

## Do uHear? Validation of uHear App for Preliminary Screening of Hearing Ability in Soundscape Studies

Zhen-Ting ONG<sup>\*</sup>; Bhan LAM<sup>\*</sup>; Kenneth OOI;  
Karn N. WATCHARASUPAT; Trevor WONG; Woon-Seng GAN

School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore<sup>1</sup>

### ABSTRACT

Studies involving soundscape perception often exclude participants with hearing loss to prevent impaired perception from affecting experimental results. Participants are typically screened with pure tone audiometry, the “gold standard” for identifying and quantifying hearing loss at specific frequencies, and excluded if a study-dependent threshold is not met. However, procuring professional audiometric equipment for soundscape studies may be cost-ineffective, and manually performing audiometric tests is labour-intensive. Moreover, testing requirements for soundscape studies may not require sensitivities and specificities as high as that in a medical diagnosis setting. Hence, in this study, we investigate the effectiveness of the uHear app, an iOS application, as an affordable and automatic alternative to a conventional audiometer in screening participants for hearing loss for the purpose of soundscape studies or listening tests in general. Based on audiometric comparisons with the audiometer of 163 participants, the uHear app was found to have high precision (98.04 %) when using the World Health Organization (WHO) grading scheme for assessing normal hearing. Precision is further improved (98.69 %) when all frequencies assessed with the uHear app is considered in the grading, which lends further support to this cost-effective, automated alternative to screen for normal hearing.

Keywords: soundscape, audiometry, listening tests

### 1. INTRODUCTION

In the normative minimum reporting requirements of soundscape studies, the hearing ability of all human participants must be explicitly reported (1). Moreover, the required computation of psychoacoustic loudness of the acoustic environment represents an “average loudness as perceived by a group of individuals with otologically normal hearing” (2). Despite the inherent requirements to determine normal hearing ability in listening experiments, there are no explicit guidelines to grade the “normality” of hearing ability even in technical directives (1, 3). Moreover, it may also be of interest to investigate the soundscape perception of persons with varying degrees of age-related hearing loss (i.e. elderly). Therefore, there is a significant impetus for an automated audiometric test that does not require expert knowledge or professional audiometric equipment for the purposes of screening ability before listening experiments.

Hearing ability is usually assessed by audiometric tests administered by a trained healthcare professional (e.g. audiologists, otolaryngologists). Since hearing impairment is a medical diagnosis, audiometric tests are manually administered by qualified personnel and governed by internationally recognised guidelines (4–6), and conducted in strict acoustic conditions with calibrated audiometric equipment (7). Beyond medical diagnoses, compliance with the stringent requirements of audiometric tests to assess hearing ability for listening experiments becomes prohibitively expensive and laborious. Nonetheless, a number of soundscape studies with listening tests have reported the hearing ability of the recruited participants based on pure tone audiometry (PTA) with professional audiometric equipment (8–13).

To overcome the barriers due to cost and availability of audiometric tests and qualified healthcare personnel, mobile health technologies have been developed and deployed, especially in the form of smartphone-based apps for preventive screening in schools and for the elderly (14, 15). A recent meta-analysis of

<sup>1</sup>{ztong, bhanlam, wooi002, karn001, trevor.wong, ewsgan}@ntu.edu.sg

<sup>\*</sup>These authors contributed equally to this work

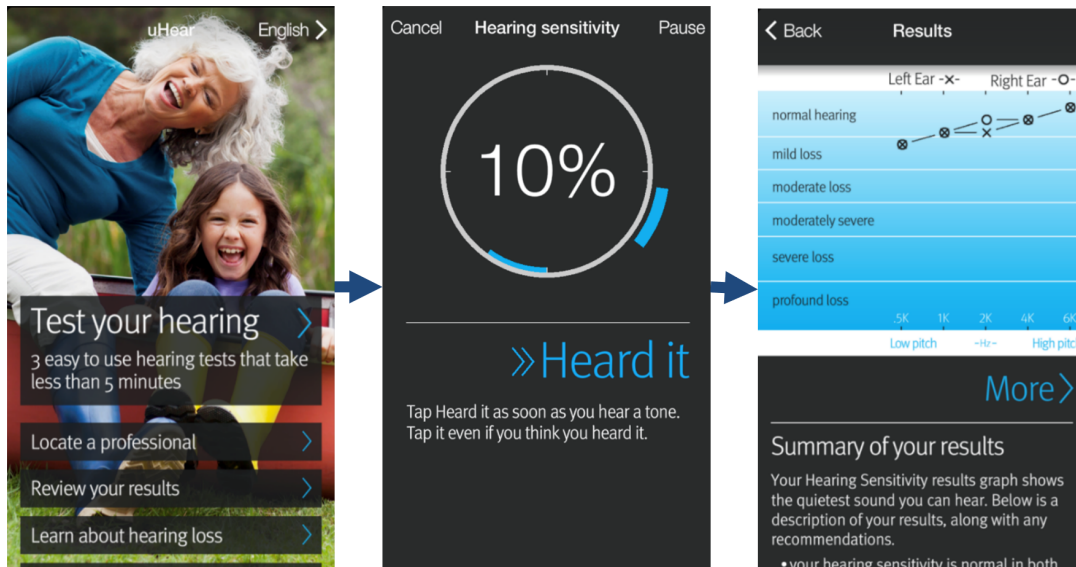


Figure 1: Procedure of pure tone audiometric test on uHear app

smartphone-based audiometry reported a respectable aggregated diagnostic sensitivity of 89 % [(83 %, 93 %), 95 % CI] and specificity of 93 % [(87 %, 97 %); 95 % CI] for PTA (16). Few apps have been scientifically validated, with the rarity of the iOS-based uHear app (uHear, Unitron – Sonova Holding AG, Stäfa, Switzerland) subjected to some rigor (17–20). Moreover, it was found that device output levels were similar across mobile phone models running the iOS platform (Apple Inc., California, USA) (21), and its bundled headphones were found to be more accurate than supra- or circumaural headphones (20), thereby increasing the availability of any iOS-based audiometry apps. To the authors’ best knowledge, there are also no published soundscape literature that screened for hearing ability using smartphone-based applications.

To this end, there is an absence of studies examining the efficacy of smartphone-based audiometric apps to primarily screen for normal hearing. Based on reported literature, this study thus examines the efficacy of the iOS-based uHear app in comparison with a “gold standard” audiometer in detecting normal hearing based on the latest World Health Organisation (WHO) grading scheme (22).

## 2. METHODOLOGY

### 2.1 Participants

This study was conducted as an auxiliary to two ongoing subjective listening experiments at Nanyang Technological University (NTU), Singapore. A total of 163 participants were recruited after first screening for any self-reported hearing loss, of which there were none. Participants were approximately evenly distributed in gender [female: 91 (55.8 %); male: 72 (44.2 %)] with a relatively young age distribution between 18 and 75 years ( $\bar{x}_{\text{age}} = 27.3$  years,  $SD_{\text{age}} = 11.4$ ,  $SE_{\text{age}} = 0.89$ ). Formal ethical approvals were obtained from the Institutional Review Board (IRB) of NTU (Ref. IRB-2020-08-035 and IRB-2021-293) for the responses collected.

### 2.2 Pure Tone Audiometry

Conventional pure tone audiometry (PTA) was conducted using a “gold standard” audiometer (AD629, Interacoustics A/S, Middelfart, Denmark) in a quiet room within the permissible ambient noise limits (6). Conventional PTA is a manual determination of threshold levels as defined by the American Speech-Language-Hearing Association (4), and is also referred to as the ascending method in ISO 8253-1, or the modified Hughson-Westlake procedure as described in the British Society of Audiology guidelines (5, 6). It is characterized by its “5 dB up” and “10 dB down” procedure in determining threshold levels. In contrast, the uHear app is an automated implementation of the modified Hughson-Westlake method with the same starting level of 30 dB HL (Hearing Level) as the conventional PTA. All the thresholds determined by the uHear app were through the same set of devices (iPhone 4S and Earpods, Apple Inc., California, USA). Instructions to conduct the automated audiometric assessment on the uHear app was presented via instructions on a graphic user interface, as shown in Figure 1.

## 2.3 Validation procedure

The audiograms obtained from both the audiometer and uHear app were graded based on the WHO hearing loss grading system (22). The grading criteria for “normal hearing” requires the hearing threshold (HT) in the better ear to be less than 20 dBHL (lowered from 26 dBHL in 2021), and that the HT in the worse ear is less than 35 dBHL. The HT is determined by the arithmetic mean of hearing levels in dB HL across 4 tested frequencies (0.5, 1, 2, and 4 kHz). Since the uHear app also assesses the 6 kHz frequency, it is included along with the 4 frequencies in the WHO grading scheme, denoted as “WHO+”. The suitability of the uHear app would be assessed based on a set of metrics (i.e. accuracy, sensitivity, specificity, and precision) in grading normal hearing via both WHO and WHO+ schemes.

The accuracy, precision, sensitivity and specificity of the uHear app in determining “normal hearing” was computed based on the binary confusion matrix with the audiometer as the reference. Positives (“1”) represent the grading of “normal hearing”, whereas negatives (“0”) indicate otherwise. The overall accuracy is determined by

$$\text{Accuracy} = (\text{TP} + \text{TN})/K, \quad (1)$$

where TP and TN are “true positives” and “true negatives”, and  $K$  is the total number of positives and negatives. The sensitivity and specificity are determined by

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN}), \quad (2)$$

and

$$\text{Specificity} = \text{TN}/(\text{FP} + \text{TN}), \quad (3)$$

respectively. Lastly, the precision is given by

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}). \quad (4)$$

## 2.4 Data analysis

Due to the interval nature of the audiogram measurements, non-parametric representations of the Bland-Altman plots were presented to investigate the agreement between both methods (23, 24). Distribution-free tests and confidence intervals (CI) were computed for the median to examine the bias, and also for the 2.5<sup>th</sup> and 97.5<sup>th</sup> percentiles to discover outliers (25). The accuracy, sensitivity, specificity, and precision of the uHear app in grading normal hearing ability was evaluated in comparison to the audiometer. To determine if there were differences between both methods in grading normal hearing, a McNemar’s test was also conducted (26). Finally, the non-parametric Wilcoxon signed rank tests were employed to determine the differences between the methods for each test frequency (27). All data analyses were conducted with the R programming language (27) on a 64-bit ARM environment. The data that support the findings of this study are openly available in NTU research data repository DR-NTU (Data) at [doi:10.21979/N9/JQDI6F](https://doi.org/10.21979/N9/JQDI6F), and the replication code is available on GitHub at [github.com/ntudsp/douHear](https://github.com/ntudsp/douHear).

# 3. RESULTS AND DISCUSSION

## 3.1 Agreement between the uHear app and the audiometer

In the non-parametric implementation of the Bland-Altman plot, the median of the paired differences between audiometer and uHear measurements is used to judge the measurement bias. The lower and upper limits of agreement (LoA) are represented by the 2.5<sup>th</sup> and 97.5<sup>th</sup> percentiles, respectively. Since no differences were lower or higher than the 95 % confidence intervals of the lower or upper tails of the respective lower and upper LoAs, no outliers were detected at each frequency across both ears, as shown in Figure 2. Moreover, the measurements appear to be symmetrically distributed about the medians.

A slight overall bias of about  $-5$  dB [ $(-5, -5)$ , 95 % CI] in the uHear app is reflected in the median across both ears when considering differences from all frequencies. The overall bias could be attributed to the  $-5$  dB bias at 1 kHz [ $(-5, -5)$ , 95 % CI] in the left ear; and  $-5$  dB bias at 1 [ $(-10, -5)$ , 95 % CI], 2 [ $(-5, -5)$ , 95 % CI] and 4 kHz [ $(-10, -5)$ , 95 % CI] frequencies in the right ear, since the equality lines were beyond the confidence intervals in those cases. Hence, there is a greater tendency to obtain a better hearing threshold using the uHear app. This leniency could lead to an increase in false positives.

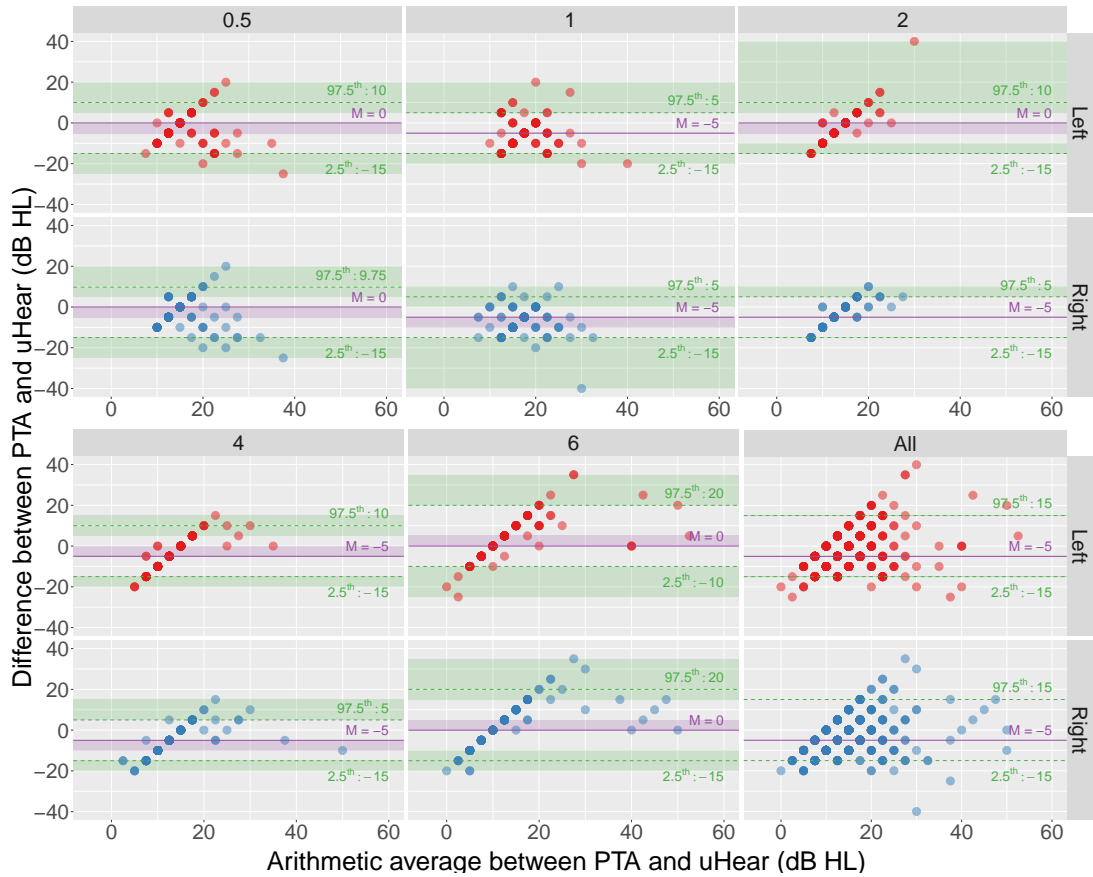


Figure 2: Bland-Altman plot of agreement between the pure tone audiometry and uHear mobile app for measuring hearing level (dB HL) across 0.5, 1, 2, 4, 6 kHz frequencies for both ears. The average bias is estimated by the median of the differences and limits of agreement by the 2.5<sup>th</sup> and 97.5<sup>th</sup> percentiles. Distribution-free confidence intervals are based on 95 % confidence level for the upper and lower percentiles (■) and the median (■).

### 3.2 Normal hearing grading performance of uHear

The normal hearing grading outcomes of all 163 participants across both WHO and WHO+ schemes are presented as confusion matrices, as shown in Figure 3. At a glance, the number of TNs and FPs were about the same across both schemes, but the number of FPs were substantially lower in the WHO+ scheme. The computed performance metrics outlined in Section 2.3 for the uHear app using both WHO and WHO+ grading schemes are summarized in Table 1. Across the metrics, the WHO+ scheme appears to be outperforming the WHO scheme, especially in terms of specificity. The low specificity across both schemes provides further evidence of the inability of the uHear app in correctly grading the presence of hearing loss (i.e. TN). Using the McNemar's test (MNT), however, the differences in sensitivity and specificity between the uHear app and the audiometer was not statistically significant in either schemes ( $p > 0.05$ ), as shown in Table 1.

Different from audiometric diagnostics that detect instances and assess the severity of hearing loss, the aim of the audiometric test for listening experiments is to screen for normal hearing as an inclusion criteria. Whereas FPs would compromise the quality of listening experiments that require participants with normal hearing, it is not detrimental to the listening experiment when an FN occurs (apart from losing potential participants). Hence, for the purpose of screening for normal hearing, precision would be the most important metric to determine the suitability of the uHear app as a low-cost replacement for audiometers, as seen from Equation (4). Although both WHO and WHO+ grading schemes are comparable in terms of precision, it is worth noting that there was also an overall change in the reference audiometer grading in the WHO (normal: 155 (95.1 %); not normal: 8) and WHO+ (normal: 152; not normal: 11) schemes.

### 3.3 Audiometric differences between uHear and the audiometer

To investigate the differences between the audiometer and uHear audiograms at each frequency, the non-parametric Wilcoxon signed rank test (WSRT) was conducted. The choice of the non-parametric WSRT

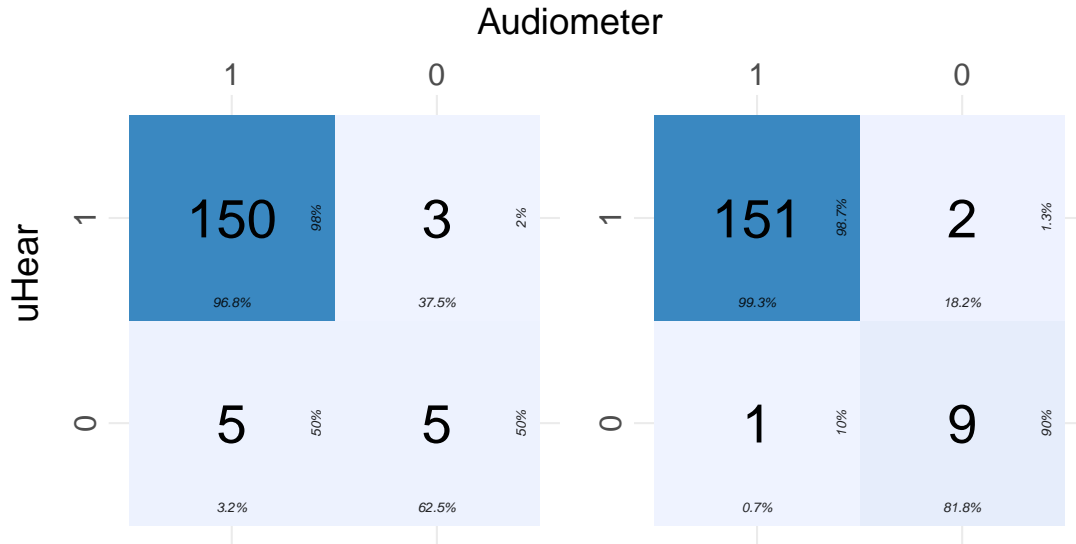


Figure 3: Confusion matrix of the uHear app against the audiometer in grading normal hearing. Normal hearing is indicated as “1”, and “0” otherwise, wherein “11”, “01”, “00”, and “10”, respectively indicates “true positive” (TP), “false negative” (FN), “true negative” (TN), and “false positive” (FP). Grading is based on the hearing threshold (HT) calculations from WHO (left) and WHO+ (right) schemes, where the latter uses HTs based on the arithmetic mean across all frequencies measured in the uHear app.

Table 1: Accuracy, sensitivity, specificity, and precision, and McNemar’s Test  $p$ -value of the WHO and WHO+ grading criteria for normal hearing. WHO+ includes the measurements at 6 kHz in the grading, which is on top of the 4 frequencies in the WHO grading (i.e. 0.5, 1, 2, and 4 kHz).

Grading	Accuracy (%; 95 % CI)	Sensitivity (%)	Specificity (%)	Precision (%)	MNT $p$ -value
WHO	95.09 (90.56, 97.86)	96.77	62.50	98.04	0.723
WHO+	98.16 (94.72, 99.62)	99.34	81.82	98.69	1.000

was motivated by the interval nature of the measurements, as well as the symmetry in the measurements as shown in Figure 2. The difference in audiograms between the audiometer and the uHear app was found to be statistically significant at all frequencies across both ears at 5 % significance level, except at 6 kHz for the right ear, as shown in Table 2.

The computed pseudomedian (28) of the differences between the audiometer and uHear measurements depicts a clear negative bias in the uHear measurements from 0.5 up to 4 kHz in both ears. Given the narrow confidence intervals, the pseudomedians were almost equivalent to the computed median of the differences in the Bland-Altman plots in Section 3.1, which confirms the symmetry of the measurements (i.e. high precision). Unsurprisingly, the strongest evidence of difference occurred at the 1 and 4 kHz tones, as also shown in Figure 2. The distinct differences also corroborates with previous investigations on the uHear and other mobile application based assessment systems (20).

Interestingly, the pseudomedians at 6 kHz in both ears were near zero and the WSRT was either weakly significant (Left ear;  $0.01 < p = 1.13 \times 10^{-3} < 0.05$ ) or not significant (Right ear;  $p = 6.82 \times 10^{-2} > 0.05$ ). Hence, the increase in precision with the addition of the 6 kHz measurement in the grading for normal hearing could be attributed to the similarity of the uHear app to the audiometer at the 6 kHz test tone. The uHear app was also found to match audiometers more closely in the mid to high frequencies ( $\geq 2$  kHz) in previous studies (20).

#### 4. CONCLUSION

A mobile-based application (uHear) was assessed for its suitability as a screening tool for normal hearing in participants of listening experiments. Due to the availability of a test frequency (i.e. 6 kHz) beyond those required by the WHO guidelines to grade normal hearing abilities, two grading schemes, namely the WHO and the modified WHO+, were investigated. Based on the assessment with 163 participants, the uHear app

Table 2: Wilcoxon signed rank tests, and the pseudomedian and its confidence intervals (dB HL; 95% confidence level) of the differences between the audiometer and uHear audiograms are computed at each frequency (0.5, 1, 2, 4 and 6 kHz) across both ears. The null hypothesis is rejected at  $p$ -value  $< 0.05$ .

Ear	Frequency (kHz)	Pseudomedian	Confidence Interval (95 %)	$p$ -value
Left	0.5	-5.0	( -5 , -5 )	$6.11 \times 10^{-7}$
	1	-7.5	( -7.5, -7.5)	$3.37 \times 10^{-19}$
	2	-5.0	( -7.5, -2.5)	$7.14 \times 10^{-7}$
	4	-5.0	( -7.5, -5 )	$1.65 \times 10^{-10}$
	6	2.5	( 0 , 5 )	$1.13 \times 10^{-3}$
Right	0.5	-5.0	( -5 , -2.5)	$2.14 \times 10^{-6}$
	1	-7.5	( -7.5, -7.5)	$4.31 \times 10^{-23}$
	2	-5.0	( -7.5, -5 )	$1.61 \times 10^{-13}$
	4	-7.5	( -7.5, -7.5)	$1.67 \times 10^{-16}$
	6	2.5	( 0 , 2.5)	$6.82 \times 10^{-2}$

was comparable with the audiometer in precision – the key metric in screening for normal hearing for listening tests – using both grading schemes. Despite higher overall accuracy, sensitivity, and precision, the proposed WHO+ grading scheme should only be used after assessing the need to include the 6 kHz test frequency based on the screening requirements.

Nevertheless, the uHear app still falls short as a low-cost diagnostic tool for assessing the severity of hearing impairment due to significant deviation from the audiometer across frequencies. Further caution should be exercised due to the low specificity (i.e. ability in detecting instances of hearing impairment). The results presented were also obtained via a singular device with a now deprecated audio interface (i.e. 3.5 mm) and should be examined for variations in newer models and modern audio interfaces in the future.

## ACKNOWLEDGEMENTS

This research is supported by the Singapore Ministry of National Development and the National Research Foundation, Prime Minister’s Office under the Cities of Tomorrow Research Programme (Award No. COT-V4-2020-1). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the view of National Research Foundation, Singapore, and Ministry of National Development, Singapore.

## REFERENCES

1. International Organization for Standardization. ISO/TS 12913-2: Acoustics — Soundscape — Part 2: Data collection and reporting requirements. Geneva, Switzerland: International Organization for Standardization; 2018.
2. International Organization for Standardization. ISO 532-1:2017 - Acoustics — Method for calculating loudness — Part 1: Zwicker method. vol. 7. Geneva: International Organization for Standardization; 2017.
3. International Telecommunication Union Radiocommunication Sector. Recommendation ITU-R BS.1116-3: Methods for the subjective assessment of small impairments in audio systems. Geneva, Switzerland: International Telecommunication Union; 2015.
4. American Speech-Language-Hearing Association. Guidelines for Manual Pure-Tone Threshold Audiometry; 2005. Available from: <http://www.asha.org/policy/GL2005-00014/>.
5. British Society of Audiology. Pure-tone air-conduction and boneconduction threshold audiometry with and without masking. Bathgate, UK: British Society of Audiology; 2018.
6. International Organization for Standardization. Acoustics — Audiometric test methods - Part 1 : Pure-tone air and bone conduction (ISO 8253-1:2010). International Organization for Standardization; 2010.
7. International Organization for Standardization. Acoustics — Reference zero for the calibration of audiometric equipment - Part 1: Reference equivalent threshold sound pressure levels for pure tones and supra-aural earphones (ISO 389-1:2017). International Organization for Standardization; 2018.
8. Hong JY, Lam B, Ong ZT, Ooi K, Gan WS, Kang J, et al. Effects of contexts in urban residential areas on

- the pleasantness and appropriateness of natural sounds. *Sustainable Cities and Society*. 2020 12;63(PG-):102475.
9. Medvedev O, Shepherd D, Hautus MJ. The restorative potential of soundscapes: A physiological investigation. *Applied Acoustics*. 2015;96:20–26. Available from: <http://dx.doi.org/10.1016/j.apacoust.2015.03.004>.
  10. Hao Y, Kang J, Wörtche H. Assessment of the masking effects of birdsong on the road traffic noise environment. *The Journal of the Acoustical Society of America*. 2016 8;140(2):978–987.
  11. Lee SC, Hong JY, Jeon JY. Effects of acoustic characteristics of combined construction noise on annoyance. *Building and Environment*. 2015 10;92:657–667.
  12. Rådsten Ekman M, Lundén P, Nilsson ME. Similarity and pleasantness assessments of water-fountain sounds recorded in urban public spaces. *The Journal of the Acoustical Society of America*. 2015;138(5):3043–3052.
  13. Axelsson Ö, Nilsson ME, Berglund B. A principal components model of soundscape perception. *The Journal of the Acoustical Society of America*. 2010 11;128(5):2836–2846. Available from: <http://asa.scitation.org/doi/10.1121/1.3493436>.
  14. World Health Organization. *Hearing screening: considerations for implementation*. Geneva, Switzerland: World Health Organization; 2021.
  15. Manus M, van der Linde J, Kuper H, Olinger R, Swanepoel DW. Community-Based Hearing and Vision Screening in Schools in Low-Income Communities Using Mobile Health Technologies. *Language, Speech, and Hearing Services in Schools*. 2021 4;52(2):568–580.
  16. Chen CH, Lin HYH, Wang MC, Chu YC, Chang CY, Huang CY, et al. Diagnostic Accuracy of Smartphone-Based Audiometry for Hearing Loss Detection: Meta-analysis. *JMIR mHealth and uHealth*. 2021 9;9(9):e28378.
  17. Saliba J, Al-Reefi M, Carriere JS, Verma N, Provencal C, Rappaport JM. Accuracy of Mobile-Based Audiometry in the Evaluation of Hearing Loss in Quiet and Noisy Environments. *Otolaryngology–Head and Neck Surgery*. 2017 4;156(4):706–711.
  18. Li LYJ, Wang SY, Wu CJ, Tsai CY, Wu TF, Lin YS. Screening for Hearing Impairment in Older Adults by Smartphone-Based Audiometry, Self-Perception, HHIE Screening Questionnaire, and Free-Field Voice Test: Comparative Evaluation of the Screening Accuracy With Standard Pure-Tone Audiometry. *JMIR mHealth and uHealth*. 2020 10;8(10):e17213.
  19. Melo IMM, Silva ARX, Camargo R, Cavalcanti HG, Ferrari DV, Taveira KVM, et al. Accuracy of smartphone-based hearing screening tests: a systematic review. *CoDAS*. 2022;34(3).
  20. Barczik J, Serpanos YC. Accuracy of Smartphone Self-Hearing Test Applications Across Frequencies and Earphone Styles in Adults. *American Journal of Audiology*. 2018 12;27(4):570–580.
  21. Patel K, Thibodeau L, McCullough D, Freeman E, Panahi I. Development and Pilot Testing of Smartphone-Based Hearing Test Application. *International Journal of Environmental Research and Public Health*. 2021 5;18(11):5529.
  22. World Health Organisation. *World Report On Hearing*. Geneva: World Health Organisation; 2021. Available from: <https://www.who.int/publications/i/item/world-report-on-hearing>.
  23. Martin Bland J, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*. 1986 2;327(8476):307–310. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0140673686908378>.
  24. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*. 1999 4;8(2):135–160.
  25. Fay MP. *asht: Applied Statistical Hypothesis Tests*; 2022. Available from: <https://cran.r-project.org/package=asht>.
  26. Kuhn M. *caret: Classification and Regression Training*; 2019. Available from: <https://github.com/topepo/caret/>.
  27. R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria: The R Foundation for Statistical Computing; 2021. Available from: <https://www.r-project.org/>.
  28. Hollander M, A Wolfe D, Chicken E. *Nonparametric Statistical Methods*. vol. 17 of Wiley Series in Probability and Statistics. Wiley; 2015.