Effect of Calibration Method on Distortion-Product Otoacoustic Emission Measurements at and Around 4 kHz

Michal L. Reuven,[1,2] Stephen T. Neely,[2] Judy G. Kopun,[2] Daniel M. Rasetshwane,[2] Jont B. Allen,[3] Hongyang Tan,[2] and Michael P. Gorga[2]

Objectives: Distortion-product otoacoustic emissions (DPOAEs) collected after sound pressure level (SPL) calibration are susceptible to standing waves that affect measurements at the plane of the probe microphone due to overlap of incident and reflected waves. These standing-wave effects can be as large as 20 dB, and may affect frequencies both above and below 4 kHz. It has been shown that forward pressure level (FPL) calibration minimizes standing-wave effects by isolating the forward-propagating component of the stimulus. Yet, previous work has failed to demonstrate more than a small difference in test performance and behavioral-threshold prediction with DPOAEs after SPL and FPL calibration. One potential limitation in prior studies is that measurements were restricted to octave and interoctave frequencies; as a consequence, data were not necessarily collected at the standing-wave null frequency. In the present study, DPOAE responses were measured with f_2 set to each participant's standing-wave frequency in an effort to increase the possibility that differences in test performance and threshold prediction would be observed for SPL and FPL calibration methods.

Design: Data were collected from 42 normal-hearing participants and 93 participants with hearing loss. DPOAEs were measured with f_2 set to 4 kHz and at each participant's notch frequency after SPL and FPL calibration. DPOAE input/output functions were obtained from −10 to 80 dB in 5 dB steps for each calibration/stimulus condition. Test performance was evaluated using clinical decision theory. Both area under receiver operating characteristic curves for all stimulus levels and cumulative distributions when L_2 = 50 dB (a level at which the best performance was observed regardless of calibration method) were used to evaluate the accuracy with which auditory status was determined. A bootstrap procedure was used to evaluate the significance of the differences in test performance between SPL and FPL calibrations. DPOAE predictions of behavioral threshold were evaluated by correlating actual behavioral thresholds and predicted thresholds using a multiple linear regression model.

Results: First, larger DPOAE levels were measured after SPL calibration than after FPL calibration, which demonstrated the expected impact of standing waves. Second, for both FPL and SPL calibration, test performance was best for moderate stimulus levels. Third, differences in test performance between calibration methods were evident at low- and high-stimulus levels. Fourth, there were small but statistically significant improvements in test performance after FPL calibration for clinically relevant conditions. Fifth, calibration method had no effect on threshold prediction.

Conclusions: Standing waves after SPL calibration have an impact on DPOAE levels. Although the effect of calibration method on test performance was small, test performance was better after FPL calibration than after SPL calibration. There was no effect of calibration method on predictions of behavioral threshold.

(Ear and Hearing 2013;34;779–788)

[1]Department of Hearing and Speech Sciences, University of Maryland, College Park, Maryland, USA; [2]Center for Hearing Research, Boys Town National Research Hospital, Omaha, Nebraska, USA; and [3]Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA.

INTRODUCTION

The purpose of this study was to determine the influence of calibration procedure on distortion-product otoacoustic emission (DPOAE) measurements. Specifically, DPOAEs were measured after standard sound pressure level (SPL) calibrations and after calibration using forward pressure level (FPL) measurements. Whereas SPL calibrations are affected by standing waves, FPL measurements are not. The frequency at which standing waves occur varies across participants, and does not always occur at a standard test frequency (such as 4 kHz). To evaluate the influence of standing-wave interactions on DPOAE measurements, data were collected at the notch frequency in SPL calibrations and at the same frequency after FPL calibrations. These measurements directly assess the extent to which standing waves influence SPL calibrations and, in turn, DPOAE measurements.

DPOAEs result from the interaction of two simultaneously presented tones that differ in frequency (f_2 and f_1, with f_2 typically about 20% higher in frequency than f_1). They are generated within the cochlea and are a result of nonlinear processes. DPOAEs are generated by outer hair cells (OHCs) as a byproduct of their electromotility (Brownell 1990). Because OHC damage results in hearing loss and DPOAEs are dependent on the integrity of the OHCs, DPOAE measurements have been used clinically to objectively evaluate auditory function (e.g., Lonsbury-Martin et al. 1993). DPOAE responses may be used to dichotomously differentiate between participants with normal hearing (NH) and those with hearing loss, which we refer to as test performance (e.g., Gorga et al. 1993, 1997; Kim et al. 1996). Because both DPOAE level and behavioral thresholds are graded responses related to OHC integrity, DPOAEs have also been used to predict threshold (Boege & Janssen 2002; Gorga et al. 2003; Rogers et al. 2010). The accuracy with which DPOAEs identify hearing status (NH versus hearing-impaired [HI]) is greatest when the responses are measured at moderate stimulus levels and at mid to high frequencies (Whitehead et al. 1995; Stover et al. 1996; Gorga et al. 1997; Johnson et al. 2010). However, even under optimal stimulus conditions, DPOAE test performance is imperfect (Kirby et al. 2011).

One source of inaccuracy in DPOAE measurements may be due to standing-wave effects when SPL calibrations are used. To account for participant ear acoustics, stimuli typically are calibrated in the ear canal (in situ) before DPOAE responses

are measured. Stimuli used for calibration are measured with a probe assembly that is located 15 to 20 mm away from the tympanic membrane (TM). During in situ SPL calibration, some of the sound energy is transmitted through the TM, and some is reflected back through the ear canal. The overlap of FPL and reflected pressure may result in pressure cancellation due to the occurrence of standing waves when stimulus frequencies are above 2 kHz because the round-trip phase shift in the ear canal forward pressure can be nearly 180° (Gilman & Dirks 1986; Siegel & Hirohata 1994; Siegel 2007; Withnell et al. 2009). Depending on the relative phase of forward and reflected waves, there may be constructive or destructive summation of pressure at the plane of the probe microphone. As a consequence, the measured level at the plane of the probe may not accurately represent the level at the TM, and is typically less than it is at the TM (Siegel 1994) and, thus, the cochlea. This leads to an underestimation of the effective level of the stimulus entering the ear. The voltage to the sound sources within the probe is increased to compensate for this null in the pressure at the probe microphone, resulting in more pressure at the TM than intended. The magnitude of the difference may be as great as 20 dB (e.g., Richmond et al. 2011). During measurements, this may lead to a larger DPOAE response than would be expected, given the intended stimulus level (Whitehead et al. 1995).

Two alternative measures that eliminate standing-wave effects during calibration are FPL and sound-intensity level (SIL). FPL calibration isolates the forward-propagating portion of the stimulus, thus eliminating interference from the reflected wave. SIL calibrations are also less susceptible to standing-wave effects (compared with SPL calibrations), as demonstrated by Neely and Gorga (1998), who measured behavioral thresholds at two different probe-insertion depths with SIL and SPL calibrations. There were depth-dependent changes in behavioral thresholds (which were assumed to be constant) for SPL calibrations, but not for SIL calibrations, demonstrating the influence of standing waves on SPL calibrations. Scheperle et al. (2008) compared DPOAE levels after SPL, FPL, and SIL calibrations using shallow and deep probe insertions. They found that probe placement resulted in more variable DPOAE levels after SPL calibrations compared with the DPOAE levels measured after FPL and SIL calibrations.

To explore whether these calibration differences affected DPOAE test performance, Burke et al. (2010) measured DPOAE input/output (I/O) functions in participants with NH and in those with hearing loss, after both SPL and FPL calibration. They then evaluated test performance (the ability of DPOAE measurements to detect normal hearing or impaired hearing) for a range of stimulus levels. At the moderate levels for which test performance was best, there were no differences in test performance between the two calibration types except at 8 kHz. Rogers et al. (2010) used the same DPOAE data to predict behavioral thresholds, using both fits to the I/O functions and a criterion DPOAE level to define DPOAE threshold. They found no differences between SPL and FPL calibration in terms of behavioral-threshold predictions except at 8 kHz. In both the test-performance and threshold-prediction studies, the effects at 8 kHz were attributed to system distortion (resulting in a situation in which system distortion was potentially incorrectly identified as a biological response). Kirby et al. (2011) used a multivariate approach to analyze the difference in DPOAE test performance with SPL and FPL calibration under a number of optimal stimulus and recording conditions. Although they observed improved DPOAE test performance with multivariate analyses, they only found a small effect of calibration method.

In the aforementioned previous work, no effort was expended in determining the frequencies at which standing-wave effects occurred; it was simply assumed that they occurred in the vicinity of 4 kHz. Richmond et al. (2011) analyzed the calibration data from the study by Burke et al. (2010) to determine the frequencies at which notches in the SPL calibration occurred as a result of standing-wave effects. As expected, they found evidence for standing-wave effects (i.e., pressure cancellation) in the SPL calibration centered around 4 kHz, with calibration errors as large as 20 dB. However, these calibration errors were dispersed in frequency over a range from 3 to 5 kHz. In the assessments of calibration effects on DPOAE test performance and threshold prediction, DPOAE measurements were restricted to octave and interoctave frequencies. As a result, the test frequencies did not always precisely align with the frequencies at which the standing-wave effects occurred in all participants. The inclusion of data from a large number of participants in their studies (Burke et al. 2010; Rogers et al. 2010; Kirby et al. 2011) makes it possible for the effects of calibration to be minimized in group statistics such as clinical decision theory and correlation analyses.

The purpose of the present study was to determine whether DPOAE test performance and threshold prediction differed between SPL and FPL calibrations when measurements were made with the higher-frequency primary (f_2) set equal to the standing-wave notch frequency. DPOAE I/O functions were measured at the notch frequency after SPL calibrations and then repeated after FPL calibrations. These I/O data were used to evaluate DPOAE test performance as a function of stimulus level and to assess the relationship between DPOAE thresholds and behavioral thresholds. By focusing measurements at the notch frequency, any differences in calibration method that were minimized in prior large-scale studies (in which measurements were restricted to widely spaced frequencies) should be evident. As a result, this study directly tested the effect of calibration method at frequencies where calibration effects occurred.

PARTICIPANTS AND METHODS

Participants

Data were collected from 42 NH participants and 93 participants with hearing loss, whose ages ranged from 11 to 75 years. Audiometric thresholds were measured at octave frequencies from 0.25 to 8 kHz and at the interoctave frequencies of 3 and 6 kHz using conventional audiometry and either insert (ER-3A; Etymōtic Research, Elk Grove Village, IL) or supra-aural (TDH-50P; Telephonics, Farmingdale, NY) earphones. Bone conduction thresholds were obtained at 0.25, 0.5, 1, 2, and 4 kHz using a Radioear B71 bone oscillator. Participants with hearing thresholds ≤ 20 dB HL from 0.25 to 8 kHz were considered to have NH. Those with thresholds > 20 dB HL at 4 kHz and at DPOAE test frequencies closest to their notch frequency between 2 and 6 kHz were classified as HI. All participants were included regardless of the magnitude of their hearing loss; however, special efforts were extended to recruit participants with mild-to-moderate degrees of loss. This emphasis was based on the view that false-negative errors (the

primary error of concern as a consequence of calibration effects) would be more likely to occur in participants with mild-to-moderate losses. All participants had normal middle-ear function based on (1) normal otoscopic inspection, (2) a normal 226-Hz tympanogram (peak-compensated static acoustic admittance between 0.3 and 2.5 mmhos and peak tympanometric pressure between −100 and +50 daPa), and (3) air–bone gaps ≤ 10 dB. DPOAEs were measured in one ear. When both ears met the inclusion criteria in NH participants, the ear with better hearing was tested. If both ears had similar hearing thresholds, the test ear was chosen randomly. When both ears met the inclusion criteria in HI participants, the ear with hearing thresholds in the 25 to 65 dB HL range between 2 and 8 kHz was tested to increase the likelihood of a DPOAE response above the noise floor. If both ears had thresholds in this range, test ear was chosen randomly. Data collection was conducted under an Institutional Review Board–approved protocol and informed consent was obtained from each participant. During data collection, participants sat on a comfortable reclining chair situated in a sound-attenuated booth. Test time, including obtaining consent, audiometric and tympanometric evaluations, and experimental data collection took approximately 60 min per participant.

Instrumentation

An ER-10C probe-microphone system (Etymōtic Research) was used to calibrate and present stimuli, and to record responses. Stimuli were generated with a 24-bit sound card (CardDeluxe; Digital Audio Labs, Chanhassen, MN). Software developed at BTHRH (EMAV v. 3.28; Neely & Liu 2011) was used for probe-source calibrations (i.e., the Thévenin-equivalent calculations) that are required for the FPL calibration, FPL conversion, and DPOAE measurements.

Calibration

The goal of this study was to evaluate the impact of calibration method on test performance and threshold predictions based on DPOAE measurements. In one case, standard in-the-ear SPL calibration was used to set stimulus levels. This represented the condition in which standing-wave interactions may occur. The SPL data were also used as part of the procedures that resulted in an alternative calibration, namely FPL.

SPL is converted to FPL with the following equation:

$$P_{+} = \frac{1}{2} P_l \bullet \left(1 + \frac{Z_0}{Z_l}\right), \quad (1)$$

where P_{+} = the forward pressure, P_l = load (ear-canal) pressure, Z_0 = characteristic impedance of the source (probe microphone), and Z_l = load (ear-canal) impedance. To solve this equation, the Thévenin-equivalent source characteristics (i.e., probe impedance and pressure) were estimated daily using a set of five brass cylindrical cavities with known acoustic impedances as described previously (Scheperle et al. 2008; Burke et al. 2010; Kirby et al. 2011). Because Burke et al. (2010) did not find an effect of temperature during probe calibration on DPOAE test performance, daily probe calibrations were performed at room temperature only. Ear-canal pressure was measured via in situ calibration for each participant. Then, ear-canal pressure (P_l), probe impedance Z_0, and probe pressure (P_s) were used to estimate ear-canal impedance (Z_l) using the equation:

$$Z_l = \frac{Z_s P_l}{P_s - P_l} \quad (2)$$

After determining the source impedance, ear-canal pressure, and ear-canal impedance, Eq. (1) was used to solve for FPL.

Notch Frequency Selection

The notch in the SPL-calibrated signal was defined as the frequency between 2 and 6 kHz at which the largest decibel difference between SPL- and FPL-calibrated signals occurred (experimental conditions above 6 kHz were not included to avoid problems due to system distortion). At the start of the study, the frequency at which the minimum intensity in the SPL-calibrated spectrum occurred, determined by visual inspection, was chosen as the notch frequency. Approximately half of the data were collected using this procedure. A second, more accurate, method was later implemented in which both the SPL- and FPL-calibrated spectra were analyzed to determine the frequency at which the largest dB difference between these two spectra occurred. The frequency location of this maximum difference was defined as the notch. The second half of data collection used this more objective (and presumably more accurate) approach. We wanted to include the data collected with the first approach, but felt that a criterion was needed to assure that it accurately represented the notch frequency. To accomplish this, the SPL/FPL calibration spectra from the initial method were compared after the fact. Data were excluded from further analyses (test performance and threshold predictions) if the notch frequency selected by the first method was not within 10% of the notch frequency selected by the second method. Figure 1 provides an example of a case for the calibration comparison at the extreme of this inclusion criterion. The FPL-SPL pressure ratio (i.e., the difference between FPL and SPL in dB) is plotted as a function of frequency. As described in the study by Richmond et al. (2011), an ideal case with 100% reflection and negligible phase shift would demonstrate a pressure ratio of −6 dB for a signal measured with FPL calibration compared with a signal measured with SPL calibration, as FPL calibration only takes into consideration the forward-propagating portion of the signal, and does not include reflected pressure. Any difference greater than −6 dB in this ratio occurs because of "standing-wave" effects in SPL calibration, and a local maximum in this difference is associated with an SPL notch. Figure 1 provides an example from one participant in whom the maximum FPL–SPL difference was about 8 dB. The notch frequencies (the frequency where the peak FPL–SPL difference occurs) chosen by both the visual-detection and objective methods are indicated by + signs, and the actual frequencies are provided in the figure legend. Even though these frequencies differed by 328 Hz, the calibration errors for the two methods differ by less than 1 dB (≈0.6 dB). Thus, the magnitude of the notch after the first procedure agreed with the magnitude selected by the more objective procedure. This approach was followed for all data collected before the implementation of the objective procedure. The example in Fig. 1 provides support for the inclusion of the data from the first procedure to determine notch frequency as long as it differed by no more than 10% from the frequency determined from a comparison of FPL and SPL calibrations.

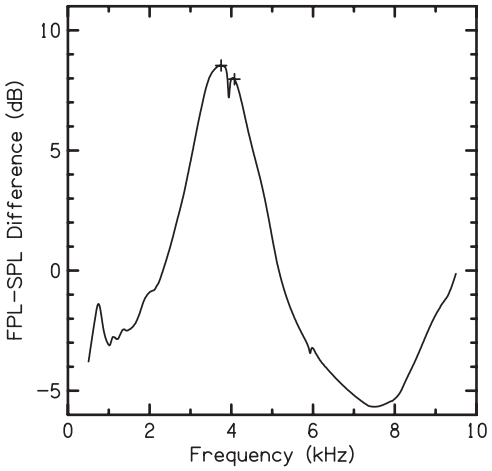

Fig. 1. FPL–SPL difference as a function of frequency for one participant. The FPL–SPL difference at the objectively selected notch ($f_{objective}$ = 3.759 kHz), is 0.6 dB greater than the FPL–SPL difference at the subjectively selected notch ($f_{visual\text{-}detection}$ = 4.078 kHz). Notch frequencies are plotted with + symbols.

Figure 2 plots the notch depth (i.e., the maximum FPL–SPL difference) as a function of frequency based on both the initial procedure and the more objective procedure for determining notch frequency. Data are provided for both NH and HI ears. These data are consistent with previous observations (Richmond et al. 2011), in that the notch frequency tends to cluster around 4 kHz, but is observed for a range of frequencies. Likewise, these data demonstrate that the depth of the notch varies across ears, ranging from about 4 to 18 or 19 dB. Together, these data justify efforts to include experiments at each participant's notch frequency.

Procedure

In situ calibration was completed in each ear using a chirp stimulus. L_1 was set according to the following formula derived by Johnson et al. (2006) and modified by Kirby et al. (2011):

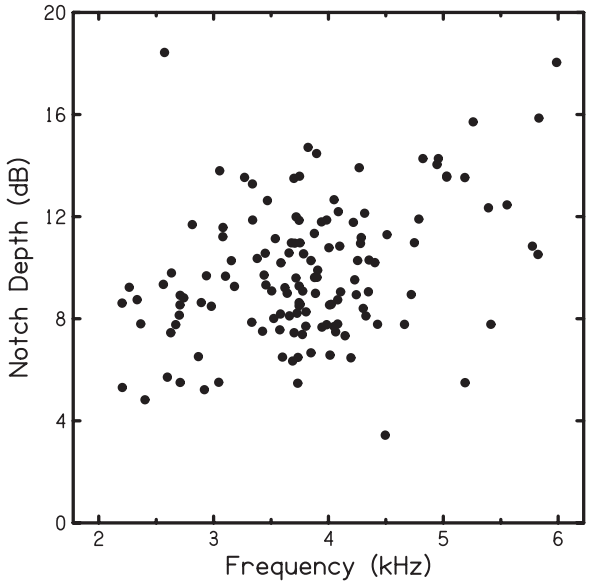

Fig. 2. Notch depth (i.e., maximum FPL-SPL difference) in dB as a function of frequency in kHz for all participants.

$$L_1 = 80 + 0.1 \bullet log_2\left(64 / f_2\right) \bullet \left(L_2 - 80\right). \quad (3)$$

The f_1 frequency was also determined using a formula derived by Johnson et al. (2006):

$$f_2 / f_1 = 1.22 + log_2\left(9.6 / f_2\right) \bullet \left(L_2 / 415\right)^2. \quad (4)$$

DPOAE responses were recorded at 4 kHz after both SPL and FPL calibrations. I/O functions were measured in 5 dB steps in descending order from 80 dB to −10 dB, or until the DPOAE response was less than 3 dB above the noise floor at two consecutive stimulus levels. Next, DPOAE I/O functions were measured with f_2 at the notch frequency using both SPL- and FPL-calibrated stimuli. To ensure that the notch frequency did not shift over time as a result of movement of the probe-microphone assembly in the ear canal, the initial 4 kHz SPL calibration was compared with a second 4 kHz SPL calibration obtained at the end of data collection. If the initial and final notch frequencies differed by more than 10%, the experimental conditions were repeated.

Each 2-second sampled DPOAE measurement was alternately stored in one of two buffers. The contents of the two buffers were summed and the level in the $2f_1 - f_2$ frequency bin was used to estimate DPOAE level (L_d). The contents of the two buffers were subtracted and the levels in the $2f_1 - f_2$ bin, along with the levels in the five bins above and below this frequency, were averaged to provide an estimate of the noise level.

DPOAE data collection continued at each L_2 until either the noise floor was ≤−25 dB SPL, the signal-to-noise ratio was >20 dB, or 210 seconds of artifact-free averaging time passed. These rules were chosen to maximize the dynamic range of the measurements. For the stimulus frequencies used in the present study, the noise levels are low; thus, averaging typically stopped on the noise-floor rule.

Analysis

Test performance was evaluated using clinical decision theory. Specifically, receiver operating characteristic (ROC) curves were constructed at each L_2 and the area under each ROC curve (A_{ROC}) was calculated to describe performance for all L_2 levels. This information was used to evaluate differences in test performance due to calibration method and to determine optimal stimulus levels for measuring DPOAEs, where optimal was defined as the stimulus level at which A_{ROC} was largest. Cumulative distributions of L_d were then constructed from the data for both NH and HI listeners in response to the optimal stimulus level for each of the four test conditions (i.e., at 4 kHz and at the notch frequency for both SPL- and FPL-calibrated stimuli) and were used to derive selected sensitivities and specificities.

Hearing threshold data, together with the DPOAE I/O data, were used to obtain multiple linear regression (MLR) coefficients that characterize the relationship between thresholds and DPOAE I/O functions for the entire group of participants. The MLR coefficients were used to predict thresholds from the DPOAE I/O measurements. Separate predictions, each with its own MLR coefficients, were made for the four different test conditions (i.e., SPL and FPL at 4 kHz, SPL and FPL at the notch frequency). Because the notch frequency was seldom exactly equal to one of the audiometric frequencies used in this study, hearing thresholds at the notch frequency were estimated as the

weighted average of measured thresholds of the audiometric frequencies below and above the notch frequency. The weight was based on the proximity of the notch frequency to the two audiometric frequencies surrounding the notch. The predictions were made using data for participants with thresholds less than or equal to 60 dB HL because most participants with thresholds greater than 60 dB HL did not produce measurable DPOAEs. Simple linear regression was used to characterize the relationship between the measured and predicted hearing thresholds, and to evaluate the success of the predictions.*This simple linear regression analysis should not be confused with the MLR analysis used for making the predictions. The accuracy of the predictions was also assessed using average root-mean-square (rms) deviation between the measured and predicted thresholds.

RESULTS

Effect of Calibration on DPOAE Level

Mean DPOAE levels as a function of L_2 are plotted in Fig. 3. The top panel displays mean I/O functions for participants with NH and the bottom panel displays mean I/O functions for those with hearing loss. Within each panel, the parameter is calibration/stimulus condition (either using SPL or FPL calibration at 4 kHz or at the notch frequency determined individually for each participant). In the interest of clarity, error bars are not provided; however, the standard deviations were similar across calibration/stimulus conditions, and, thus, were averaged to provide an overall estimate of the variability in the data. Averaged across all conditions, the standard deviations were 4.9 and 4.4 dB for NH and HI participants, respectively. Thus, there was little difference between participant groups. The average noise floor across test conditions is also not displayed, but averaged −27.2 dB SPL for NH participants and −25.9 dB SPL for HI participants. Variability in noise level was low (SD was 2.2 and 1.6 dB for NH and HI participants, respectively). Both the mean noise levels and their standard deviations are the result of the measurement-based stopping rule that continued averaging until the noise floor was ≤−25 dB SPL.

For both NH and HI participants, the I/O functions measured after SPL calibration at 4 kHz (filled circles) and at the notch frequency (open circles) were nearly identical for all L_2 levels. In NH participants, mean I/O functions measured after FPL calibration at 4 kHz (filled squares) and at the notch frequency (open squares) were shifted to the right of the mean I/O functions measured after SPL calibration at most L_2 levels. The shape of these functions, however, was not dependent on calibration/stimulus condition. In HI participants, there was a similar shift in mean DPOAE levels once the response exceeded the noise floor. However, in NH participants, there was a larger difference between I/O functions at 4 kHz and at the notch frequency with FPL calibration than in HI participants for the same stimulus conditions (filled squares compared with open squares). Because the notch frequency was usually close to 4 kHz, we expected to see little or no difference in L_d for the two frequencies for either FPL or SPL calibration. Thus, the separation in the I/O functions for 4 kHz and the notch frequency after FPL calibration in NH participants was unexpected. The reason for this effect is presently unknown.

* Correlation coefficients and p values were obtained from the simple linear regression.

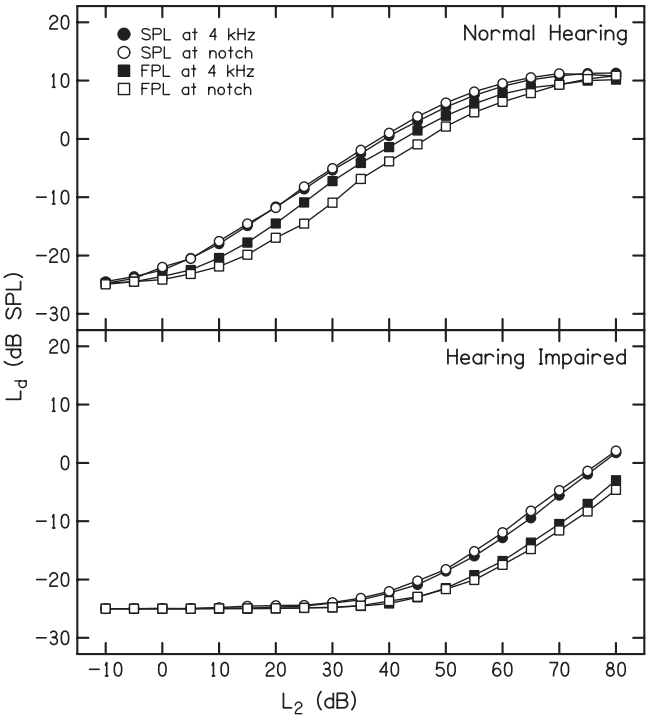

Fig. 3. Mean L_d in dB SPL as a function of L_2 in dB (both SPL and FPL) for normal-hearing (top panel) and hearing-impaired (bottom panel) participants. Circles represent data collected after SPL calibration at 4 kHz (filled) and at the notch frequency (open); squares represent data collected after FPL calibration at 4 kHz (filled) and at the notch frequency (open).

The differences between the I/O functions in Fig. 3 (with larger L_d after SPL calibration compared with L_d after FPL calibration) are the result of underestimated stimulus levels during SPL calibration, especially at the notch frequency, a condition for which the impact of standing waves is greatest. Because of destructive summation of forward and reflected waves at the plane of the probe microphone, the level measured at the probe was less than the level at the eardrum and, therefore, the level entering the cochlea. The DPOAE measurement software compensated for this difference by increasing the voltage to the loudspeakers to achieve nominal requested levels. These standing-wave effects did not occur during FPL calibration. As a consequence, the actual level presented to the ear was greater after SPL calibrations, compared with FPL calibrations. Because L_d increases as stimulus level increases, L_d measured after SPL calibration was larger than L_d measured after FPL calibration. When measurements were made at the notch frequency, the differences in stimulus levels between SPL and FPL were greatest, thus resulting in larger differences between measured L_d levels.

Test Performance

Figure 4 plots A_{ROC} as a function of L_2 for each calibration/stimulus condition. An A_{ROC} of 0.5 represents test performance at chance level, and an A_{ROC} of 1.0 represents perfect test performance. Filled circles represent A_{ROC} when L_2 was set after SPL calibrations, and open squares represent A_{ROC} after FPL calibrations. The top panel displays A_{ROC} at 4 kHz and the bottom panel displays A_{ROC} at the individually determined notch frequency. Test performance improved as L_2 increased for all test conditions up to moderate levels, after which it decreased. These

general trends are consistent with the results from previous studies (e.g., Whitehead et al. 1995; Stover et al. 1996; Burke et al. 2010; Johnson et al. 2010), and are a consequence of the fact that as stimulus level increases (regardless of calibration method) more NH participants produce responses, which drives down the false-positive rate. At high stimulus levels, some HI participants (especially those with milder losses) produce responses, which drives up the false-negative rate. Moderate stimulus levels result in conditions for which both the false-positive and false-negative rates are minimized, thus resulting in the largest A_{ROC}.

The observation of better test performance at low stimulus levels after SPL calibration is a consequence of the fact that the stimulus level, on average, was about 10 dB greater for this condition, compared with when measurements were made after FPL calibration. This difference in stimulus level is indirectly evident in the I/O functions of Fig. 3; a 10-dB rightward horizontal shift of the I/O functions after SPL calibration would cause them to overlap with the functions after FPL calibration. Therefore, when SPL calibration was used at low L_2 levels, more NH participants produced L_d levels above the noise floor and were correctly classified as having NH for all the comparisons shown in Fig. 4. Stimulus calibration effects also explain why better test performance was obtained after FPL calibration at high L_2 levels compared with test performance after SPL calibration. Larger L_d levels were produced after SPL calibration in both NH and HI participants. This result has

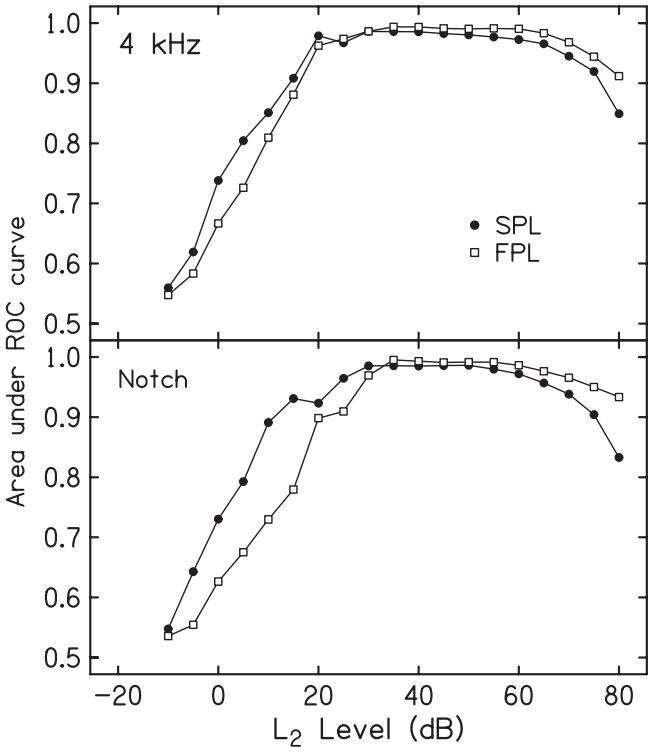

Fig. 4. A_{ROC} as a function of L_2. Filled circles represent A_{ROC} calculated from data collected after SPL calibration, and open squares represent A_{ROC} calculated from data collected after FPL calibration. The top panel displays A_{ROC} when DPOAEs were measured at 4 kHz and the bottom panel displays A_{ROC} when DPOAEs were measured at each participant's notch frequency.

no impact on the false-positive rate because NH participants were already producing large responses and are still correctly identified as having NH. However, the higher stimulus level, presented as a consequence of standing-wave interactions after SPL calibration, has the effect of eliciting responses from some HI participants, thus causing an increase in the false-negative rate. Evidence in support of this view can be seen in the bottom panel of Figure 3, showing larger L_d levels for stimulus conditions after SPL calibration in HI participants, compared with L_d when FPL calibrations were used. Therefore, at high L_2 levels, more HI participants produced responses and were incorrectly classified as having normal hearing when SPL calibration was used compared with when FPL calibration was used. Even after FPL calibration, A_{ROC} decreased at the highest levels for the same reasons, because, at high stimulus levels, some HI participants produced responses.

Over a range of moderate levels (approximately 30 to 55 or 60 dB), there was a small difference in test performance between SPL- and FPL-calibrated stimuli at both 4 kHz and at the individually determined notch frequency. A bootstrap procedure with replacement (1000 samples) was used to assess the significance of these differences in A_{ROC} when L_2 = 50 dB (Fox 2008). The results of this analysis are summarized in Table 1, where mean A_{ROC} and 95% confidence intervals are provided. While the differences were small for both 4 kHz and for the notch frequency, they were statistically significant.

Figure 5 shows mean cumulative distributions of L_d for NH and HI participants. Just as in the case in which A_{ROC}s were compared, this L_2 was chosen because optimal test performance was achieved at this stimulus level (as well as others; see Fig. 4) and because it is a level commonly used to measure DPOAE responses in the clinic. The top panel provides data for conditions after SPL calibration and the bottom panel provides data for the same stimulus conditions after FPL calibration. Thick lines represent cumulative distributions at 4 kHz, and thin lines represent cumulative distributions at the notch frequency. Solid lines represent cumulative distributions for ears with normal hearing and dashed lines represent cumulative distributions for ears with hearing loss. Any hit rate (sensitivity) and it's associated false-positive rate (1 minus the specificity) can be determined from these distributions. The top panel of Fig. 5 demonstrates how to derive false-positive rates for a selected hit rate, in this case 95%. A horizontal line is drawn from the 95th percentile on the *y* axis to the point at which it intersects with the distribution of responses from HI participants. A vertical line is extended downward from this point until it intersects with the distribution of responses from NH participants. The percentile at which the vertical line intersects with the NH distribution represents the false-positive rate that would be expected if the criterion value that resulted in a 95% hit rate was used. A similar approach can be used in which a desired false-positive rate (5% in this example) is selected and the associated hit rate is then determined, which is demonstrated in the bottom panel of Fig. 5. A horizontal line is drawn from the fifth percentile on the *y* axis to the point at which it intersects with the distribution of responses from NH participants. A vertical line is extended upward from this point until it intersects with the distribution of responses from HI participants. The percentile at which the vertical line intersects with the distribution of responses from HI participants represents the hit rate that would be expected if the criterion value resulting in a 5% false-positive rate was used.

Following the above approach, false-positive rates were determined when the hit rates were fixed at 90% and 95%, and

TABLE 1. Mean AROC at L_2 = 50 dB, based on a bootstrap procedure with 1000 samples

	SPL at 4 kHz	FPL at 4 kHz	SPL at Notch	FPL at Notch
A_{ROC} at L_2 = 50	98.0 (0.06)	99.0 (0.04)	98.6 (0.05)	99.1 (0.03)

Values in parentheses represent the 95% confidence intervals. A_{ROC} after FPL calibration exceeded A_{ROC} after SPL calibration for both stimulus frequency conditions. The differences were small but statistically significant.

hit rates were determined when the false-positive rates were set to 5% and 10%. A bootstrap procedure with replacement (1000 samples) was used to derive mean values and confidence intervals. These values are provided in Table 2. There were small differences in hit and false-alarm rates after SPL and FPL calibrations favoring FPL calibrations for all but one of the eight comparisons in Table 2. Specifically, with sensitivity fixed at either 90% or 95%, the false-alarm rates were lower with FPL calibration than with SPL calibration, and with false-positive rates fixed at either 5% or 10%, the hit rates were higher with FPL calibration than with SPL calibration. Except for the false-alarm rates when the hit rate was fixed at 90% (a condition for which there was no difference in false-alarm rates between SPL and FPL), the seven other comparisons differed significantly. This result was not found previously (Burke et al. 2010), which may be a consequence of the fact that the equation used to calculate FPL calibration has been improved since the earlier work. The cumulative distributions provided in Fig. 6 may provide support for this view. Note that there was less variability in notch depth when the characteristic impedance was set equal to the surge impedance (the present study), compared with when it was set equal to the impedance of the calibration tube as it was in the study by Burke et al. (2010). This may account for the differences in test performance between the present study and that by Burke et al.

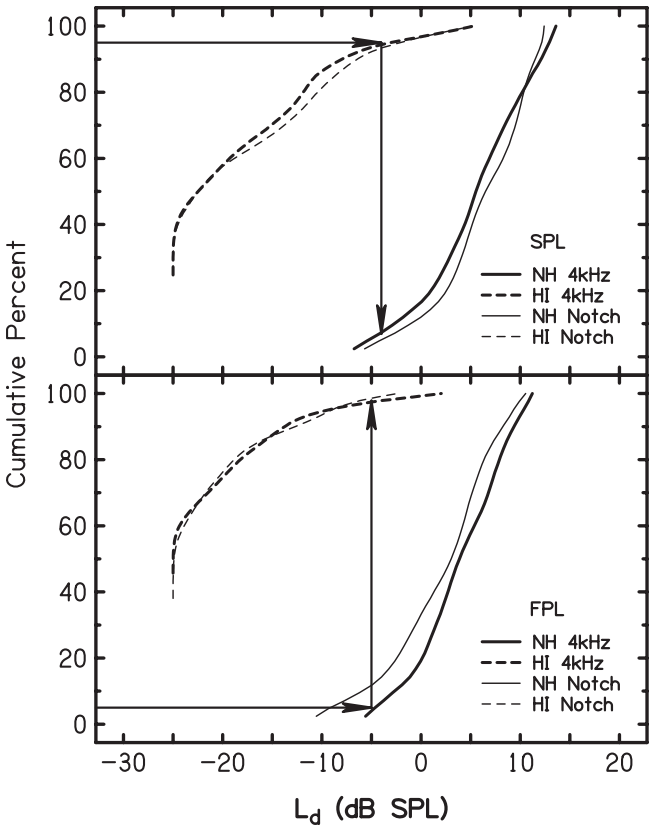

Fig. 5. Mean cumulative distributions of distortion product otoacoustic emission levels from NH (solid lines) and HI (dashed lines) participants after SPL and FPL calibration. Data are shown for the conditions in which L_2 = 50 dB (for both SPL and FPL). Thick lines represent responses collected at 4 kHz and thin lines represent responses collected with f_2 equal to the notch frequency. Results after SPL and FPL calibrations are shown in top and bottom panels, respectively. For the purpose of illustration, the false-positive rate was derived when a 95% hit rate was selected after SPL calibrations for the 4 kHz condition. Similarly, the bottom panel illustrates the case when the hit rate was determined when a 5% false-positive rate was selected.

Threshold Prediction

Figure 7 shows the relation between measured behavioral thresholds and predicted thresholds from DPOAE data. Each panel displays threshold prediction using one of the four calibration/stimulus conditions (SPL and FPL calibration at 4 kHz and at the notch frequency). The associated correlations, rms errors, and number of participants are included in each panel. Cases in which the behavioral threshold exceeded 60 dB HL were not included in this analysis because measureable DPOAE responses would not be expected for participants with this degree of hearing loss. The numbers of participants included for analysis for each condition were 116, 116, 111, and 112 for SPL at 4 kHz, FPL at 4 kHz, SPL at the notch frequency and FPL at the notch frequency, respectively. Correlation coefficients for all conditions ranged from $r = 0.83$ to $r = 0.88$, and all were statistically significant ($p < 0.001$). Average rms errors for each condition ranged from 10.5 to 11.8 dB. These results demonstrate that there was a negligible difference in the strength of the relationship between predicted thresholds based on DPOAE data and behavioral thresholds when either SPL or FPL calibration was used before measuring DPOAEs at 4 kHz or at the notch frequency.

DISCUSSION

The purpose of the present study was to examine effects of standing waves on stimulus calibration by assessing the ability of DPOAEs to differentiate between NH and HI participants (test performance) and to predict behavioral threshold. Previous research has not found a difference in test performance or threshold prediction after SPL and FPL calibration, but that work was restricted to measurements at octave and interoctave frequencies, including 3, 4, and 6 kHz. Richmond et al. (2011) demonstrated that although standing-wave effects occur most often near 4 kHz, they are distributed over the frequency range of 3 to 5 kHz. Thus, the inability of previous studies to demonstrate an influence of calibration method may have been a consequence of the fact that standing waves in participants did not occur at a test frequency. To optimize conditions in which a difference in test performance and threshold prediction might occur between calibration

TABLE 2. Derived false-positive rates from fixed hit rates of 90 and 95% and derived hit rates from fixed false-positive rates of 10 and 5%

	SPL at 4 kHz	FPL at 4 kHz	SPL at Notch	FPL at Notch
Hit rate (%)	Derived false alarm rate (%)			
90	2.4 (0.15)	0.0 (0.00)	0.0 (0.00)	0.0 (0.00)
95	12.0 (0.32)	0.0 (0.00)	9.6 (0.28)	4.8 (0.20)
False alarm rate (%)	Derived hit rate (%)			
10	94.5 (0.15)	97.8 (0.09)	95.6 (0.14)	97.9 (0.09)
5	94.5 (0.15)	97.8 (0.09)	94.5 (0.15)	96.8 (0.11)

When the hit rate was set to 90% and measurements were made at the notch frequency, there was no difference in the false-alarm rates for SPL and FPL calibrations. For all other comparisons, the differences in performance were statistically significant and favored FPL calibrations.

methods, DPOAE I/O functions were measured with f_2 set to 4 kHz and to each participant's notch frequency (i.e., the frequency at which the maximum destructive summation of pressure occurred as a result of standing waves). The results of this study are summarized with the following observations:

1. Larger L_d levels at equivalent nominal stimulus levels were produced after SPL calibration than after FPL calibration, as expected from the influence of standing-wave nulls in SPL calibrations, but not FPL calibrations.
2. Test performance was best at moderate stimulus levels with both SPL and FPL calibration.
3. There were differences in test performance between calibration methods at both high- and low-stimulus levels, but these stimulus conditions would not be used in the clinic.
4. There were small but statistically significant FPL advantages at moderate stimulus levels for conditions that might be used clinically.
5. There was no effect of calibration method on threshold prediction.

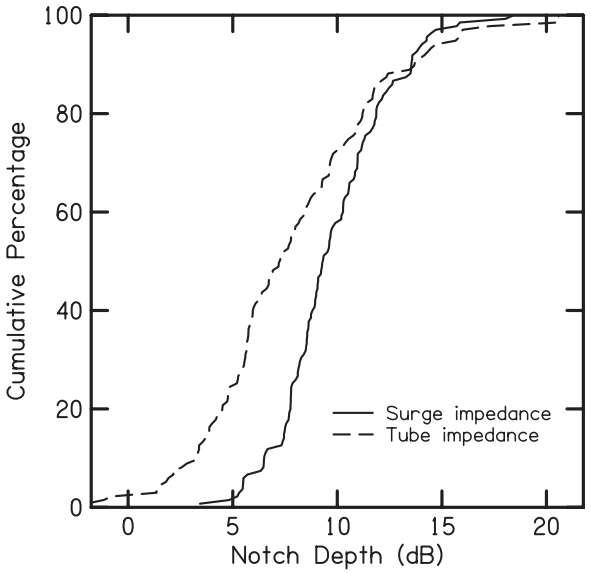

Fig. 6. Cumulative distributions of the notch depth in dB. Dashed line represents data from the study by Burke et al. (2010), in which tube impedance was used. Solid line represents data from the present study, in which surge impedance was used.

Evidence of Standing-Wave Effects

Richmond et al. (2011) demonstrated that standing waves can cause calibration differences as large as 20 dB. Our comparison of stimulus levels, inferred from the horizontal shifts in the I/O functions (Fig. 3) after SPL and FPL calibration, also demonstrated level differences due to standing waves. Because FPL calibration only takes into consideration only the forward-propagating portion of a signal, a stimulus calibrated with FPL should be 6 dB less than a signal calibrated with SPL. Deviations from this 6 dB difference likely result from standing-wave effects. Figure 1 provides an example from one participant in which the FPL–SPL difference was 8 dB at approximately 4 kHz. If it were not for cancellation of pressure at the plane of the probe microphone during SPL calibration (due to standing waves), the difference between SPL and FPL would have been −6 dB across all frequencies. Thus, standing waves exerted an influence on calibration in this representative example, which was largest near 4 kHz. Figure 2 summarizes the distribution of notch frequency and notch depth in our sample of participants. As with the data reported by Richmond et al., these data show that notch frequency varies over a range from 2.1 to 6 kHz, although the distribution clusters around 4 kHz. The notch depth varies over a range from 4 to 19 dB.

The mean DPOAE I/O functions shown in Fig. 3 demonstrate the predicted effect of standing waves after SPL calibration. Because of the destructive summation of pressure that results from standing waves during SPL calibration, the stimulus level measured at the probe microphone resulted in an underestimation of effective stimulus level. To compensate, voltage level at the receivers in the probe-microphone system was increased. FPL calibrations are not affected by standing waves at the plane of the calibration microphone so no compensation to increase level was needed. As a result, the level presented after SPL calibration exceeded the level presented after FPL calibration, and the magnitude of this difference was equivalent to the magnitude of the standing-wave effect. Thus, we predicted that the measured L_d for equivalent, nominal L_2 levels would be larger for cases in which SPL calibration was used compared with cases when measurements were made after FPL calibration. In fact, indirect evidence of this effect is provided in Fig. 3, as the DPOAE I/O functions obtained after FPL calibration are shifted to the right of those obtained after SPL calibration. Like the example shown in Fig. 1, the I/O functions in Fig. 3 serve to validate the FPL calibration approach and demonstrate the expected consequence of standing-wave effects after SPL calibrations.

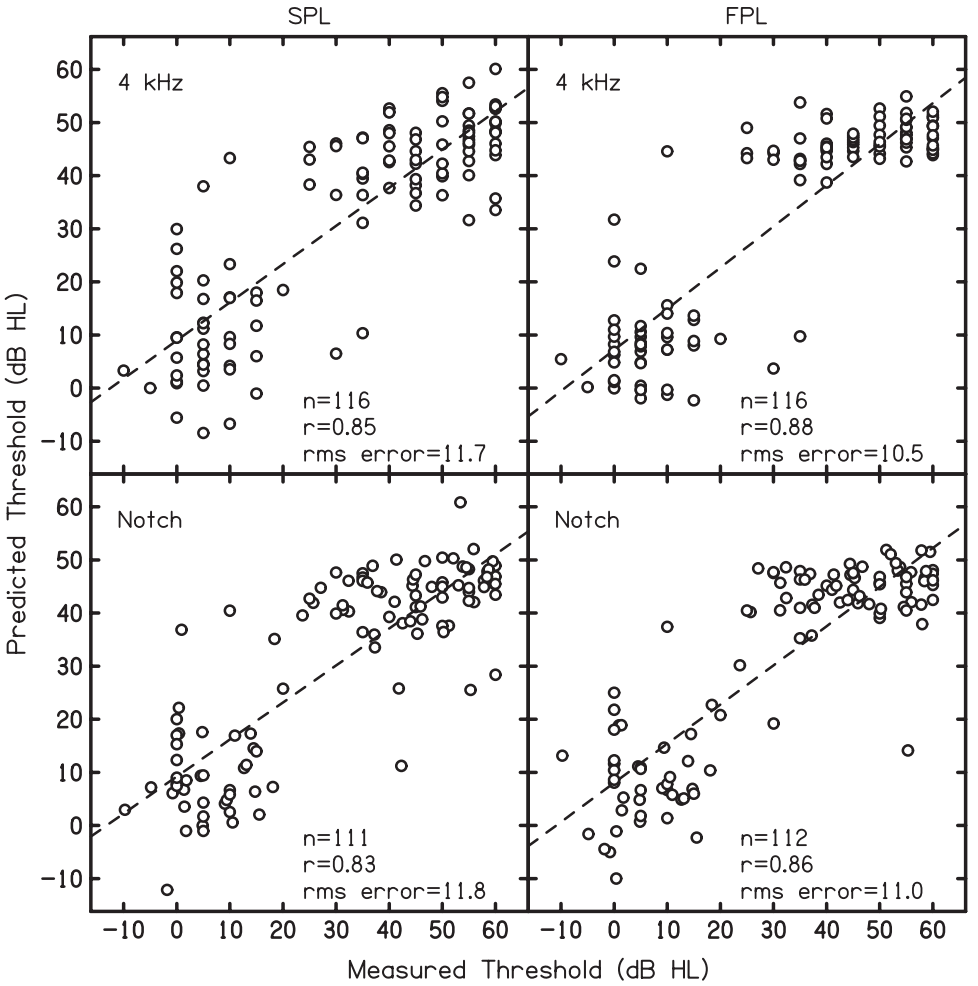

Fig. 7. Predicted threshold (dB HL) as a function of measured behavioral threshold (dB HL) with SPL and FPL calibration at 4 kHz and at the notch frequency. Dashed lines are simple linear regression fits that characterize the relationship between the predicted and measured thresholds. Correlations, numbers of participants and rms errors (dB) are provided in each panel.

Test Performance

Figure 4 plots A_{ROC} as a function of stimulus level for both SPL and FPL calibrations for 4 kHz and the individually determined notch frequency. Regardless of calibration or stimulus, test performance was best at moderate stimulus levels, which is consistent with previous findings (e.g., Whitehead et al. 1995; Stover et al.1996; Burke et al. 2010; Johnson et al. 2010). At low levels, test performance was poor for all calibration procedures but was better after SPL calibration than it was after FPL calibration. At high levels, test performance decreased from the maximum achieved at moderate levels, but was better after FPL calibration than after SPL calibration. These general trends are expected because low-level stimuli (regardless of the approach to calibration) result in high false-positive rates and high-level stimuli increase false-negative rates. Because DPOAE levels were greater after SPL calibration than after FPL calibration, with SPL calibration more NH participants were classified correctly at low stimulus levels (a decrease in the false-positive rate) and some HI participants were classified incorrectly at high stimulus levels (an increase in the false-negative rate). These level-dependent trends are consistent with the view that greater stimulus levels were presented after SPL calibration, presumably as a consequence of standing-wave effects. Notably, neither low-level nor high-level stimuli are used clinically because these levels do not produce optimal test performance as described by A_{ROC}.

At moderate stimulus levels, which are typically used in the clinic, only small differences in test performance were found, but results after FPL calibration resulted in slightly greater, statistically significant, A_{ROC} compared with when SPL calibration was used (Table 1). To supplement the summary provided in Fig. 4, we constructed cumulative distributions of DPOAE responses collected at 50 dB after both SPL and FPL calibrations in both NH and HI participants. We then selected two hit rates (90% and 95%) and determined the associated false-alarm rates, and then selected two false-alarm rates (5% and 10%) and determined the associated hit rates (Fig. 5 and Table 2). At 50 dB, there were statistically significant improvements in test performance with FPL calibration compared with SPL calibration. These results differ from those reported by Burke et al. (2010), where differences in test performance were not observed. The differences between the present findings and those reported by Burke et al. may be the result of small improvements in the FPL calculation since the collection of data in the earlier study (see Fig. 6 and its associated discussion). For example, the characteristic impedance Z_0, which was previously assumed to be the same in all ears, is now set equal to

the "surge" component of the load impedance Z_l, which is slightly different for each ear (Rasetshwane & Neely 2012).

Threshold Prediction

Predicted behavioral thresholds were calculated with an MLR analysis that used the entire DPOAE I/O function. Although there were differences in mean DPOAE level with calibration type (see Fig. 3), correlations between actual and predicted thresholds were similar for both SPL and FPL calibration at 4 kHz and at the notch frequency. Thus, there was no evidence that calibration method improved predictions of behavioral threshold. This result may not be important, as DPOAEs are not being used to predict behavioral thresholds clinically. The impact of differences in calibration method might be more evident in other applications, where a 10 dB difference in sound level might have a larger impact, such as when making real-ear measurements of hearing-aid output.

CONCLUSIONS

In summary, our findings indicate that there are standing-wave effects with SPL stimulus calibration. These effects were evident at 4 kHz and at each participant's notch frequency. Their impact was such that test performance after FPL calibration either equaled or exceeded the performance after SPL calibrations. Although the differences were small, they were statistically significant. Therefore, it may be of value to use FPL calibration as an alternative to SPL calibration for DPOAE measurements in the clinic. In addition, there may be other circumstances in which stimulus levels are measured in closed ear canals. For these other applications, the impact of standing waves on measured levels may be large enough to warrant the use of FPL calibration.

ACKNOWLEDGMENTS

The authors thank Colleen Gibilisco for her assistance with participant recruitment and Shapelle Freudenberg for her contributions to the experimental design and to data collection.

This work was supported by the National Institutes of Health (National Institute on Deafness and Other Communication Disorders grants T35 DC8757, R01 DC2251, R01 DC8318, P30 DC4662).

Address for correspondence: Michael P. Gorga, Boys Town National Research Hospital, 555 North 30th Street, Omaha, NE 68131, USA. E-mail: michael.gorga@boystown.org

Received November 27, 2012; accepted April 24, 2013.

REFERENCES

Boege, P., & Janssen, T. (2002). Pure-tone threshold estimation from extrapolated distortion product otoacoustic emission I/O-functions in normal and cochlear hearing loss ears. *J Acoust Soc Am*, *111*, 1810–1818.

Brownell, W. E. (1990). Outer hair cell electromotility and otoacoustic emissions. *Ear Hear*, *11*, 82–92.

Burke, S. R., Rogers, A. R., Neely, S. T., et al. (2010). Influence of calibration method on distortion-product otoacoustic emission measurements: I. test performance. *Ear Hear*, *31*, 533–545.

Fox, J. (2008). *Applied Regression Analysis and Generalized Linear Models*, (2nd ed.). Thousand Oaks, CA: Sage, Chap. 21.

Gilman, S., & Dirks, D. D. (1986). Acoustics of ear canal measurement of eardrum SPL in simulators. *J Acoust Soc Am*, *80*, 783–793.

Gorga, M. P., Neely, S. T., Bergman, B., et al. (1993). Otoacoustic emissions from normal-hearing and HI participants: Distortion product responses. *J Acoust Soc Am*, *93*(4), Pt. 1, 2050–2060.

Gorga, M. P., Neely, S. T., Dorn, P. A., et al. (2003). Further efforts to predict pure-tone thresholds from distortion product otoacoustic emission input/output functions. *J Acoust Soc Am*, *113*, 3275–3284.

Gorga, M. P., Neely, S. T., Ohlrich, B., et al. (1997). From laboratory to clinic: A large scale study of distortion product otoacoustic emissions in ears with normal hearing and ears with hearing loss. *Ear Hear*, *18*, 440–455.

Johnson, T. A., Neely, S. T., Garner, C. A., et al. (2006). Influence of primary-level and primary-frequency ratios on human distortion product otoacoustic emissions. *J Acoust Soc Am*, *119*, 418–428.

Johnson, T. A., Neely, S. T., Kopun, J. G., et al. (2010). Clinical test performance of distortion-product otoacoustic emissions using new stimulus conditions. *Ear Hear*, *31*, 74–83.

Kim, D. O., Paparello, J., Jung, M. D., et al. (1996). Distortion product otoacoustic emission test of sensorineural hearing loss: Performance regarding sensitivity, specificity and receiver operating characteristics. *Acta Otolaryngol*, *116*, 3–11.

Kirby, B. J., Kopun, J. G., Tan, H., et al. (2011). Do "optimal" conditions improve distortion product otoacoustic emission test performance? *Ear Hear*, *32*, 230–237.

Lonsbury-Martin, B. L., McCoy, M. J., Whitehead, M. L., et al. (1993). Clinical testing of distortion-product otoacoustic emissions. *Ear Hear*, *14*, 11–22.

Neely, S. T. & Liu, Z. (2011). *EMAV: Otoacoustic Emission Average. Tech Memo No. 17*. Omaha, NE: Boys Town National Research Hospital.

Neely, S. T., & Gorga, M. P. (1998). Comparison between intensity and pressure as measures of sound level in the ear canal. *J Acoust Soc Am*, *104*, 2925–2934.

Rasetshwane, D. M., & Neely, S. T. (2012). Measurements of wide-band cochlear reflectance in humans. *J Assoc Res Otolaryngol*, *13*, 591–607.

Richmond, S. A., Kopun, J. G., Neely, S. T., et al. (2011). Distribution of standing-wave errors in real-ear sound-level measurements. *J Acoust Soc Am*, *129*, 3134–3140.

Rogers, A. R., Burke, S. R., Kopun, J. G., et al. (2010). Influence of calibration method on distortion-product otoacoustic emission measurements: II. threshold prediction. *Ear Hear*, *31*, 546–554.

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.

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. (2007). Calibrating otoacoustic emission probes. In M. S. Robinette & T. J. Glattke (Eds.), *Otoacoustic Emissions: Clinical Application* (3rd ed.). (pp. 403–427) New York, NY: Thieme Medical Publishers, Inc.

Siegel, J. H., & Hirohata, E. T. (1994). Sound calibration and distortion product otoacoustic emissions at high frequencies. *Hear Res*, *80*, 146–152.

Stover, L., Gorga, M. P., Neely, S. T., et al. (1996). Toward optimizing the clinical utility of distortion product otoacoustic emission measurements. *J Acoust Soc Am*, *100* (2), Pt. 1, 956–967.

Whitehead, M. L., McCoy, M. J., Lonsbury-Martin, B. L., et al. (1995). Dependence of distortion-product otoacoustic emissions on primary levels in normal and impaired ears. I. Effects of decreasing L2 below L1. *J Acoust Soc Am*, *97*, 2346–2358.

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.