AMERICAN ACADEMY OF OTOLARYNGOLOGY–HEAD AND NECK SURGERY FOUNDATION

Otolaryngology–Head and Neck Surgery
2026, Vol. 00(00) 1–10

DOI: 10.1002/ohn.70194
http://otojournal.org
WILEY

Apple Hearing Test Feature for the AirPods Pro 2: Accuracy, Reliability, and Time-Efficiency

Megan Kruger, PhD[1,2,3], Vinaya Manchaiah, PhD[1,2,3,4,5,6], and De Wet Swanepoel, PhD[1,2,3,4]

Abstract

Objective. To evaluate the accuracy, test-retest reliability, and time-efficiency of Apple's Hearing Test Feature (HTF) compared to reference standard pure-tone audiometry (PTA).

Study Design. Cross-sectional validation study.

Setting. Single-center study at a university clinic. PTA was performed in a sound-treated booth. HTF testing occurred in a quiet room.

Methods. A sample of 25 adults (mean age 50.1 years [SD 14.2]; 68% female) with self-reported mild-to-moderate hearing loss participated. Each contributed 16 thresholds, yielding 400 comparisons. Participants underwent PTA by an audiologist, followed by two independent HTF assessments (start and end of session) using Apple AirPods Pro 2 paired with an iPhone 13. Outcomes included threshold accuracy versus PTA, test-retest reliability, and test duration.

Results. Across 400 comparisons, 86.5% of HTF thresholds were within 10 dB HL of PTA. Root mean square deviation (RMSD) values ranged from 3.3 to 7.9 dB HL (left ear) and 5.8 to 9.7 dB HL (right ear), meeting minimally acceptable accuracy (≤10 dB HL). Test-retest was reliable, 84.1% of thresholds within 5 dB HL and 96.6% within 10 dB HL. Desired reliability (≤6 dB HL) was met at all frequencies except 250 Hz (left ear), which met minimum acceptable level. HTF was significantly faster (median 5.5 minutes) than PTA (10.0 minutes; $P < .001$).

Conclusion. Apple's HTF demonstrated clinically acceptable accuracy and reliability, with improved time-efficiency compared to PTA. Findings support its potential for consumer-led hearing monitoring and OTC hearing aid self-fitting. Further research should assess inter-device reliability and integration with Apple's Hearing Aid Feature.

Keywords

Apple Hearing Test Feature, hearing aids, hearing loss, mobile audiometry, over-the-counter

Received July 17, 2025; accepted February 24, 2026.

Hearing loss affects more than 1.5 billion people worldwide and nearly 1 in 4 US adults, yet it often remains undiagnosed and untreated.[1,2] Untreated hearing loss can contribute to poorer communication outcomes, social isolation, cognitive decline, and reduced quality of life.[3-6] Timely intervention, often involving hearing aids, help mitigate these negative consequences.[1,7] The reference standard for determining hearing thresholds and the foundation for hearing aid fitting has always been pure-tone audiometry (PTA), conducted by trained audiologists in sound-treated environments.[1] However, access to conventional hearing care is frequently limited by cost, geographic barriers, and shortages of trained professionals.[8-10]

In response, there has been a growing shift toward the development and validation of mobile and automated hearing assessments that can be self-administered outside of traditional clinical settings.[11-15] Smartphone-based tests, such as the digits-in-noise (DIN) test, used by the World Health Organization's (WHO's) hearWHO application (app), allow individuals to screen their hearing independently using readily available devices.[16] PTA-based apps, such as the Mimi Hearing Test, have demonstrated the ability to approximate air-conduction thresholds when used with consumer headphones, and

[1]Department of Speech-language Pathology and Audiology, University of Pretoria, Pretoria, South Africa
[2]Virtual Hearing Lab, a Collaborative initiative between the Department of Otolaryngology-Head and Neck Surgery, University of Colorado School of Medicine, Aurora, Colorado, USA
[3]Department of Speech-Language Pathology and Audiology, University of Pretoria, Pretoria, South Africa
[4]Department of Otolaryngology–Head and Neck Surgery, University of Colorado School of Medicine, Aurora, Colorado, USA
[5]UCHealth Hearing and Balance, University of Colorado Hospital (UCHealth), Aurora, Colorado, USA
[6]Department of Speech and Hearing, School of Allied Health Sciences, Manipal Academy of Higher Education, Manipal, India

Corresponding Author:
De Wet Swanepoel, PhD, Department of Speech-Language Pathology and Audiology, University of Pretoria, Lynnwood Road & Roper Street, Room 3-5, Pretoria, South Africa.
Email: dewet.swanepoel@up.ac.za

have previously been integrated into Apple's Health app.[17]

A recent scoping review highlighted that mobile and automated PTA tools can offer accuracy, reliability, and time-efficiency comparable to conventional PTA.[18] Advances such as machine learning and probabilistic modeling have further enhanced these tools, offering versatility, innovative features, and cost-effectiveness beyond traditional methods.[18] Within known limits, these technologies can support task-shifting, self-care, telehealth, and clinical integration.[18] As a result, hearing assessments are increasingly being embedded into consumer electronics, especially into over-the-counter (OTC) hearing aids, helping to expand access to hearing healthcare beyond conventional clinical environments.

A recent example is Apple's Hearing Test Feature (HTF), an integrated, FDA-cleared, pure-tone hearing test, available through iOS/iPadOS for use with AirPods Pro 2.[19] The HTF uses a Gaussian process classification algorithm with Bayesian active learning, allowing it to efficiently estimate hearing thresholds without relying on fixed frequency-step testing.[19] This approach builds on probabilistic audiometry models, previously validated by Barbour and colleagues.[20] Upon completing the test, users can initiate self-fitting of a FDA-approved OTC hearing aid through Apple's Hearing Aid Feature (HAF). This creates an end-to-end, self-managed OTC hearing care solution integrated with consumer technologies for those with perceived mild-to-moderate hearing loss.[19]

While some early validation data is available from Apple[19] that was used for regulatory clearance, the accuracy, test-retest reliability, and efficiency of Apple's HTF have not yet been independently assessed. This study therefore evaluated the accuracy, test-retest reliability, and time-efficiency of Apple's HTF against the reference standard PTA in a controlled simulated home-use environment.

Method

Study Design

The study used a cross-sectional experimental design. Ethical approval for this study was obtained from the Faculty of Humanities at the University of Pretoria, South Africa (Ref: HUM032/1124).

Participants

Twenty-five participants were recruited through purposive sampling using digital advertisements on social media platforms. Inclusion criteria required participants to be 18 years or older, self-report mild to moderate hearing loss, have no history of chronic or recurrent outer or middle ear pathology, and be current iPhone users. All participants provided informed consent.

While a formal priori power analysis was not conducted, consistent with the descriptive, validation-focused design of the study, the sample size aligns with precedent in prior mobile audiometry validation research.[20,21] Each participant contributed 16 threshold comparisons (8 frequencies × 2 ears), resulting in 400 data points. This volume of data enabled precise estimation of accuracy and test-retest reliability.

To justify the adequacy of the sample size, we calculated 95% confidence intervals (CIs) for key performance metrics. The proportion of thresholds within 10 dB HL was 86.5% (95% CI: 82.8%-89.5%). Mean Root Mean Square Deviation (RMSD) values were 5.56 dB HL (95% CI: 4.34-6.69) for the left ear and 7.22 dB HL (95% CI: 6.22-8.35) for the right ear. Test-retest reliability showed strong agreement, with mean Spearman correlation coefficients of $\rho = 0.91$ (95% CI: 0.58-0.98) for the left ear and $\rho = 0.87$ (95% CI: 0.44-0.98) for the right ear. These indicators suggest that the sample size was sufficient to produce stable, interpretable estimates for method comparison, as evidenced by narrow confidence intervals around key performance metrics and alignment with accepted standards in prior validation studies.

Data Collection

Data were collected during a single scheduled session at the Department of Speech-Language Pathology and Audiology, University of Pretoria, South Africa. Each participant underwent a comprehensive audiological evaluation performed by a certified audiologist, including otoscopic examination, tympanometry, and pure-tone air-conduction testing. Audiometry duration was recorded.

Following reference standard audiometry, participants were provided with Apple AirPods Pro 2 and an iPhone 13, along with printed Apple instructions on "How to take a hearing test with your Apple AirPods Pro 2." The Apple HTF was completed independently in a quiet room. No additional instruction or coaching was provided by the researcher. The duration of the test was recorded. After all outcome measures were completed, each participant then completed a second HTF at the end of the same session to assess test-retest reliability.

Efforts to reduce bias included standardized instructions, consistent equipment, and assessment of test-retest reliability.

Material and Apparatus

Reference Standard PTA

Otoscopic examination was performed using a Welch Allyn diagnostic otoscope and tympanometry was conducted using a Maico Diagnostics touchTymp (Maico Diagnostics GmbH). Pure-tone air-conduction thresholds were obtained using a calibrated GSI Audiostar Pro diagnostic audiometer (Grason-Stadler Inc.) with RadioEar DD65 v2 headphones (Demant A/S), conducted in a double-walled soundproof booth. Calibration adhered to International Organization for Standardization (ISO)

389-1:2017 and American National Standards Institute (ANSI) S3.6-2018 standards. Testing was administered by a certified audiologist in accordance with ISO 8253-1:2010 guidelines. Thresholds were recorded at standard octave frequencies from 250 to 8000 Hz, and test duration was recorded using a digital stopwatch.

Apple HTF

The Apple HTF was conducted using Apple AirPods Pro 2 (2nd generation) paired with an iPhone 13 running iOS 18.1. The iPhone was configured with a US Apple ID to access the feature, as it was not yet available in South Africa. Participants followed printed Apple instructions "How to take a hearing test with your Apple AirPods Pro 2." The Apple HTF was accessed by navigating to: Settings > AirPods > Hearing Health > Take a Hearing Test. Self-testing was conducted in a quiet room, rather than a soundproof booth, to simulate typical home environments. Frequency-specific thresholds were automatically generated by the device for both ears. These thresholds can be used to activate Apple Hearing Assistance Features, such as the HAF, which customizes and amplifies sound through the AirPods Pro 2 for individuals with mild-to-moderate hearing loss. However, the HAF was not evaluated in this study. The duration of the Apple HTF was recorded using a digital stopwatch. Each participant completed the Apple HTF twice, once at the beginning and once at the end of the session, to allow assessment of test-retest reliability.

Outcome Measures

Accuracy of Apple HTF

Accuracy was assessed by comparing pure-tone air-conduction thresholds obtained from reference standard PTA to those obtained via the Apple HTF. Thresholds were compared across frequencies to evaluate the agreement between the 2 methods.

Test-Retest Reliability of Apple HTF

Test-retest reliability of the Apple HTF was assessed by comparing thresholds obtained from 2 tests conducted at the beginning and end of the same session. Threshold differences across frequencies were analyzed to assess consistency.

Time-Efficiency of Apple HTF

To assess time-efficiency, the duration of the Apple HTF was compared to that of reference standard PTA.

Data Analysis

All statistical analyses were performed using IBM SPSS Statistics (Version 30.0.0.0) and figures were created using RStudio (Version 2023.06.2). Descriptive statistics (means, standard deviations, and percentages) were used to summarize test durations, hearing thresholds and agreement ranges. Wilcoxon signed-rank tests were used to compare test durations and threshold differences due to data violating the assumption of normality. Spearman's rank correlation coefficients were computed to assess the linear relationship between thresholds across methods (accuracy) and across sessions (test-retest reliability). Root Mean Squared Deviation (RMSD) was calculated as the square root of the average of the squared differences between methods (reference standard PTA vs Apple HTF) or sessions (Apple HTF Test 1 vs Test 2). RMSD was used to quantify agreement, with ≤6 dB HL considered desired accuracy/reliability and ≤10 dB HL considered minimally acceptable.[22,23] Statistical significance was set at $P < .05$. One participant did not complete the Apple HTF due to unclassifiable results and was excluded from analysis.

Results

Participant Characteristics

A total of 25 participants (68% female) were included in the study, with a mean age of 50.1 years (14.2 SD). The majority (88%) had no prior experience using hearing aids. Mean PTA thresholds at 0.5, 1, 2, and 4 kHz, measured using reference standard PTA, were 19.7 dB HL (14.8 SD) for the right ear and 19.2 dB HL (14.5 SD) for the left ear.

Accuracy of Apple HTF

Mean thresholds obtained via reference standard audiometry and Apple HTF are shown in **Table 1** and **Figure 1**. Thresholds were similar between the 2 methods, with 60.7% of thresholds within 5 dB HL, 86.5% within 10 dB HL, and only 13.5% exceeding 10 dB HL. For the left ear, significant differences were observed at 250, 6000, and 8000 Hz, where the Apple HTF yielded slightly better (lower) thresholds at 250 Hz and slightly poorer (higher) thresholds at the higher frequencies. For the right ear, significant differences were observed at 250, 1000, 6000, and 8000 Hz, following a similar pattern to the left ear, although the additional difference at 1000 Hz (where the Apple HTF produced slightly better thresholds) was not observed in the left ear. **Figure 2A** shows threshold differences across frequencies. Overall, the median differences and distribution of hearing threshold differences remained consistent across frequencies, showing no substantial variability. Strong positive correlations were observed between reference standard PTA and Apple HTF thresholds (**Figure 3**). Spearman's rank correlation coefficients (ρ) ranged from 0.72 to 0.98 across test frequencies, indicating a strong relationship between the two methods.

To further quantify accuracy, RMSD values were calculated (**Table 1** and **Figure 4A**). For the left ear, RMSD ranged from 3.3 dB HL (1000 Hz) to 7.9 dB HL

Table 1. Comparison of Reference Standard Pure-Tone Audiometry and Apple Hearing Test Feature (n = 48 Ears, 24 Participants)

	Reference standard audiometry mean (SD) in dB HL	Apple HTF mean (SD) in dB HL	Mean difference[a] (SD)	*P*-value[b]	Threshold Correspondence % (n) 0-5 dB HL	Threshold correspondence % (n) 6-10 dB HL	Threshold correspondence % (n) 11-15 dB HL	Threshold correspondence % (n) ≥ 16 dB HL	RMSD in dB HL[c]
Left ear									
250 Hz	13.0 (9.1)	9.7 (9.4)	2.8 (5.3)	.026[d]	62.5% (15)	29.2% (7)	8.3% (2)		5.9
500 Hz	16.4 (10.4)	16.3 (9.2)	−0.2 (4.5)	.940	70.8% (17)	25.0% (6)	4.2% (1)		4.4
1000 Hz	16.6 (15.0)	15.1 (14.9)	1.2 (3.2)	.114	95.8% (23)	4.2% (1)			3.3
2000 Hz	19.8 (19.6)	19.5 (20.2)	−0.7 (5.5)	.769	79.2% (19)	4.2% (1)	16.7% (4)		5.4
3000 Hz	21.8 (19.7)	23.3 (22.6)	−2.6 (6.9)	.096	45.8% (11)	37.5% (9)	16.7% (4)		7.3
4000 Hz	24.0 (22.1)	24.7 (21.9)	−2.0 (7.8)	.094	50.0% (12)	33.3% (8)	8.3% (2)	8.3% (2)	7.9
6000 Hz	28.2 (20.2)	29.3 (19.6)	−2.9 (5.2)	.017[d]	66.7% (16)	20.8% (5)	12.5% (3)		5.9
8000 Hz	33.0 (23.3)	33.6 (20.7)	−3.2 (6.4)	.017[d]	37.5% (9)	45.8% (11)	16.7% (4)		7.0
Right ear									
250 Hz	15.8 (10.8)	11.3 (10.5)	4.1 (4.3)	<.001[d]	75% (18)	20.8% (5)		4.2% (1)	5.9
500 Hz	18.0 (12.9)	16.3 (13.5)	1.4 (6.8)	.291	58.3% (14)	25.0% (6)	16.7% (4)		6.8
1000 Hz	19.0 (15.1)	16.2 (16.3)	2.8 (5.2)	.019[d]	54.2% (13)	41.7% (10)	4.2% (1)		5.8
2000 Hz	20.0 (18.2)	20.5 (18.7)	−0.9 (7.5)	.763	75.0% (18)	8.3% (2)	8.3% (2)	8.3% (2)	7.4
3000 Hz	20.0 (21.0)	23.1 (23.2)	−3.5 (9.3)	.084	41.7% (10)	29.2% (7)	20.8% (5)	8.3% (2)	9.7
4000 Hz	21.8 (22.5)	23.3 (22.8)	−2.0 (6.1)	.067	66.7% (16)	20.8% (5)	12.5% (3)		6.3
6000 Hz	22.4 (22.8)	27.3 (21.4)	−6.7 (7.2)	<.001[d]	41.7% (10)	25.0% (6)	20.8% (5)	12.5% (3)	9.7
8000 Hz	31.6 (24.2)	33.5 (24.9)	−3.1 (5.5)	.012[d]	50.0% (12)	41.7% (10)	8.3% (2)		6.2

Abbreviations: HTF, Hearing Test Feature; RMSD, Root Mean Square Deviation; SD, standard deviation.

[a]Threshold differences were calculated as reference standard audiometry minus Apple HTF, such that positive values reflect better hearing sensitivity with Apple HTF.

[b]*P*-values are based on Wilcoxon signed-rank tests comparing thresholds between test methods at each frequency.

[c]RMSD was calculated as the square root of the mean of squared threshold differences between reference standard audiometry and Apple HTF at each frequency. RMSD reflects the magnitude of deviation regardless of direction. Lower values indicate closer agreement between the two testing methods.

[d]$P < .05$ considered statistically significant.

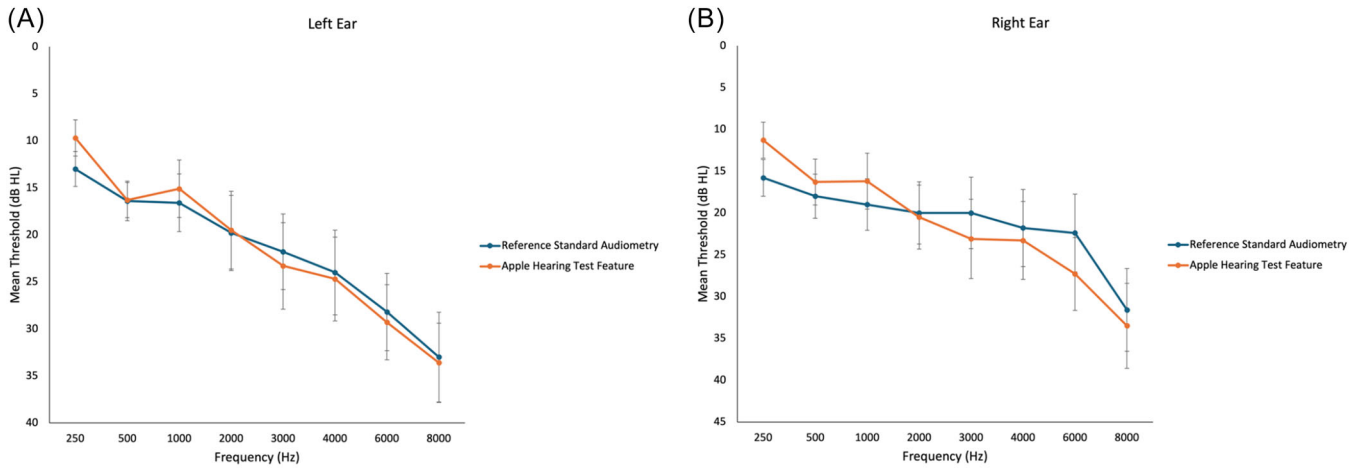

Figure 1. Mean hearing thresholds measured via reference standard pure-tone audiometry and the Apple Hearing Test Feature at each frequency (250-8000 Hz), displayed for (A) left ear and (B) right ear.

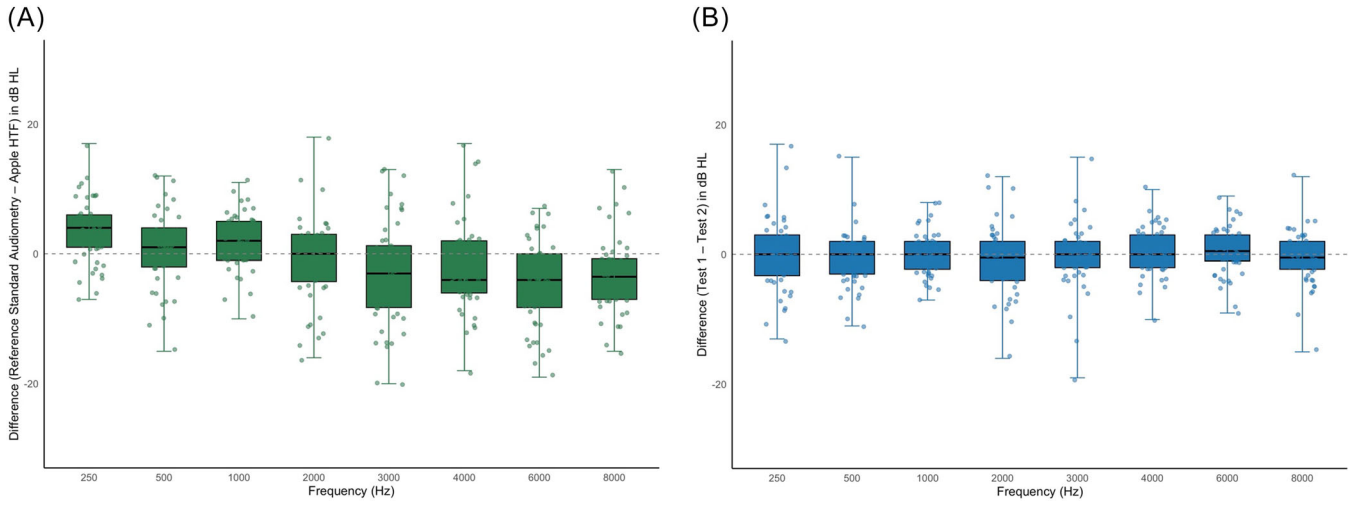

Figure 2. Boxplots illustrating threshold differences across frequencies 250-8000 Hz (n = 48 ears, 24 participants). (A) Differences between reference standard pure-tone audiometry and the Apple HTF. (B) Test-retest differences using the Apple HTF.

(4000 Hz). Desired accuracy (≤6 dB HL) was achieved for 5 out of 8 frequencies and minimal acceptable accuracy (≤10 dB HL) at all frequencies. For the right ear, RMSD values ranged from 5.8 dB HL (1000 Hz) to 9.7 dB HL (3000 and 6000 Hz), meeting desired accuracy at two out of eight frequencies and minimal acceptable accuracy at all.

Test-Retest Reliability of Apple HTF

Test-retest results for the Apple HTF (250-8000 Hz) are presented in **Table 2** and **Figure 5**. Thresholds were highly consistent across sessions with 84.1% of thresholds within 5 dB HL, 96.6% within 10 dB HL, and only 3.4% exceeding 10 dB HL. For the left ear, mean differences ranged from −1.1 dB HL (6000 Hz) to 0.7 dB HL (8000 Hz); for the right ear, from −0.6 dB HL (500 Hz) to 1.5 dB HL (2000 Hz). No significant differences were found between sessions at any frequency. **Figure 2B** shows threshold differences between Test 1 and Test 2. Overall, strong positive correlations were observed between thresholds obtained from Test 1 and Test 2 for both ears (**Figure 6**), with correlation coefficients ranging from 0.60 to 0.98 across frequencies, indicating strong correspondence.

Root Mean Square Deviation (RMSD) values between Test 1 and Test 2, presented in **Table 2** and **Figure 4B**, also confirmed high test-retest reliability. For the left ear, RMSD values ranged from 3.6 dB HL (1000 Hz) to 6.5 dB HL (250 Hz), while for the right ear, values ranged from 3.0 dB HL (1000 Hz) to 4.7 dB HL (2000 Hz). All frequencies for both ears met the minimal acceptable reliability threshold of ≤10 dB HL, and all frequencies achieved the desired reliability criteria of ≤6 dB HL, except for 250 Hz in the left ear.

Time-Efficiency of Apple HTF

The mean duration for reference standard audiometry (250-8000 Hz), conducted by an audiologist, was

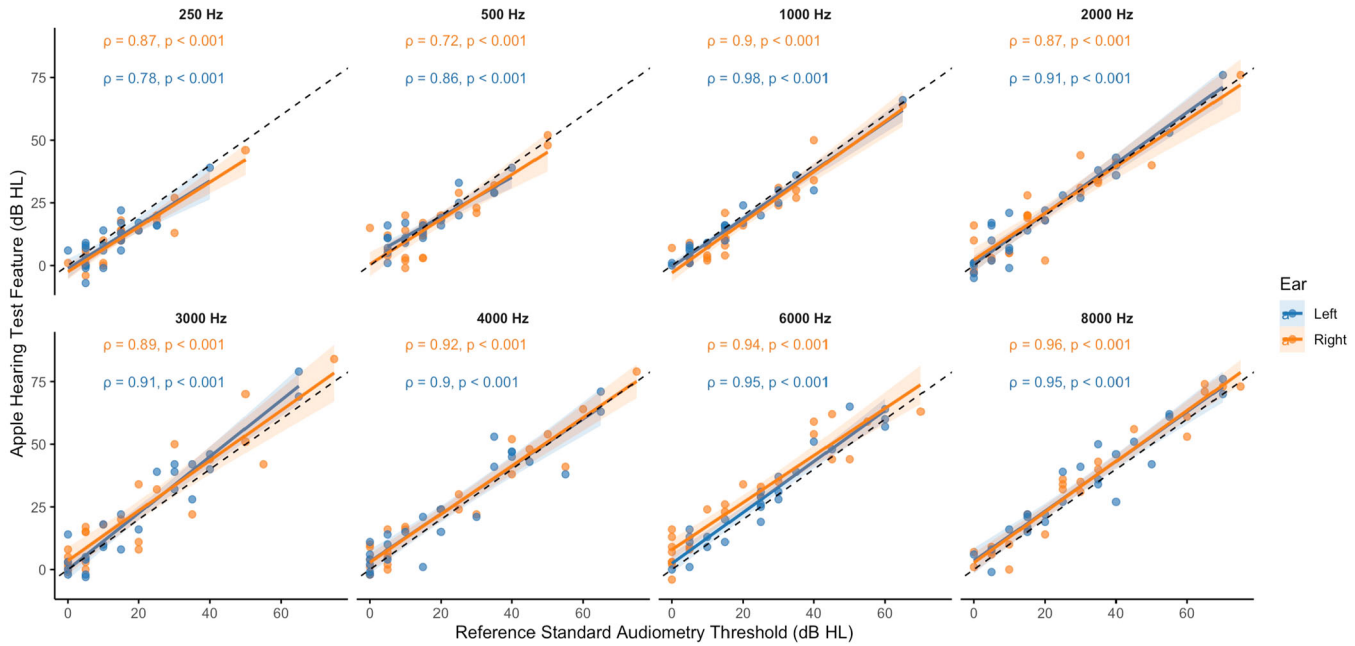

Figure 3. Correlation between reference standard pure-tone audiometry and the Apple Hearing Test Feature at 250-8000 Hz for left and right ears (n = 48 ears, 24 participants).

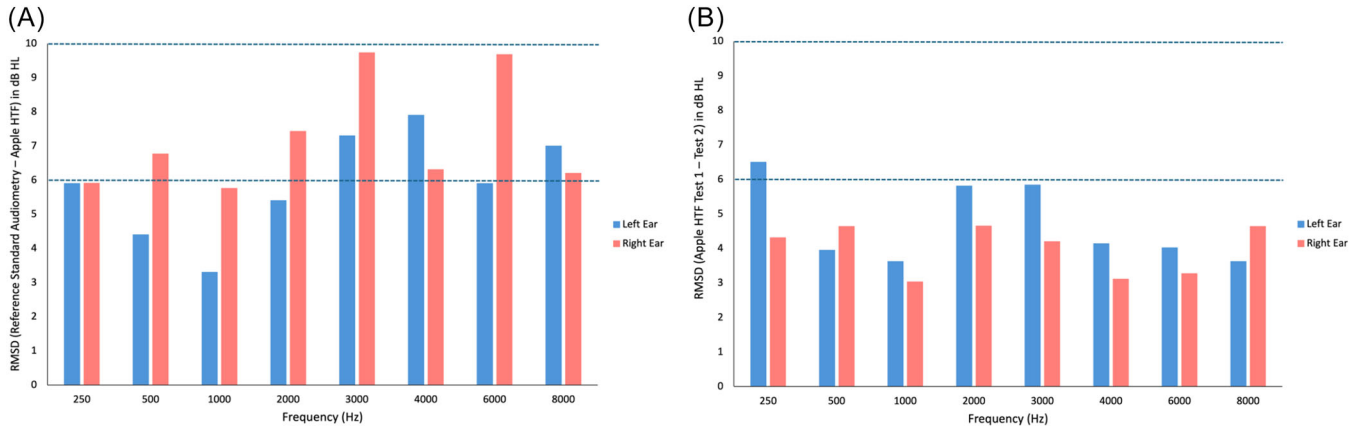

Figure 4. RMSD across 250-8000 Hz for left and right ears (n = 48 ears, 24 participants). (A) Reference standard pure-tone audiometry versus Apple HTF. (B) Test 1 versus Test 2 using Apple HTF. RMSD, Root mean square deviation.

9.8 minutes (1.6 SD; ranging from 7.0 to 14.0 minutes). Participants completed the Apple HTF twice. Test 1 had a mean duration of 5.6 minutes (1.1 SD; ranging from 4.0 to 9.0 minutes), and Test 2 averaged 5.7 minutes (1.0 SD; ranging from 4.4 to 8.0 minutes). Compared to reference standard audiometry, the Apple HTF was significantly shorter in duration (median = 5.5 vs 10.0 minutes), $Z = -4.17$, $P < .001$ (Wilcoxon signed-rank test). On average, participants completed the Apple HTF 4.3 minutes faster than the reference standard audiometry (1.9 SD; ranging from –0.3 to 9.0 minutes).

Discussion

This study evaluated Apple's HTF for accuracy, test-retest reliability, and time-efficiency versus reference standard PTA. The HTF demonstrated accurate thresholds, with all frequencies achieving at least minimally acceptable criteria with reference standard PTA.[22,23] Test-retest reliability met criteria for both minimal and desired reliability at all frequencies except for 250 Hz in the left ear.[22,23] Notably, the HTF was significantly faster than reference standard audiometry. These findings align with Apple's clinical validation study as well as prior research on mobile and automated audiometry, which has shown good agreement with reference standard PTA while offering efficiency gains and the potential for scalability for broader access.[14,15,18,19,24]

Apple's clinical validation study evaluated the HTF against reference standard audiometry in a sample of 201 participants.[19] Apple reported a median absolute deviation (MAD) of 1.81 dB HL for the four-frequency pure-

Table 2. Test-Retest Reliability of Apple Hearing Test Feature (n = 48 Ears, 24 Participants)

	Test 1 mean (SD) in dB HL	Test 2 mean (SD) in dB HL	Mean difference[a] (SD)	*P*-value[b]	Threshold correspondence % (n) 0-5 dB HL	Threshold correspondence % (n) 6-10 dB HL	Threshold correspondence % (n) 11-15 dB HL	Threshold correspondence % (n) ≥ 16 dB HL	RMSD in dB HL[c]
Left ear									
250 Hz	9.7 (9.4)	9.7 (8.1)	0.0 (6.6)	.964	66.7% (16)	20.8% (5)	8.3% (2)	4.2% (1)	6.5
500 Hz	16.3 (9.2)	14.7 (10.2)	1.5 (3.7)	.105	83.3% (20)	16.7% (4)			4.0
1000 Hz	15.1 (14.9)	15.1 (15.8)	–0.0 (3.7)	.961	87.5% (21)	12.5% (3)			3.6
2000 Hz	19.5 (20.2)	19.0 (19.9)	0.4 (5.9)	.936	70.8% (17)	25.0% (6)		4.2% (1)	5.8
3000 Hz	23.3 (22.6)	22.7 (21.5)	0.6 (5.9)	.931	83.3% (20)	8.3% (2)	4.2% (1)	4.2% (1)	5.8
4000 Hz	24.7 (21.9)	25.3 (21.9)	–0.6 (4.2)	.410	87.5% (21)	12.5% (3)			4.1
6000 Hz	29.3 (19.6)	30.5 (20.6)	–1.1 (4.0)	.125	83.3% (20)	16.7% (4)			4.0
8000 Hz	33.6 (20.7)	32.9 (20.3)	0.7 (3.6)	.443	87.5% (21)	12.5% (3)			3.6
Right ear									
250 Hz	11.3 (10.5)	11.5 (10.6)	–0.2 (4.4)	1.000	75.0% (18)	20.8% (5)	4.2% (1)		4.3
500 Hz	16.3 (13.5)	17.0 (13.3)	–0.6 (4.7)	.566	87.5% (21)	4.2% (1)	8.3% (2)		4.6
1000 Hz	16.2 (16.3)	16.2 (17.2)	0 (3.1)	.735	95.8% (23)	4.2% (1)			3.0
2000 Hz	20.5 (18.7)	19.0 (20.3)	1.5 (4.5)	.053	75.0% (18)	20.8% (5)	4.2% (1)		4.7
3000 Hz	23.1 (23.2)	22.9 (23.6)	0.2 (4.3)	.504	87.5% (21)	8.3% (2)	4.2% (1)		4.2
4000 Hz	23.3 (22.8)	23.6 (21.9)	–0.3 (3.2)	.613	91.7% (22)	8.3% (2)			3.1
6000 Hz	27.3 (21.4)	27.1 (20.6)	0.2 (3.3)	.925	91.7% (22)	8.3% (2)			3.3
8000 Hz	33.5 (24.9)	33.0 (25.8)	0.5 (4.7)	.442	91.7% (22)		8.3% (2)		4.6

Abbreviations: RMSD, root mean square deviation; SD, standard deviation.

[a]Threshold differences were calculated as Test 1 minus Test 2, such that positive values reflect better hearing sensitivity during the second test. Mean difference values were rounded to one decimal place; values appearing as 0.0 may represent small but non-zero differences.

[b]*P*-values are based on Wilcoxon signed-rank tests comparing thresholds between test methods at each frequency. $P < .05$ considered statistically significant.

[c]RMSD was calculated as the square root of the mean of squared threshold differences between Test 1 and Test 2 using the Apple HTF at each frequency. RMSD reflects the magnitude of deviation regardless of direction. Lower values indicate closer agreement between the two testing methods.

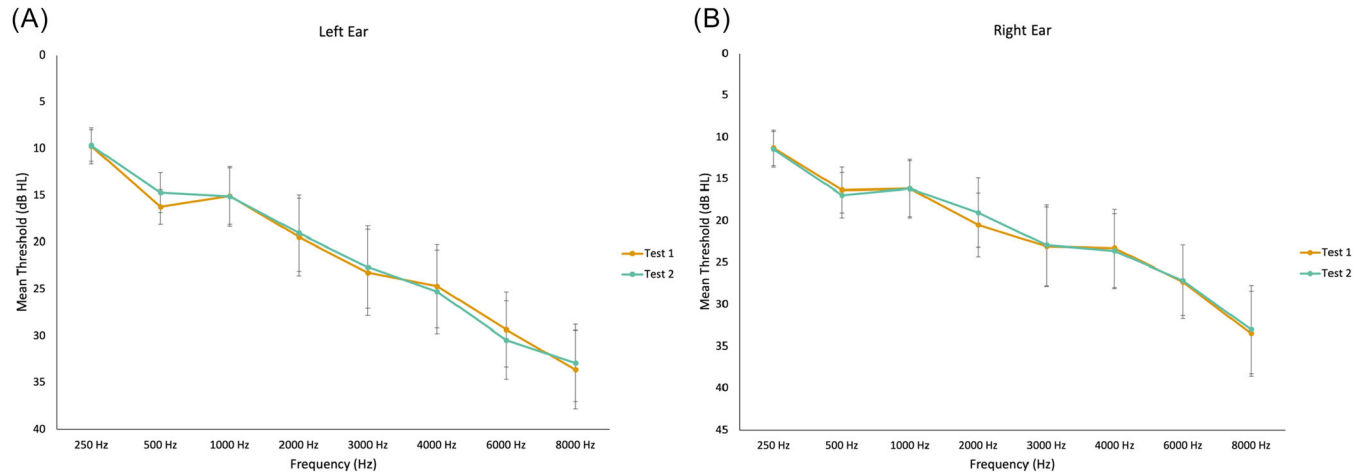

Figure 5. Mean hearing thresholds measured using the Apple Hearing Test Feature at Test 1 and Test 2 for each frequency (250-8000 Hz), displayed separately for (A) left ear and (B) right ear.

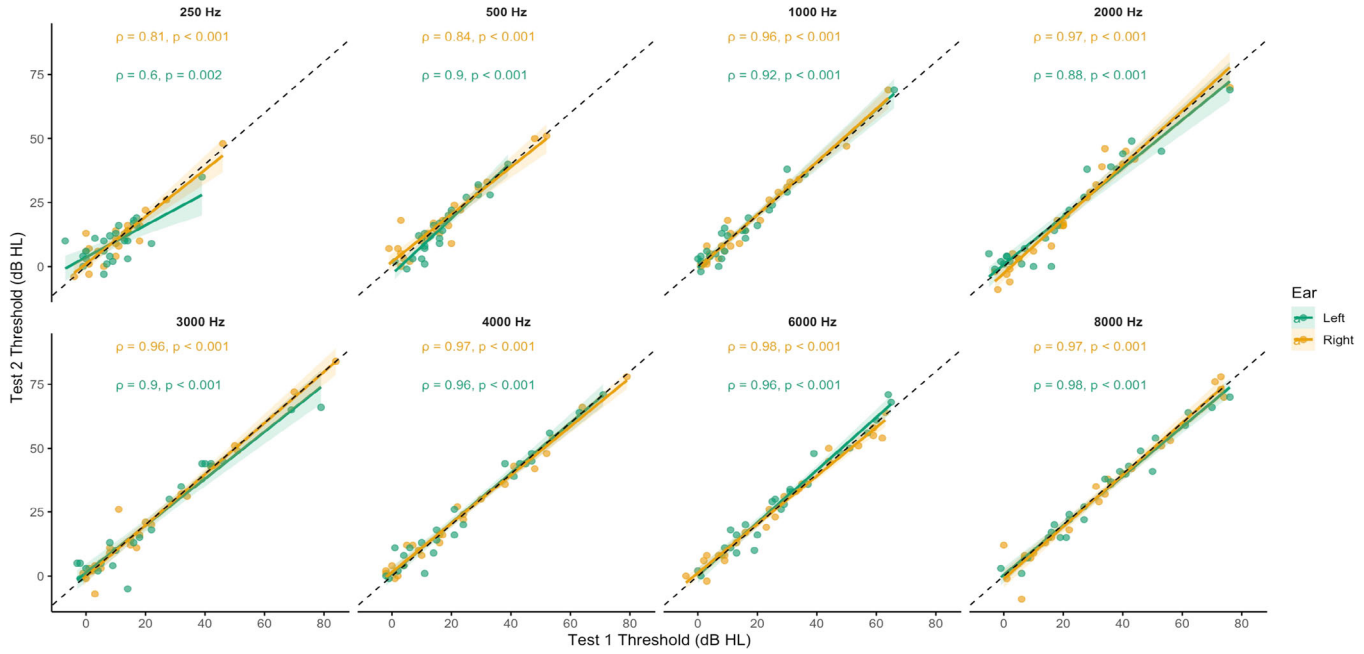

Figure 6. Correlation between test and retest Apple Hearing Test Feature thresholds at 250-8000 Hz for left and right ears (n = 48 ears, 24 participants).

tone average and 1.75 dB HL for the eight-frequency pure-tone average, with high correlations ($r = 0.974$ and $r = 0.979$, respectively) between HTF and reference thresholds.[19] While our study used mean differences, standard deviation, and RMSD to assess agreement, the consistency of findings across both studies provides converging evidence for the clinical accuracy of the HTF.[19] Since Apple's internal validation data are not publicly accessible, the present study provides the first independent validation of the HTF under simulated home-use conditions, strengthening transparency and external validity. Moreover, our study extends Apple's findings by providing RMSD-based accuracy metrics, frequency-specific correlation analyses, dedicated test-retest reliability assessments, and time-efficiency comparisons—elements not reported in Apple's internal clinical documentation.

In the present study, the overall mean difference between HTF and reference standard PTA was −1.0 dB HL (SD 6.0), comparable to meta-analyses of mobile and automated audiometry, including Oremule et al[15] (1.36 dB HL, 95% CI: 0.07-2.66) and Mahomed et al[14] (0.4 dB HL, SD 6.1). This is consistent with broader findings by Wasmann et al,[18] who found only 22% of automated approaches achieved strict accuracy (RMSD ≤ 6 dB HL), while most met minimal acceptable

accuracy (RMSD ≤ 10 dB HL), as observed in our study.[22,23]

Notably, in our study, reduced agreement was mainly at higher frequencies (6000-8000 Hz), similar to patterns observed in Apple's internal clinical validation study[19] and in the meta-analyses by Mahomed et al[14] which reported the largest mean differences at 8000 Hz. Though relevant for speech understanding in noise,[25] our results remain within clinically acceptable limits.[22,23] Other mobile audiometry studies reported greater variability at lower frequencies.[26] Berampu et al[26] highlighted key factors that could potentially influence validity, including background noise, transducer type, calibration, device hardware, user attention, ear canal conditions, and user errors.

Our test-retest results showed an average difference of 0.2 dB HL (SD 4.4) across ears and frequencies, slightly better than the 0.3 dB HL (SD 6.9) reported for automated audiometry in the meta-analysis by Mahomed et al[14] and Swords et al[13] similarly reported high test-retest correlation in app-based audiometry, even in cases where accuracy was more variable. In our study, nearly all frequencies met desired reliability criteria (RMSD ≤ 6 dB HL), demonstrating consistent HTF performance across tests that exceeded what was observed for accuracy.[22,23]

Unlike many mobile audiometry tools that have been validated primarily for screening purposes,[17,27,28] the integration of Apple's HTF into consumer electronics and its role to inform Apple's self-fitting OTC HAF represents a novel advancement.[19] Integrating consumer electronics with software-based medical-grade hearing tests opens new opportunities for adults to self-monitor their hearing status, and also support hearing aid fitting for persons with perceived mild-to-moderate hearing loss. Furthermore, the direct linkage with Apple's HAF offers the possibility of a comprehensive, self-managed hearing solution, supporting models of self-care within hearing healthcare.[15,18]

Certain limitations of Apple's HTF should be considered. The feature does not assess thresholds at 125 Hz and lacks bone conduction testing, which limits its ability to differentiate between conductive and sensorineural hearing loss. Additionally, it does not include automated flagging of asymmetrical hearing loss or incorporate masking protocols in cases of potential cross-hearing, which may delay appropriate medical referral. From a study design perspective, only a single pair of AirPods Pro 2 was used across participants, so inter-device variability was not evaluated, an area for future research. Although the sample size aligns with precedent in prior validation studies, its sample size may still limit broader generalizability, underscoring the need for replication in larger and more diverse samples. Finally, since all participants had self-reported mild-to-moderate hearing loss, consistent with the intended user group for the HTF and Apple's FDA-cleared HAF, the performance of the HTF in individuals with more substantial degrees of hearing loss remains unknown and should be examined in future studies.

Conclusion

Apple's HTF demonstrated accuracy and reliability within established acceptable criteria, and significantly improved time-efficiency compared to reference standard PTA. These findings support its potential for self-monitoring hearing status and as part of a self-fitting OTC hearing aid pathway. Nevertheless, awareness of its limitations, including the absence of bone conduction testing, is important to ensure appropriate use and referral when needed. Further studies should investigate the inter-device reliability of the HTF, its performance across different user environments, and the usability and effectiveness of its integration with Apple's HAF as part of a complete OTC hearing care solution.

Author Contributions

Megan Kruger, contributed to the study design, data collection, data analysis, interpretation of results, and drafting and revision of the manuscript; **Vinaya Manchaiah**, contributed to the study design, data collection, data analysis, interpretation of results, and drafting and revision of the manuscript; **De Wet Swanepoel**, contributed to the study design, data collection, data analysis, interpretation of results, and drafting and revision of the manuscript.

Disclosures

Competing interests: None.

Funding source: None.

Data Availability Statement

Data are not publicly available due to ethical restrictions but may be provided upon reasonable request to the corresponding author.

ORCID iD

Megan Kruger https://orcid.org/0000-0002-7187-9281
Vinaya Manchaiah https://orcid.org/0000-0002-1254-8407
De Wet Swanepoel https://orcid.org/0000-0001-8313-1636

References

1. World Health Organization. World Report on Hearing; 2021. https://www.who.int/publications/i/item/world-report-on-hearing
2. Orji A, Kamenov K, Dirac M, et al. Global and regional needs, unmet needs and access to hearing aids. *Int J Audiol*. 2020;59(3):166-172. doi:10.1080/14992027.2020.1721577
3. Huddle MG, Goman AM, Kernizan FC, et al. The economic impact of adult hearing loss: a systematic review. *JAMA Otolaryngol Head Neck Surg*. 2017;143(10):1040-1048. doi:10.1001/jamaoto.2017.1243
4. Nordvik Ø, Laugen Heggdal PO, Brännström J, et al. Generic quality of life in persons with hearing loss: a systematic literature review. *BMC Ear Nose Throat Disord*. 2018;18(1):1. doi:10.1186/s12901-018-0051-6
5. Olusanya BO, Neumann KJ, Saunders JE. The global burden of disabling hearing impairment: a call to action.

Bull World Health Organ. 2014;92(5):367-373. doi:10.2471/BLT.13.128728

6. Livingston G, Huntley J, Sommerlad A, et al. Dementia prevention intervention and care. *Lancet*. 2020;396(10248):413-446.https://doi.org/10.1016/S0140-6736(20)30367-6
7. Ferguson MA, Kitterick PT, Chong LY, et al. Hearing aids for mild to moderate hearing loss in adults. *Cochrane Database Syst Rev*. 2017;9(9):012023. doi:10.1002/14651858.CD012023.pub2
8. Jenstad L, Moon J. Systematic review of barriers and facilitators to hearing aid uptake in older adults. *Audiol Res*. 2011;1(1):e25. doi:10.4081/audiores.2011.e25
9. Knoetze M, Manchaiah V, Mothemela B, et al. Factors influencing hearing help-seeking and hearing aid uptake in adults: a systematic review of the past decade. *Trends Hear*. 2023;27:23312165231157255. doi:10.1177/23312165231157255
10. Vestergaard Knudsen L, Öberg M, Nielsen C, et al. Factors influencing help seeking, hearing aid uptake, hearing aid use and satisfaction with hearing aids: a review of the literature. *Trends Amplif*. 2010;14(3):127-154. doi:10.1177/1084713810385712
11. Bright T, Pallawela D. Validated smartphone-based apps for ear and hearing assessments: a review. *JMIR Rehabil Assist Technol*. 2016;3(2):e13. doi:10.2196/rehab.6074
12. Irace AL, Sharma RK, Reed NS, et al. Smartphone-based applications to detect hearing loss: a review of current technology. *J Am Geriatr Soc*. 2021;69(2):307-316. doi:10.1111/jgs.16985
13. Swords C, Twumasi E, Fitzgerald M, et al. A multicenter validity study of four smartphone hearing test apps in optimized and home environments. *Laryngoscope*. 2024;134(6):2864-2870. doi:10.1002/lary.31256
14. Mahomed F, Swanepoel DW, Eikelboom RH, et al. Validity of automated threshold audiometry: a systematic review and meta-analysis. *Ear Hear*. 2013;34(6):745-752. https://journals.lww.com/ear-hearing/fulltext/9000/validity_of_automated_threshold_audiometry__a.99522.aspx
15. Oremule B, Abbas J, Saunders G, et al. Mobile audiometry for hearing threshold assessment: a systematic review and meta-analysis. *Clin Otolaryngol*. 2024;49(1):74-86. doi:10.1111/coa.14107
16. De Sousa KC, Smits C, Moore DR, et al. Global use and outcomes of the hearWHO mHealth hearing test. *Digital Health*. 2022;8:20552076221113204. doi:10.1177/20552076221113204
17. Moazzami C, Gagnon C, Bertrand L, et al. The emerging future of mobile audiometry: a prospective validation study of the mimi hearing test application. *Otol Neurotol*. 2024;45(7):740-744.https://doi.org/10.1097/MAO.0000000000004229
18. Wasmann JW, Pragt L, Eikelboom R, et al. Digital approaches to automated and machine learning assessments of hearing: scoping review. *J Med Internet Res*. 2022;24(2):e32581. doi:10.2196/32581
19. Apple. Using AirPods Pro 2 with iPhone and iPad to Help Protect, Assess, and Assist Hearing. https://www.apple.com/au/health/pdf/Hearing_Health_Features_on_AirPods_Pro_2_October_2024.pdf
20. Barbour DL, Howard RT, Song XD, et al. Online machine learning audiometry. *Ear Hear*. 2019;40(4):918-926.https://doi.org/10.1097/AUD.0000000000000669
21. Swanepoel DW, Mngemane S, Molemong S, et al. Hearing assessment—reliability, accuracy, and efficiency of automated audiometry. *Telemed e-Health*. 2010;16(5):557-563. doi:10.1089/tmj.2009.0143
22. ASHA. Guidelines for manual pure-tone threshold audiometry [Guidelines]; 2005. https://www.asha.org/policy/gl2005-00014/
23. Margolis RH, Glasberg BR, Creeke S, et al. AMTAS®: Automated method for testing auditory sensitivity: validation studies. *Int J Audiol*. 2010;49(3):185-194. doi:10.3109/14992020903092608
24. Mahomed F, Swanepoel DW, Eikelboom RH, et al. Validity of automated threshold audiometry: a systematic review and meta-analysis. *Ear Hear*. 2013;34(6):745-752.https://doi.org/10.1097/01.aud.0000436255.53747.a4
25. Raphael LJ, Borden GJ, Harris KS. *Speech science primer: Physiology, acoustics, and perception of speech*. 5th ed. Lippincott Williams & Wilkins; 2007.
26. Berampu RW, Adriztina I, Sofyan F, et al. Accuracy and pitfalls in the smartphone-based audiometry examination. *Iran J Otorhinolaryngol*. 2024;36(2):421-431. doi:10.22038/IJORL.2024.71187
27. Al-Abri R, Al-Balushi M, Kolethekkat A, et al. The accuracy of IOS device-based uhear as a screening tool for hearing loss: a preliminary study from the middle east. *Oman Med J*. 2016;31(2):142-145. doi:10.5001/omj.2016.27
28. Szudek J, Ostevik A, Dziegielewski P, et al. Can u hear me now? Validation of an iPod-based hearing loss screening test. *J Otolaryngol Head Neck Surg*. 2012;41(Suppl 1):S78-S84. doi:10.2310/7070.2011.110089