Pure-Tone Audiometry With Forward Pressure Level Calibration Leads to Clinically-Relevant Improvements in Test–Retest Reliability

Judi A. Lapsley Miller,[1] Charlotte M. Reed,[2] Sarah R. Robinson,[1,3] and Zachary D. Perez[2]

Objectives: Clinical pure-tone audiometry is conducted using stimuli delivered through supra-aural headphones or insert earphones. The stimuli are calibrated in an acoustic (average ear) coupler. Deviations in individual-ear acoustics from the coupler acoustics affect test validity, and variations in probe insertion and headphone placement affect both test validity and test–retest reliability. Using an insert earphone designed for otoacoustic emission testing, which contains a microphone and loudspeaker, an individualized in-the-ear calibration can be calculated from the ear-canal sound pressure measured at the microphone. However, the total sound pressure level (SPL) measured at the microphone may be affected by standing-wave nulls at higher frequencies, producing errors in stimulus level of up to 20 dB. An alternative is to calibrate using the forward pressure level (FPL) component, which is derived from the total SPL using a wideband acoustic immittance measurement, and represents the pressure wave incident on the eardrum. The objective of this study is to establish test–retest reliability for FPL calibration of pure-tone audiometry stimuli, compared with in-the-ear and coupler sound pressure calibrations.

Design: The authors compared standard audiometry using a modern clinical audiometer with TDH-39P supra-aural headphones calibrated in a coupler to a prototype audiometer with an ER10C earphone calibrated three ways: (1) in-the-ear using the total SPL at the microphone, (2) in-the-ear using the FPL at the microphone, and (3) in a coupler (all three are derived from the same measurement). The test procedure was similar to that commonly used in hearing-conservation programs, using pulsed-tone test frequencies at 0.5, 1, 2, 3, 4, 6, and 8 kHz, and an automated modified Hughson-Westlake audiometric procedure. Fifteen adult human participants with normal to mildly-impaired hearing were selected, and one ear from each was tested. Participants completed 10 audiograms on each system, with test-order randomly varied and with headphones and earphones refitted by the tester between tests.

Results: Fourteen of 15 ears had standing-wave nulls present between 4 and 8 kHz. The mean intrasubject SD at 6 and 8 kHz was lowest for the FPL calibration, and was comparable with the low-frequency reliability across calibration methods. This decrease in variability translates to statistically-derived significant threshold shift criteria indicating that 15 dB shifts in hearing can be reliably detected at 6 and 8 kHz using FPL-calibrated ER10C earphones, compared with 20 to 25 dB shifts using standard TDH-39P headphones with a coupler calibration.

Conclusions: These results indicate that reliability is better with insert earphones, especially with in-the-ear FPL calibration, compared with a standard clinical audiometer with supra-aural headphones. However, in-the-ear SPL calibration should not be used due to its sensitivity to standing waves. The improvement in reliability is clinically meaningful, potentially allowing hearing-conservation programs to more confidently determine significant threshold shifts at 6 kHz—a key frequency for the early detection of noise-induced hearing loss.

[1]Mimosa Acoustics, Champaign, Illinois, USA; [2]Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA; and [3]Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA.

Key words: Calibration, Forward pressure level, Hearing conservation, Noise-induced hearing loss, Permanent threshold shift, Pure-tone audiometry, Significant threshold shift.

(Ear & Hearing 2018;XX;00–00)

INTRODUCTION

An abiding problem in detecting significant threshold shifts (STS) in hearing-conservation programs (HCPs) is poor test–retest reliability at 6 kHz, which is also a frequency sensitive to noise-induced hearing loss (NIHL; Humes et al. 2005). The use of insert earphones (probes) designed for otoacoustic emission (OAE) testing, which contain a miniature microphone in addition to the speakers, allows for individualized in-the-ear (in situ) calibrations that can improve the accuracy of presented stimulus levels. However, due to standing waves in the ear canal, the sound pressure level (SPL) measured at the microphone can have a deep minimum above 3 kHz that can be up to 20 dB different from the stimulus level at the eardrum (Siegel 1994). Calibrating the target stimulus level using the forward pressure level (FPL) rather than the total SPL circumvents these standing-wave errors, and improves both accuracy (test validity) and test–retest reliability (Withnell et al. 2009, 2014). The present study was undertaken to determine the clinical relevance of this improvement for individual listeners assessed in HCPs for noise-induced changes to hearing levels. In this study, we measured test–retest reliability for a prototype audiometer with an ER10C earphone (containing both a microphone and speakers) calibrated three ways, in comparison to a standard audiometer with coupler-calibrated TDH-39P supra-aural headphones (widely used in HCPs).

Significant Threshold Shift

In the clinical context of an HCP, high test–retest reliability is desired because it translates to smaller STS criteria. Smaller criteria mean smaller amounts of hearing loss can be reliably detected without necessarily increasing the number of false positives needing follow-up.

STS, in general, is defined as the change in threshold that is "significant" such that there has been a real change in hearing. How "significant" is determined and its specific value varies across HCPs and has changed over time. For instance, STS criteria can be determined and compared by simulation and statistical analysis, and by database analysis from real-world HCPs. This includes consideration of follow-up audiograms that confirm persistence of a change in hearing or not, and consideration of improvements in hearing, both of which can be used to estimate the false-positive rate (Royster & Royster 1986; Dobie 2005; Schlauch & Carney 2007). Our preferred approach is

0196/0202/2018/XXXX-00/0 • Ear & Hearing • Copyright © 2018 Wolters Kluwer Health, Inc. All rights reserved • Printed in the U.S.A.

Copyright © 2018 Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.

<zdoi; 10.1097/AUD.0000000000000555>

to derive STS criteria from a statistical analysis of test–retest variability on a group of non–noise-exposed ears (along with a cost–benefit analysis of detecting true hearing loss versus dealing with many false positives). In our previous studies, we have used a statistical definition of STS based on either the standard error of measurement or the intrasubject SD of a non–noise-exposed group, which allows comparison across studies (Lapsley Miller et al. 2004, 2006; Marshall et al. 2009). With certain assumptions, the intrasubject SD is equivalent to the SE of measurement (McMillan et al. 2013), both of which are commonly used as measures of reliability. Once a statistically-determined STS has been established, HCP programs can determine with confidence which frequencies or frequency averages should be monitored depending on the costs involved. The problem of an inflating false-positive rate when multiple comparisons are made must also be considered (Schlauch & Carney 2007, 2011; Konrad-Martin et al. 2016). It is important to determine the frequency or frequencies most likely to show hearing loss with sufficient reliability to make sensitive and clinically-meaningful interpretations of hearing health.

A distinction should be made between STS criteria, determined statistically, and "standard threshold shifts" used by Occupational Safety and Health Administration, National Institute for Occupational Safety & Health, the Department of Defense and other standard bodies, where factors other than reliability also come into play (i.e., cost–benefit analyses). These standards are not set in stone, but may change over time as new research comes to light.

Headphone and Earphone Calibration in Current Practice

Coupler-calibrated supra-aural headphones are the most common method of delivering the tonal stimuli for pure-tone audiometry (PTA) in HCPs. Coupler-calibrated insert earphones for audiometry, such as the ER-2 and ER-3A (Etymōtic Research, Elk Grove Village, IL), have been widely available for some time. These earphones have only a speaker and no microphone in the probe assembly, which means they too can only be calibrated in a coupler and not in-the-ear. They are not commonly used in HCPs, presumably because their advantages do not yet outweigh their disadvantages. Supra-aural headphones are easier to fit than insert earphones because they do not need to be inserted into the ear canal (with the concomitant risks of cerumen impaction, etc.), but their limitations have long been known. These include poor noise exclusion, lower interaural attenuation, and potential ear-canal collapse (Killion & Villchur 1989). TDH-39P headphones (Telephonics, Farmingdale, NY) are widely used in HCPs, despite known poor characteristics at 6 kHz due to a strong peak in the frequency-response function when measured in a coupler, which does not correlate to the response when coupled to a human ear (Rudmose 1964).

Both headphone and insert earphone transducers are calibrated in an artificial ear (also known as an ear simulator or coupler), which represents an average adult ear, using the reference equivalent threshold SPL (RETSPL) for that earphone (ANSI 2003b). Such a calibration cannot account for variations in acoustic transmission for an individual ear. Thus, the stimulus level output in an individual ear may differ from target. Additionally, and more so for insert earphones, the stimulus level is dependent on the enclosed volume between the earphone and eardrum, which depends on insertion depth. Without further knowledge of the acoustics of the individual ear and earphone insertion, stimulus calibration cannot be improved.

In-the-Ear Calibration Using OAE Earphones

Using an OAE insert earphone, which contains both a microphone and speakers, the stimulus level may be set based on a coupler SPL (CPL) calibration (Zebian et al. 2012) or an in-the-ear SPL calibration (Siegel 2002) or FPL. In an in-the-ear SPL calibration, a stimulus such as a wideband chirp or tone is output into the ear canal, and the response of the ear canal and middle ear is measured at the microphone. The earphone voltage is then adjusted to ensure the measured stimulus level matches the target level. At frequencies below about 2 kHz, the total SPL measured at the microphone may be used to accurately set the target pressure level at the eardrum. However, above 3 to 4 kHz, standing waves between the earphone and the eardrum can cause differences of up to 20 dB between the SPL at the eardrum and the SPL measured at the microphone (Siegel 1994). Alternatively, the stimulus may be calibrated using an in-the-ear FPL calibration, which ensures that the target stimulus level matches the actual stimulus level incident on the eardrum (Scheperle et al. 2008, 2011; Lewis et al. 2009; McCreery et al. 2009; Withnell et al. 2009; Souza et al. 2014; Withnell et al. 2014), aside from small acoustic losses in the ear canal (Abur et al. 2014). As for SPL, the FPL also has the units of dB re 20 μPa (i.e., dB SPL).

The relationship between SPL, FPL, and ear-canal standing waves may be understood as follows. The total sound pressure measured at the microphone is composed of forward (toward the eardrum) and reverse (away from the eardrum) pressure waves. The forward and reverse pressure waves can interfere constructively or destructively depending on their relative phases. At frequencies where the two waves are out of phase and are therefore interfering with each other destructively, a deep null is seen in the SPL measured at the microphone. But there is no such null at the eardrum. These frequencies of maximum interference are called standing-wave frequencies. The specific frequencies where the standing waves occur depend on the distance between the microphone and the eardrum.* For example, a standing wave at 6 kHz corresponds to a distance of about 14 mm between the microphone and the point of the largest eardrum reflection. If the stimulus level is set based on the total SPL at the microphone, the actual stimulus level near the standing-wave frequency will be much higher than the target level. For PTA measurements, this could result in hearing thresholds that seem much lower than they truly are. The solution is to isolate the forward-going pressure component at the microphone, thus avoiding standing-wave effects in the stimulus calibration and better representing the true stimulus incident on the eardrum.

To calculate the forward pressure component as a function of frequency, the wideband acoustic immittance (WAI) is measured for each probe insertion (Scheperle et al. 2008; Withnell et al. 2009; Feeney et al. 2013; Allen et al. 2016). WAI measurements require a Thévenin calibration, using acoustic cavities with

* Below 8 kHz, insert-earphone measurements are typically only affected by the lowest standing-wave frequency, corresponding to a distance of one-quarter wavelength. The one-quarter wavelength standing wave occurs in a tube that is closed at one end and open at the other. For an insert-earphone configuration, the eardrum represents the closed end of the tube, while the earphone represents the open end, due to the earphone sound source. The distance between the eardrum and the probe is not well defined for in-the-ear measurements due to the angled eardrum, and eardrum delay (Puria & Allen 1998) may make the earphone–eardrum distance appear longer, decreasing the standing-wave frequency.

Copyright © 2018 Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.

known WAI properties to determine the relationship between the voltage and the sound output of the earphone (Allen 1986). The probe need not undergo a full Thévenin calibration for each test, but instead only run if a quick probe verification test does not pass with the desired accuracy. Given a Thévenin calibration, WAI, the SPL, and the FPL may be directly calculated from a single in-the-ear calibration measurement (Scheperle et al. 2008; Withnell et al. 2009). This process takes only seconds. Once the FPL is derived, the earphone voltage can be adjusted so that the FPL matches the target level.

The use of WAI to separate the forward pressure component from the total sound pressure has been verified in cylindrical cavities (Lewis et al. 2009; Scheperle et al. 2011) and for real-ear measurements (McCreery et al. 2009). FPL calibration is less sensitive to earphone insertion depth than many other calibration methods (Souza et al. 2014). Removing standing-wave effects by using FPL calibration improves the accuracy (i.e., validity) of pure-tone audiometric thresholds compared with SPL calibration (Lewis et al. 2009; McCreery et al. 2009; Withnell et al. 2009, 2014). Withnell et al. (2014) also demonstrated that test–retest reliability was higher for FPL and CPL calibration of earphones, compared with in-the-ear SPL calibration, but did not report their results as a function of frequency.

Adding WAI (Allen et al. 2016) and OAE tests to HCPs (Lapsley Miller & Marshall 2007), which require insert earphones with microphones and speakers, may motivate PTA testing using the same apparatus, such that one earphone and insertion may be used for all tests. A key question therefore is which calibration method should be used for the PTA stimulus, and, in particular, which method provides the best test–retest reliability. Reliability in HCPs is of primary concern because the main purpose of PTA testing in HCPs is to monitor individuals longitudinally, looking for changes in hearing.

Table 1 outlines the key validity and reliability errors for headphone and insert earphones, for each calibration method.

In this study, we examine the effect of calibration method on PTA test–retest reliability from 0.5 to 8 kHz at audiometric frequencies. The aim was to establish if the anticipated improvement in reliability with FPL calibration was large enough to have clinical significance, that is whether it produced smaller STS criteria that would allow smaller changes in hearing to be reliably detected. A prototype audiometer with ER10C insert earphones that allows for FPL, SPL, and CPL calibrations was compared with an audiometer in standard clinical use in HCPs. The clinical system used TDH-39P supra-aural headphones with CPL calibration. The key statistic for comparison was the intrasubject SD, which is a measure of test–retest reliability. This statistic can be converted to an STS criterion, which provides a way to compare real-world performance for detecting changes in hearing thresholds in individual ears across systems and calibration types.

MATERIALS AND METHODS

Participants

Seventeen adults were screened for eligibility and 15 completed the study (10 women, 5 men; age 18 to 30 years). Participants were screened for age (between 18 and 40 years), normal otoscopy, normal tympanogram with tympanometric peak pressure within 50 daPa of 0 daPa, and no recent noise exposure. Participants' hearing thresholds were screened using the Benson audiometer (see Equipment section) in the default test mode. They could have a moderate hearing loss up to 50 dB HL, providing there was a known etiology and the loss had been unchanging in recent times. Thirteen of the 15 participants had hearing within normal limits of ≤20 dB HL; 1 participant had a hearing loss of 35 and 40 dB HL at 6 and 8 kHz, respectively, and another participant had a hearing loss of 25 dB at 6 kHz.

If both ears passed screening criteria, 1 ear was chosen at random as the test ear (7 left ears, 8 right ears). One person did not continue because a stable probe fit was not achieved in either ear. One person was withdrawn from the study because they were unable to achieve stable thresholds on many trials across both test systems even after reinstruction. All testing was conducted at the Massachusetts Institute of Technology. The experimental protocol was approved by Massachusetts Institute of Technology's institutional review board and conducted in compliance with regulations and ethical guidelines for experimentation with human subjects. Participants provided informed consent and were paid for their participation in the experiment.

Equipment

Tympanometry was conducted with a GSI 37 Auto Tymp portable screening tympanometer (Grason-Stadler, Eden Prairie, MN).

Two audiometers were compared in this study—a Benson CCA-200mini audiometer and an OtoStat-HCP prototype audiometer. The "standard" audiometer was a Benson CCA-200mini with TDH-39P earphones and supra-aural MX-41/AR cushions (Benson Medical, Minneapolis, MN), connected to a Lenovo laptop running CCA-200mini, Version 7.11 software. The BAS-200 bioacoustic simulator was used for daily biologic calibration checks (Benson Medical Instruments Co., Reference Note 1). The Benson system was set to run automated audiograms using the modified Hughson-Westlake method using pulsed tones (typically used in HCPs because of a higher incidence of tinnitus in these noise-exposed populations). The screening audiogram was run using the default settings. For the test

TABLE 1. Main sources of validity and reliability errors from the calibration method used

Device	Calibration	Accuracy/Validity Error	Reliability Error
Supra-aural Earphone	Coupler	Individual ear variations from average ear	Variations in headset placement can cause variations in stimulus level
Insert Earphone	Coupler	Individual ear variations from average ear	Variations in earphone insertion depth and angle can cause variation in stimulus level
Insert Earphone	In-the-ear sound pressure level	Standing waves can result in calibration errors up to 20 dB from target	Variations in earphone insertion depth and angle moves frequency of standing-wave null
Insert Earphone	In-the-ear forward pressure level	Individual ear variations are minimized and standing-wave errors are eliminated	Less affected by changes in probe depth and angle on repeated measures

Copyright © 2018 Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.

conditions, to allow for comparison, certain settings were modified from their defaults. "Historical starting levels" and "variable starting levels," which modify the stimulus level of the first presentation depending on the participant's previous responses, were disabled. Instead, all tests started from 40 dB HL. "Terminate stimulus on response" was also disabled so that the pulsed tones played for their entire duration. The maximum level was set to 70 dB HL and the minimum level set to −10 dB HL.

The "experimental" audiometer was a prototype OtoStat-HCP (Rev C hardware) system (Mimosa Acoustics, Champaign, IL) connected to a Dell desktop computer running prototype OtoStation software (v1.0.1.12103) with ER10C insert earphones and foam eartips (Etymōtic Research, Elk Grove Village, IL). The ER10C earphone consists of a probe assembly with one microphone channel (used for SPL and FPL calibration) and two speaker channels, one of which was used for the stimulus delivery. The same system was used to perform WAI tests using Mimosa Acoustics' OtoStat MEPA module (which was identical to the commercial OtoStat 2.0 version). The PTA module was configured as an automated Type IV audiometer using the modified Hughson-Westlake method with pulsed tones, similar to that currently used in military HCPs (ANSI 2003a, 2010; ANSI/ASA 2009; Department of Defense, Reference Note 2). Marshall's automated "CLIN" procedure was used as a guideline (Marshall & Hanna 1989; Marshall et al. 1996). The target stimulus levels were set using the FPL calibration, and the equivalent coupler, hearing, and sound pressure levels were all derivable from the same WAI measurement.

To enable fair comparisons, the Benson audiometric thresholds were converted from dB HL to dB re 20 µPa by removing the RETSPL adjustment for TDH-39 earphones (ANSI 2010). Note that this transformation has no effect on variability. All subsequent analyses and comparisons are presented in dB re 20 µPa for all calibration methods.

Audiometric test frequencies were 1, 2, 3, 4, 6, 8, and 0.5 kHz, presented in that order.

Procedure

Screening and test sessions were run in 1 day, within a 3-hr block, with the participant taking breaks as needed. One experienced research assistant did all the testing in a double-walled sound-treated booth, under supervision from an audiologist. The audiologist made all determinations about each participant's auditory status. Before testing began each day, the audiometer underwent a functional check by a normal-hearing person and a biologic check with the bioacoustic simulator as per the user manual to check the output was free from distorted or unwanted sounds and that the levels were correct (Benson Medical Instruments Co., Reference Note 1). The OtoStat probe calibration was verified in the OtoStat cavity set using the manufacturer protocol (Mimosa Acoustics, Reference Note 3).

Trial-Pair Blocks • Each block consisted of one test on the OtoStat and one test on the Benson audiometer. The order was randomly determined for each block. After each audiogram, the earphone or headset was removed from the participant and refitted. The tester was instructed to achieve a good fit or placement, but was not asked to achieve either an identical or different fit to previous blocks. A good placement for the insert earphone was to insert the foam tip deeply, preferably past the tragus such that ½ to ¾ of the tip was in the ear canal. Three ear-tip sizes were available. Ten complete trial-pair blocks were obtained for each participant. If the headset or earphone moved or fell out, the earphone or headset was refitted and the audiogram was repeated. A foam earplug was placed in the contralateral nontest ear when testing with the OtoStat audiometer to make it comparable with testing on the standard audiometer where headphones covered both ears.

The tester was not made aware of the role of standing waves for in-the-ear calibration, and did not deliberately try to enhance differences by adjusting probe depth to ensure that standing-wave nulls were at test frequencies.

RESULTS

For each ear, three alternative calibrations for the same audiogram for the OtoStat+ER10C probe system were compared with the audiogram for the coupler-calibrated Benson system with TDH-39P headphones. For the OtoStat system, the forward and microphone pressure responses in each ear canal were calculated along with the equivalent volume in the ear canal, and a coupler pressure response was derived for the ER10C probe. The mean and intra- and intersubject SDs for hearing thresholds were then calculated for each ear and across all ears for all calibration types and systems.

In-the-Ear Pressure Responses

The 10 in-the-ear pressure response spectra from a broadband chirp stimulus from the OtoStat are shown in Figure 1 for each of the 15 ears. These pressure responses are used to calibrate the stimulus level presented to the ear during the threshold testing so that it meets the desired target level. Below 500 Hz, almost all the forward-going pressure is reflected from the eardrum, such that the reflected pressure has approximately the same magnitude and phase as the forward pressure at the probe microphone; thus, the FPL is half the SPL (i.e., 6 dB lower). At higher frequencies, the relationship between FPL and SPL becomes more complicated and depends more on individual ear characteristics and the residual ear-canal length (Lewis et al. 2009; McCreery et al. 2009). The chirp stimulus is output with constant voltage across frequency so the overall shape of the pressure response spectra, with increased response around 4 kHz, reflects the ER10C probe characteristics. Variations from this pattern are due to individual ear characteristics.

Standing-wave nulls are apparent for many participants between 4 and 8 kHz, indicated by sharp dips in the SPL spectra (gray lines). In comparison, FPL response spectra show a smooth response through those same frequency regions (e.g., consider participant 10 where the probe depth varies with each insertion, which moves the standing-wave null frequency around between 5 and 7 kHz for the SPL responses, whereas the FPL responses are consistent with repeated insertions).

The equivalent ear-canal volume is the volume of air between the end of the ER10C probe tip and the tympanic membrane, and can be estimated from the WAI measurement. It can be used to interpret variability in probe fit and depth and to identify acoustic leaks (Groon et al. 2015). Larger ear-canal volumes result in lower overall pressure and smaller volumes result in higher pressure, for the same output voltage. Table 2 shows the equivalent volume (mean and SD) over the 10 measurements for each ear, calculated by the OtoStat software. In general, ear-canal volumes were smaller than that assumed by a 2-cm^3 artificial ear coupler. Participant 12 had an acoustic leak on 4 of the

Copyright © 2018 Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.

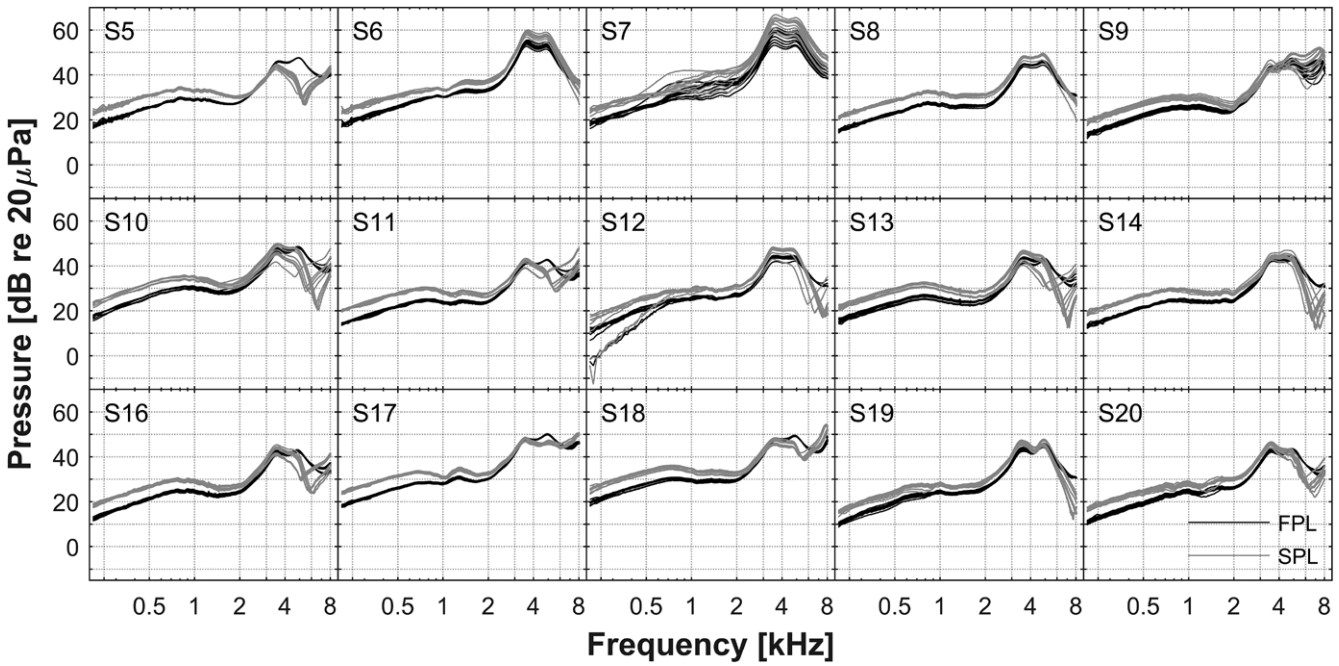

Fig. 1. Plotted for the 15 ears are the 10 pressure responses for the chirp stimulus used for the OtoStat in-the-ear calibrations. The probe was refitted in the ear canal between each measurement. Sound pressure level (SPL; in gray) measured at the probe microphone. Forward pressure level (FPL; in black) derived from the SPL.

10 measurements. Variations in the probe depth with repeated insertions move the pressure spectra up and down for the same ear (e.g., participant 7 showed large variations in probe depth, whereas participant 8 showed a consistent depth with refitting; this is also reflected in the lower SD for equivalent volume for participant 8 in Table 2).

Figure 2 plots the maximum absolute difference between any two calibrations (frequency-by-frequency, for each individual ear) from Figure 1, separately for SPL and FPL calibrations. This comparison represents the maximum error attributable to the calibration if two audiograms are compared (e.g., a baseline and an annual audiogram in a HCP) for these ears. In the frequency region where standing waves are likely to affect adult ears (i.e., above 2 kHz), the maximum differences between any two measurements across participants ranged from 5.4 to 25 dB for SPL calibrations. When considering individual ears, the maximum difference was always less for FPL calibrations compared with SPL calibrations, ranging from 4.5 to 10.5 dB. Below 2 kHz, the maximums were typically lower than 3 dB, other than for participant 12 where there was an acoustic leak.

TABLE 2. Mean and SD for equivalent ear-canal volume (cm^3), which is the volume between the end of the ER10C probe tip and the tympanic membrane, estimated by the OtoStat software

ID	M	SD
5	1.49	0.12
6	1.47	0.13
7	1.41	0.27
8	0.81	0.05
9	1.14	0.13
10	1.40	0.11
11	0.95	0.06
12	1.65 (9.51)	0.44 (11.50)
13	0.76	0.08
14	1.16	0.11
16	1.22	0.06
17	1.35	0.07
18	1.08	0.18
19	2.32	0.77
20	1.45	0.12

Participant 12 had an acoustic leak on 4 of the 10 measurements, which resulted in abnormally large and variable equivalent volumes when those measurements were included (values in brackets).

Post hoc, thresholds using a coupler calibration for the OtoStat ER10C earphone were calculated by: (1) measuring the pressure response to the calibration stimulus using a BK 4157 (Brüel & Kjær, Nærum, Denmark) artificial ear with the DB 2012 ear canal extension, (2) subtracting the in-the-ear SPL calibration for each individual measurement from the coupler pressure, and (3) adding this difference to the measured hearing thresholds in dB SPL. By removing the effect of the individual in-the-ear adjustment and replacing it with the average-ear calibrated level, referred to here for convenience as CPL, we obtained an intrasubject comparison for the ER10C insert earphones to the clinical audiometer with coupler-calibrated TDH-39P headphones.

PTA Measurements

For the 2 audiometers, 3 calibration methods, 7 test frequencies, and 10 repeats, the intersubject means and SDs (Fig. 3A) and the mean intrasubject SDs (MISD; Fig. 3B) were calculated for all 15 ears. To further investigate the intrasubject variation, the individual intrasubject SDs for each ear were also plotted (Fig. 4).

The mean hearing thresholds differed by up to approximately 10 dB across systems and calibration methods, with the best agreement at 1 kHz. Comparing FPL- and SPL-calibrated results, the differences seen at the lower frequencies can be

Copyright © 2018 Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.

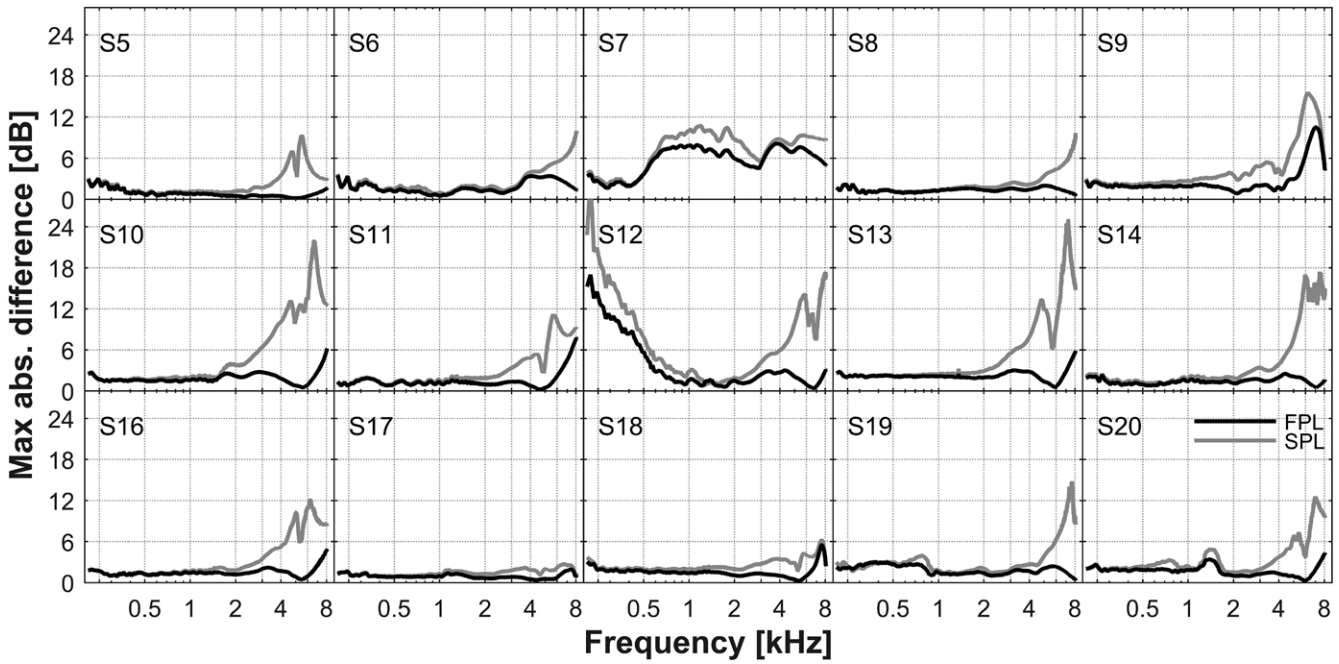

Fig. 2. The maximum absolute difference for each ear (i.e., worst case scenario) between any two pressure responses from Figure 1, separately for sound pressure level (SPL; in gray) and forward pressure level (FPL; in black), for the OtoStat system. This indicates the maximum error attributable to calibration when comparing two audiograms.

predicted from the in-the-ear pressure responses in Figure 1, which differ by 6 dB. Mean thresholds for the higher frequencies are much closer in value, except at 6 kHz where there is a 3-dB difference. This is the frequency region most affected by standing-wave nulls, as can be seen in Figure 1. For the two CPL-calibrated comparisons (which used different earphones on different equipment and were calibrated with different couplers), there was disparity in hearing thresholds at 0.5 and 6 kHz of 9.3 and 10.9 dB, respectively. Comparing CPL calibration to the SPL and FPL for the ER10C probe, CPL calibration generated lower thresholds at higher frequencies for these ears.

All the OtoStat measurements with the ER10C probe across all calibration methods were less variable than those made in the same session on the same ears with the standard audiometer using TDH-39P headphones (Fig. 3B). Of note is the poor test–retest variability for the TDH-39P headphones at 6 kHz. Within the OtoStat results, all three calibration methods had similar MISDs of around 2 dB SPL below 4 kHz, but above 4 kHz they diverged. FPL calibration showed the least variability at 6 and 8 kHz, and to a lesser extent at 4 kHz, and this variability was similar to lower frequencies. For in-the-ear SPL calibration, the frequencies affected by standing-wave nulls (6 to 8 kHz) had higher test–retest variability (as can be predicted from Figs. 1 and 2). CPL calibration for the ER10C also showed consistent, low variability across frequencies, only slightly higher than FPL calibration at 4 to 8 kHz.

DISCUSSION

Hearing Threshold Reliability for the ER10C Earphone for Three Calibration Methods

The primary result (Fig. 3B) shows for the OtoStat system and ER10C probe that intrasubject variability for FPL calibration remains around 2 dB across the tested frequency range 0.5 to 8 kHz, whereas for the identical audiograms (but with different calibrations applied), the intrasubject variability for the CPL and SPL calibrations show increased variability at 4 to 8 kHz. The only difference is due to the calibration assumed in the analysis, leading to the conclusion that FPL calibration results in higher test–retest reliability for PTA measurements.

For the OtoStat+ER10C measurements below 3 kHz, as expected, FPL calibration did not provide additional reliability to SPL or CPL calibration. Above 3 kHz, the lower variability for FPL was due in part to (a) more accurate calibration through regions with standing waves (because the SPL SDs are higher than the FPL SDs, especially at 6 kHz), and (b) in-the-ear calibration (because the CPL SDs are slightly higher than the in-the-ear FPL SDs).

In-the-ear SPL calibration is not in common use for PTA; clinical PTA systems with insert earphones all use a coupler calibration by necessity (because there is no microphone). The current results (and those already in the literature) reinforce the notion that in-the-ear SPL calibration should not be used for PTA because an up to 20 dB error in threshold estimation is larger than the size of clinically-meaningful threshold shifts.† In an HCP where measurements are repeated over time, each probe insertion can result in different fits and depths each time, which can move the frequency and depth of the standing-wave null (e.g., participants 10, 12, 14, and 20 in Fig. 1). By comparison, when calibrating with FPL, the calibrations are much more consistent at the standing-wave null frequency regions, because

† In-the-ear SPL calibration, however, is the most common form of calibration for clinical OAE systems. These standing-wave calibration errors are a source of error at higher frequencies, but until now there has been no clinically-viable alternative (Siegel 2002). FPL calibration has been shown to decrease OAE variability (Scheperle et al. 2008), and we anticipate that this will also translate into clinically-meaningful improvements in significant-OAE-shift criteria.

Copyright © 2018 Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.

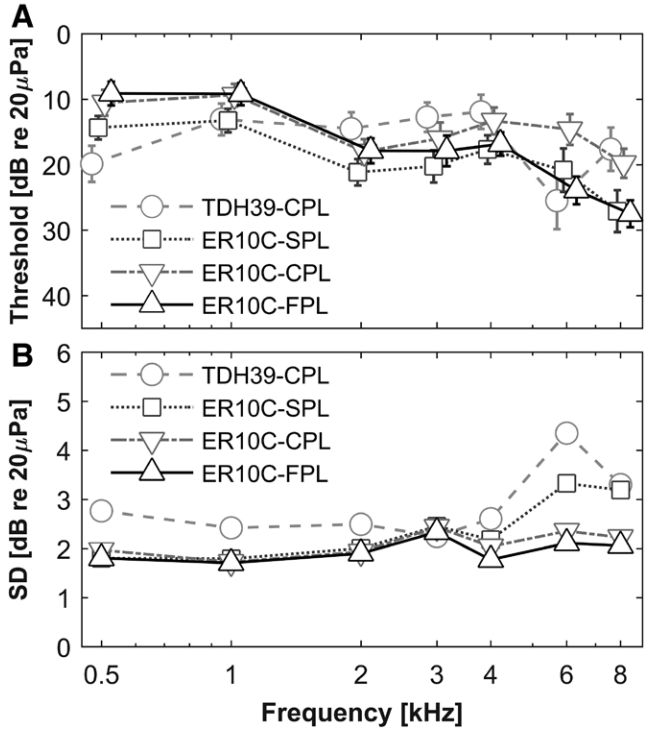

Fig. 3. A, Group mean hearing thresholds and 1 SD error bars (frequencies are slightly offset to allow a better view of the error bars) and (B) mean intra-subject SD. Benson audiometer with THD-39P headphones and a coupler calibration (TDH39-CPL light gray circles with dashed lines), OtoStat with ER10C earphones with SPL calibration (ER10C-SPL dark gray squares with dotted lines), OtoStat with a coupler calibration (ER10C-CPL gray triangles with dot-dashed lines), and OtoStat with FPL calibration (ER10C-FPL black triangles with solid lines). CPL indicates coupler sound pressure level; FPL, forward pressure level; SPL, sound pressure level.

the interfering reverse-traveling wave is eliminated from the pressure response. The effect of standing-wave variability on hearing threshold variability in individual ears can be examined in Figure 4, where it is notable that participants 10, 12, 14, and 20 showed more variability for SPL calibrations compared with FPL calibrations around 6 to 8 kHz. For example, participant 14 had variation of 15 dB at 4 kHz with repeated fits (Fig. 1) which translated to a higher MISD for SPL calibration compared with FPL and CPL calibration (Fig. 4).

Test–retest reliability was only ~0.2 dB lower for FPL compared with CPL calibration for the ER10C (Fig. 3B and Table 3). This may in part be due to the careful and consistent probe fits by the experienced tester. We would expect much more variation in real-world settings, especially when repeated tests are done by different testers and the depth of probe fit and angle could vary over a much wider range. The effect of CPL calibration is most notable for participant 7 where probe depth varied considerably across measurements (this is seen in the variable pressure responses in Fig. 1 and in the equivalent volume SD in Table 2). Note that in this case, the standing-wave null is above the measurement range of 8 kHz (presumably because the ear canal is short) so the change in the null frequency with probe depth is not visible. Both the in-the-ear SPL and FPL calibrations could compensate for the variation in stimulus level due to varying probe depth (outside of the standing-wave null region), whereas the CPL calibration could not, and the hearing threshold SDs at 4 and 6 kHz (Fig. 4) were elevated for CPL in comparison.

There are further gains to be had by using FPL calibration over CPL calibration, because validity for individual ears may improve for those whose residual ear-canal volumes are larger or smaller than the average ear assumed in a coupler calibration. For ears that diverge from an average ear, reliability will not be necessarily affected if probe depth remains constant, but the measured threshold may differ from the true threshold. Furthermore, perforations can cause all earphones to overestimate the amount of loss, but insert earphones produce a larger amount of error than supra-aural earphones (Voss et al. 2000). This results from the larger proportional increase in the area enclosed by the earphone. We anticipate that FPL calibration would provide more accurate thresholds over CPL calibration for people with perforated eardrums.

For those ear canals where repeated probe insertions tend to the same insertion depth, there will be little difference in validity or reliability for CPL calibration compared with FPL calibration. Participants 5, 6, 8, and 17 showed little variation in pressure responses in Figure 1, and as expected, this translated into stable and similar hearing threshold SDs for all OtoStat calibrations in Figure 4.

Hearing Threshold Reliability for the ER10C Insert Earphone Compared With the TDH-39P Headphone

The motivation for the experiment was to compare a new system with one that is standardly used, of which calibration is only one factor. (The version of the OtoStat hardware in use did not have the capability of using supra-aural headphones for stimulus delivery.) To help bridge the gap when considering calibration specifically, the CPL analysis for the ER10C allowed direct comparison to SPL and FPL (above). In comparison to the Benson audiometer with the TDH-39P headphone, the OtoStat+ER10C system showed lower variability across all frequencies and all calibration methods, except 3 kHz where it is about the same for all systems (Fig. 3B). Here the comparison is based on a separate measurement on the same ears made on the Benson audiometer with TDH-39P headphones, interleaved with the OtoStat measurements, so there is the potential for factors unrelated to calibration to produce differences. The lower variability for the OtoStat (regardless of calibration method), compared with the Benson audiometer, may be due to (a) increased noise attenuation provided by the ER10C foam ear tips (affecting lower frequencies), (b) more variability from headphone versus earphone placement (affecting higher frequencies), (c) methodological differences in the automated algorithms, although every attempt was made to ensure parameters were equivalent, and (d) the known poor performance of the TDH-39P headphones at 6 kHz (Rudmose 1964).

Comparison to Other Studies

Historically, there have been a number of studies assessing test–retest reliability for clinical and industrial PTA (i.e., using a variation of the modified Hughson-Westlake method) using headphones (including Hickling 1966; Marshall & Gossman 1982; Stelmachowicz et al. 1988; Marshall & Hanna 1989), summarized in Table 3, which show a tendency for higher variability at higher frequencies.

Marshall et al.'s (1996) study is relevant to the present study due to our implementation of their procedure for the automated PTA measurements. The comparisons of reliability are

Copyright © 2018 Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.

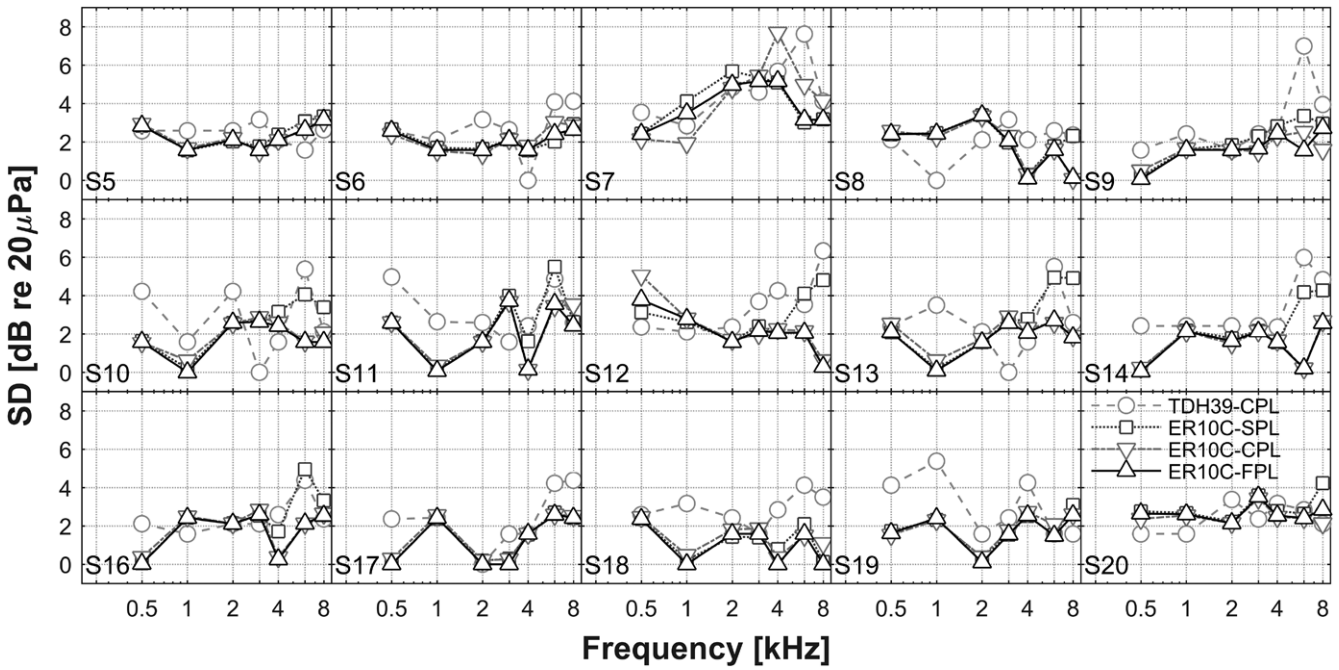

Fig. 4. Hearing threshold variability represented by the intrasubject SDs for each participant, audiometric frequency, system, and calibration type. Benson audiometer with THD-39P headphones and a coupler calibration (TDH39-CPL light gray circles with dashed lines), OtoStat with ER10C earphones with SPL calibration (ER10C-SPL dark gray squares with dotted lines), OtoStat with a coupler calibration (ER10C-CPL gray triangles with dot-dashed lines), and OtoStat with FPL calibration (ER10C-FPL black triangles with solid lines). CPL indicates coupler sound pressure level; FPL, forward pressure level; SPL, sound pressure level.

therefore of interest (as well as providing a benchmark and methodology), albeit just for 1 kHz. Across 7 studies, Marshall et al. reported an unweighted average intrasubject SD of 2.8 dB at 1 kHz for clinical audiometric protocols and achieved 3.1 dB at 1 kHz in their study.

Three longitudinal studies in military settings have produced some examples of the reliability achievable in real-world HCPs (Lapsley Miller et al. 2004, 2006; Marshall et al. 2009). Of interest is the (previously unpublished) comparison of manual versus automated audiometry in the same participants (Marshall et al. 2009), showing poorer performance at 6 kHz with manual audiometry (possibly because participants were fitting their own headphones). Regardless, all showed signs of poor reliability above 4 kHz. Of interest in these studies was the comparison with OAEs (both absolute levels and changes over time). It appeared that OAEs were more sensitive at picking up the early stages of NIHL but this may be because the poor high-frequency PTA reliability masked shifts of 15 to 20 dB. Thus, the question as to whether changes in OAEs predate changes in hearing for physiological rather than methodological reasons is still open. For a larger review of earlier studies in the validity and reliability of industrial audiometry, see Dobie (1983).

There are a handful of studies looking at reliability of coupler-calibrated ER-3A (Etymōtic Research) insert earphones compared with headphones (TDH-49 or TDH-50), but there are either methodological differences or the statistical results are not directly comparable with the present study. In general, little-to-no difference in test–retest reliability was found when comparing insert earphones to headphones in these studies (Larson et al. 1988; Lindgren 1990; Stuart et al. 1991).

Comparing insert earphones with different calibration methods, recently, Withnell et al. (2014) compared CPL, SPL, and FPL calibration for PTA (using a different Mimosa Acoustics system). Here, we have recalculated their test–retest reliability in a form comparable with the current results (Table 3). Tests for the first and last (fifth) days were chosen, but comparisons among other days were similar. Their results were consistent with the current results, although they found less of a difference between CPL and FPL calibration.

Comparisons across studies for reliability are problematic because it is difficult to ensure that every variable is the same or at least comparable. Such factors include the time elapsed between test and retest, the naivety and motivation of the participants, differences in methodology, and the experience of the tester. A strength of the present study was the intrasubject comparison across two systems and three calibration methods, albeit with a relatively small N.

Implications for STS Criteria Used in HCPs

Although 15 ears are too low to provide reliable clinical criteria (McMillan & Hanson 2014), because the comparisons are intrasubject, we can consider how a decrease in intrasubject variability might translate to a clinically-meaningful improvement in an STS criterion. An STS criterion is used to determine whether an individual has experienced a change in hearing between one test and another (e.g., a baseline and a follow-up). STS criteria may be derived statistically from the SE of measurement (Ghiselli 1964; Lapsley Miller et al. 2006; McMillan 2014) or MISD which is statistically equivalent (McMillan et al. 2013). The process involves creating a confidence interval around zero-change that signifies the amount of variability to be expected when there has been no real change in hearing. Any change that is larger than the confidence interval is then considered to be statistically significant and unlikely to be due to chance. Here we applied this method, as described in Lapsley Miller et al. (2006), by using a

Copyright © 2018 Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.

TABLE 3. Comparisons of current test–retest reliability statistics to some previous studies, listed by number of ears, duration between test and retest, manual or automated PTA (all using modified Hughson-Westlake procedures with 5-dB step-size), the specific earphone transducer, the calibration method, the reliability statistic reported (SE_{MEAS} or MISD), the individual test frequencies and frequency averages (kHz), and additional notes

							Frequency (kHz)									
Study	Ears	Duration	PTA	Earphone	Calibration	Reliability	0.5	1	2	3	4	6	8	Ave. 2, 3, 4	Ave. 4, 6	Note
Hickling (1966)	30	Within session	Manual	TDH-39	Coupler	MISD		2.3	2.2			3.5	3.4			
	?	1 day	Manual	TDH-39	Coupler	MISD		3.3	3.3			5.2	5.0			Includes data from an earlier study
Marshall and Gossman (1982)	20	Within session	Manual	TDH-49	Coupler	SE_{MEAS}	2.9	3.2	3.9	3.2	3.4	6.3	5.5			Numbers from Marshall and Hanna (1989)
Stelmachowicz et al. (1988)	40	1 month	Unstated	TDH-39	Coupler	SE_{MEAS}	2.1	2.1	2.4		2.7		3.2			
Summary by Marshall et al. (1996)		Various	Manual	Various	Coupler	SE_{MEAS}		2.8								Summarizing 7 previous studies including those above
Marshall et al. (1996)	18	Within session	Auto	THD-50	Coupler	MISD		3.1								
Lapsley Miller et al. (2004)	106	1 year	Manual	TDH-50	Coupler	SE_{MEAS}	3.6	3.0	2.7	2.9	3.3	4.5	4.8	2.0		
Lapsley Miller et al. (2006)	56	20 min–2 days	Manual/auto	TDH-39/ TDH-50P	Coupler	SE_{MEAS}		2.8	2.1	3.4	3.8	5.5		2.2		
Marshall et al. (2009)	86	1–2 days	Auto	TDH-49	Coupler	SE_{MEAS}	3.7	3.2	3.4	3.7	4.3	5.8				Unpublished data from this study
Marshall et al. (2009)	64	1–2 days	Manual	TDH-49	Coupler	SE_{MEAS}	4.0	3.6	3.4	3.6	3.9	4.5		2.7		
Withnell et al. (2014)	40	~5 days	Auto	ER10C	Coupler	SE_{MEAS}	2.3	2.7	2.1	3.0	2.9	3.2				Recalculated in terms of SE_{MEAS} using tests 1 and 5
	40	~5 days	Auto	ER10C	ITE SPL	SE_{MEAS}	2.3	3.4	2.1	3.3	3.8	5.3				Recalculated in terms of SE_{MEAS} using tests 1 and 5
	40	~5 days	Auto	ER10C	ITE FPL	SE_{MEAS}	2.3	3.1	2.0	3.1	3.0	2.9				Recalculated in terms of SE_{MEAS} using test 1 and 5
Present study	15	Within session	Auto	TDH-39	Coupler	MISD	2.8	2.4	2.5	2.2	2.6	4.3	3.3	1.7	2.6	
	15	Within session	Auto	ER10C	Coupler	MISD	2.0	1.7	1.9	2.4	2.0	2.4	2.2	1.6	1.8	Probe calibrated using BK 4157 coupler
	15	Within session	Auto	ER10C	ITE SPL	MISD	1.8	1.8	2.0	2.5	2.2	3.3	3.2	1.6	2.0	
	15	Within session	Auto	ER10C	ITE FPL	MISD	1.8	1.7	1.9	2.3	1.7	2.1	2.0	1.4	1.6	

MISD, mean intrasubject SD; PTA, pure-tone audiometry; SE_{MEAS}, SE of measurement.

Copyright © 2018 Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.

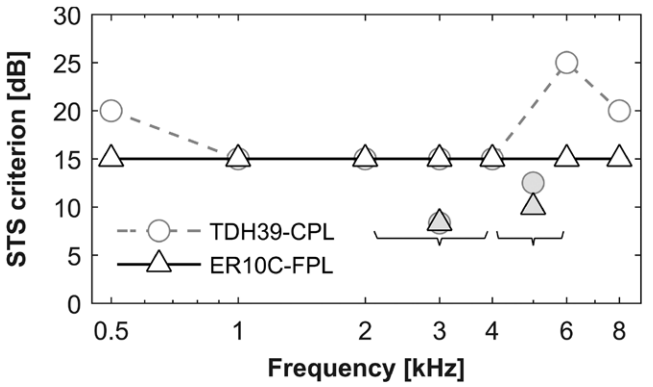

Fig. 5. Comparisons of statistically-derived STS criteria based on the mean intrasubject SDs. Benson audiometer with TDH-39P earphones (TDH39-CPL light gray circles with dashed lines) and OtoStat with ER10C earphones and FPL calibration (ER10C-FPL black triangles with solid lines). The individual shaded symbols above the braces indicate the STS criteria based on averaged frequencies at 2, 3, and 4 kHz and 4 to 6 kHz. CPL indicates coupler sound pressure level; FPL, forward pressure level; STS, significant threshold shifts.

99% confidence interval and assuming a normal variation (i.e., z = 2.56), thus multiplying the MISD by $\sqrt{2} \times 2.56$. The $\sqrt{2}$ is needed because the difference between two tests is being considered (Beattie 2003; Lapsley Miller et al. 2006).‡ The resulting STS criterion is applied to an individual test result at a specific frequency or frequency-average by considering whether the difference between the baseline and follow-up test is greater than or equal to (i.e., inclusive) to the criterion.

The resulting STS criteria for the present study are shown in Figure 5 for the Benson audiometer with CPL calibration and for the OtoStat with FPL calibration (i.e., the best and worst cases). First, these results show why 6 kHz, although measured in most HCPs, is not typically used for detecting/defining STS cases. It is simply too unreliable with a change in hearing thresholds of at least 25 dB needed before being able to reliably differentiate the shift from chance. If a more typical 15 dB STS criterion was used, the HCP would be overwhelmed by false positives. However, a 15-dB shift in hearing thresholds could be reliably detected at 6 and 8 kHz if insert earphones with FPL calibration were used. With NIHL showing up in higher frequencies first, it has long been a limitation in HCPs that testing above 4 kHz has not had sufficient reliability to enable the detection of the early stages of hearing loss.

The small difference in the MISD for OtoStat thresholds based on FPL and CPL calibrations did not translate into a difference in the STS criteria (not plotted), in part because the 5-dB step-size used in audiometric testing is much larger than the small differences in the underlying SDs.

These STS statistics are based on a laboratory study done under good conditions and, as Dobie (1983) discusses, real-world industrial audiometry has much lower reliability. The choice of STS criteria, first and foremost, should be based on the actual statistically-determined variability for that site and population. From there, the costs involved with false positives need to be considered and traded-off against detecting true changes in hearing, especially when considering multiple comparisons. For these reasons, some industrial audiometry regulations have moved from detecting STS at multiple individual frequencies to detecting a shift of the average hearing thresholds at 2, 3, and 4 kHz (typically with a shift of 10 dB needed) (Dobie 1983). This reduces the ability to detect some hearing losses, but the false-positive rate from multiple comparisons is reduced (Dobie 1983).

The current results suggest that an alternative to the average at 2, 3, and 4 kHz could be to look for average shifts at 4 and 6 kHz if FPL calibration was used to improve reliability. This potentially would increase test sensitivity to detecting NIHL without increasing the false-positive rate.

Using the average of 4 to 8 kHz might not be as sensitive to NIHL because NIHL often displays as a "noise-notch" at 3 to 6 kHz, with relatively normal hearing still apparent at 8 kHz. Noise notches can be over-estimated when using earphones with poor calibration at 6 kHz, such as the TDH-39P when calibrated with an IC 303 coupler (Schlauch & Carney 2011, 2012). The noise-notch phenomenon has also been questioned as being incidental to NIHL (McBride & Williams 2001), perhaps being an artifact of a poor standardization of dB HL at 6 kHz. It is possible the use of in-the-ear FPL calibration, which provides improved validity and reliability around 6 and 8 kHz, could shed light on this issue.

Other Implications for HCPs

These results and others (Schlauch & Carney 2011, 2012) demonstrate that the TDH-39P earphones that are ubiquitous in HCPs are not ideal. Although there are some practical benefits such as perceived easier fitting of headphones (which does not necessarily translate into actual consistency of fit across measurements), this is outweighed by poorer reliability. In general, insert earphones are more reliable than supra-aural earphones at higher audiometric frequencies; however, Occupational Safety and Health Administration still recommends using the latter. Over the next few years, we anticipate that a number of HCPs will enhance their capabilities by including OAE and WAI testing, as well as calibrating using FPL. These tests require the use of an insert earphone and the obvious step is to use insert earphones for PTA too, ideally using a system that can measure all three automatically in one test session, with one probe fit and calibration, and the same equipment.

With FPL-calibrated insert earphones, STS criteria and frequencies used in HCP standards can be re-examined to improve sensitivity to small changes in hearing and reducing false positives. This would involve some medium-scale reliability field studies in both noise-exposed and quiet control participants, where more time elapses between tests, testers are less experienced, and testers varied at retest are considered.

Audiometric Zero for FPL-Calibrated PTA

For clinical use, audiometric zero (i.e., the RETSPL) for PTA with FPL calibration must be established, ideally with a multisite study with a group of people with otologically normal ears. The addition to these studies of people with ears with large (e.g., due to perforations) and small (e.g., due to Down

‡ There is a complication when applying this method to audiometric data because the clinical audiometric resolution (step-size) of 5 dB is typically larger than the MISD. To use this number clinically, it must be rounded up to the next audiometric step-size. Typically, STS criteria are inclusive, so an additional audiometric step-size is added. The audiometric step-size is 5 dB for individual frequencies, 2.5 dB for two-frequency averages, and 1.67 dB for three-frequency averages. The reader may calculate STS criteria using different assumptions using the SDs presented in Table 3.

Copyright © 2018 Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.

Syndrome) ear-canal volumes would establish if FPL also reduced variability and improved accuracy in thresholds measured in these ears. Establishing audiometric zero is necessary before FPL calibration of PTA stimuli can become widely available on clinical systems.

Implications for Other Applications

These results showing a benefit to using FPL calibration apply not only to HCPs but also to other applications where accurate and repeatable measurements are needed at higher frequencies. This includes ototoxicity monitoring and hearing-aid fitting. Serial monitoring for ototoxic effects on the inner ear depends on reliable detection of small changes in hearing often at high to very high frequencies (Fausti et al. 1998, 1999). Another important application where FPL calibration may show an advantage is in documenting hearing threshold differences between ears, which is a metric used to assess whether a person may have an acoustic tumor. These tumors may affect high-frequency thresholds first, and 8 kHz thresholds have been found to be important for this assessment (Mangham 1991; Schlauch et al. 1995).

For hearing-aid fitting, where the gain settings are determined by in situ measurements in the ear canal, the issue is determining the correct gain for frequencies at the standing-wave nulls (McCreery et al. 2009). If SPL at the probe microphone is used, there is risk of overamplification at standing-wave frequencies and a risk of damaging remaining hearing. The effect of standing-wave nulls on accurate gain settings for higher frequencies is becoming more of an issue as manufacturers increase the upper frequency of hearing aids to improve speech understanding. One alternative is to use measurements near the tympanic membrane; however, these are clinically more difficult. FPL measurements provide an easier and safer method.

Summary

The current results are limited due to the small sample size, the laboratory test conditions, and the experienced tester. Larger-scale field studies are needed to demonstrate whether the FPL advantage holds (or even improves) in real-world settings, and whether the detection of smaller higher-frequency STSs allows HCPs to detect NIHL earlier. Factors that could show larger differences among the calibration and earphone options are whether people being tested fit their own headphones and insert earphones (with the likelihood of the headset being poorly placed and the earphone being only shallowly inserted) versus being fitted and adjusted by a technician. FPL calibration might show more of an advantage in the self-fitting case because it is more invariant to probe depth, and automated feedback can be given to individuals who have a shallow self-insertion with an acoustical leak.

These results indicate that coupler calibration for insert earphones is quite reliable too—more so than in-the-ear SPL calibration. Existing HCPs that are interested in only doing PTA measurements could use coupler-calibrated insert earphones instead of supra-aural headphones to improve reliability (although they will not gain improvement in test validity). FPL calibration will be most advantageous for HCPs that are interested in adding WAI and OAE testing, because the FPL information is gained from the WAI test with no additional testing needed and both validity and reliability will be improved. The WAI test provides the calibration information needed for both the OAE and PTA tests, will show if the probe is fitted properly (Groon et al. 2015), and also provides an evaluation of middle ear status (e.g., as summarized in Allen et al. 2016).

ACKNOWLEDGMENTS

The authors thank Pat Jeng for advice on experimental design; Lynne Marshall for advice on the experimental protocols and for reviewing an earlier version of the manuscript; Bill Ahroon (U.S. Army Aeromedical Research Laboratory) for the loan of an Army-issue audiometer; Laurie Heller for statistical advice; Rob Withnell for sharing data; and to Kurt Yankaskas, Program Officer for Noise Induced Hearing Loss at the Office of Naval Research, for his support.

This article was supported by small business innovation research awards to Mimosa Acoustics from the Office of the Secretary of Defense under the contract number N00014-15-C-0046 and the Defense Health Program under the contract number W81XWH-16-C-0185. Portions of this article were presented at the 43rd Annual AAS Scientific and Technology Conference of the American Auditory Society, Scottsdale, AZ. The content of this report is solely the responsibility of the authors and does not necessarily represent the official views of the Department of Defense or the US Government.

J.L.M. and C.M.R. designed the experiment. C.M.R. and Z.D.P. performed the experiment at the Research Laboratory of Electronics at Massachusetts Institute of Technology. J.L.M. analyzed the data. J.L.M. and S.R.R. wrote the article.

The authors have no conflicts of interest to disclose.

Address for correspondence: Judi A. Lapsley Miller, Mimosa Acoustics, 335 Fremont Street, Champaign, IL 61820, USA. E-mail: judi@mimosaacoustics.com

Received January 5, 2017; accepted December 21, 2017.

REFERENCES

Abur, D., Horton, N. J., Voss, S. E. (2014). Intrasubject variability in power reflectance. *J Am Acad Audiol*, *25*, 441–448.

Allen, J. B. (1986). Measurement of eardrum acoustic impedance. In J. B. Allen, J. L. Hall, A. E. Hubbard, et al. (Eds.), *Peripheral Auditory Mechanisms* (pp. 44–51). New York, NY: Springer-Verlag.

Allen, J. B., Robinson, S. R., Lapsley Miller, J. A., et al. (2016). Middle-ear reflectance: Concepts and clinical applications. In A. T. Cacace, E. de Kleine, A. G. Holt, *et al*. (Eds.), *Scientific Foundations of Audiology: Perspectives from Physics, Biology, Modeling, and Medicine* (pp. 1–40). San Diego, CA: Plural.

ANSI. (2003a). *Maximum Permissible Ambient Noise Levels for Audiometric Test Rooms (ANSI S3.1-1999 (R2003))*. New York, NY: American National Standards Institute.

ANSI. (2003b). *Method for Coupler Calibration of Earphones (ANSI S3.7-1995 (R2003))*. New York, NY: American National Standards Institute.

ANSI. (2010). *Specifications for Audiometers (ANSI S3.6-2010)*. New York, NY: American National Standards Institute.

ANSI/ASA. (2009). *Methods for Manual Pure-Tone Threshold Audiometry (ANSI S3.21–2004 R2009)*. New York, NY: American National Standards Institute.

Beattie, R. C. (2003). Distortion product otoacoustic emissions: Comparison of sequential versus simultaneous presentation of primary-tone pairs. *J Am Acad Audiol*, *14*, 471–484.

Dobie, R. A. (1983). Reliability and validity of industrial audiometry: Implications for hearing conservation program design. *Laryngoscope*, *93*, 906–927.

Dobie, R. A. (2005). Audiometric threshold shift definitions: Simulations and suggestions. *Ear Hear*, *26*, 62–77.

Fausti, S. A., Henry, J. A., Hayden, D., et al. (1998). Intrasubject reliability of high-frequency (9-14 kHz) thresholds: Tested separately vs. following conventional-frequency testing. *J Am Acad Audiol*, *9*, 147–152.

Fausti, S. A., Henry, J. A., Helt, W. J., et al. (1999). An individualized, sensitive frequency range for early detection of ototoxicity. *Ear Hear*, *20*, 497–505.

Feeney, M. P., Hunter, L. L., Kei, J., et al. (2013). Consensus statement: Eriksholm workshop on wideband absorbance measures of the middle ear. *Ear Hear*, *34*(Suppl 1), 78S–79S.

Ghiselli, E. E. (1964). *Theory of Psychological Measurement*. New York, NY: McGraw-Hill.

Copyright © 2018 Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.

Groon, K. A., Rasetshwane, D. M., Kopun, J. G., et al. (2015). Air-leak effects on ear-canal acoustic absorbance. *Ear Hear*, *36*, 155–163.
Hickling, S. (1966). Studies on the reliability of auditory threshold values. *J Auditory Res*, *6*, 39–46.
Humes, L. E., Joellenbeck, L. M., Durch, J. S. (Eds.). (2005). *Noise and Military Service: Implications for Hearing Loss and Tinnitus*. Washington, DC: National Academies.
Killion, M. C., & Villchur, E. (1989). Comments on "Earphones in Audiometry" [Zwislocki *et al*., J. Acoust. Soc. Am. 83, 1688-1689 (1988)]. *J Acoust Soc Am*, *85*, 1775–1779.
Konrad-Martin, D., Poling, G. L., Dreisbach, L. E., et al. (2016). Serial monitoring of otoacoustic emissions in clinical trials. *Otol Neurotol*, *37*, e286–e294.
Lapsley Miller, J. A., & Marshall, L. (2007). Otoacoustic emissions as a preclinical measure of NIHL and susceptibility to NIHL. In M. S. Robinette, T. J. Glattke (Eds.), *Otoacoustic Emissions: Clinical Applications* (pp. 321–341). New York, NY: Thieme.
Lapsley Miller, J. A., Marshall, L., Heller, L. M. (2004). A longitudinal study of changes in evoked otoacoustic emissions and pure-tone thresholds as measured in a hearing conservation program. *Int J Audiol*, *43*, 307–322.
Lapsley Miller, J. A., Marshall, L., Heller, L. M., et al. (2006). Low-level otoacoustic emissions may predict susceptibility to noise-induced hearing loss. *J Acoust Soc Am*, *120*, 280–296.
Larson, V. D., Cooper, W. A., Talbott, R. E., et al. (1988). Reference threshold sound-pressure levels for the TDH-50 and ER-3A earphones. *J Acoust Soc Am*, *84*, 46–51.
Lewis, J. D., McCreery, R. W., Neely, S. T., et al. (2009). Comparison of in-situ calibration methods for quantifying input to the middle ear. *J Acoust Soc Am*, *126*, 3114–3124.
Lindgren, F. (1990). A comparison of the variability in thresholds measured with insert and conventional supra-aural earphones. *Scand Audiol*, *19*, 19–23.
Mangham, C. A. (1991). Hearing threshold difference between ears and risk of acoustic tumor. *Otolaryngol Head Neck Surg*, *105*, 814–817.
Marshall, L., & Gossman, M. A. (1982). Management of ear-canal collapse. *Arch Otolaryngol*, *108*, 357–361.
Marshall, L., & Hanna, T. E. (1989). Evaluation of stopping rules for audiological ascending test procedures using computer simulations. *J Speech Hear Res*, *32*, 265–273.
Marshall, L., Hanna, T. E., Wilson, R. H. (1996). Effect of step size on clinical and adaptive 2IFC procedures in quiet and in a noise background. *J Speech Hear Res*, *39*, 687–696.
Marshall, L., Lapsley Miller, J. A., Heller, L. M., et al. (2009). Detecting incipient inner-ear damage from impulse noise with otoacoustic emissions. *J Acoust Soc Am*, *125*, 995–1013.
McBride, D. I., & Williams, S. (2001). Audiometric notch as a sign of noise induced hearing loss. *Occup Environ Med*, *58*, 46–51.
McCreery, R. W., Pittman, A., Lewis, J., et al. (2009). Use of forward pressure level to minimize the influence of acoustic standing waves during probe-microphone hearing-aid verification. *J Acoust Soc Am*, *126*, 15–24.
McMillan, G. P. (2014). On reliability. *Ear Hear*, *35*, 589–590.
McMillan, G. P., & Hanson, T. E. (2014). Sample size requirements for establishing clinical test-retest standards. *Ear Hear*, *35*, 283–286.
McMillan, G. P., Reavis, K. M., Konrad-Martin, D., et al. (2013). The statistical basis for serial monitoring in audiology. *Ear Hear*, *34*, 610–618.
Puria, S., & Allen, J. B. (1998). Measurements and model of the cat middle ear: Evidence of tympanic membrane acoustic delay. *J Acoust Soc Am*, *104*, 3463–3481.
Royster, J. D., & Royster, L. H. (1986). Audiometric data base analysis. In E. H. Berger, W. D. Ward, J. C. Morrill, et al. (Eds.), *Noise and Hearing Conservation Manual*. Akron, OH: American Industrial Hygiene Association.
Rudmose, W. (1964). Concerning the problem of calibrating TDH-39 earphones at 6 kHz with a 9A coupler. *J Acoust Soc Am*, *36*, 1049–1049.
Scheperle, R. A., Goodman, S. S., Neely, S. T. (2011). Further assessment of forward pressure level for in situ calibration. *J Acoust Soc Am*, *130*, 3882–3892.
Scheperle, R. A., Neely, S. T., Kopun, J. G., et al. (2008). Influence of in situ, sound-level calibration on distortion-product otoacoustic emission variability. *J Acoust Soc Am*, *124*, 288–300.
Schlauch, R. S., & Carney, E. (2007). A multinomial model for identifying significant pure-tone threshold shifts. *J Speech Lang Hear Res*, *50*, 1391–1403.
Schlauch, R. S., & Carney, E. (2011). Are false-positive rates leading to an overestimation of noise-induced hearing loss? *J Speech Lang Hear Res*, *54*, 679–692.
Schlauch, R. S., & Carney, E. (2012). The challenge of detecting minimal hearing loss in audiometric surveys. *Am J Audiol*, *21*, 106–119.
Schlauch, R. S., Levine, S., Li, Y., et al. (1995). Evaluating hearing threshold differences between ears as a screen for acoustic neuroma. *J Speech Hear Res*, *38*, 1168–1175.
Siegel, J. H. (1994). Ear-canal standing waves and high-frequency sound calibration using otoacoustic emission probes. *J Acoust Soc Am*, *95*, 2589–2597.
Siegel, J. H. (2002). Calibrating otoacoustic emission probes. In M. S. Robinette, T. J. Glattke (Eds.), *Otoacoustic Emissions: Clinical Applications* (pp. 416–441). New York, NY: Thieme.
Souza, N. N., Dhar, S., Neely, S. T., et al. (2014). Comparison of nine methods to estimate ear-canal stimulus levels. *J Acoust Soc Am*, *136*, 1768–1787.
Stelmachowicz, P. G., Beauchaine, K. A., Kalberer, A., et al. (1988). The reliability of auditory thresholds in the 8- to 20-kHz range using a prototype audiometer. *J Acoust Soc Am*, *83*, 1528–1535.
Stuart, A., Stenstrom, R., Tompkins, C., et al. (1991). Test-retest variability in audiometric threshold with supraaural and insert earphones among children and adults. *Audiology*, *30*, 82–90.
Voss, S. E., Rosowski, J. J., Merchant, S. N., et al. (2000). Middle ear pathology can affect the ear-canal sound pressure generated by audiologic earphones. *Ear Hear*, *21*, 265–274.
Withnell, R. H., Jeng, P. S., Waldvogel, K., et al. (2009). An in situ calibration for hearing thresholds. *J Acoust Soc Am*, *125*, 1605–1611.
Withnell, R. H., Jeng, P. S., Parent, P., et al. (2014). The clinical utility of expressing hearing thresholds in terms of the forward-going sound pressure wave. *Int J Audiol*, *53*, 522–530.
Zebian, M., Hensel, J., Fedtke, T., et al. (2012). Equivalent hearing threshold levels for the Etymotic Research ER-10C otoacoustic emission probe. *Int J Audiol*, *51*, 564–568.

REFERENCE NOTES

Benson Medical Instruments Co. (2007). *User's manual CCA-200 and CCA-200mini Version 6.00*.
Department of Defense. (2010). Department of Defense Instruction 6055.12: DOD Hearing Conservation Program (HCP). *USD/AT&L*.
Mimosa Acoustics. (2016). *OtoStat 2.1 with OtoStation user manual v2.1*.

Copyright © 2018 Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.