattention model well since it considers the influence of the miss rate and could even estimate the miss rate at the end of a track. However, as Audiffren and Bresciani (2022) explain, the QUEST+ procedure would no longer be robust if the listener behavior was modeled differently (e.g., with a non-logistic psychometric function like a beta distribution that violates the basic assumptions of the QUEST+ procedure). For the short-term inattention listener, an increase in inattention leads to an increase in the variance of threshold estimates. Since the initial PF in the short-term inattention model (indicated in Eqs. 2.2) does not exhibit a fixed and stationary miss rate, this mismatch to the model assumed by the QUEST+ procedure results in a high uncertainty of the threshold estimates. In comparison to GRaBr, this produces a significantly larger RMSE and a lower normalized efficiency at least for the short-term inattentive observer. This

suggests that the GRaBr is a better choice than the QUEST+ procedure for unsupervised psychophysical tests, e.g., using a smartphone.

MLP & UML: The MLP procedure estimates the hearing threshold precisely for the FC listener, however it yields severe overestimates for the MC and NC listener. Green (1995) reported that the MLP procedure produced a standard deviation of nearly 5 dB if the N was smaller than 20 trials, and 2.5 dB after 50 trials. Our results are roughly in line with Green (1995). Green (1993) indicated (without proof) that the MLP procedure was more efficient than the SIAM procedure. This was later questioned by Shepherd et al. (2011) who disagreed with Green's statement. In our study, the comparison of the rate of convergence between the SIAM and MLP procedures (cf. Fig. 6) indicates that the MLP procedure has a smaller standard deviation for the three simulated listeners given the same number of trials than the SIAM procedure. Hence, our data support Green's (1993) assertion that MLP is more efficient than SIAM. Moreover, for the MC and NC listeners, the normalized efficiency of the procedure is much lower and the bias is larger than for the model-free procedures considered here (e.g., GRaBr). Therefore, the use of the MLP procedure for smartphone experiments is not encouraged since its performance is greatly affected by the status of the listeners.

Several researchers (Green, 1995; Gu & Green, 1994; Lecluyse & Meddis, 2009; Leek, 2001; Leek et al., 2000; Shepherd et al.; 2011) highlight that the MLP is not a robust procedure and try to assess why the MLP procedure deviates from the expected advantageous high efficiency and fast convergence. On one hand, Green (1995) assumed that inattentive participants produce unreliable results. In support of that, Gu and Green (1994) reported that inattention occurring in an early trial (especially before the 5th trial) makes the measurement inaccurate. Our simulations show that the inattention had less effect on the threshold estimates when N is larger than 5 trials, indicating that in the adaptive strategy of the MLP procedure, the early trials have more weight/importance than the late trials. Furthermore, Leek et al. (2000) performed human experiments to validate the MLP procedure and found out that it was difficult and costly for listeners to maintain a concentration. Leek et al. (2000) also point out that the MLP procedure is inappropriate for tasks in which the listener model is not based on a fixed psychometric function. On the other hand, Lecluyse and Meddis (2009) did not attribute the poor performance of the MLP to inattentive listeners but rather argued that the

adaptive procedure itself results in poor performance. They demonstrated that the adaptive strategy of the MLP procedure is somehow self-reinforcing which prevents a regression to the true threshold, leading to permanently false estimates. Lecluyse and Meddis (2009) further suggested that those false estimates do not disappear even if the number of trials becomes larger. Our data support that both Green (1995) and Lecluyse and Meddis (2009) are correct in their statements: while the different inattentive observer models clearly lead to a significant bias and reduced efficiency (in line with Green (1995)), the standard deviation for the SIUD procedure exhibits a much steeper decline with an increasing number of trials for the initial trials than for a larger number of trials (cf. Fig. 2.6) where most other procedures show a steeper slope. This supports the assumption by Lecluyse and Meddis (2009). In conclusion, the low reliability of the MLP procedure appears to originate both from the inattentive participants and the procedure itself.

The optimized procedure UML significantly surpasses the original MLP in terms of robustness and efficiency, which is in agreement with the previous study (Shen & Richards, 2012). Since UML is specially designed to solve the shortcoming of inattention, the better performance of UML is expected. As a consequence, when choosing an adaptive procedure for mobile devices, where listeners are highly likely to be distracted, UML appears to be better suited in comparison to MLP. However, with increasing inattentiveness, the model assumptions are increasingly violated which leads to a decrease in efficiency and a slight increase in bias, especially for the non-stationary inattention case. In comparison to GRaBr, UML shows a poorer performance in these conditions as UML typically does not incorporate the short-term inattention model with the ‘unusual’ psychometric function when designing the adaptive procedure while GRaBr, as a model-free procedure, is relatively less sensitive to the inattention model. Hence, GRaBr appears to be preferable in cases where stable attention of the subject cannot be secured.

SIAM: SIAM is considered to be less robust and efficient than the other adaptive procedures in most cases. SIAM was introduced by Kaernbach (1990) who demonstrated that the SIAM procedure was reliable and robust for the FC listener by considering the response bias. Shepherd et al. (2011) reported that the SIAM procedure was efficient because a single interval task is employed which consumes less time per

trial than if nIFC procedures are used. However, our simulations indicate that the SIAM procedure yields a large bias for inattentive listeners and is generally not as efficient as other procedures, especially for the groups of MC and NC listeners. The main reason is that the SIAM procedure according to Kaernbach (1990) assumes that both the hit and false-alarm rates assume maximum values of 1 which is reflected in the payoff matrix that controls the adaptive track. This assumption, however, is only true for the FC listener, whereas the maximum hit and false-alarm rates for the MC and NC listener in our inattentive model are reduced to be less than 1, i.e., smaller than for the FC listener. If these values are known a priori, a modified payoff matrix might be employed which may avoid the observed bias. However, such a-priori knowledge is usually not available during the time of testing. As a consequence, SIAM appears to be a viable procedure for the FC listener in the laboratory but does not appear to be an unbiased, efficient, and robust threshold estimation method as soon as an MC or NC listener is assumed.

SIUD & GRaBr: Lecluyse and Meddis (2009) examined the reliability of the SIUD and MLP procedure for the FC listener and found that the threshold estimates did not differ, even though the MLP procedure yielded a larger standard deviation. Our experiments are in agreement with Lecluyse and Meddis (2009) insofar as there is no difference in the threshold estimates between the SIUD and MLP procedures. When considering the bias for the MC and NC listener, the SIUD procedure is more robust than the SIAM procedure and comparable to the MLP procedure whereas the rate of convergence and the normalized efficiency are superior for the SIUD procedure ($p < 0.001$). This compromised performance of the MLP and SIAM procedures for the MC and NC listener appears to be due to the assumption of a fixed psychometric function during the threshold estimation process which is not met unless for an FC listener as assumed by the procedures. The SIUD procedure, on the other hand, is model-free which helps to overcome the inattention to some extent based on the numerical data.

An efficiency comparison against SIUD with other adaptive procedures (e.g., MLP) was not carried out in Lecluyse and Meddis (2009), which motivated parts of our study. We observe that the MLP procedure is generally less efficient than SIUD for all levels of inattention and false alarm rate which points towards an advantage of the SIUD over the MLP procedure for practical applications where the attentiveness of the listener is not assured.

One difference between the SIUD procedure and the other procedures established so far (e.g., MLP, APTA, SIAM, UML, and QUEST+) is the psychophysical task (0, 1, or 2 tones detected with one of the two tones being presented at a level 10 dB above the other tone) which is intuitive and easy for naïve subjects to perform. Hence, this procedure is attractive for use in smartphone applications which motivated its usage in the GRaBr procedures (see below). Lecluyse and Meddis (2009) pointed out that there is little training effort needed which was supported by Shepherd et al. (2011). Rather, for the single-interval-yes-no task, Gu and Green (1994) showed a significant difference between the naïve and experienced listeners employing the MLP procedure. Leek et al. (2000) further confirmed this finding. Amitay et al. (2006) validated that more trials were required for naïve listeners applying the MLP procedure for threshold assessment since too few trials produced a very large variance.

The GRaBr procedure is very similar to the SIUD procedure during the initial part of the adaptive track. Once the threshold region is approached, however, GRaBr makes the sound level difference between the two tones adaptive with the effect that more cue tones are presented near the threshold in GRaBr than in SIUD (see Fig. 2.3). This provides more information per trial about the psychometric function close to the threshold than if a fixed 10-dB difference is employed. In addition, GRaBr treats the responses gained from presenting each of both tones in a similar way, i.e., the audibility of the cue tone is exploited to steer the adaptive level placement. In contrast, in the SIUD, the audibility of the cue tone is only evaluated in sham trials which causes additional time effort and hence a reduction in efficiency in comparison to GRaBr. Therefore, the GRaBr procedure was expected to be more efficient than the SIUD procedure. This was confirmed by the normalized efficiency estimated in our simulations (cf. Fig. 2.6).

Choosing model-free non-parametric procedures or model-based parametric procedures is highly dependent on the context and the availability of previous information about the performance of the subjects. If the shape of the psychometric function is roughly known a priori, model-based procedures appear advantageous because they can assess the target threshold more quickly and efficiently than non-parametric procedures. However, estimating or assuming the parameters may be problematic, and any inconsistent response behavior of the subject might lead to

“unforgiving” slow convergence and a bias in the resulting threshold estimate. Moreover, inaccurate parameter choices in model-based procedures may produce incorrect estimates (Audiffren & Bresciani, 2022). Hence, model-free procedures are often preferred by experimenters since they usually are insensitive to lapses and to “unusual” shapes of the psychometric functions that might not match the expectation. Moreover, due to their robustness against the slope of the underlying psychometric function, they are also quite independent from the step size chosen (albeit the product of step size and slope of the psychometric function is a scale-invariant parameter which has some effect on the efficiency of the procedures). Hence, in agreement with some previous studies, e.g., Smits et al. (2022), the current study shows that model-free procedures (such as GRaBr) typically do not perform worse than parametric procedures and can yield high robustness against comparatively drastic changes of the psychometric functions due to inattention that are considered here.

Taken together, the GRaBr procedure is recommended for the (hearing) threshold detection with potentially inattentive observers due to its high efficiency and robustness against inattention as well as due to the psychophysical task employed (i.e., the graded response of 0, 1, or 2 tones detected) which is supposed to be easily employed by naïve listeners.

2.4.2 Influence of inattention and false alarm rate

In our numerical experiments, the impact of three independent factors (i.e., type and level of inattention, and level of false alarm rate) is systematically investigated. Previously, Green (1995) examined the two factors miss rate and false alarm rate on only one model-based procedure MLP and found that both factors impacted the accuracy of MLP. We extended the scope by assessing the influence of inattention and false alarm rates on six additional adaptive procedures including three model-free procedures. Furthermore, we implemented a short-term inattention model, which differs from Green (1995).

- Level of inattention and false alarm rate: The robustness and efficiency of adaptive procedures tend to diminish with increasing levels of inattention or false alarm rates, as evidenced by our simulations (e.g., Fig. 2.4 and 2.5). To mitigate these effects, it is essential for experimenters to monitor and ensure participant attention, as

highlighted by Leek et al. (2000), and to maintain a low false alarm rate. In scenarios where distractions or high false alarm rates are anticipated, or where close supervision is impractical—such as in smartphone-based remote hearing assessments—consideration should be given to adaptive procedures that are less sensitive to these factors, like GRaBr, or to the use of model-free approaches.

- Type of inattention: A short-term inattention model is introduced that is motivated by the possible distraction from external events. It differs from the well-known long-term inattention model of sustained inattention proposed by Green (1995). This short-term inattention model has a comparable effect on adaptive procedures to the long-term inattention model at the same level of inattention, which is expected. While Green (1995) mainly models the inattention process on the overall psychometric function, we, however, explicitly model the inattention for a single trial and focus on the research question of whether different procedures are differently sensitive to this "unusual", newly introduced single-trial-PF. Please note that in the long-term inattention model, p('yes'|inattention) is fixed at p_{min}. However, more generally, p('yes'|inattention) can be assigned to any probability in the short-term inattention model, allowing greater flexibility in characterizing inattention events.

A direct comparison of the effect of both types of inattention models on the different adaptive procedures is difficult since the same parameter values in both models lead to slightly different long-term psychometric functions (see Fig. 2.2). Table 2.3 therefore compares the normalized efficiency values from Fig. 2.5 for those parameter combinations that exhibit the same equivalent expected PF for the short-term model with the respective PF of the long-term model (see section Methods - Inattention model). Even though the differences for the various adaptive procedures are small, there are some statistically significant differences in the normalized efficiency between the long- and short-term inattention model for most adaptive procedures ($p < 0.0001$), indicating that the short-term inattention has a larger negative impact on the performance of the procedures (e.g., MLP) if compared on the bases of the same "effective" psychometric function.

- Structural stability of tracking procedures: the tracking procedures considered here differ in their stability against non-stationary lapses of attention. It may

be interesting to investigate in more detail how sensitive the adaptive track is to an incidence of inattention during the trials immediately following the incidence, and how many trials it would take for the adaptive track to return to the neighborhood of the ground-truth threshold. Note, however, that a detailed micro-analysis of all procedures employed that would uncover the exact mathematical reason for the (in)stability of the respective tracking method is beyond the scope of the current paper. One way to achieve this might be to model the tracking procedures as Markov chains (e.g., Kollmeier et al., 1988), where the dynamics of the flow of level distributions across trials may be analyzed by eigenvalues of the respective transition matrices.

Table 2.3. Comparison on the normalized efficiency between the long-term inattention model and the corresponding (comparable) short-term inattention model (derived given the equivalent expected PF in Eqs. 2.3) in terms of t value and the level of significance for p value, implied via t tests.

		p_{min}	SIUD	GRaBr	APTA	QUEST	MLP	UML	SIAM
MC	Long-term	0.05	0.0ns	0.2****	-0.9****	0.2****	2.3****	-14.5****	-0.4****
	Short-term	0							
MC	Long-term	0.1	0.2****	0.2****	-0.7****	-5.9****	0.5****	-14.4****	-0.5****
	Shot-term	0.05							
NC	Long-term	0.1	0.1****	0.3****	-1.4****	-8.1****	4.4****	-0.8****	-0.1****
	Short-term	0							

2.4.3 Limitations

One possible limitation of the current study is that we only set up the true threshold to a single value, i.e., 15 dB in combination with a fixed slope of the psychometric function and a uniform distribution of starting levels (range of 10 dB) that all approximate a realistic experiment with human observers. Even though these parameters are highly interrelated and are expected to have only a marginal effect on the main outcomes of our study (see below), a larger variation of these parameters might be considered in future studies, e.g., randomly drawn threshold and slope parameters could be considered (Shen & Richards, 2012). This might also avoid possible misjudgments about the value of adaptively fixating the step size: If the distribution of initial levels (in relation to the distance from the true threshold) is too narrow, it could happen that

procedures that adaptively determine the step size would perform very differently from procedures with pre-specified step sizes because the initial trials of a track would exhibit too little variations across repetitions of the Monte-Carlo simulations.

However, within the given numerical limits the simulations are assumed to be shift-invariant with respect to the true threshold and scale-invariant with respect to the product of the (initial) step size of the respective procedures and the slope of the underlying psychometric function. Hence, even if threshold and slope parameters are changed, the simulation results will not change as long as these invariants are still in place. A change in initial step size and in the width of the distribution of initial levels would therefore be the only parameters that will produce a slight, but notable change in the simulation results. Please note that the effects on initial trials may require further investigation, as the ultimate goal of this procedure is to determine thresholds using a minimum number of trials. As they only have an impact on the first few trials of the simulation, the main outcomes of the simulations are expected to be unchanged. This assumption is based on findings by Kollmeier et al. (1988), which demonstrate that the influence of the starting parameters on the distribution of levels in an adaptive track vanishes quickly, especially after the first reversal in the track. Hence, a systematic variation of the starting level parameters is expected to yield too few effects to be of interest in the current study, given the already large number of parameters and versions that are being reported on in the current study.

A similar argument holds for the number of trials which is restricted to 50 in our study. It could be expanded to larger values (e.g., 100 and 200 as in Audiffren & Bresciani, 2022). However, the convergence of the procedures considered here was already observed for the 50 trials such that no new information is expected for longer runs. In addition, in practical experiments, the limited measurement time should be distributed to more, but shorter tracks rather than to fewer, but longer tracks in order to average out individual track-to-track variability of the “true” threshold (Kollmeier et al., 1988).

Finally, the simulations employed and discussed here need to be supplemented by experimental data with real human subjects preferably with a variation of the type and level of inattention to validate the simulations performed here. Even though it is

difficult to quantify inattention in daily life, future studies will have to systematically study the effect of limited cognitive resources (including attention) on the outcome of psychophysical experiments.

2.5 Conclusion

Inattentiveness of the observer—simulated here with a long- and short-term inattentive behavior model and a moderately- and non-concentrated observer in comparison to a fully concentrated observer—exhibits a major influence on the robustness and the efficiency of the various adaptive psychoacoustic procedures employed here. Most of these have been well-established for well-controlled laboratory conditions in the past. As a consequence, adaptive tracking procedures that have been validated in laboratory studies for the fully concentrated observer cannot be simply transferred to non-laboratory situations with several possible sources of distraction, e.g., smartphone experiments in the real world.

The short-term inattentive observer—which has been introduced here to reflect typical disturbances during real-life conditions using smartphones as a measurement tool—provides a significantly different challenge for the robustness and efficiency of the adaptive psychophysical tracking methods studied here to the well-known long-term inattentive observer if compared at the same rate of inattention or the same shape of the long-term psychometric function. In addition, the false alarm rate significantly influences the robustness and efficiency of adaptive procedures. Generally, as the false alarm rate increases, both robustness and efficiency decrease for most adaptive procedures.

The different psychophysical adaptive tracking methods vary considerably with respect to their robustness against inattentiveness. Overall, the newly introduced method GRaBr optimizes the baseline method SIUD and shows at least comparable performance with some of the latest model-based adaptive procedures, e.g., QUEST+. GRaBr provides relatively high normalized efficiency and high robustness against the different conditions of user inattentiveness. This is due to the design of the psychophysical task, which uses a graded response with 0, 1, or 2 tones detected in a trial, and its simple yet ‘forgiving’ tracking algorithm that considers only the most

recent response history of the adaptive track. Hence, the GRaBr procedure appears to be recommendable both for well-controlled in-lab hearing assessments and for psychophysical measurements using mobile devices in real life (e.g., smartphones).

2.6 Appendix: hybrid inattention model

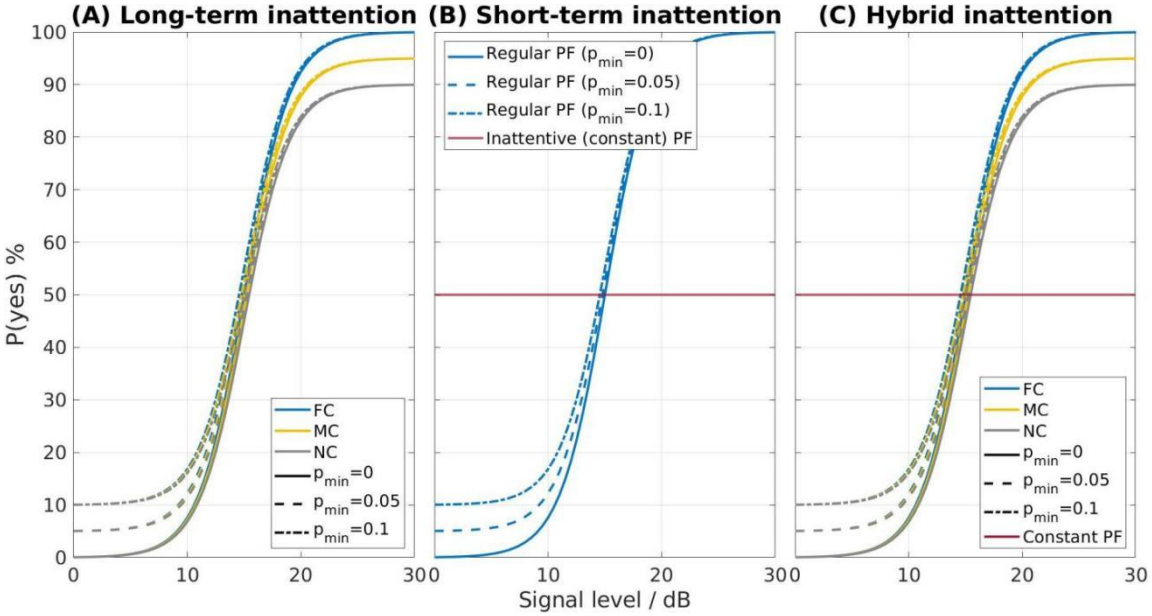

Fig. 2.A1: A) long- and B) short-term attention model (replicated from Fig. 2.2 in the main corpus for comparison) C) observer model for the hybrid inattention model (i.e., a listener is distracted both by the long- and short-term inattention simultaneously)

We evaluated a hybrid inattentive listener model that comprises both the long- and short-term inattention model, where the psychometric function is shown in Fig. 2.A1 (C). Again, for the hybrid inattentive listener, three levels of inattention and three levels of false alarm rate are adjusted so that they can be compared with the other types of inattentive listeners. More specifically, a hybrid FC, MC, and NC listener responds randomly in 0%, 10%, and 20% trials and exhibits a pmax of 1, 0.95, and 0.9, respectively. Subsequently, simulations with this model and the seven adaptive procedures were performed. Results are now presented in Fig. 2.A2, where different point shapes represent three types of inattentive listeners.

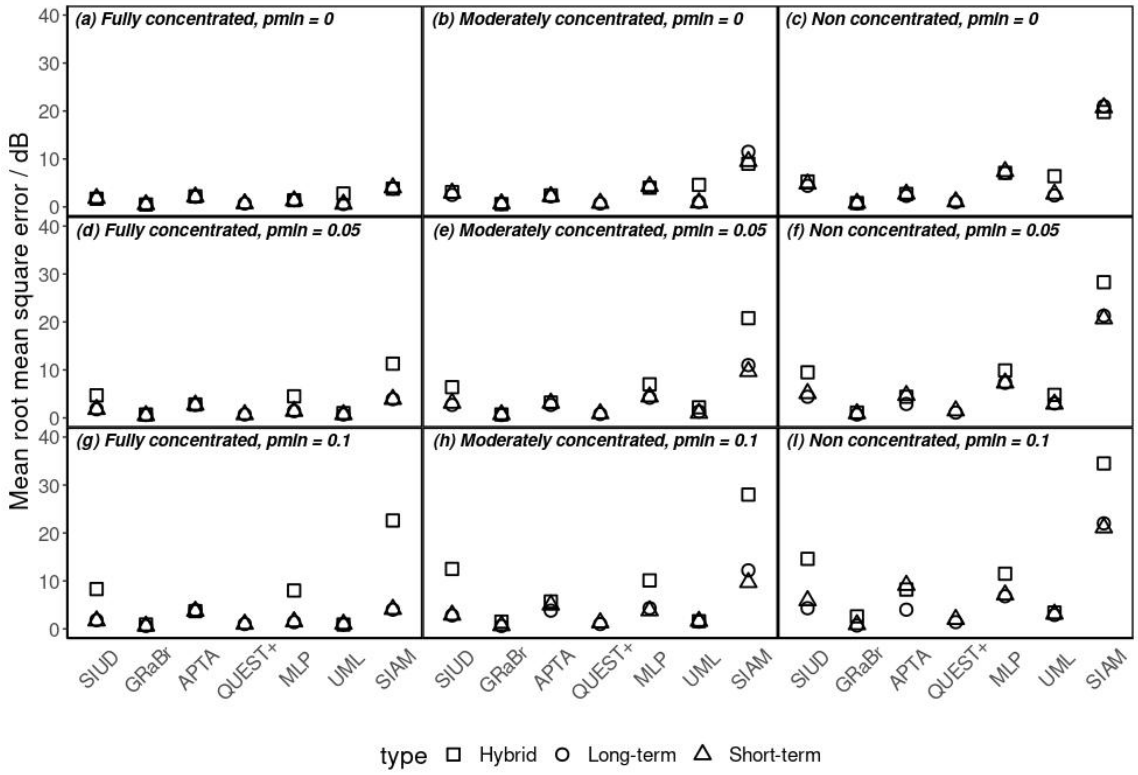

Fig. 2.A2: Mean root-mean-square error (RMSE) of the threshold estimate for a track length of 50 for three types of inattentive listeners.

As expected, the hybrid inattentive listener exhibited a larger threshold estimation error than the other two types of inattentive listener, while the main effect of the false-alarm rate and the level of inattention on the performance for the different adaptive procedures stays the same as for the main two listener types. Specifically, our proposed GRaBr procedure yields a relatively small RMSE for all types of inattentive listeners among all adaptive procedures.

3 Influence of supervision[2]

Abstract

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

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

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

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

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

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

---

[2] This section is a formatted reprint of

Xu, C., Schell-Majoor, L., & Kollmeier, B. (2024). Development and verification of non-supervised smartphone-based methods for assessing pure-tone thresholds and loudness perception. International Journal of Audiology, 1—11. https://doi.org/10.1080/14992027.2024.2424876

3.1 Introduction

Although the clinical routine audiometry tests (e.g., tone audiometry and speech audiometry) are highly valid and reliable to evaluate hearing ability, their practical drawbacks in terms of time consumption and costs for healthcare providers are not negligible (Colsman et al., 2020). Hence, employing a smartphone to conduct listening tests—at least for simple routine cases where no medical supervision is required—might be a cost-effective alternative (Swanepoel et al., 2014; 2015; Van der Aerschot et al., 2016) and has attracted significant interest from healthcare providers and researchers. The current study aims at validating this approach by comparing non-supervised threshold and supra-threshold tests to classical laboratory-based audiometric assessments in a controlled way.

While the validity and reliability of smartphone tone audiometry apps have been demonstrated in a number of studies (see details in the supplementary materials 3.S1), further investigation appears necessary due to potential limitations in the procedures currently employed:

a) Nearly all of the current smartphone apps use a modified Hughson-Westlake (Hughson et al., 1944) procedure which is widely adopted by clinicians due to its simple administration, little patient training, and easy implementation. However, if administered in a self-paced, unsupervised way due to occasional inattentiveness of the listeners, this procedure might be inaccurate and might overestimate the true threshold according to Lecluyse and Meddis (2009) and Xu et al. (2023). The present study therefore adopts the single-interval-up-and-down procedure SIUD, proposed by Lecluyse and Meddis (2009), to assess air-conduction pure-tone audiometry on a smartphone and compares the acquired results with the laboratory-based measurements.

b) Another limitation of smartphone apps for measuring individual audiograms is their reliance on low ambient noise conditions and the precise calibration of participants' headphones and smartphones, which may not always be guaranteed (Guo et al., 2021; Zhao et al., 2022). This problem is largely circumvented by using supra-threshold tests that consider a larger dynamic range, such as the categorical loudness scaling (CLS) test (Brand & Hohmann, 2002), which appears to be more resistant to issues related to missing calibration and external noise. Consequently, supra-threshold auditory

measures have gained considerable attention in recent years for hearing screening through smartphone applications.

The assessment of loudness growth with increasing level or stimulus bandwidth is of clinical interest, e.g., determining the recruitment phenomenon and for fitting hearing devices (Kollmeier & Hohmann, 1995; Oetting et al., 2016; Koppun et al., 2022). Individual loudness perception is commonly measured employing the categorical loudness scaling (CLS) technique and quantified with a monotonic loudness growth function (Brand & Hohmann, 2002; Oetting et al., 2014). The task of the CLS requires participants—based on their loudness perception—to select the descriptors from an 11-point scale, e.g., 'very soft', 'soft', 'medium', 'loud', and 'very loud' with four (unnamed) intermediate and two limit categories 'not heard' and 'too loud'. The CLS is a supra-threshold listening test that has been included in the 'auditory profile' (i.e., a comprehensive and well-specified set of audiological test procedures described in Van Esch et al., 2013) and has also recently been proposed for usage in machine-learning-supported auditory profiles by Saak et al. (2022; 2024).

The standardized adaptive procedure to perform CLS measurements (i.e., Adaptive Categorical Loudness Scaling, ACALOS) was introduced by Brand and Hohmann (2002) and standardized in ISO 16832 (2006). CLS has a broad application in clinical audiology, not only as a diagnostic tool but also to fit hearing aids or cochlea implants. For diagnostic purposes, an increase in loudness growth with stimulus level—clinically termed as recruitment phenomenon and assumed to be due to dysfunctional outer hair cells (Hallpike & Hood, 1959; Buus & Florentine, 2002)—can well be characterized by CLS (e.g., Kollmeier & Hohmann, 1995; Launer, 1995; Rasetshwane et al., 2015). Jürgens et al. (2011) proposed to estimate the hearing loss attributable to outer hair cells (OHC) by applying CLS and concluded that CLS could be a measure of auditory nonlinearity. Further diagnostical applications of CLS were described, e.g., by Shiraki et al. (2022) as a means to better characterize patients with certain patterns in Bekesy audiometry and by Erinc et al. (2022) and Hébert et al. (2013) as a means to better characterize patients with tinnitus and hyperacusis.

With respect to using CLS as a tool for hearing device fitting, many studies have demonstrated that individualized loudness compensation for narrowband signals can

lead to a better-individualized treatment with hearing devices (see Kollmeier & Hohmann, 1995, Kollmeier & Kießling, 2018, Oetting et al., 2018, and Fereczkowski et al., 2023 for hearing aids and Müller-Deile et al., 2021 for cochlea implants). However, despite the theoretical advantage of CLS for characterizing supra-threshold functional hearing deficits and individually fitting hearing devices, its usage for clinical purposes has been limited, e.g., due to:

a) Time constraints in clinical settings that may be a barrier to the usage of more sophisticated methods beyond the minimum set of clinical routine procedures (Colsman et al., 2020). However, self-administered, smartphone-based procedures may eventually supplant traditional methods, reducing the time-intensive burden currently placed on professional audiologists.

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

On the other hand, a strong argument for the clinical use of CLS arises from the recent discovery of individually strongly varying loudness summation across frequency and across ears by Oetting et al. (2016): They reported significant individual variations in loudness perception for binaural broadband signals among participants with the same hearing thresholds. This resulted in lower levels required for hearing impaired (HI) listeners (with a pure-tone average (PTA) > 20 dB HL, PTA: average thresholds of 0.5, 1, 2, and 4 kHz) to reach ‘medium loud’ for broadband signals than for normal hearing (NH) listeners (PTA <= 20 dB HL) when narrowband gain compensation was applied. Thus Oetting et al. (2016) recommended that broadband and binaural loudness scaling should be included for hearing-aid fitting to avoid over-amplification in bilateral fitting prescribed on monaural fitting rules. Van Beurden et al. (2018) confirmed the results of Oetting et al. (2016) using more test participants with a broader range of hearing loss.

They found large individual variations in HI listeners for binaural broadband signals and confirmed that binaural loudness summation can not accurately be predicted based on hearing thresholds. In this study, we therefore employed not only narrowband signals presented unilaterally, but also broadband signals presented bilaterally for both NH and HI listeners.

Even though CLS is an applicable and useful measurement for clinical diagnostics and assessment of hearing loss compensation as introduced above (e.g., Rasetshwane et al., 2015; Fultz et al., 2020), it is not yet available for a smartphone or any other mobile device. There is only one study published so far that introduced a remote CLS measurement on a laptop and compared it with the laboratory setting (Kopun et al., 2022). However, they did not examine the test persons via smartphone and did not include HI participants. Furthermore, Kopun et al. (2022) only included 5 participants for the validation study. One possible obstacle to self-controlled CLS measurement in an unrestricted environment is the influence of background noise (which might cause a bias at low stimulus levels that might be confused with a recruitment phenomenon) or any inattention effect of the participant (as simulated in Xu et al., 2023). Hence, in this study, one of our objectives was to examine the plausibility and validity of the smartphone-based app for CLS measurement under different degrees of control in experimental settings.

Taken together, the following research questions were addressed by our study by performing three sub-experiments (i.e., Exp 1: pure-tone audiometry reported in the supplementary materials 3.S1; Exp 2: adaptive categorical loudness scaling; Exp 3: binaural and spectral loudness summation) that all employ normal-hearing and hearing-impaired listeners and compare laboratory situations with self-steered, smartphone-based setups:

- Are the results of the smartphone-based categorical loudness scaling (and pure-tone audiometry, see supplementary materials 3.S1 for details) quantitatively comparable to a laboratory-based assessment?

- Which factors influence the differences between smartphone-based and laboratory-based measurements?

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

3.2 Materials and methods

3.2.1 Subject groups

21 normal hearing (NH, aged between 20 and 35 years; 7 males, 14 females) and 16 hearing impaired listeners (HI, aged between 67 and 88 years; 11 males, 5 females) participated in the study and received a financial compensation of 12 euros per hour. The participants in the NH group are mainly members of the working group and students of the university. The HI listeners were recruited via the database of Hörzentrum Oldenburg gGmbH (cf. Table 3.S1 in the supplementary materials for the means and standard deviations of their hearing thresholds). The mild-to-moderately impaired listeners with sensorineural hearing loss exhibited pure-tone averages (PTA: average thresholds of 0.5, 1, 2, and 4 kHz) varying between 26.3 and 42.5 dB HL. The PTAs for better ears of HI listeners averaged 31.8 (± 5.3) dB HL while the mean and maximum PTA difference across ears were less than 2 dB HL and 10 dB HL, respectively. NH listeners yield thresholds at or below 15 dB HL for all frequencies between 250 Hz and 4 kHz. All participants did not have any previous experience with smartphone hearing tests. The study was approved by the research ethics committee of the Universität Oldenburg (Drs. EK/2022/011).

3.2.2 Test conditions

A repeated-measures experimental design was employed that mainly varied the degree of supervision in three conditions (cf. Table 3.1). Condition I was a fully-supervised, manual measurement as reference. Condition III was a non-supervised assessment. Condition II was semi-supervised, i.e., the experiment ran automatically under the control of the same adaptive procedure as for condition III while the test examiner was available on request for questions without having access to the log data.

Table 3.1. Experimental design for the three conditions employed that differed in the degree of supervision.

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

a Test supervisor available on request for general questions without having access to the log data

Furthermore, the same calibrated Android smartphone (OnePlus Nord N10 5G 128 GB, Google Chrome installed) with its own built-in sound card was provided to all participants in condition III. In all three conditions the same HDA200 headphone was employed in a sound-attenuated booth. All conditions were calibrated employing 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. The target level for calibration was 80 dB SPL.

3.2.3 Adaptive categorical loudness scaling

Adaptive categorical loudness scaling (ACALOS; see Brand and Hohmann (2002) and ISO 16832 (2006) for details) requires users to rate their individual loudness perception, elicited by each stimulus, using a categorical scale with 11 values (see Introduction and bottom corner of Fig. 3.1b for the user interface). Participants could select both the main categories (denoted by words) and the intermediate categories (denoted by inverted trapezoids). The responses are mapped to the 50-point categorical units (CU) scale according to Heller (1985). In the first of two phases (‘dynamic range estimation’), ACALOS starts at 65 dB presentation level which is increased and decreased in an interleaved manner to obtain a rough estimate of the dynamic range between 0 CU and 50 CU. In the second phase (‘presenting and re-estimation’), the

individual loudness function is then fine-tuned by presenting stimuli at 5 levels estimated from the first phase corresponding to the categorical loudness of 5, 15, 25, 35, and 45 CU in a randomized order and by re-estimating the loudness function and its dynamic range as a basis for a possible repetition of the second phase.

The 'BTUX' method introduced by Oetting et al. (2014) was used to fit a loudness growth function to the individual data, allowing the derivation of descriptive parameters: hearing threshold level (HTL), corresponding to 2.5 CU; median loudness level (MLL), corresponding to 25 CU; and uncomfortable loudness level (UCL), corresponding to 50 CU. Additionally, the most comfortable loudness (MCL), defined as the sound level at 20 CU (Van Esch et al., 2013), and the dynamic range (DR), calculated as the difference between UCL and HTL, were determined.

The narrowband stimuli were one-third-octave-band low-noise noises (Kohlrausch et al., 1997) centered at 0.25, 1, and 4 kHz (later referred to as LNN250, LNN1000, and LNN4000, respectively). The broadband stimulus was uniformly exciting noise (UEN17) with equal energy in each of the 17 critical frequency bands, defined in Zwicker (1961). All stimuli (i.e., three narrowband and one broadband stimuli) were presented monaurally for both ears. In addition, LNN1000 and UEN17 were played bilaterally. The duration for all signals was 1 s with 50 ms rise and fall ramps.

3.2.4 Procedures

To mitigate learning effects, participants were familiarized with the measurement procedure. An initial training session was conducted using a randomly selected stimulus from each condition. Subsequently, three main sessions were performed: Condition I (reference) first, followed by Conditions II and III in a random order (see Table 3.1). Within each session, stimuli were presented in a random order.

3.2.5 Smartphone application design

Fig. 3.1 illustrates the smartphone application employed. The web-app was developed using the Flask (version 1.1.2) framework in Python (Python Software

Foundation, version 3.10.6), while the database was based on SQLite3 (version 3.37.2). Both frameworks are open source.

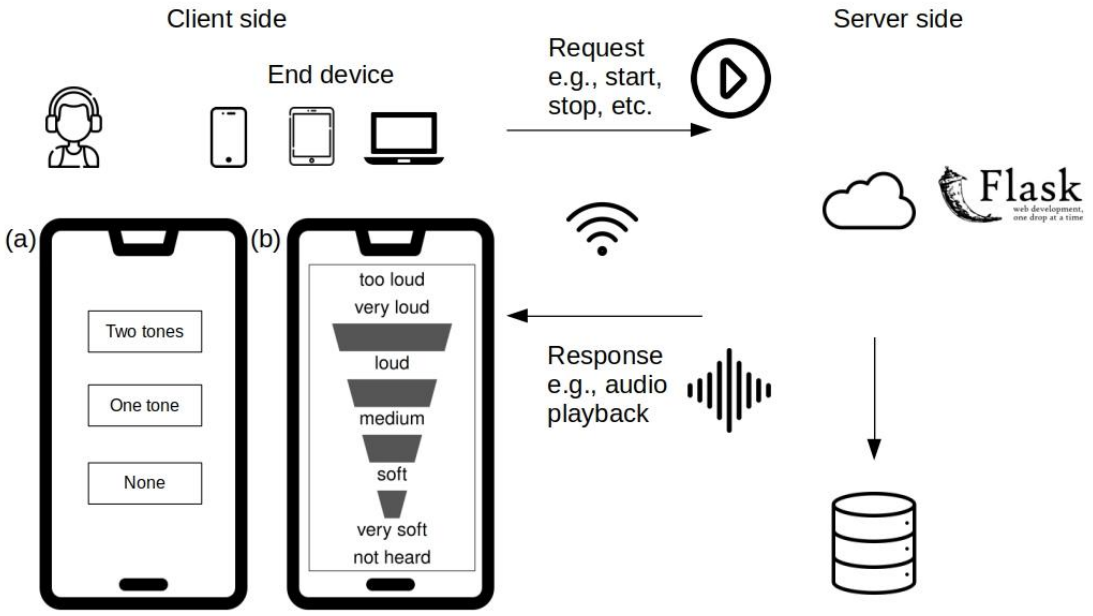

Fig. 3.1. Overview of the smartphone application. On the client side, the user interface of two assessments (i.e., (a) pure-tone audiometry and (b) categorical loudness scaling) is shown. On the server side, the web application framework 'Flask' is available for processing requests from a listener. The measurement data is not stored locally but in the cloud database.

A non-supervised listening test on a smartphone followed the sequence below: The listener first registered an account and signed in to the dashboard, which displayed some general instructions in text format, e.g., study background, user consent, and test environments. After selecting which measurement to perform, the listener was presented with specific guidelines for the chosen listening test, i.e., 'How many sounds do you hear (none/one/two)?' for the audiogram measurement according to Lecluyse and Meddis (2009) or 'How loud do you judge the sound you heard?' for loudness scaling. After clicking the 'start' button, the stimuli were automatically presented to the listener. The response data were sent to the server via WLAN and stored in the cloud database. Based on the incoming response, the server prepared the adjusted stimulus (in this case, mainly adjusting the sound levels for both listening tests) and played it back to the

listener. The listener was redirected to the dashboard once the listening test was completed. No data were stored locally on the smartphones; instead, they were primarily stored on the server.

3.2.6 Data analysis

Psychophysical parameters

For the categorical loudness scaling experiment, loudness functions as defined in Brand and Hohmann (2002) and Oetting et al., (2014) were employed (cf. Eq. 3.1), which consist of two linear parts and one transition region using a Bezier fit:

$$F(L) = \begin{cases} 25CU + m_{low}(L - L_{cut}) \text{ for } L \leq L_{15} \\ bez(L, L_{cut}, L_{15}, L_{35}) \text{ for } L_{15} < L < L_{35} \\ 25CU + m_{high}(L - L_{cut}) \text{ for } L \geq L_{35} \end{cases} \quad (3.1)$$

where m_{low} and m_{high} denote the slope value of the low and high linear part, L_{cut} is the intersection level of the two linear parts, L_{15} and L_{35} are the levels of the 'soft' and 'loud' category respectively, and bez is a quadratic smoothing function between L_{15} and L_{35}. The Pearson correlation coefficient (R), root mean square error (RMSE), and bias of levels for each category (in total 11 categories) are calculated. For binaural loudness summation, the level difference for equal loudness (LDEL) is calculated as:

$$LDEL = L_b - L_l \quad (3.2)$$

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

$$LDEL = L_{LNN} - L_{UEN17} \quad (3.3)$$

where LLNN and LUEN17 denote the level for low-noise narrowband noise and UEN17 broadband noise at the same category unit, respectively. All algorithms for experimental data fitting were developed in MATLAB R2021a (The MathWorks, Inc., Natick, MA).

Statistical analysis

A mixed-design ANOVA was applied using hearing loss (two levels: NH/HI) as a between-subject factor, condition (three levels: I/II/III), and frequency (three levels: 0.25, 1, and 4 kHz) as within-subject factors. Furthermore, a post-hoc analysis among conditions using a pair-wise t-test was carried out, where the p value was corrected with 'Bonferroni'. In the post-hoc analysis, condition I was set up as a reference group. If p value < 0.05, the difference between two conditions was considered as being statistically significant, while if p value >= 0.05, the difference was not significant. The 'Tidyverse' package (Wickham et al., 2019) developed in the software environment 'R' (R Foundation for Statistical Computing) was employed for the statistical analysis of the mixed-design ANOVA and the post-hoc analysis.

3.3 Results

3.3.1 Experiment II: adaptive categorical loudness scaling

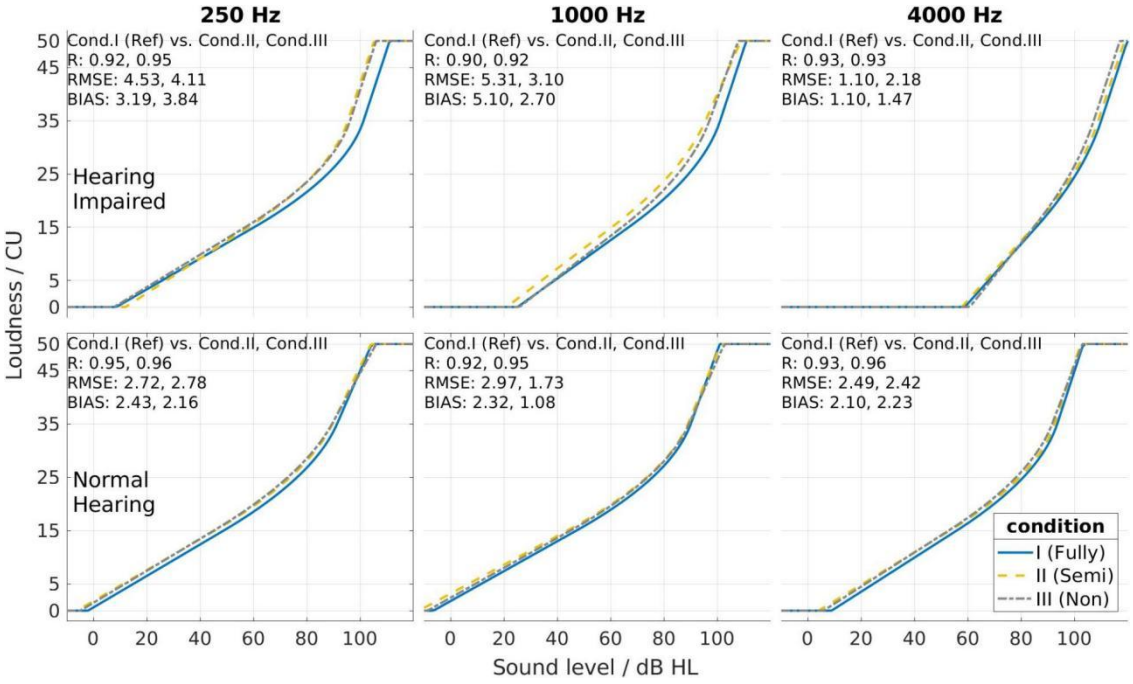

Fig. 3.2. Average loudness growth function (i.e., loudness in CU as a function of sound level in dB HL) of the three experimental conditions employed (condition I = fully-supervised; II = semi-supervised; III = non-supervised) for HI (upper row) and NH (bottom row) listeners at 0.25 kHz (left column), 1 kHz (middle column), and 4 kHz (right column). The Pearson correlation coefficients (R), root mean square errors

(RMSE), and biases between two conditions II and III against I (reference) of levels for each category units are provided in the upper left corner of each sub-figure.

Fig. 3.2 plots the average loudness function of three conditions for HI and NH listeners at 0.25, 1, and 4 kHz frequencies. For all frequencies and listener groups, the average loudness functions of conditions II and III were consistent with condition I. The average loudness functions of HI listeners generally showed steeper growth than NH listeners, especially at 4 kHz, which could be explained by the 'loudness recruitment', as mentioned above. HI listeners exhibited a significant increase in the slope of the loudness function with an increase in frequency which was not observed in NH listeners.

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

Five descriptive parameters (i.e., HTL, MCL, UCL, MLL, and DR) of three conditions for HI and NH listeners at 0.25, 1, and 4 kHz are shown in Fig. 3.3. The median descriptive parameters for all three frequencies and both listener groups in conditions II and III were close to the condition I. Moreover, the median levels of the five descriptive parameters did not change with an increase in frequency for NH listeners. The median levels of HTL increased while DR decreased with an increase in frequency for HI listeners. The IQRs of HTL and DR were larger for HI listeners compared to NH listeners.

The statistical analysis of the differences across conditions and groups is detailed in the supplemental material 3.S2. Taken together, while for most cases the five parameters did not differ between the reference condition I and the less supervised conditions II and III, respectively, statistically significant differences only existed in a

few groups, suggesting that these significant differences might not be systematic differences but rather random differences. In addition, the magnitudes of most differences between the three conditions were less than 5 dB, indicating that the differences might not be clinically relevant. As we always measured condition I first, the sequence or training effect might at least partially explain such a difference.

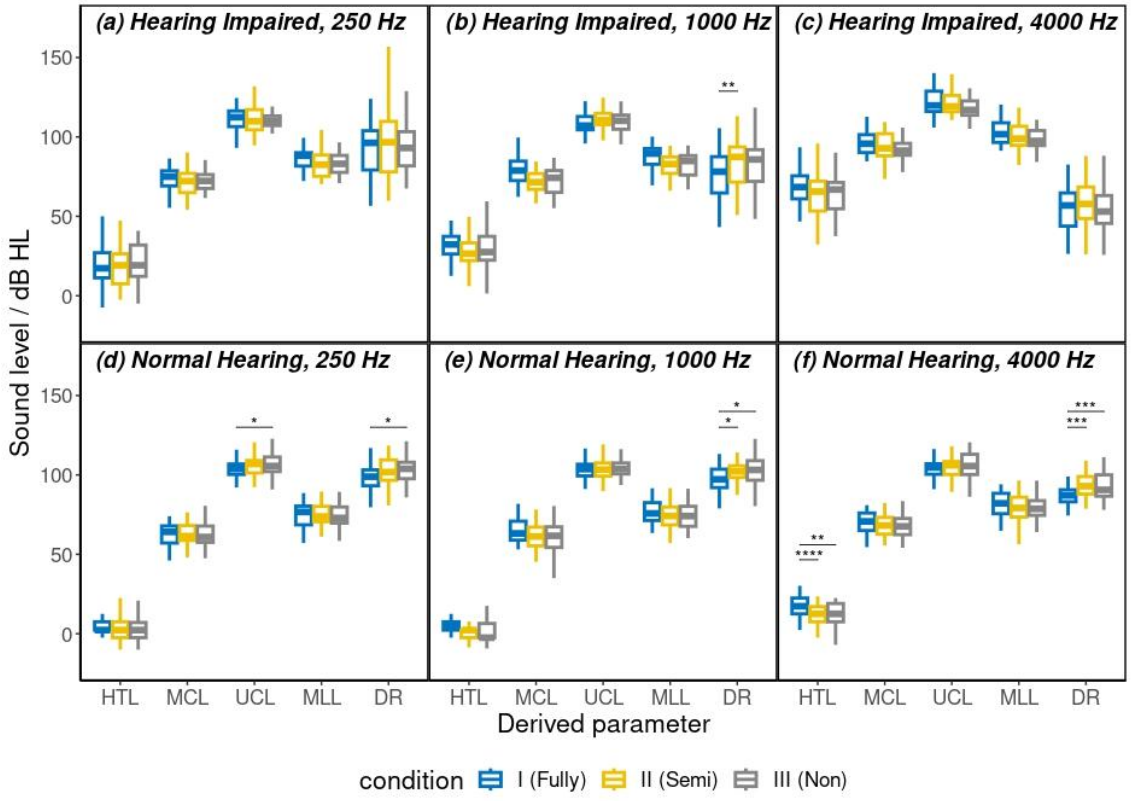

Fig. 3.3. Five descriptive and intuitive parameters (in dB HL) derived from the loudness function of three conditions (Fully = fully-supervised; Semi = semi-supervised; Non = non-supervised) for HI and NH listeners at 0.25, 1, and 4 kHz frequencies. HTL: hearing threshold level (2.5 CU); MCL: most comfortable loudness level (20 CU); UCL: uncomfortable loudness level (50 CU); MLL: median loudness level (25 CU); DR: dynamic range (UCL-HTL). The medians, 25%, 75% percentiles, and interquartile ranges (IQR) are given in the respective bar-and-whiskers plot. The ends of the whiskers describe values within 1.5*IQR of the 25% and 75% percentiles. In case of statistically significant differences, the level of significance is labeled with stars above the lines.

3.3.2 Experiment III: binaural and spectral loudness summation

Binaural loudness summation

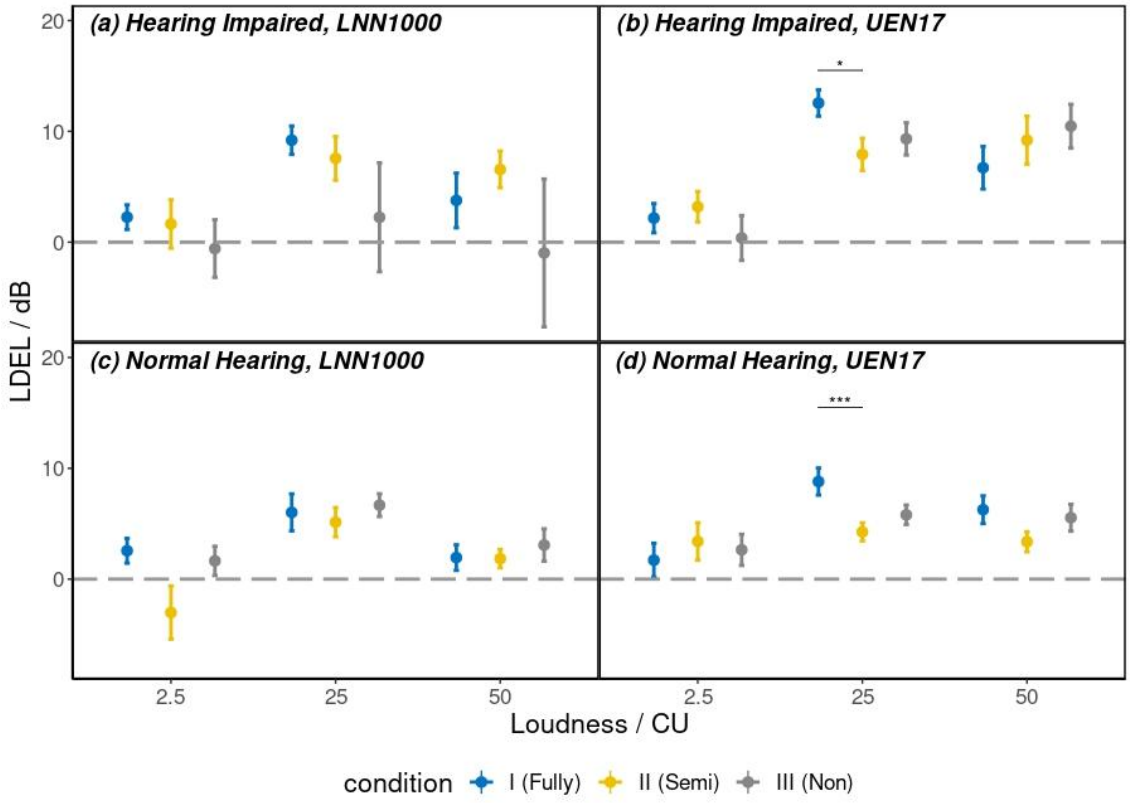

Fig. 3.4. Mean and standard deviation (denoted by whiskers) of level difference for equal loudness (LDEL, in dB) between binaural and monaural (left ear) presentation for equal loudness at 2.5, 25, and 50 CU using narrowband noise (LNN1000) and UEN17 broadband noise, respectively, for HI (upper row) and NH (bottom row) listeners. Conditions I, II, and III are differentiated with three colors (Fully = fully-supervised; Semi = semi-supervised; Non = non-supervised). Grey dashed line: 0 dB. LNN1000: one-third-octave-band centered at 1 kHz low-noise noise; UEN17: uniformly exciting noise at 17 critical bands.

Mean and standard deviation of the level differences for equal loudness (LDELs) as a function of loudness in CU of HI and NH participants for LNN1000 and UEN17 among three conditions are shown in Fig. 3.4. In most cases, the mean LDELs of conditions III and II were in agreement with those of condition I. It is notable that the standard deviation of LDEL of the condition III for LNN1000 at 25 and 50 CU for HI

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

The results of the statistical analysis are provided in details in the supplementary material 3.S2. In general, the LDEL of conditions II and III did not differ from condition I. However, a significant difference occurred in some pairs, i.e., between conditions I and II at 25 CU for both NH ($p < 0.05$) and HI ($p < 0.001$) stimulated by the UEN17 broadband signal. Even though these differences were statistically significant, the mean values of the differences were roughly 6 dB. Thus, similar to the results above, the statistically significant differences might not be clinically relevant.

Spectral loudness summation

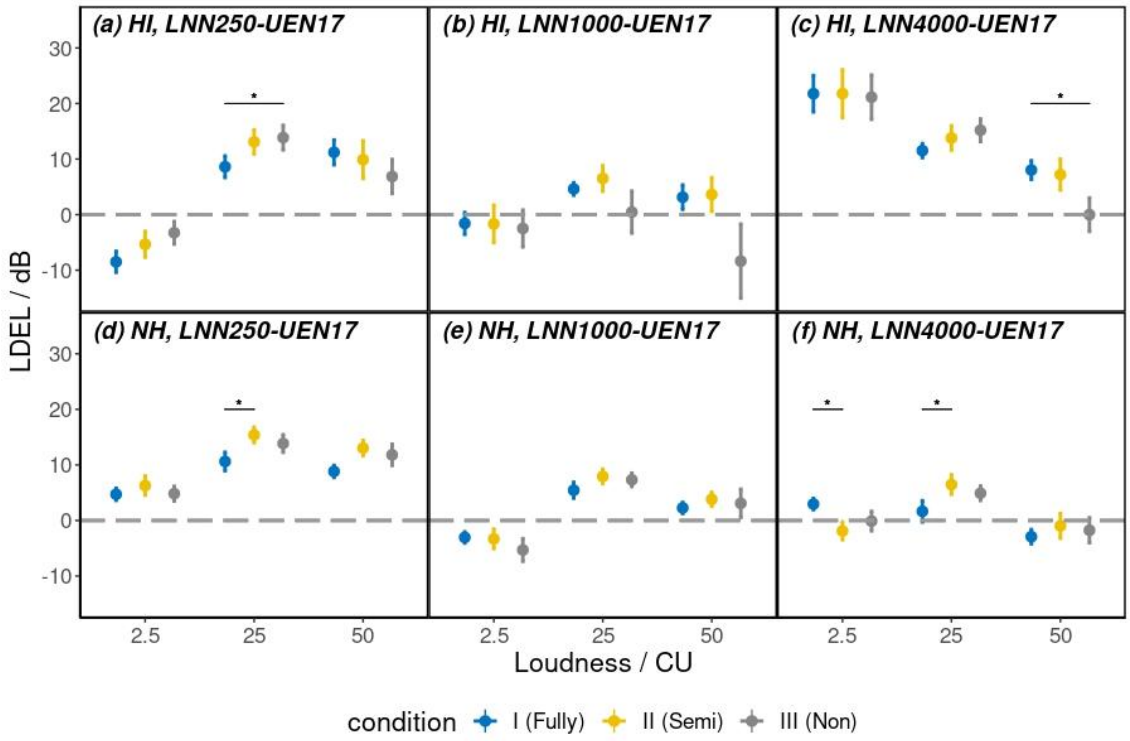

Fig. 3.5. Mean and standard deviation (denoted by whiskers) of level difference for equal loudness (LDEL) between three narrowband stimuli (LNN250, left; LNN1000, middle; LNN4000, right) and one broadband stimulus (UEN17) for equal loudness at

2.5, 25, and 50 CU for HI (upper row) and NH (bottom row) listeners across three conditions (Fully = fully-supervised; Semi = semi-supervised; Non = non-supervised). Grey dashed line: 0 dB. All signals were presented monaurally on the left ear. LNN250, 1000, 4000: one-third-octave-band centered at 0.25, 1, and 4 kHz low-noise noise; UEN17: unified excitation noise at 17 critical bands.

Fig. 3.5 shows LDEL (with error bars) of three conditions as a function of loudness in CU between LNN250 and UEN17 (left), LNN1000 and UEN17 (middle), and LNN4000 and UEN17 (right) for HI (upper) and NH (bottom) listeners. Generally, the mean difference of LDEL between conditions II and I, and between III and I was small with values smaller than 10 dB. For HI listeners, the mean LDELs at 25 and 50 CU were greater than 0 while lower than 0 at 2.5 CU concerning the comparison between LNN250 and UEN17. However, the mean LDELs of NH listeners were larger than 0 at three CU. Comparing the LDELs between LNN1000 and UEN17, both NH and HI listeners exhibited a negative LDEL at 2.5 CU while positive at 25 and 50 CU for three conditions with one exception of the HI listener for the condition III at 50 CU. Regarding the mean LDEL difference between LNN4000 and UEN17, NH and HI participants showed a substantial difference: the mean LDELs of HI listeners were always positive, while NH listeners were around 0.

The statistical analysis is provided in the supplemental material 3.S2. It reveals that all four factors (i.e., hearing loss (HI/NH), condition (I, II, and III), comparison (LNN250-UEN17, LNN1000-UEN17, LNN4000-UEN17), and loudness (2.5, 25, and 50 CU)) had a significant effect ($p < 0.05$) on LDELs. For HI participants, there was only a significant difference between conditions I and III on LDEL at 25 CU in comparison pairs of LNN250-UEN17 and at 50 CU of LNN4000-UEN17 ($p < 0.05$). For NH listeners, only the difference in LDEL between conditions I and II was significant at 25 CU of LNN250-UEN17, and at 2.5 and 25 CU of LNN4000-UEN17 ($p < 0.05$).

3.4 Discussion

Performing pure-tone audiometry and categorical loudness scaling on a properly calibrated smartphone, with controlled ambient noise and a precise adaptive procedure, was experimentally shown to be feasible. The test outcomes align closely with laboratory measurements in most cases. Supervision does not significantly affect the results obtained here, making non-supervised automated tests essentially equivalent to fully-supervised manual ones.

These smartphone-based tests are accessible for both normal-hearing and hearing-impaired (HI) individuals, who can self-administer the tests with ease if familiar with the general procedure. On top of the commonly employed unaided ACALOS measurement, i.e., narrowband signals presented unilaterally, broadband stimuli for binaural presentation on a smartphone were employed with a similar finding of minimal differences compared to lab tests. Using these stimuli for adaptive categorical loudness scaling could support future fine-tuning of non-linear hearing aids via smartphone.

3.4.1 Pure tone audiometry

Given the extensive literature on conducting pure-tone audiometry in remote settings, including via smartphones (as reviewed in the supplementary material 3.S1), it is unsurprising that Experiment I revealed no significant differences across conditions with varying degrees of supervision (see supplementary materials 3.S1). Consistent with previous validation studies conducted under similar clinical, acoustically controlled, and distraction-sparse conditions, our findings align with those reporting only a small mean (signed) difference across conditions, typically within 5 dB (e.g., Thai-Van et al., 2022). A key distinction in our study is the use of the SIUD method (Lecluyse & Meddis, 2009), which simplifies the task for naïve participants by requiring them to count the number of sounds heard (0, 1, or 2) rather than detect a single tone in quiet. This method also employs an adaptive bracketing procedure by presenting two tone levels, which is designed to be more robust against both short- and long-term attentional lapses compared to the traditional modified Hughson-Westlake procedure. The robustness of this approach is particularly enhanced when using the ‘GRaBr’ variant, which adaptively narrows the level difference between the upper and lower presentation levels (see Xu et al., 2023). However, since we did not specifically assess the efficiency or robustness of these test procedures in this study, empirical validation of the theoretical

advantages of the SIUD and GRaBR methods over the classical Hughson-Westlake procedure remains an open question.

3.4.2 Adaptive categorical loudness scaling

To our knowledge, there is no study so far evaluating categorical loudness scaling on a smartphone. Our experimental results provide the first evidence that it is plausible and valid to perform non-supervised CLS measurement on a smartphone both for NH and HI listeners. In addition, there is only one study so far, i.e., Kopun et al. (2022), which evaluated the CLS measurement on a laptop remotely in comparison to a clinical database. This is comparable with the comparison between conditions II and I in our study on the group level. Kopun et al. (2022) reported that for NH participants (N = 5), the mean signed difference averaged across categories was 5.9 and 4.9 at 1 and 4 kHz, respectively. The mean signed difference of our study is much smaller, i.e., 2.3 and 2.1 for 1 and 4 kHz. First, the fitting of the loudness function might play a role. Kopun et al. (2022) simply calculated the median level of each category to describe the individual loudness function without fitting the data to a 2-segment linear function. Second, the outliers were not removed, leading to non-monotonic loudness growth. This contrasts to our study where we fitted the individual responses based on the method introduced in Oetting et al. (2014) to obtain an individual monotonic loudness function. Third, the test environment might make an impact. We conducted all experiments in a sound-attenuated booth to eliminate the influence of environmental noise. Kopun et al. (2022), however, did in-lab measurements at a sound-treated booth while remote laptop measurements at home. Although Kopun et al. (2022) attempted to control and check the noise level between runs in the remote measurements, the fluctuating environmental noise might influence the loudness judgment during the run. Fourth, Kopun et al. (2022) used a different calibrated headphone (i.e., Sennheiser HD 280 Pro). Lastly, the time gap between conditions II and I in Kopun et al. (2022) ranged from 2 years 6 months to 2 years 9 months while our time gap was less than a day. Overall, these differences not only in the experimental setup but also in the data processing would explain why our study exhibits a higher reproducibility than the earlier study, indicated by a smaller mean signed difference.

The descriptive parameters (i.e., HTL, MLL, UCL, and DR) of our study measured with a smartphone for NH listeners match quite well with the reference values reported in Oetting et al. (2016). The mean difference of the 4 parameters between Oetting et al. (2016) (N = 9) and our results is less than 2 dB at 0.25 kHz while lying within one standard deviation at 1 and 4 kHz. Furthermore, our measured MLLs and DRs are quite consistent with the empirical values for young NH listeners (N = 11) and HI listeners (N = 70) provided by Sanchez-Lopez et al. (2021). The median MCLs and DRs of NH listeners reported by Sanchez-Lopez et al. (2021) were 70 and 97.5 dB HL at low frequencies, and 75 and 92.5 dB HL at high frequencies while the median MCLs and DRs of listeners measured in the current study were 73.5 and 103.5 dB HL for low frequencies, and 78.7 and 90.6 dB HL for high frequencies. The difference between Sanchez-Lopez et al. (2021) and our study is around 5-6 dB and relatively small. Comparing the HI listeners of Sanchez-Lopez et al. (2021), most of our measured parameters for both low and high frequencies stay within the 25% and 75% percentile range of Sanchez-Lopez et al. (2021) except for MCLs at high frequencies. One possible reason might be different high frequency measurements: we only measured 4 kHz while Sanchez-Lopez et al. (2021) measured 2, 4, and 6 kHz and averaged the values of MCL. Another explanation could be that individual (within-subject) preference for MCLs might vary. Overall, the descriptive parameters measured by a smartphone show good consistency with the empirical values reported in the literature for both NH and HI listeners.

The three conditions differing in degree of supervision with calibrated hardware appear not to systematically influence the results of CLS in terms of both loudness growth functions and derived parameters (as shown in Fig. 3.2 and revealed by the mix-designed ANOVA), implying that we could let the participants test themselves on a smartphone for the CLS test, which meets our expectations. One reason to explain the results might be that the task for loudness judgment is rather intuitive and natural based on the feedback from our participants. In addition, CLS is a supra-threshold measurement, which is expected to be less prone to influence by factors such as hardware and environment. Unlike some other speech-related tasks, e.g., the speech-in-noise test or listening effort test which are rather cognitively demanding, the CLS task does not involve speech comprehension, and, therefore, should be rather robust without any additional assistance from experimenters.

3.4.3 Binaural and spectral loudness summation

Level differences for equal loudness (LDELs)—that quantify the binaural and spectral loudness summation—mostly do not show differences between the standard in-lab and smartphone measurements. This indicates that the smartphone measurements could detect the binaural and spectral loudness summation as well as the assessment conducted in a laboratory. However, the ANOVA and post-hoc t-tests revealed that supervision significantly influenced LDEL in certain groups for spectral loudness summation. Since these significant differences mainly occur between the reference condition (always measured first) and one of the less supervised conditions, they are most likely due to training or adaptation effects rather than the type of test supervision. This suggests that either a proper familiarization phase should be implemented prior to data collection or the testing conditions should be randomized to avoid order effects.

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

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

showed that LDELs were approximately 10 dB and decreased with lower loudness, which is in line with our results.

Loudness scaling and loudness matching appear to be the two main tools to assess loudness summation for practical applications. Van Beurden et al. (2021) compared the two measurement procedures and concluded that both procedures provided valid and reliable results. Loudness scaling, on one hand, provides information on the entire loudness range. It requires a simple categorical judgment task, which is quite intuitive even for the elderly and naïve participants while loudness matching is less intuitive and needs more instructions for the listeners who have to "equalize apples and pears", i.e., are forced to judge two differently perceived stimuli as being equal in one domain which is a challenge for inexperienced persons. On the other hand, loudness scaling might be more time-consuming than loudness matching. Even though we do not systematically compare the two methods on a smartphone, we prefer to apply loudness scaling on mobile devices since the feedback from our participants indicates that it is rather straightforward and easy to measure while using an acceptable measurement time.

3.4.4 Individual variability

Significant individual variations in loudness perception among hearing-impaired listeners have been observed in conventional laboratory assessments (e.g., Oetting et al., 2016; Van Beurden et al., 2018; 2021), indicating that even for listeners with a similar PTA, the range of individual uncomfortable loudness levels or LDEL can vary by as much as 20 dB. This motivated the introduction of modern fitting concepts like 'true loudness fitting' (Oetting et al., 2016; 2018) that aim at partially compensating for the large individual differences in binaural spectral loudness summation in hearing-impaired listeners. A similar finding could be derived from our data (please see Figs. 3.S2 and 3.S3 in the supplementary materials, where individual LDELs are plotted as a function of PTA for normal-hearing and hearing-impaired listeners). Given this high variability and the limited predictability from other measures like the PTA, it is desirable to measure individual loudness growth and loudness summation for improving individual hearing aid fitting (Oetting et al., 2018) with an easy-access method like the mobile-device-based test described here.

3.4.5 Limitations and outlook

Our current study only considers conducting the smartphone measurements in a sound-treated booth in order to eliminate any effects of the environment on the measurement outcome (e.g., distraction or background noise). It is worthwhile to consider experiments outside the booth while still ensuring the quality of the audiometric data. A possible solution could be monitoring the real-time noise level during the measurement as Kopun et al. (2022), Swanepoel et al. (2014; 2015), Maclennan-Smith et al. (2013), and Serpanos et al. (2022) did. Another approach for out-of-booth measurement could be using noise cancellation earphones (e.g., Clark et al., 2017). Moreover, in the current study, only participants with mild-to-moderate hearing loss were considered.

The headphone employed here is a professional audiometric headphone (Sennheiser HDA200), which appears to be expensive and not publicly accessible. Van der Aerschot et al. (2016) recommended that affordable headphones, e.g., Sennheiser HD202 could be applied for pure-tone audiometry assessment. The true wireless stereo (TWS) earbuds for pure-tone audiometry introduced by Guo et al. (2021) could also be considered as a daily-accessible alternative to the audiology headphone.

In our current study, we calibrated the smartphone output accurately in order to eliminate the influence of calibration and make it comparable to the standard laboratory measurement. However, in everyday life, the smartphone is normally not calibrated. How to treat the uncalibrated mobile device and additional hardware in non-laboratory setups remains a challenge. Kisić et al. (2022), for instance, proposed that human speech might be an appropriate and stable test signal for microphone calibration while Scharf et al. (2024) considered the whistling sound of a 0.33 l beer bottle as a rough calibration signal. On the other hand, most of the (diagnostic) parameters from loudness scaling derived here are rather independent from an exact absolute presentation level calibration as they primarily consider level difference measures (like the dynamic range DR or the level difference for equal loudness LDEL) or show a common individual calibration offset that can easily be compensated if sufficient reference data for similar individual cases are available.

3.5 Conclusions

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

- **Experiment I (Pure-tone Audiometry using the SIUD procedure)** indicates that the method of supervision does not influence the measurement outcome.

- **Experiment II (Adaptive CLS)** reveals that supervision does not affect the outcome values of categorical loudness scaling (i.e., the derived loudness growth functions of NH and HI listeners). The bias between smartphone and in-lab loudness function is small while the 5 intuitive parameters (i.e., HTL, MCL, MLL, UCL, and DR) of smartphone CLS do not differ from the standard CLS assessment. Note that for most of these parameters, no calibration of the mobile device was required. .

- **Experiment III (binaural and spectral loudness summation)** implies that binaural and spectral loudness summation can be derived by employing a smartphone in a way consistent with lab experiments. Furthermore, the individual variations of HI listeners in loudness summation at high (i.e., uncomfortable) levels for binaural broadband signals are considerably large, which is not predictable from the average audiogram. Therefore, incorporating binaural broadband signals into loudness perception assessments is a desirable step for optimizing hearing aid fittings, as it can enhance the outcomes for aided listeners (Oetting et al., 2016; 2018).

In conclusion, both audiometric tests considered here can be used for non-supervised smartphone-based hearing examination and are expected to yield very similar results as being conducted in a controlled laboratory experiment.

4 Influence of ambient noise[3]

Abstract

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

---

[3] This section is a formatted reprint of

Xu, C., Schell-Majoor, L., & Kollmeier, B. (2024). Feasibility of efficient smartphone-based threshold and loudness assessments in typical home settings. medRxiv, 2024-11. https://doi.org/10.1101/2024.11.19.24317529

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

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). 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.

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 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.

4.2 Methods

4.2.1 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).

4.2.2 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.

4.2.3 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.

4.2.4 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. 4.1), 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. 4.1. 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.

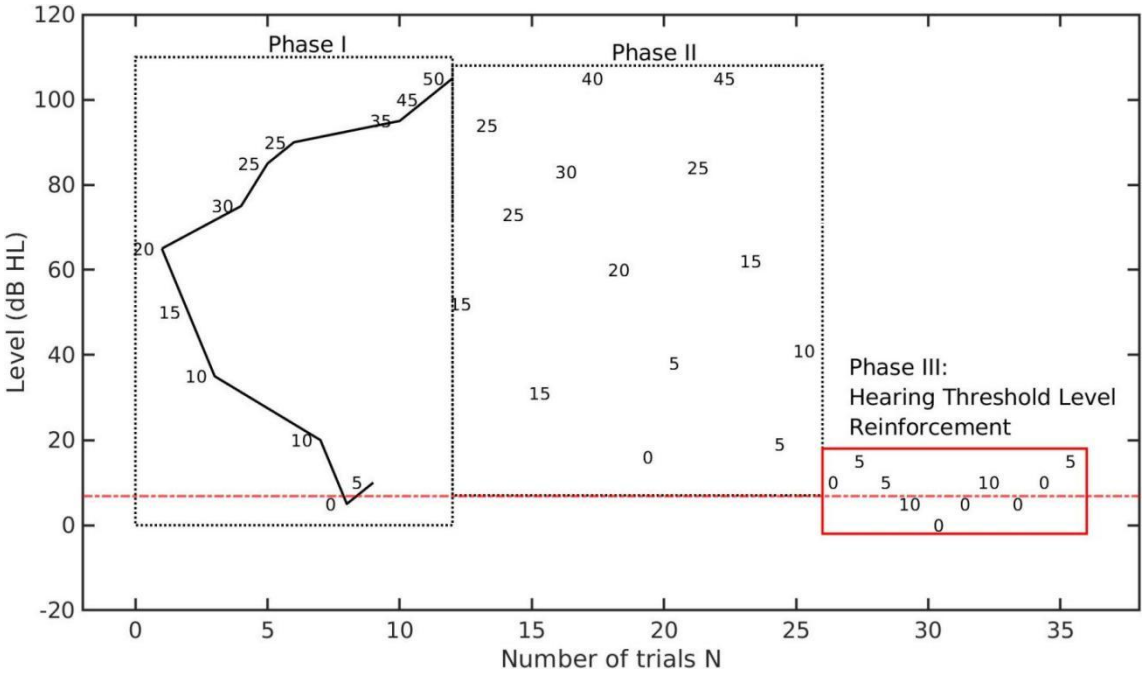

Fig. 4.1. An example track of the reinforced adaptive categorical loudness scaling (rCALOS), where the level (in dB HL) is plotted as a function of the number of trials N. The listener's response (in categorical units (CU)) is annotated with numbers between 0 ('not heard') and 50 ('too loud'). Left dotted 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.

4.2.5 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_{cut}, 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.

4.2.6 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 4.S1.

4.3 Results

4.3.1 Noise level measurements

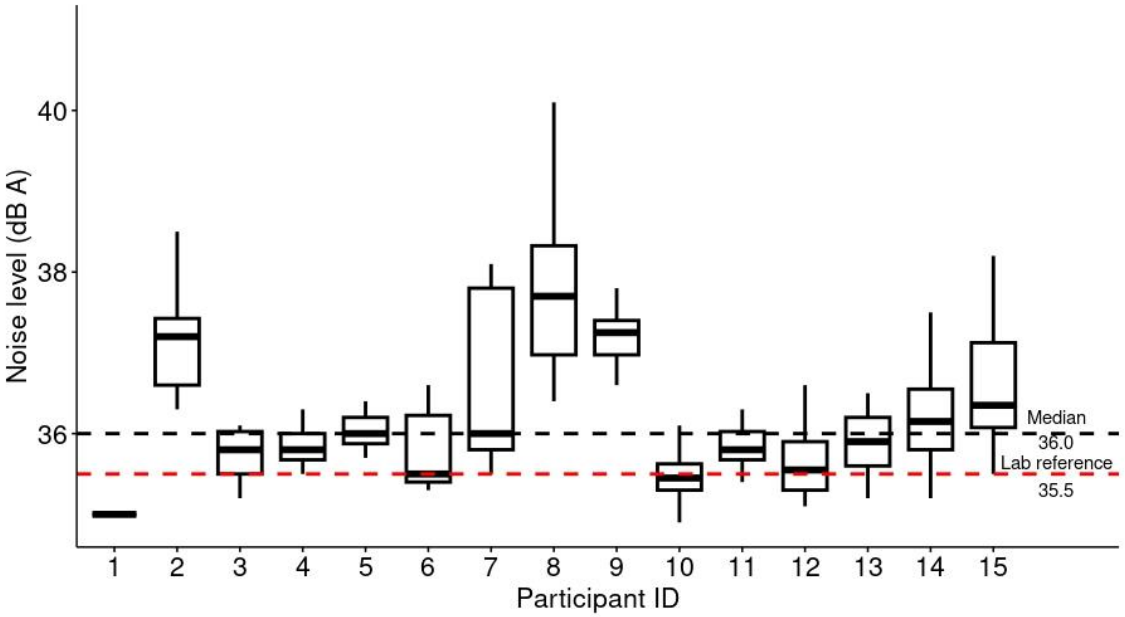

Fig. 4.2. Ambient noise level (in dB A) measurement across participants (N = 15). Medians, 25th and 75th percentiles, and interquartile ranges (IQR) are visualized in the box-plot while the end of the whiskers denotes the minimum and maximum, indicating the 5th and 95th percentiles respectively. Red dashed line: lab reference (i.e., ambient noise level measured within a booth). Black dashed line: median value across subjects.

Fig. 4.2 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.

4.3.2 Validation experiment

GRaBr

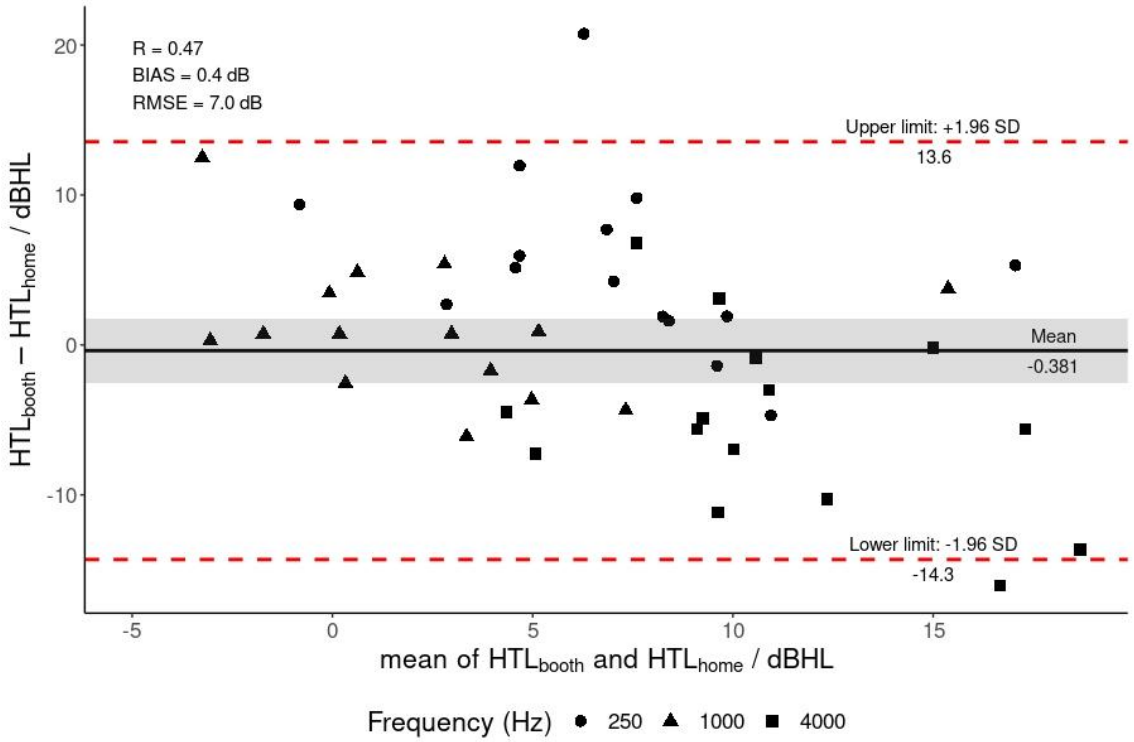

Fig. 4.3. Bland-Altman plot of hearing threshold levels (HTL) in dB HL of frequencies at 0.25, 1, and 4 kHz (represented with circle, triangle, and rectangle, respectively) measured inside the booth (i.e., HTLbooth) using the standard audiometry and at home (i.e., HTLhome) using the GRaBr procedure. Red dashed lines: 95% level of agreement; Black solid line: bias between the two measurement environments; Grey shaded rectangle area: 95% confidence interval of the bias. The correlation coefficient (R), bias (BIAS), and root mean squared error (RMSE) are provided in the top-left corner.

Fig. 4.3 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. 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 HTLbooth and HTLhome 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 Fig. 4.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.

rACALOS

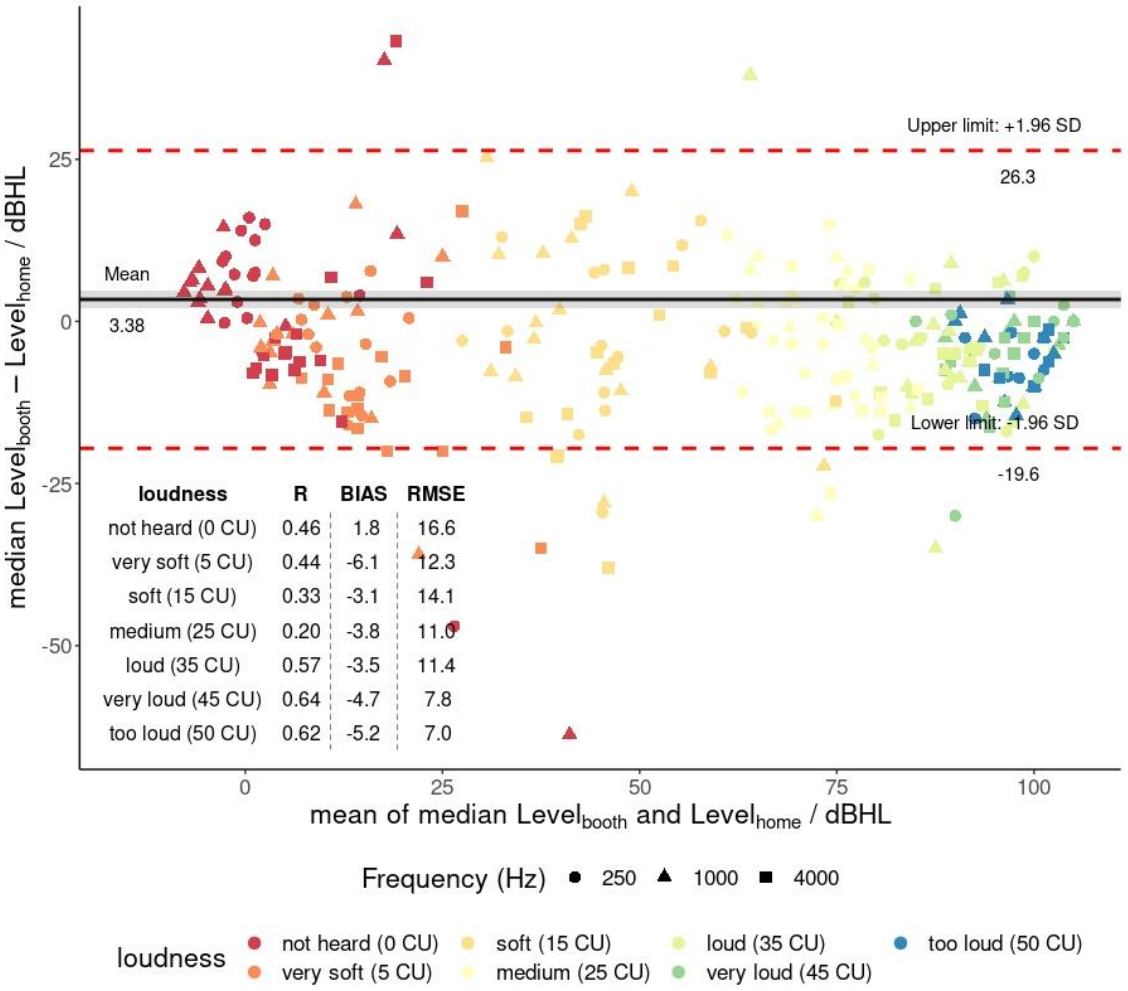

Fig. 4.4. Bland-Altman plot of median levels assigned to each CU (denoted with different colors) for three frequencies (represented with different shapes) for comparing two test environments, i.e., inside the booth using a standard CLS procedure and at home using the rACALOS procedure for each participant. 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. 4.3 for an explanation of the Bland-Altman plot and Supplementary Materials 4.S1 for its statistical implication.

Fig. 4.4 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. 4.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.

4.3.3 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 4.S2 and Table 4.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 4.S3 and Table 4.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. 4.1).

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).

4.3.4 Accuracy of HTL estimation for the rACALOS procedure

Computer simulations

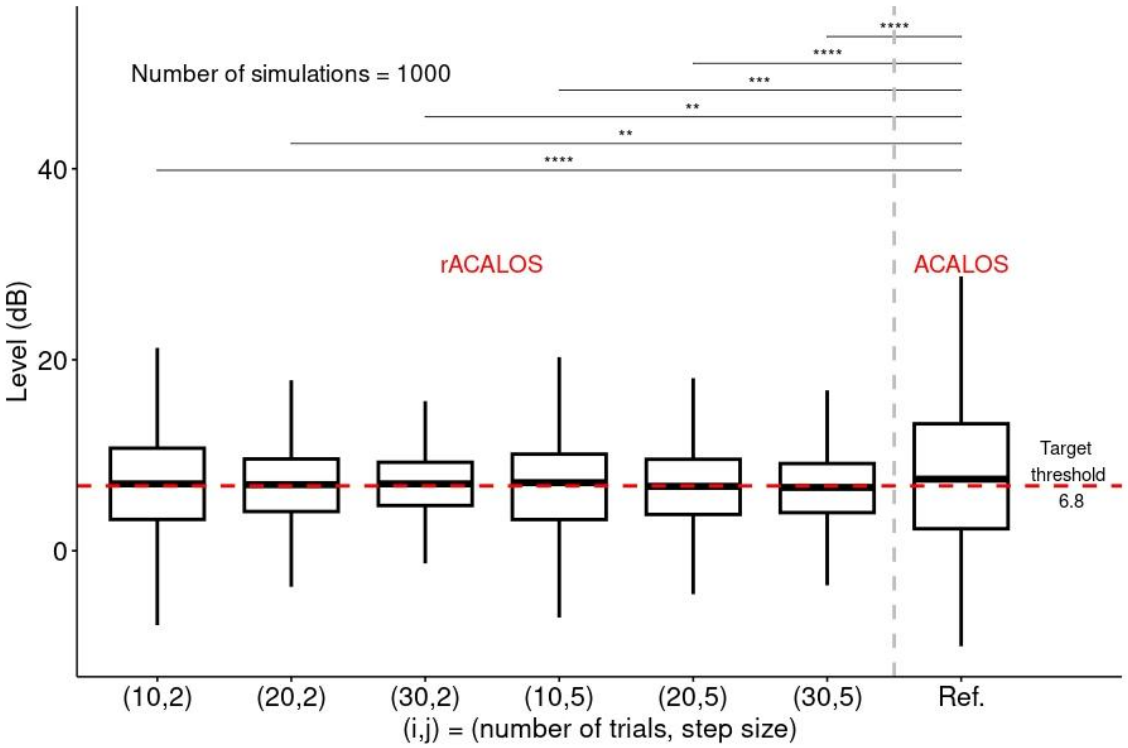

Fig. 4.5. Estimated hearing thresholds in dB (i.e., level of the loudness growth function corresponding to 2.5 CU) of rACALOS variants (to the left of the vertical dashed line) using reinforcement at the hearing threshold level obtained with Monte-Carlo simulations (N = 1000 runs) in comparison to the 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. 4.2 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. 4.5. 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.

Behavioral experiments

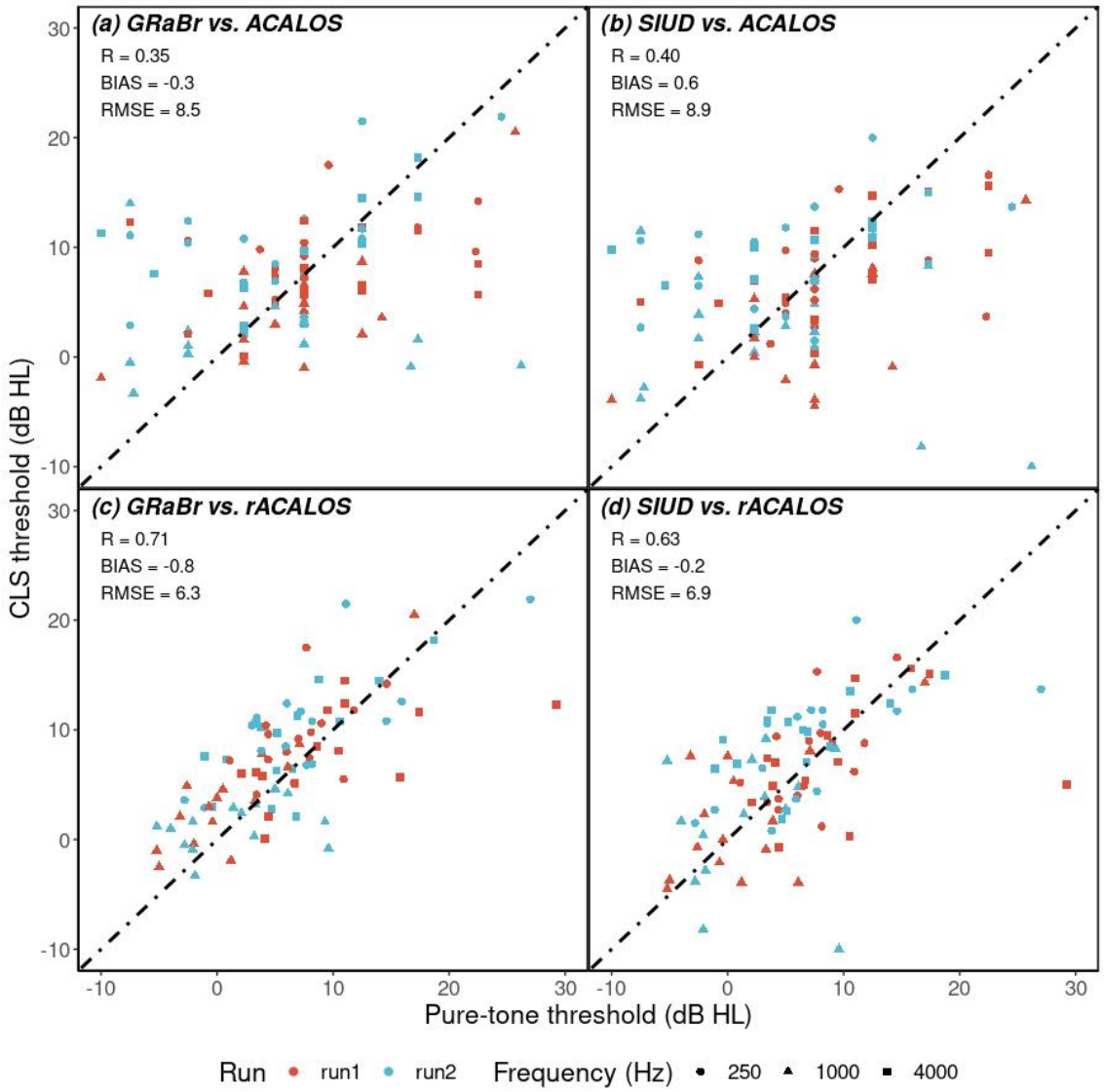

Fig. 4.6. Scatter plot for comparison between pure-tone (abscissa) and CLS (ordinate) thresholds in dB HL of N = 15 individual listeners. Frequency is labeled with different shapes while the run is denoted with different colors (run1: red, run2: blue). A set of statistical metrics (R, Bias, and RMSE) 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. 4.6. 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 ($p < 0.05$).

4.4 Discussion

4.4.1 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).

4.4.2 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 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.

4.4.3 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.

4.4.4 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. 4.1). 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 4.1 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-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 4.1. 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	CLS method	N	R (spearman)	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**

4.4.5 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 4.1 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.

4.4.6 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.

4.5 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.

5 General discussion and conclusions

Chapter 1 of this thesis provides a general introduction to the challenges, limitations, and background of smartphone-based hearing tests. Chapters 2, 3, and 4 systematically evaluate the influence of various factors—such as inattention, supervision methods, and ambient noise—on the accuracy, reliability, and efficiency of these tests. The current chapter discusses two key topics: (1) the validity, test-retest reliability, and efficiency of smartphone-based listening tests, and (2) the selection of appropriate listening tests for mobile platforms. Subsequently, two possible applications of smartphone-based listening tests are presented: determining auditory profiles and establishing a national hearing health cohort. Finally, the chapter concludes with a discussion of the study's limitations and potential directions for future research.

- Validity, test-retest reliability, and efficiency of smartphone-based listening tests

A. Validity

Both smartphone-based auditory tests—pure-tone audiometry and categorical loudness scaling—demonstrate strong validity when compared to standard listening tests (see Chapters 3 and 4 for details). Specifically, the threshold differences at all three tested frequencies between smartphone-based and standard audiograms are within 5 dB for both normal-hearing and hearing-impaired listeners, as discussed in Chapter 3. Furthermore, Chapter 4 highlights that the smartphone-based pure-tone audiometry maintains high validity, with threshold differences of less than 2 dB, even when conducted in a home environment with ambient noise. Overall, the results of these validation experiments align with previous studies (Swanepoel et al., 2014; Yousuf Hussein et al., 2016; Hazan et al., 2022). This consistency can be attributed to the use of model-free, precise adaptive procedures for threshold estimation, rigorous device calibration, and effective control of environmental noise.

Moreover, the differences between the smartphone-based and standard adaptive categorical loudness scaling (ACALOS) methods, measured at 0.25, 1, and 4 kHz, are less than 5 dB for both normal-hearing and hearing-impaired listeners. These

differences are not statistically significant, as reported in Chapters 3 and 4, indicating good validity. Chapter 3 details the validation experiments conducted in a sound-treated booth, while Chapter 4 presents validation results obtained in a home environment. Notably, the validity observed in this study is significantly better than that reported in previous studies, such as Kopun et al. (2022). In our study, the systematic bias between the smartphone-based and standard ACALOS methods, used to quantify validity, is approximately 2–3 dB. This bias is smaller than that reported by Kopun et al. (2022), who observed a bias exceeding 5 dB. Several factors may contribute to this improved validity. First, the test environments in our study were generally quieter, with ambient noise levels approximately 10 dB lower than those reported in earlier studies. Second, we employed the novel and optimized approach for fitting loudness growth functions introduced by Oetting et al. (2014), which was not used in most previous studies. This method effectively removes outliers, ensuring an individual, monotonic loudness function. These advancements likely enhance the accuracy of smartphone-based ACALOS measurements, resulting in higher validity compared to previous research.

B. Test-retest reliability

Chapter 4 presents the test-retest reliability results for smartphone-based pure-tone audiometry and ACALOS tests. Both smartphone-based listening tests demonstrated relatively high reliability. For normal-hearing participants, the intraclass correlation coefficient (ICC) values between test and retest measurements for the smartphone-based audiograms at 0.25, 1, and 4 kHz were generally greater than 0.6, aligning with findings in the existing literature. This consistency can be attributed to controlled experimental conditions, including the use of an accurate adaptive procedure, precise device calibration, and controlled ambient noise levels. The smartphone-based ACALOS test in this study exhibited significantly higher reliability compared to earlier studies (e.g., Kopun et al., 2022; Rasetshwane et al., 2015; Fultz et al., 2020). This improvement is evidenced by mean signed differences of less than 5 dB in the current study between test and retest measurements across all three test frequencies for normal-hearing listeners whereas other studies (e.g., Kopun et al., 2022) have reported mean signed differences that are significantly greater than 5 dB. Similar to the validation experiments discussed earlier, two key factors likely contribute to this increased reliability: reduced environmental noise and the application of an optimized fitting

method. However, since this study evaluates test-retest reliability only in normal-hearing listeners, future research should consider including hearing-impaired listeners.

C. Efficiency

Chapter 2 compares the efficiency of smartphone-based pure-tone audiometry across different simulated listeners and adaptive approaches. Audiogram measurements using the GRaBr adaptive procedure were significantly more efficient than standard clinical audiogram measurements, as evidenced by a higher normalized efficiency index and lower convergence rates (see Xu et al., 2024a for details). The clinical audiogram procedure demonstrates high efficiency only under ideal conditions, such as when participants are fully attentive or produce low false alarm rates. However, in more practical scenarios—such as mobile-device-based listening tests—GRaBr proves to be more robust and efficient, as it is less influenced by variations in participant behavior. Overall, we recommend the use of GRaBr for smartphone-based pure-tone audiometry (see subsequent sections for the selection of smartphone-based audiogram measurements). However, a direct comparison of efficiency between smartphone-based and standard ACALOS tests has not yet been conducted. Future studies should address this gap to evaluate the efficiency of smartphone-based ACALOS tests.

- Selection of listening tests for mobile testing

First, we recommend integrating GRaBr into smartphone-based audiogram testing within the Virtual Hearing Clinic (VHC) if an audiogram is the desired outcome. Traditionally, the audiogram has been considered as a fundamental tool for detecting hearing loss and, to a lesser extent, for supporting audiogram-based hearing aid fittings (Kollmeier & Kiessling, 2018). Moreover, it facilitates the generation of audiogram-based auditory profiles, which are essential for profile-based hearing device fittings (Sanchez-Lopez et al., 2022; Wu et al., 2022)."

However, the lack of calibration and the influence of external noise could pose challenges for conducting audiogram testing within the VHC. As a result, relative measures, such as speech-in-noise tests, might be more suitable for this setting. Additionally, the audiogram is a threshold determination method that inherently assumes the auditory system to be linear, making it potentially less appropriate for

functional hearing tests aimed at assessing everyday hearing abilities. Consequently, speech-in-noise tests are often regarded as a better alternative, particularly for self-screening hearing tests (further details are provided below). Clinically, it is also unnecessary to measure thresholds at all audiometric frequencies independently, given the high degree of mutual dependency and redundancy across frequencies. In this context, parametric audiogram estimation methods (e.g., Schlittenlacher et al., 2018) are particularly appealing, as they efficiently estimate the individual audiogram in a minimal amount of time (e.g., less than 5 minutes). However, if threshold measurements at specific frequencies are required, the GRaBr method can be employed.

GRaBr, introduced in Chapter 2 and validated with human participants in Chapter 4, is a model-free adaptive procedure that is both efficient and robust to inattention. In contrast, many model-based adaptive procedures (e.g., the maximum likelihood procedure) are highly sensitive to the lapse rate of the psychometric function. In scenarios such as smartphone-based listening tests, where participants' attention cannot be reliably monitored or the false alarm rate is likely to be high, GRaBr offers a clear advantage over model-based methods. Furthermore, our findings show that GRaBr significantly outperforms the baseline single-interval up-and-down (SIUD) method in robustness, efficiency, and reliability (see Chapters 2 and 4 for details). As a result, among model-free adaptive procedures, GRaBr is the preferred choice over the SIUD method. Overall, we recommend adopting GRaBr for smartphone-based audiogram measurements if a classical audiogram measure with independent frequency measurement is desired.

Second, we propose integrating smartphone-based loudness assessments using the reinforced adaptive categorical loudness scaling (rACALOS) approach within the framework of the Virtual Hearing Clinic. As an advanced auditory assessment method, loudness tests not only provide information about basic audibility but also yield critical supra-threshold parameters, such as loudness discomfort levels and dynamic range. These parameters are pivotal for loudness-based hearing aid fittings (Kollmeier & Kiessling, 2018). Additionally, recent studies have increasingly utilized categorical loudness scaling (CLS) to derive auditory profiles (Saak et al., 2022; Sanchez-Lopez et al., 2018; 2020). Considering its substantial utility in auditory research, we recommend incorporating CLS into smartphone-based auditory assessments.

The rACALOS approach, developed and validated in Chapter 4, integrates threshold measurements directly into the CLS procedure. This novel, unified method significantly enhances the accuracy of hearing threshold estimation compared to the original ACALOS method, aligning more closely with standard audiometric thresholds (see Chapter 4 for a detailed comparison). Furthermore, rACALOS improves time efficiency by combining two auditory assessment tools into a single protocol. Its user interface remains identical to that of ACALOS, eliminating the need for additional participant training for those already familiar with the original method. The rACALOS is relatively insensitive to calibration errors and can effectively serve as a calibration-robust, relative measure. Specifically, several rACALOS outcomes, such as the dynamic range and the slope of the loudness growth functions, are unaffected by calibration as they represent relative quantities. However, other parameters, such as the medium loudness level L_{25}, are influenced by calibration. A consistency check of individually measured parameters (e.g., L_{25} with an unknown calibration offset) against expected values derived from calibration-independent parameters (e.g., dynamic range or slope) is expected to enhance robustness against calibration errors. Additionally, the dependence of rACALOS measurements across frequencies is high. Hence, parametric approaches (e.g., Schlittenlacher & Moore, 2020) for estimating equal-loudness contours across frequencies could be employed to enhance measurement efficiency. Given these advantages, we advocate for the adoption of rACALOS in smartphone-based CLS measurements.

Third, speech tests, such as the Matrix Sentence Test (see Kollmeier et al., 2015), could be integrated into the Virtual Hearing Clinic, although they are not addressed in this thesis. These tests assess the speech recognition threshold (SRT) in individuals with hearing loss by systematically varying the speech-to-noise ratio (SNR), thereby evaluating speech understanding in noisy environments (Akeroyd et al., 2015). Previous studies have demonstrated the feasibility of conducting the digits-in-noise test (Potgieter et al., 2016) and the matrix sentence test (Saak et al., 2024) on smartphone platforms. Notably, Saak et al. (2022) derived 13 distinct auditory profiles based on the outcomes of speech tests in combination with only few other audiometric tests. Moreover, Saak et al. (2024) compared different user interfaces for smartphone-based matrix sentence tests. Additionally, speech tests are generally independent of device calibration, particularly when used with smartphones and headphones (Almufarrij et al., 2022).

Given their critical role in assessing speech comprehension and generating auditory profiles (Saak et al., 2022), as well as their independence from calibration requirements, integrating speech tests into the Virtual Hearing Clinic would be a valuable addition.

- Towards a useful tool to determine auditory profiles and to establish a national hearing health cohort

This section explores two key applications of smartphone-based listening tests developed within the Virtual Hearing Clinic: generating auditory profiles and building a national hearing cohort.

Auditory profiles have recently garnered significant attention in the auditory research field. An auditory profile is defined as a participant-specific variable that relates to data-driven, precise classification of individuals, particularly within the hearing-impaired population. Contrary to the simple and superficial audiogram-based classification in audiology, which categorizes individuals as normal hearing or as having mild, moderate, severe, or profound hearing loss, auditory profiles provide a more nuanced approach. This aligns with the principles of precision audiology, particularly when suprathreshold measures are utilized. Traditionally, auditory profiles have been derived primarily from audiometric data, such as the well-known "Bisgaard profiles" (Bisgaard et al., 2010). However, recent advancements have incorporated supra-threshold parameters into auditory profile generation (e.g., Saak et al., 2022; Sanchez Lopez et al., 2018; 2020; 2022; Van Esch et al., 2013). According to Saak et al. (2022), accurately classifying a participant into one of the auditory profiles may be both sufficient and efficient using only a few measures.

The smartphone-based listening tests described in this study enable the determination of various auditory profiles outlined in the literature by capturing only a few key auditory parameters from individual participants. These auditory profiles hold promise for several critical applications. Participants can be accurately classified into distinct auditory profiles despite the use of only a few imprecise measures. This helps to support profile-based hearing device fitting, as emphasized in recent work by Sanchez Lopez et al. (2022). By leveraging the capabilities of mobile technology, this approach

is expected to bridge the gap between advanced auditory diagnostics and accessible, individualized hearing care solutions. Furthermore, auditory profiles can be linked with genetic profiles to enhance understanding of the causes and consequences of hearing loss, enabling more precise diagnostics and tailored treatments (Hochmuth et al., 2024).

Smartphone-based listening tests offer easy access for participants and can efficiently gather data from a large sample, resulting in a comprehensive mobile database. Traditional data collection in clinical or laboratory settings, as described in previous studies (e.g., OHHR – The Oldenburg Hearing Health Repository; Jafri et al., 2024), is typically time-intensive and costly. As a result, obtaining a substantial dataset often requires several years. Additionally, research-based measures may significantly extend measurement time per patient beyond routine assessments. In contrast, smartphone-based data collection provides a cost-effective alternative, reducing the need for specially trained professionals and specialized equipment by leveraging mobile devices. Participants can perform self-administered listening tests remotely from their homes, making the process more efficient (see Chapter 4). Furthermore, performing a preliminary self-driven classification of the auditory profile based on a few key parameters is also efficient. Importantly, the validity and reliability of the data collected via smartphones could match that of traditional clinical methods (see Chapters 3 and 4). Overall, this novel approach enhances hearing screening by making it more efficient. By providing easy access for participants, it enables data collection outside traditional clinical or laboratory settings, such as in home environments. This accessibility not only improves the ecological validity of assessments—for instance, through ecological momentary assessment (EMA)—but also facilitates the investigation and monitoring of long-term hearing status in longitudinal studies, considering that speech perception can vary daily (Kuhlmann et al., 2023).

- Limitations and future work

Chapters 2, 3, and 4 specifically examine the effects of inattention, supervision, and ambient noise, respectively, on the performance of smartphone-based listening tests. Beyond the major factors discussed before, several other factors significantly impact smartphone-based hearing assessments. First, unlike the calibrated equipment used in acoustics labs, participants' smartphones are typically uncalibrated, leading to potential

inaccuracies in auditory stimulus presentation. The variety of mobile devices and earphones among subjects necessitates a general approach for remote calibration (e.g., Scharf et al. (2024) proposed using resonating bottles as a method for microphone calibration in mobile audiological testing). Second, financial incentives are worth noting: Bianco et al. (2021) reported that rewards enhance participant engagement in remote hearing assessments, especially for demanding auditory tasks. Third, headphone placement can affect the accuracy of smartphone-based tests. Unlike in-lab tests where audiologists ensure proper headphone placement, smartphone-based tests lack expert assistance (Paquier et al., 2016; Sørensen et al., 2023). Unfortunately, these factors were not systematically investigated, as they are beyond the scope of the thesis. Instead, our focus was directed toward the major influencing variables. Below, we outline additional listening tests that could enhance the diagnostics module of the Virtual Hearing Clinic.

Additional listening tests, such as listening effort assessments (e.g., the Adaptive Categorical Listening Effort Scaling [ACALES]; Krueger et al., 2017), could be incorporated and validated in the Virtual Hearing Clinic (VHC) to quantify the attentional resources required for listening tasks, which may be relevant for mobile listening tests. Currently, the user interface for input is a button on a smartphone. Considering an automatic speech recognition (ASR) based approach (as detailed in Ooster et al., 2020) might be beneficial, allowing participants to speak their answers directly into the app instead of clicking a button. The speech-based interface is highly convenient and user-friendly, particularly for individuals who cannot read, such as children, visually impaired individuals, or those who are illiterate. Furthermore, a text-to-speech (TTS) technique could be employed to generate test stimuli, as TTS-generated stimuli are believed to provide equal intelligibility among words (see more details in Ibelings et al., 2022). According to Ibelings et al. (2022), using TTS systems offers several benefits: first, it can significantly simplify the process of developing speech tests. Second, it is more cost-effective, as it eliminates the need to hire a (professional) speaker or purchase recording equipment. Although ASR and TTS techniques provide notable advantages for the development and implementation of speech-in-noise tests (Polspoel et al., 2024), they fall outside the scope of this thesis. Therefore, they are not considered in the current study but should be explored in future research.

Moreover, an appropriate combination of auditory tests is essential for classifying participants into specific auditory profiles. Efficiently assigning participants to these profiles could be achieved by leveraging a few key parameters derived from the combination of (mobile) auditory tests. Proposing such a method is significant because the auditory profile typically encompasses more comprehensive information, including audibility, loudness perception, and speech comprehension, than single listening tests. Additionally, profile-based hearing aid fitting could provide more robust treatment and intervention for hearing loss, becoming one of the future focal points in the auditory field.

Last but not least, the development and validation of a treatment recommendation module for the Virtual Hearing Clinic may be explored in future work, as this thesis primarily focuses on the diagnostic module. The open Master Hearing Aid (openMHA), which includes various real-time hearing aid signal processing algorithms, that could be integrated into the current smartphone-based application to simulate a virtual hearing aid for participants, compensating for hearing loss (Kayser et al., 2022). Additionally, AI-based algorithms for speech enhancement or hearing aid outcome prediction could be incorporated in future iterations (Schädler et al., 2020).

Finally, we have demonstrated that smartphone-based listening tests exhibit good validity, high test-retest reliability, and excellent efficiency. The minimal set of tests for mobile auditory assessment should include speech-in-noise tests and reinforced adaptive categorical loudness scaling tests. However, inattentive participants can compromise the robustness and efficiency of these tests. Notably, supervision does not significantly impact the performance of smartphone-based listening tests. When ambient noise levels are sufficiently low, the results from smartphone-based tests are comparable to those obtained in laboratory settings. Overall, this approach represents a valuable foundation for future applications, such as providing fitting parameters for the treatment recommendation module of the Virtual Hearing Clinic (VHC).

Reference List

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

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

Amitay, S., Irwin, A., Hawkey, D. J., Cowan, J. A., and Moore, D. R. (2006). "A comparison of adaptive procedures for rapid and reliable threshold assessment and training in naive listeners," The Journal of the acoustical society of America 119(3), 1616–1625.

Audiffren, J., & Bresciani, J. P. (2022). Model Based or Model Free? Comparing Adaptive Methods for Estimating Thresholds in Neuroscience. Neural Computation, 34(2), 338-359.

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.

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

Bisitz, T., and Silzle, A. (2011). "Automated pure-tone audiometry software tool with extended frequency range," in Audio Engineering Society Convention 130, Audio Engineering Society.

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

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

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

Brand, T., and Kollmeier, B. (2002). "Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests," The Journal of the Acoustical Society of America 111(6), 2801–2810.

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.

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

Cameron, E. L., Tai, J. C., and Carrasco, M. (2002). "Covert attention affects the psychometric function of contrast sensitivity," Vision research 42(8), 949–967.

Christensen, J. H., Pontoppidan, N. H., Rossing, R., Anisetti, M., Bamiou, D. E., Spanoudakis, G., ... & Ecomomou, A. (2019). Fully synthetic longitudinal real-world data from hearing aid wearers for public health policy modeling. Frontiers in Neuroscience, 13, 850.

Christensen, J. H., Saunders, G. H., Havtorn, L., & Pontoppidan, N. H. (2021). Real-world hearing aid usage patterns and smartphone connectivity. Frontiers in Digital Health, 3, 722186.

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

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

Corona, A. P., Ferrite, S., Bright, T., & Polack, S. (2020). Validity of hearing screening using hearTest smartphone-based audiometry: performance evaluation of different response modes. International journal of audiology, 59(9), 666-673.

Cruickshanks, K. J., Nondahl, D. M., Fischer, M. E., Schubert, C. R., & Tweed, T. S. (2020). A novel method for classifying hearing impairment in epidemiological studies of aging: The Wisconsin age-related hearing impairment classification scale. American Journal of Audiology, 29(1), 59-67.

De Sousa, K. C., Moore, D. R., Smits, C., & Swanepoel, D. W. (2021). Digital technology for remote hearing assessment—current Status and future directions for consumers. Sustainability, 13(18), 10124.

Doll, R. J., Veltink, P. H., and Buitenweg, J. R. (2015). “Observation of time-dependent psychophysical functions and accounting for threshold drifts,” Attention, Perception, & Psychophysics 77(4), 1440–1447.

D'Onofrio, K. L., & Zeng, F. G. (2022). Tele-audiology: Current state and future directions. Frontiers in Digital Health, 3, 788103.

Dubno, J. R., Eckert, M. A., Lee, F. S., Matthews, L. J., & Schmiedt, R. A. (2013). Classifying human audiometric phenotypes of age-related hearing loss from animal models. Journal of the Association for Research in Otolaryngology, 14, 687-701.

Elberling, C. (1999). Loudness scaling revisited. Journal of the American Academy of Audiology, 10(05), 248-260.

Erinc, M., & Derinsu, U. (2022). Behavioural and Electrophysiological Evaluation of Loudness Growth in Clinically Normal Hearing Tinnitus Patients with and without Hyperacusis. Audiology and Neurotology, 27(6), 469-477.

Fereczkowski, M., & Neher, T. (2023). Predicting Aided Outcome With Aided Word Recognition Scores Measured With Linear Amplification at Above-conversational Levels. Ear and Hearing, 44(1), 155-166.

Fultz, S. E., Neely, S. T., Kopun, J. G., & Rasetshwane, D. M.(2020). Maximum expected information approach for improv-ing efficiency of categorical loudness scaling.Frontiers in Psy-chology, 11,32–63. https://doi.org/10.3389/fpsyg.2020.578352

Gathman, T. J., Choi, J. S., Vasdev, R. M., Schoephoerster, J. A., & Adams, M. E. (2023). Machine Learning Prediction of Objective Hearing Loss With Demographics, Clinical Factors, and Subjective Hearing Status. Otolaryngology–Head and Neck Surgery.

Gescheider, G. A. (2013). Psychophysics: the fundamentals. Psychology Press.

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

Gieseler, A., Tahden, M. A., Thiel, C. M., Wagener, K. C., Meis, M., & Colonius, H. (2017). Auditory and non-auditory contributions for unaided speech recognition in noise as a function of hearing aid use. Frontiers in psychology, 8, 219.

Grassi, M., and Soranzo, A. (2009). "Mlp: A matlab toolbox for rapid and reliable auditory threshold estimation," Behavior research methods 41(1), 20–28.

Green, D. M. (1990). "Stimulus selection in adaptive psychophysical procedures," The Journal of the Acoustical Society of America 87(6), 2662–2674.

Green, D. M. (1993). "A maximum-likelihood method for estimating thresholds in a yes–no task," The Journal of the Acoustical Society of America 93(4), 2096–2105.

Green, D. M. (1995). "Maximum-likelihood procedures and the inattentive observer," The Journal of the Acoustical Society of America 97(6), 3749–3760.

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics (Vol. 1, pp. 1969-2012). New York: Wiley.

Gu, X., & Green, D. M. (1994). Further studies of a maximum-likelihood yes–no procedure. The Journal of the Acoustical Society of America, 96(1), 93-101.

Guo, Z., Yu, G., Zhou, H., Wang, X., Lu, Y., and Meng, Q. (2021). "Utilizing true wireless stereo earbuds in automated pure-tone audiometry," Trends in Hearing 25, 23312165211057367.

Hall, J. L. (1981). Hybrid adaptive procedure for estimation of psychometric functions. The Journal of the Acoustical Society of America, 69(6), 1763-1769.

Hallpike, C. S., & Hood, J. D. (1959). Observations upon the neurological mechanism of the loudness recruitment phenomenon. Acta Oto-laryngologica, 50(3-6), 472-486.

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.

Heller, O. (1985). Hörfeldaudiometrie mit dem Verfahren der Kategorienunterteilung (KU). Psychologische Beitrage.

Herbert, N., Keller, M., Derleth, P., Kühnel, V., & Strelcyk, O. (2022). Optimised adaptive procedures and analysis methods for conducting speech-in-noise tests. International Journal of Audiology, 1-11.

Hochmuth, S., Koifman, S., Warzybok-Oetjen, A., Avan, P., Kollmeier, B., & Radeloff, A. (2024). Das Projekt PREciSion audiology for AGE-related hearing loss (PRESAGE): Verbesserung der Diagnose von vorzeitigem altersbedingten Hörverlust. Laryngo-Rhino-Otologie, 103(S 02).

Hughson, W., Westlake, H. et al. (1944). "Manual for program outline for rehabilitation of aural casualties both military and civilian," Trans Am Acad Ophthalmol Otolaryngol 48(Suppl), 1–15.

Hébert, S., Fournier, P., & Noreña, A. (2013). The auditory sensitivity is increased in tinnitus ears. Journal of Neuroscience, 33(6), 2356-2364.

Ibelings, S., Brand, T., & Holube, I. (2022). Speech Recognition and Listening Effort of Meaningful Sentences Using Synthetic Speech. Trends in Hearing, 26, 23312165221130656.

IEC 60645-1, 2002. Electroacoustics - Audiometric Equipment - Part 1: Equipmentfor Pure-tone Audiometry. Standard of the International ElectrotechnicalCommission, Geneva, Switzerland.

Irace, A. L., Sharma, R. K., Reed, N. S., & Golub, J. S. (2021). Smartphone-based applications to detect hearing loss: a review of current technology. Journal of the American Geriatrics Society, 69(2), 307-316.

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

Jafri, S., Berg, D., Buhl, M., Vormann, M., Saak, S., Wagener, K. C., Thiel, C., Hildebrandt, A., & Kollmeier, B. (2024). OHHR – The Oldenburg Hearing Health Repository [Dataset] (1.0.0.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.14177903

Jones, P. R. (2018). QuestPlus: a matlab implementation of the QUEST+ adaptive psychometric method. Journal of Open Research Software, 6(1).

Jürgens, T., Kollmeier, B., Brand, T., & Ewert, S. D. (2011). Assessment of auditory nonlinearity for listeners with different hearing losses using temporal masking and categorical loudness scaling. Hearing Research, 280(1-2), 177-191.

Kaernbach, C. (1990). "A single-interval adjustment-matrix (siam) procedure for unbiased adaptive testing," The Journal of the Acoustical Society of America 88(6), 2645–2655.

Kam, A. C. S., Sung, J. K. K., Lee, T., Wong, T. K. C., & van Hasselt, A. (2012). Clinical evaluation of a computerized self-administered hearing test. International Journal of audiology, 51(8), 606-610.

Kayser, H., Herzke, T., Maanen, P., Zimmermann, M., Grimm, G., & Hohmann, V. (2022). Open community platform for hearing aid algorithm research: open Master Hearing Aid (openMHA). SoftwareX, 17, 100953.

Kinkel, M. (2007). The new ISO 16832 'Acoustics–loudness scaling by means of categories'. In 8th EFAS Congress/10th Congress of the German Society of Audiology, Heidelberg.

Kirkwood, B. R., & Sterne, J. A. (2010). Essential medical statistics. John Wiley & Sons.

Kisić, D., Horvat, M., Jambrošić, K., & Franček, P. (2022). The Potential of Speech as the Calibration Sound for Level Calibration of Non-Laboratory Listening Test Setups. Applied Sciences, 12(14), 7202.

Klein, S. A. (2001). "Measuring, estimating, and understanding the psychometric function: A commentary," Perception & psychophysics 63(8), 1421–1455.

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., & Hohmann, V. (1995). Loudness estimation and compensation for impaired listeners employing a categorical scale. Advances in hearing research, 441-453.

Kollmeier, B., & Kiessling, J. (2018). Functionality of hearing aids: State-of-the-art and future model-based solutions. International journal of audiology, 57(sup3), S3-S28.

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.

Kollmeier, B., Warzybok, A., Hochmuth, S., Zokoll, M. A., Uslar, V., Brand, T., & Wagener, K. C. (2015). The multilingual matrix test: Principles, applications, and comparison across languages: A review. International journal of audiology, 54(sup2), 3-16.

Kollmeier, B., Warzybok, A., Saak, S., Xu, C., & Schell-Majoor, L. (2023). Psychoacoustics with limited resources: How smartphone-based hearing tests change hearing research. ISAAR, Nyborg, Denmark.

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.

Krueger, M., Schulte, M., Brand, T., & Holube, I. (2017). Development of an adaptive scaling method for subjective listening effort. The Journal of the Acoustical Society of America, 141(6), 4680-4693.

Kuhlmann, I., Angonese, G., Thiel, C., Kollmeier, B., & Hildebrandt, A. (2023). Are there good days and bad days for hearing? Quantifying day-to-day intraindividual speech perception variability in older and younger adults. Journal of Experimental Psychology: Human Perception and Performance, 49(11), 1377.

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.

Kushalnagar, R. (2019). Deafness and hearing loss. Web accessibility: A foundation for research, 35-47.

Lansbergen, S., & Dreschler, W. A. (2022). Classification of hearing aids into feature profiles using hierarchical latent class analysis applied to a large dataset of hearing aids. Ear and hearing.

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.

Leek, M. R. (2001). “Adaptive procedures in psychophysical research,” Perception & psychophysics 63(8), 1279–1292.

Leek, M. R., Dubno, J. R., He, N. J., & Ahlstrom, J. B. (2000). Experience with a yes–no single-interval maximum-likelihood procedure. The Journal of the Acoustical Society of America, 107(5), 2674-2684.

Lenatti, M., Moreno-Sánchez, P. A., Polo, E. M., Mollura, M., Barbieri, R., & Paglialonga, A. (2022). Evaluation of machine learning algorithms and explainability techniques to detect hearing loss from a speech-in-noise screening test. American Journal of Audiology, 31(3S), 961-979.

Luengen, M., Garrelfs, C., Adiloğlu, K., Krueger, M., Cauchi, B., Markert, U., ... & Schultz, C. (2021). Connected Hearing Devices and Audiologists: The User-Centered Development of Digital Service Innovations. Frontiers in Digital Health, 3, 739370.

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.

Manning, C., Jones, P. R., Dekker, T. M., and Pellicano, E. (2018). "Psychophysics with children: Investigating the effects of attentional lapses on threshold estimates," Attention, Perception, & Psychophysics 80(5), 1311–1324.

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.

Meddis, R., and Lecluyse, W. (2011). "The psychophysics of absolute threshold and signal duration: a probabilistic approach," The Journal of the Acoustical Society of America 129(5), 3153–3165.

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

Mellor, J. C., Stone, M. A., & Keane, J. (2018a). Application of data mining to "big data" acquired in audiology: Principles and potential. Trends in hearing, 22, 2331216518776817.

Mellor, J., Stone, M. A., & Keane, J. (2018b). Application of data mining to a large hearing-aid manufacturer's dataset to identify possible benefits for clinicians, manufacturers, and users. Trends in hearing, 22, 2331216518773632.

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

Mok, B. A., Viswanathan, V., Borjigin, A., Singh, R., Kafi, H., & Bharadwaj, H. M. (2023). Web-based psychoacoustics: Hearing screening, infrastructure, and validation. Behavior Research Methods, 1-16.

Moore, B. C., & Schlittenlacher, J. (2023). Diagnosing Noise-Induced Hearing Loss Sustained During Military Service Using Deep Neural Networks. Trends in Hearing, 27, 23312165231184982.

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

Oetting, D., Hohmann, V., Appell, J. E., Kollmeier, B., & Ewert, S. D. (2016). Spectral and binaural loudness summation for hearing-impaired listeners. Hearing Research, 335, 179-192.

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.

Paglialonga, A., Tognola, G., & Pinciroli, F. (2015). Apps for hearing science and care. American Journal of Audiology, 24(3), 293-298.

Paquier, M., Koehl, V., & Jantzem, B. (2016). Effect of headphone position on absolute threshold measurements. Applied Acoustics, 105, 179-185.

Peng, Z. E., Buss, E., Shen, Y., Bharadwaj, H., Stecker, G. C., Beim, J. A., ... & Waz, S. (2020, December). Remote testing for psychological and physiological acoustics: Initial report of the p&p task force on remote testing. In Proceedings of Meetings on Acoustics (Vol. 42, No. 1). AIP Publishing.

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.

Pentland, A. P. (1980). Maximum likelihood estimation: The best PEST. Percept. Psychophys., 28, 377-379.

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

Polspoel, S., Moore, D. R., Swanepoel, D. W., Kramer, S. E., & Smits, C. (2024). Global access to speech hearing tests. medRxiv, 2024-06.

Potgieter, J. M., Swanepoel, D. W., Myburgh, H. C., Hopper, T. C., & Smits, C. (2016). Development and validation of a smartphone-based digits-in-noise hearing test in South African English. International journal of audiology, 55(7), 405-411.

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.

Rinderknecht, M. D., Ranzani, R., Popp, W. L., Lambercy, O., and Gassert, R. (2018).“Algorithm for improving psychophysical threshold estimates by detecting sustained inattention in experiments using pest,” Attention, Perception, & Psychophysics 80(6), 1629–1645.

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.

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

Saak, S., Kothe, A., Buhl, M., & Kollmeier, B. (2024). Comparison of user interfaces for measuring the matrix sentence test on a smartphone. International Journal of Audiology, 1-13.

Saberi, K., and Green, D. M. (1997). “Evaluation of maximum-likelihood estimators in nonintensive auditory psychophysics,” Perception & psychophysics 59(6), 867–876.

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., Bianchi, F., Fereczkowski, M., Santurette, S., & Dau, T. (2018). Data-driven approach for auditory profiling and characterization of individual hearing loss. Trends in hearing, 22, 2331216518807400.

Sanchez-Lopez, R., Dau, T., & Whitmer, W. M. (2022). Audiometric profiles and patterns of benefit: a data-driven analysis of subjective hearing difficulties and handicaps. International Journal of Audiology, 61(4), 301-310.

Sanchez-Lopez, R., Fereczkowski, M., Neher, T., Santurette, S., & Dau, T. (2020). Robust data-driven auditory profiling towards precision audiology. Trends in hearing, 24, 2331216520973539.

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

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

Schlittenlacher, J., & Moore, B. C. (2020). Fast estimation of equal-loudness contours using Bayesian active learning and direct scaling. Acoustical Science and Technology, 41(1), 358-360.

Schlittenlacher, J., Turner, R. E., & Moore, B. C. (2018). Audiogram estimation using Bayesian active learning. The Journal of the Acoustical Society of America, 144(1), 421-430.

Schädler, M. R., Hülsmeier, D., Warzybok, A., & Kollmeier, B. (2020). Individual aided speech-recognition performance and predictions of benefit for listeners with impaired hearing employing FADE. Trends in Hearing, 24, 2331216520938929.

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

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

Shen, Y., and Richards, V. M. (2012). "A maximum-likelihood procedure for estimating psychometric functions: Thresholds, slopes, and lapses of attention," The Journal of the Acoustical Society of America 132(2), 957–967.

Shen, Y., Dai, W., & Richards, V. M. (2015). A MATLAB toolbox for the efficient estimation of the psychometric function using the updated maximum-likelihood adaptive procedure. Behavior research methods, 47, 13-26.

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.

Shepherd, D., Hautus, M. J., Stocks, M. A., and Quek, S. Y. (2011). "The single interval adjustment matrix (siam) yes–no task: an empirical assessment using auditory and gustatory stimuli," Attention, Perception, & Psychophysics 73(6), 1934–1947.

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

Smits, C., Festen, J. M., Swanepoel, D. W., Moore, D. R., & Dillon, H. (2022). The one-up one-down adaptive (staircase) procedure in speech-in-noise testing: Standard error of measurement and fluctuations in the track. The Journal of the Acoustical Society of America, 152(4), 2357-2368.

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

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.

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

Sørensen, C. B., Gyldenlund Pedersen, R., Nielsen, J., Sidiras, C., Schmidt, J. H., & Pedersen, E. R. (2023). User-operated audiometry–an evaluation of expert vs. non-expert headphone placement. International journal of audiology, 62(10), 938-945.

Taylor, M., and Creelman, C. D. (1967). "Pest: Efficient estimates on probability functions," The Journal of the Acoustical Society of America 41(4A), 782–787.

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

Treutwein, B. (1995). "Adaptive psychophysical procedures," Vision research 35(17), 2503–2522.

Treutwein, B., and Strasburger, H. (1999). "Fitting the psychometric function," Perception & psychophysics 61(1), 87–106.

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.

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

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

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

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

Wasmann, J. W. A., Lanting, C. P., Huinck, W. J., Mylanus, E. A., van der Laak, J. W., Govaerts, P. J., ... & Barbour, D. L. (2021). Computational audiology: new approaches to advance hearing health care in the digital age. Ear and hearing, 42(6), 1499.

Watson, A. B. (2017). QUEST+: A general multidimensional Bayesian adaptive psychometric method. Journal of Vision, 17(3), 10-10.

Watson, A. B., & Pelli, D. G. (1983). QUEST: A Bayesian adaptive psychometric method. Perception & psychophysics, 33(2), 113-120.

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

Wichmann, F. A., and Hill, N. J. (2001). "The psychometric function: I. fitting, sampling, and goodness of fit," Perception & psychophysics 63(8), 1293–1313.

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.

Wu, M., Christiansen, S., Fereczkowski, M., & Neher, T. (2022). Revisiting auditory profiling: Can cognitive factors improve the prediction of aided speech-in-noise outcome?. Trends in Hearing, 26, 23312165221113889.

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. https://doi.org/10.1080/14992027.2024.2424876

Xu, C., Schell-Majoor, L., & Kollmeier, B. (2024c). Feasibility of efficient smartphone-based threshold and loudness assessments in typical home settings. medRxiv, 2024-11.

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

Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands (Frequenzgruppen). J. Acoust. Soc. Am. 33:248. doi: 10.1121/1.1908630

Affidavit

I hereby declare that I have developed and written the enclosed dissertation entirely on my own and have not used outside sources without declaration in the text. Any concepts or quotations applicable to these sources are clearly attributed to them. This dissertation has not been submitted in the same or a substantially similar version, not even in part, to any other authority for grading and has not been published elsewhere. This is to certify that the printed version is equivalent to the submitted electronic one. I am aware of the fact that a misstatement may have serious legal consequences.

I also agree that my thesis can be sent and stored anonymously for plagiarism purposes. I know that my thesis may not be corrected if the declaration is not issued.

Oldenburg, February 11, 2025

Chen Xu

Chen Xu

Doctoral Student

Harlingerstr. 15b, D-26121, Oldenburg, Germany

chen.xu@uni-oldenburg.de | github.com/chenxu1995 | gitlab.uni-oldenburg.de/dima9572 | linkedin.com/in/cx1995

Education

University of Oldenburg *Oldenburg, Germany*

PhD Medical Physics *10.2020 - present*

- Advisor: Prof. Dr. Dr. Birger Kollmeier
- Structured doctoral degree program: Neurosensory Science and Systems
- Dissertation: Crucial Elements of a Virtual Hearing Clinic on Mobile Devices: Psychophysics, Diagnostic Parameter Estimation, and Validation
- Research interests: mobile health application, machine learning, psychophysical and hearing research

Technical University of Munich *Munich, Germany*

Master of Science *10.2017 - 09.2020*

- Advisor: Prof. Dr. Bernhard Wolfrum
- Master program: Electrical Engineering and Information Technology
- Master thesis: Data-Driven Error Detection with Hybrid EEG-fNIRS Measurements

Northwestern Polytechnical University *Xi'an, China*

Bachelor of Engineering *09.2013 - 06.2017*

- Majors in Automation and Control Theory

Technical Skills

Programming:	Python, R, Matlab, C++, Octave, Bash
Web Programming:	HTML, CSS, JavaScript, Flask, SQLite
Software & Tools:	SSH, Git, Zsh, Docker, Office, Latex
Data Science Packages:	**Python:** NumPy, Pandas, Keras, Tensorflow, Scikit-Learn, Jupyter, etc. **R:** dplyr, ggplot2, tidyr, tidyverse, etc.
Statistics & Machine learning:	Statistical Analysis, Linear/Logistic Regression, Clustering, etc.
Electrical Engineering:	Raspberry Pi
Language:	Chinese (native), English (CET6 & DAAD-Test B2-C1), German (DSH2)

Professional Experience

2020-now	**Research & Teaching Assistant**, Dept. of Medical Physics and Acoustics, University of Oldenburg
2020	**Research Assistant**, Max Planck Institute of Psychiatry
2020	**Research Assistant**, Helmholtz Zentrum München
2020	**Semester Project**, TUM data innovation lab, BMW IT center
2019-2020	**Working Student**, Siemens AG

Publications

Published

Xu, C., Hülsmeier, D., Buhl, M., & Kollmeier, B. (2024). How Does Inattention Influence the Robustness and Efficiency of Adaptive Procedures in the Context of Psychoacoustic Assessments via Smartphone? Trends in Hearing. 2024;28. doi: 10.1177/23312165241288051

Xu, C., Schell-Majoor, L., & Kollmeier, B. (2024). Development and verification of non-supervised smartphone-based methods for assessing pure-tone thresholds and loudness perception. International Journal of Audiology, 1–11. doi: 10.1080/1

4992027.2024.2424876.

In Review

Xu, C., Schell-Majoor, L., & Kollmeier, B. (2024). Feasibility of efficient smartphone-based threshold and loudness assessments in typical home settings. Manuscript submitted for publication in Trends in Hearing.

In Prep

Xu, C., Schell-Majoor, L., & Kollmeier, B. (2023a). Predict standard audiogram from a loudness scaling test employing unsupervised, supervised, and explainable machine learning techniques. Manuscript in preparation.

Xu, C., Schell-Majoor, L., & Kollmeier, B. (2023b). Derive a robust and optimal feature set for standard audiogram prediction. Manuscript in preparation.

Xu, C., Schell-Majoor, L., & Kollmeier, B. (2023c). Comparison of auditory profiles using manifold learning and intrinsic measures. Manuscript in preparation.

Awards

2025 **Congress Scholarships**, European Federation of Audiology Societies, Vienna
2025 **Travel Grant**, Precision Digital Therapeutics Master Class, Singapore-ETH Centre, Singapore
2024 **Travel Grant**, Graduate School Science, Medicine and Technology - OLTECH
2023 **Travel Grant**, Graduate School Science, Medicine and Technology - OLTECH
2020 **Starting Stipends**, Collaborative Research Centre SFB 1330 Hearing Acoustics (HAPPAA)
2019 **Swiss-European Mobility Programme**, École polytechnique fédérale de Lausanne
2018, 2019 **Scholarships for TUM international students** , Bavarian government

Presentations

Contributed Oral Presentations

Xu, C., Schell-Majoor, L., & Kollmeier, B. (2025). Reinforced categorical loudness scaling (rCLS) – An efficient procedure for self-administered simultaneous assessment of hearing thresholds and loudness perception. In 17th European Federation Audiology Societies Congress, Vienna, Austria.

Xu, C., Schell-Majoor, L., & Kollmeier, B. (2024). Predict standard audiogram from a loudness scaling test employing unsupervised, supervised, and explainable machine learning techniques. In Proc. "Fortschritte der Akustik - DAGA'24", Hannover, Germany.

Xu, C., Schell-Majoor, L., & Kollmeier, B. (2023a). Development and verification of self-supervised smartphone-based methods for assessing pure-tone audiometry and loudness growth function. In 16th European Federation Audiology Societies Congress, Sibenik, Croatia.

Xu, C., Schell-Majoor, L., & Kollmeier, B. (2023b). Smartphone-based hearing tests for a Virtual Hearing Clinic: Influence of ambient noise on the absolute threshold and loudness scaling at home. In Virtual Conference on Computational Audiology VCCA June 29-30, online.

Kollmeier, B., Warzybok, A., Saak, S., **Xu, C.**, & Schell-Majoor, L. (2023). Psychoacoustics with limited resources: How smartphone-based hearing tests change hearing research. International Symposium on Auditory and Audiological Research ISAAR, Nyborg, Denmark.

Contributed Poster Presentations

Xu, C., Schell-Majoor, L., & Kollmeier, B. (2024). Towards a robust and optimum prediction of audiometric profiles from non-audiometric features. In Audiological Research Cores in Europe (ARCHES), Leuven, Belgium.

Xu, C., Hülsmeier, D., Buhl, M., & Kollmeier, B. (2022). How Robust and Efficient Are Different Adaptive Hearing Threshold Procedures for Use With Mobile Devices. In Audiological Research Cores in Europe (ARCHES), Amsterdam, The Netherlands.

Seminar Talks

Xu Chen (2024, May). Comparison between model-based and model-free adaptive procedures in terms of the inattentive listener using smartphones. Online Audiology Journal Club, Univ of Washington (UW), Seattle, USA.

Xu Chen (2023, October). How does the ambient noise influence the smartphone-based hearing tests? Online Audiology Journal Club, the Univ of Hong Kong (HKU), Hong Kong, China.

Teaching Experience

2021-2023 **Physiological, psychological, and audiological acoustics**, Teaching Assistant *Oldenburg*

Outreach & Professional Development

Service and Outreach

2022 **Hearing4all Summer School**, Conference Organizer *Visselhövede*
2022 **SFB 1330 PhD Students' Retreat**, Organizer *Wardenburg*
2022 **10 Year Anniversary of the Oldenburg Medical School (UMO)** , Demonstrator *Oldenburg*
2020 **German Academic Exchange Service (DAAD) 'my research diary'**, Volunteer *Oldenburg*

Development

2024 **Summer School on Machine Learning & Numerics for Acoustics**, Oldenburg
2024 **Mediterranean Machine Learning Summer School (M2L)**, Milan
2024 **EuADS Summer School – Generative AI**, Luxembourg
2024 **Docker for Neuroscience**, Oldenburg
2023 **Advanced Topics and Publications in Hearing Research**, Hvar
2023 **Mobile Health in Communication, Perception, and Mobility**, Oldenburg
2023 **Research Data Management**, Oldenburg
2023 **Winter School "Hearing Acoustics"** , Oldenburg
2022 **Hearing4all International Symposium**, Hannover

Professional Memberships

Student Membership of the International Society of Audiology (ISA)
Membership of the European Association for Data Science (EuADS)
Member of the Computational Audiology Network Special Interest Group (CAN-SIG)
Member of the European Acoustics Association (EAA) Young Acousticians Network (YAN)

Peer Review

Ear and Hearing

Certifications

2022 **Audiometry for Intermediates**, Interacoustics A/S
2021 **Responsive Website Basics: Code with HTML, CSS, and JavaScript**, Coursera
2020 **Recommendation Systems on Google Cloud**, Coursera
2020 **Introduction to Digital Health**, Coursera