Validation of a Self-Administered Audiometry Application: An Equivalence Study

Jonathon P. Whitton, AuD; Kenneth E. Hancock, PhD; Jeffrey M. Shannon, AuD; Daniel B. Polley, PhD

**Objectives/Hypothesis:** To compare hearing measurements made at home using self-administered audiometric software against audiological tests performed on the same subjects in a clinical setting

**Study Design:** Prospective, crossover equivalence study

**Methods:** In experiment 1, adults with varying degrees of hearing loss (N = 19) performed air-conduction audiometry, frequency discrimination, and speech recognition in noise testing twice at home with an automated tablet application and twice in sound-treated clinical booths with an audiologist. The accuracy and reliability of computer-guided home hearing tests were compared to audiologist administered tests. In experiment 2, the reliability and accuracy of pure-tone audiometric results were examined in a separate cohort across a variety of clinical settings (N = 21).

**Results:** Remote, automated audiograms were statistically equivalent to manual, clinic-based testing from 500 to 8,000 Hz (P < .02); however, 250 Hz thresholds were elevated when collected at home. Remote and sound-treated booth testing of frequency discrimination and speech recognition thresholds were equivalent (P ≤ 5 × 10^-5). In the second experiment, remote testing was equivalent to manual sound-booth testing from 500 to 8,000 Hz (P ≤ .02) for a different cohort who received clinic-based testing in a variety of settings.

**Conclusion:** These data provide a proof of concept that several self-administered, automated hearing measurements are statistically equivalent to manual measurements made by an audiologist in the clinic. The demonstration of statistical equivalence for these basic behavioral hearing tests points toward the eventual feasibility of monitoring progressive or fluctuant hearing disorders outside of the clinic to increase the efficiency of clinical information collection.

**Key Words:** Telehealth, mobile health, audiology, automation, diagnostics, audiometry, speech, auditory processing.

**Level of Evidence:** 2b.

**Laryngoscope, 00:000–000, 2016**

**INTRODUCTION**

Diagnosing and treating hearing loss (HL) represents a formidable public health challenge that is expected to worsen in the near future. An estimated 48 million individuals in the United States currently live with HL, and this number is expected to nearly double over the next 20 years due to changing population demographics.1,2 Audiological service providers in the United States are already in high demand, and the mismatch between the capacity of Audiologists and the number of patients who need assessment and treatment is expected to grow to unsustainable levels in the coming decade.2,3 Widening the lens beyond the United States, of the estimated 360 million people worldwide who live with disabling HL, most live in regions where audiological services are scarce.4

To address the supply and demand mismatch in the U.S. hearing services market, the efficiency or size of the current cadre of hearing healthcare providers would have to increase by a factor of two over the next two decades.2,3 There is no indication that this gap in patient care will be filled by a doubling of trained personnel.2,3 Logically, this points toward the central importance of increasing clinical efficiency as the most plausible means of solving the impending supply and demand dilemma. Some savings are likely to be found in developing more efficient standards for the diagnosis and management of HL. Audiologic behavioral tests represent a direct measure of hearing ability and currently require substantial time costs from doctorally trained clinicians because they are performed manually, making this form of information collection a reasonable target for increased efficiency.

Audiologic behavioral tests can be divided into three classes: 1) absolute detection thresholds, 2) feature discrimination thresholds, and 3) speech recognition testing. As the gold standard test of hearing sensitivity, measurements of absolute detection thresholds (i.e., the pure tone audiogram) are the most commonly performed
audiological behavioral tests. The latter two classes of behavioral tests are employed in some cases to obtain supplementary information to pure tone audiograms (e.g., multilevel word recognition testing5,6) but also when characterizing disorders that have little or no correlation with basic audibility, such as auditory neuropathy spectrum disorder and auditory processing disorder.7–14 The administration procedures for all three classes of behavioral tests are standardized to ensure comparable results across testing facilities, making them amenable to automation and improved clinical efficiency.3,15,16

Margolis and Morgan3 estimated that by automating 80% of pure tone audiograms currently performed by clinicians, each audiologist could see an additional 139 patients annually (1.5 million patients total), which would only partially close the gap between supply and demand in the United States for basic hearing services. But the Margolis and Morgan capacity estimates were based on the need for testing to be performed in available sound-treated booths with specialized equipment. The processing power of smartphones and tablet computers that are now ubiquitous in both developed and emerging nations17,18 make it feasible to distribute applications that could perform sophisticated, automated audiological testing with consumer-grade hardware outside of sound-treated booths. If remote and clinic-based audiology tests were ever to be considered interchangeable for the eventual purposes of differential diagnosis or monitoring response to treatment, it must first be established whether remote and clinic-based tests are statistically equivalent. This level of rigorous quantitative scrutiny has not been applied to automated, remote hearing assessment.19–26 This was the purpose of the present study.

We programmed an interactive application for tablet computers (Surface Pro 2, Windows 8.1 operating system; Microsoft Corporation, Reading WA). Using consumer-grade headphones model AE2i; Bose Corporation, Framingham, MA, this application provided a means to measure pure-tone audiograms, frequency discrimination thresholds, and speech-in-noise recognition scores from participants of varying ages and hearing abilities. We chose these three measures because each test represents one of the three classes of audiological behavioral tests (absolute detection, feature discrimination, and speech recognition) and thus affords a proof of concept for automated, remote testing across the behavioral test battery. The accuracy of these remote tests was assessed by comparing home-based results to measurements made in the same subjects by an audiologist in a clinical sound booth. Our aim was to test whether the results of home-based, self-administered audiology tests were equivalent to the same tests administered by an audiologist in a sound-treated booth.

MATERIALS AND METHODS

Participants

All procedures were approved by the Human Studies Committee at Massachusetts Eye and Ear Infirmary and the Committee on the Use of Humans as Experimental Subjects at the Massachusetts Institute of Technology in Cambridge, Massachusetts. Informed consent was obtained from each participant. For experiment 1, 19 individuals (7 female) participated in the study, presenting with various degrees of hearing loss (Fig. 1A), ranging in age from 25 to 82 years, and reporting a variety of auditory complaints (Supp. Fig. S1). Two subjects only completed the first two time points of the study. For experiment 2, we recruited 21 additional subjects to test the repeatability and generalization of the audiometric findings when clinic-based testing was performed by various ear, nose, and throat (ENT) and audiology clinics (see Supporting Methods for more detail). Most of the participants in experiment 2 reported chronic tinnitus and presented with hearing thresholds that ranged from normal hearing to profound HL.

Procedures

For experiment 1, participants were asked to complete pure-tone audiometric, frequency discrimination, and speech-in-noise recognition testing in a sound-treated room under the care of a licensed audiologist on two separate visits (clinic visits 1 and 2). Between clinic visits, participants were given a tablet

Fig. 1. (A, left) Air-conduction thresholds from 19 subjects collected in the clinic (red = right ear, blue = left ear). (A, right) Distribution of PTAs (.5–2 kHz) in the sample plotted according to American Academy of Otolaryngology–Head and Neck Surgery recommendations.48 (B, top) Study design. (B, bottom) Average ambient noise measurements made at the subjects’ homes (gray line) and in the sound-treated clinical booth (black line) with the tablet computer. Actual sound-treated clinical booth noise (black broken line) measured with a high-quality microphone and signal analyzer revealed that the measurement noise floor of the table exceeded all clinic and some home ambient-noise measurements (see Supporting Methods for details). HL = hearing level; PTA = pure tone average; SPL = sound pressure level.
was employed according to the optimized masking approach\textsuperscript{29} when detection thresholds between ears differed by 35 dB or greater based on the reported interaural attenuation for the TDH-\textcopyright 39 headphone (Telephones; Farmingdale, NY).\textsuperscript{30} For remote testing, participants interacted with an automated tablet audiogram application using the Bose AE2i consumer-grade circumaural head-phones (Framingham, MA) (see Supporting Methods and Supp. Fig. S2 for description of equipment calibration). These head-phones do not provide active noise cancellation. Although the equipment and reliance on computer guidance were different from the standard approach for threshold measurements, the testing algorithm\textsuperscript{28} followed the same rules as a clinician. Specifically, tones were presented for 1 second with an interstimulus interval of 3 to 7 seconds. Responses (indicated by virtual button press) were considered hits if they occurred within a response window of 2.5 seconds. The tone level was decreased by 10 dB following hits and increased by 5 dB following misses. Threshold was defined as the lowest level that evoked a hit response on two of three ascending runs, or three of six runs if there was no concordance after three runs.

**Frequency Discrimination and Speech-in-Noise Thresholds.** For both clinic and remote testing, frequency discrimination thresholds were measured diotically through an interactive two-alternative forced-choice software interface. The center frequency was roved around 2,000 Hz, and the threshold was adaptively measured using standard psychophysical procedures (see Supporting Methods for details).\textsuperscript{31}

Speech-in-noise thresholds were estimated using two lists of 35 monosyllabic words from the Northwestern University 6 lists\textsuperscript{32} presented diotically at 70 dB HL. Our software allowed participants to initiate trials wherein they heard a female speaker produce monosyllabic words while six-talker babble played in the background. Participants were then cued to respond, and their voices were recorded via the tablet microphone. Recorded responses were transmitted to Massachusetts Eye and Ear Infirmary computer servers and scored offline. For clinic-based testing, an audiologist listened to their responses through the talk back on the audiometer and scored whether or not each word was correct. Thresholds were computed using the Spearman–Kärber equation.

**Statistical Testing**

We quantified the accuracy of automated home-based hearing assessments by performing statistical equivalence tests using the two one-sided testing (TOST) procedure.\textsuperscript{27} The TOST requires a defined clinical equivalency margin (i.e., difference in measurement that would not be clinically significant). The shorthand equivalency margin for clinic-based pure tone audiometry is considered ± 10 dB,\textsuperscript{33–36} based on American National Standards Institute specifications for equipment variation and the testing resolution used with most audiometers.\textsuperscript{37} The clinical margin of equivalence can also be empirically defined from the test–retest difference for any diagnostic measure. Using this approach, we conservatively defined equivalency as an audiometric difference value that fell within the measurement margin of error (80% confidence), as defined by a previous audiometric reliability study\textsuperscript{38} and our own clinical test–retest dataset for speech and frequency discrimination. Cases where the 90% confidence interval for accuracy (clinic vs. home) falls within the clinical equivalency margin provide a visual proxy for the TOST equivalence hypothesis. For all TOST statistical tests are listed in Supporting Table SIII. We tested for differences in reliability between remote and clinic-based testing by comparing test–retest differences for each condition using mixed-model analysis of variance for multiple comparisons and two-sample t tests for paired comparisons. We also compared test–retest variances for remote and clinic-based measurements with two-sample F-tests. Data were log transformed for statistical testing when they were not normally distributed (normality tested with one-sample Kolmogorov–Smirnov tests). A post hoc power analysis indicated the study was adequately powered (β = 0.2) to test for equivalence (α = 0.025) for all reported comparisons except for ≤ 250 Hz and ≥ 12,500 Hz pure tone thresholds. Importantly, a result was only considered statistically significant if criterion significance level (P ≤ 0.025) was met for all measurement conditions. For example, a home-based audiometric threshold was not considered equivalent to a clinic-based threshold unless significance was met for both left and right ears during both the first and second home to clinic comparisons. We viewed this method as the most conservative approach to data analysis.

**RESULTS**

**Automated, Remote Pure-Tone Audiometry Is Largely Equivalent to Clinic-Based Testing**

We measured pure tone thresholds at home and in the clinic from subjects with a wide range of hearing thresholds (Fig. 1A), ages (Supp. Fig S1A), histories of auditory impairment, and computer tablet experience (Supp. Fig. S1B–C). The mean differences between home and clinic testing were small and fell within the clinical equivalency margin from 500 Hz to 8,000 Hz (Fig. 2A and Supp. Table SIV, N = 19 participants). This finding was repeatable, with the same pattern emerging from the home versus clinic comparisons at tests 1 and 2 in both left and right ear comparisons. However, very low-frequency thresholds (≤ 250 Hz) were slightly but consistently elevated when collected at home. Low-frequency threshold elevation could be attributed to elevated levels of low-frequency background noise in the home environment (Fig. 1B bottom) and was only observed in subjects with thresholds in the normal range (Supp. Fig. S3). We next compared the test–retest reliability of home versus clinic audiogram measurements. The difference scores for each testing environment are plotted.
in Figure 2B and demonstrate considerable overlap between measurements, with no repeatable significant differences in means \((P \geq 0.44\), group and group \(\times\) frequency interaction terms) or variances \((P \geq 0.09\), Supp. Table SI).

**Automated, Remote Measurement of Frequency Discrimination and Word Recognition in Noise Is Equivalent to Clinic-Based Measurements**

We did not expect a slight threshold elevation at very low frequencies to have any influence on discrimination and recognition abilities that are measured at sound levels well above ambient room noise. Indeed, we observed that home tests were statistically equivalent to the clinic-based measurements for suprathreshold tests of frequency discrimination (Fig. 3A) and speech recognition in noise (Fig. 3B) across both testing repetitions (home 1 vs. clinic 1 and home 2 vs. clinic 2, \(N = 19\) participants). We next compared the test–retest reliability of home versus clinic frequency discrimination and speech-recognition-in-noise threshold measurements. The difference scores for each testing environment are plotted in Figure 3A–B (right side) and reveal considerable overlap between measurements, with no statistical difference in means or variances (frequency discrimination, \(P = 0.6\) means and \(P = 0.6\) variances; speech recognition in noise, \(P = 0.09\) means and \(P = 0.5\) variances).

**Age Predicts Absolute Perceptual Scores But Not Remote Testing Accuracy**

Hearing loss is most prevalent in middle-aged and older adults.\(^1\) Thus any method that is intended to
increase the efficiency and accessibility of hearing assessment must be amenable to implementation in this population. While older adults are adopting mobile technologies at higher rates, they still lag behind younger adults. Based on this discrepancy, one could speculate that older adults might be less able to accurately self-administer hearing measures using mobile devices and applications. We enrolled a diverse sample in order to assess the predictive power of participant age on test results (Supp. Fig S1A). Subject age could explain a significant amount of the variability in audiometric thresholds and speech recognition in noise thresholds (accounting for 38% and 52% of the variability respectively, \( P < 0.005 \) for both). However, neither age nor degree of HL was a significant factor when comparing measurements made in clinical settings versus unsupervised testing at home (\( R^2 \leq 0.17, P \geq 0.12 \) for all associations). Furthermore, accuracy was not worse for any of the audiological tests when individuals who did not own tablets were compared to tablet owners (\( P \geq 0.67 \) for all comparisons). Within our sample, there was no evidence that participant age, hearing status, or tablet ownership conferred a disadvantage in diagnostic accuracy on any of the automated, remote audiological tests that were administered.

**Generalization of the Home Testing Approach to a Different Sample**

As a final step, we conducted a second experiment wherein we tested the generalizability of remote testing in a separate sample of patients (\( N = 21 \)) who had previously received clinical testing from other audiology and ENT clinics (see Supporting Methods). As an example, Figure 4A shows that an audiogram collected at a Midwestern ENT clinic from a patient who presented with sudden sensorineural HL in her left ear (solid lines) was captured with a high degree of accuracy using tablet-based software at that patient’s home 2 days later (broken lines). When comparing clinical-based audiograms from all patients' medical records to automated, remote test results, we observed equivalency from 500 Hz to 8,000 Hz, replicating the results from experiment 1 (Fig. 4B).

**DISCUSSION**

There are three classes of behavioral audiological tests that are used for hearing assessment (absolute detection, feature discrimination, and speech recognition). We chose a single test from each class and assessed whether measurements made from home with an automated tablet application using consumer-grade headphones were statistically equivalent to manual measurements made by an audiologist in the clinic. Absolute detection thresholds collected remotely were equivalent to clinical measures across frequencies that convey speech information (500–8,000 Hz). Frequency discrimination and speech-recognition thresholds were equivalent when measured remotely and in the clinic.

Validation studies for automated, remote hearing testing have predominately focused on the sensitivity and specificity of audiograms or speech-in-noise recognition tests to identify individuals with elevated detection thresholds. One exception was a study that reported the differences between tone detection thresholds obtained with an iOS application in patients’ homes versus manual measurements made in the clinic. Although statistical equivalency was not tested in that study, the findings were qualitatively similar to the data described here.

Whereas the behavioral tests examined in this study were as accurate when performed remotely by an application as when an audiologist manually collected them in the clinic, several obstacles prevent the realization of interchangeable remote and clinic-based test results: 1) When hearing sensitivity was normal, accuracy of remote audiogram measurements was reduced for 250 Hz tones, likely as a consequence of...
environmental noise contamination. 2) The results of air-conducted tests must be interpreted with caution in the absence of initial otoscopic examination. 3) Without coupling bone-conduction to air-conduction measurements, the data from remote audiograms do not afford clinicians the necessary information to characterize type in addition to severity of HL. We do not believe that these obstacles are insurmountable. Active 41 and passive 42 noise-reduction techniques reduce low-frequency threshold elevations measured outside of sound-treated booths and are incorporated in consumer-grade circumaural and in-ear headphones. Additionally, smartphone compatible equipment exists to remotely obtain otoscopic images and transmit these data to clinicians for review. 43 Although the widespread release of Google Glass (Mountain View, CA) has been suspended, it served as an example of consumer-grade hardware that stimulated the cochlea via bone-conduction pathways. 44,45 The experiments reported here suggest that, without further hardware solutions, the remote testing approach could serve as a means to monitor patients with known pathology or as an initial screening wherein normal scores would be interpretable, but measured loss would provide only screening level information. If remote testing were ever to move beyond screening to provide true diagnostic-grade measurements, combined hardware and software solutions will need to reduce ambient low-frequency noise levels below approximately 0 dB HL at the tympanic membrane 44,47 and provide information concerning external and middle ear transmission.

CONCLUSION
Automated, unsupervised audiometry is a feasible, accurate, and reliable approach for the measurement of tone detection thresholds, frequency discrimination abilities, and word recognition in noise scores. These three behavioral tests represent each major class of audiologic behavioral measures. This study provides a proof of concept that automated, remote testing in relatively quiet environments can provide equivalent accuracy and reliability to clinic-based measures across a battery of audiometric behavioral tests.

Acknowledgment
We thank M. Ravicz for assistance in performing acoustic measurements. We thank Drs. A. Remenschneider and E. Kozin for providing feedback on an earlier version of this article.

AUTHOR CONTRIBUTIONS
J.W., K.H., and D.P. designed the experiments; K.H. programmed the software; J.W. and J.S. performed the behavioral experiments; J.W. analyzed the data; and J.W. and D.P. wrote the article.

BIBLIOGRAPHY
6. Jerger J, Jerger S. Diagnostic significance of PB word functions. Arch Oto-
27. Killon MC, Wilber LA, Gudmunsen GI. Insert earphones for more inter-


