Comparing the Accuracy and Speed of Manual and Tracking Methods of Measuring Hearing Thresholds

Gayla L. Poling,[1] Theresa J. Kunnel,[2] Sumitrajit Dhar[1]

Objectives: The reliability of hearing thresholds obtained using the standard clinical method (modified Hughson-Westlake) has been the focus of previous investigation given the potential for tester bias (Margolis et al., 2015). In recent years, more precise methods in laboratory studies have been used that control for sources of bias, often at the expense of longer test times. The aim of this pilot study was to compare test-retest variability and time requirement to obtain a full set of hearing thresholds (0.125 – 20 kHz) of the clinical modified Hughson-Westlake (manual) method with that of the automated, modified (single frequency) Békésy tracking method (Lee et al., 2012).

Design: Hearing thresholds from 10 subjects (8 female) between 19 to 47 years old (mean = 28.3; SD = 9.4) were measured using two methods with identical test hardware and calibration. Thresholds were obtained using the modified Hughson-Westlake (manual) method and the Békésy method (tracking). Measurements using each method were repeated after one-week. Test-retest variability within each measurement method was computed across test sessions. Results from each test method as well as test time across methods were compared.

Results: Test-retest variability was comparable and statistically indistinguishable between the two test methods. Thresholds were approximately 5 dB lower when measured using the tracking method. This difference was not statistically significant. The manual method of measuring thresholds was faster by approximately 4 minutes. Both methods required less time (~ 2 mins) in the second session as compared to the first.

Conclusion: Hearing thresholds obtained using the manual method can be just as reliable as those obtained using the tracking method over the large frequency range explored here (0.125 – 20 kHz). These results perhaps point to the importance of equivalent and valid calibration techniques that can overcome frequency dependent discrepancies, most prominent at higher frequencies, in the sound pressure delivered to the ear.

Key words: Behavioral audiometry, Comparative study, Hearing loss, Normal hearing, Threshold assessment.

(Ear & Hearing 2016;37;e336–e340)

INTRODUCTION

Significant advancement toward finding a standard method for obtaining hearing thresholds occurred when Hughson and Westlake (1944) suggested starting at a level high enough to be audible, attenuating in 5 *or* 10 dB steps until the tone became inaudible, and then increasing the presentation level until the tone was audible again. Further constraints were imposed on the clinical method by Carhart and Jerger (1959) by specifying decrements and increments to be in 10- and 5-dB steps, respectively, and by reinforcing the need to ascertain thresholds based on ascending runs only. While this modified Hughson–Westlake procedure has since been the method of choice for clinical audiometry, there have been previous explorations seeking alternate methods to augment accuracy as well as promote automation.

The motivation to develop automated techniques for audiometry has come primarily from the need to reach the population with hearing loss (Lin et al. 2011) with greater efficiency than can be afforded by the current work force in audiology (Margolis & Morgan 2008) and the burgeoning interest in telehealth (Mahomed et al. 2013). The urgent need to increase access to audiological assessment has to be counterbalanced against the accuracy of the methods available to automate threshold estimation. Of note, Zhou and Green (1995) investigated several sources of variability of pure-tone thresholds, specifically investigating the effects of standing waves in the ear canal, the impedance of the headphone used for measurement, and inherent variability in threshold estimation due to differences in the psychometric function of signal detection at various frequencies. In brief, these authors confirmed that standing waves contributed more to inconsistencies in threshold estimation at higher, rather than lower, frequencies. However, this source of variability was balanced by the use of low impedance headphones thereby mitigating the formation of standing waves and by the steeper psychometric functions at higher frequencies.

The accuracy and efficiency of various psychophysical methods for measuring hearing thresholds has also been examined with an interest in applying them in the laboratory for minimizing bias as well as to achieve precision in threshold estimation in large populations (reviewed by Margolis et al. 2015). In brief, a derivative of a tracking procedure first proposed by Levitt (1971) is often used in laboratory studies, as is a general two-alternate forced choice (2AFC) paradigm. Green (1993) developed a yes-no single interval task that resulted in similar accuracy but significant timesaving compared with a more traditional 2AFC method (Leek et al. 2000). Most recently, Margolis et al. (2010, 2011) have developed and validated a single interval yes-no protocol with feedback and catch trials to automatically obtain air and bone conduction thresholds. Furthermore, we reported (Lee et al. 2012) a large set of threshold data obtained using an automated tracking procedure designed after Levitt (1971). In short, thresholds were determined by the listener controlling the attenuation of the stimulus using a push button, a single frequency at a time. In this brief report, we compare thresholds obtained using this method to those obtained with the clinical modified Hughson–Westlake (manual) procedure. Notably, the same signal delivery hardware and calibration procedure were used to obtain both sets of data allowing a direct comparison of their accuracy and efficiency across approximately the entire human hearing range (0.125 and 20 kHz).

The two primary research questions addressed in this preliminary investigation included (1) Can comparable test–retest variability be achieved between the clinical modified Hughson–Westlake (manual) method and an automated, modified (single

[1]The Roxelyn and Richard Pepper Department of Communication Sciences and Disorders, Northwestern University, Evanston, Illinois, USA; and [2]Loyola University's College of Arts and Sciences, Loyola University of Chicago, Chicago, Illinois, USA.

<zdoi; 10.1097/AUD.0000000000000317>

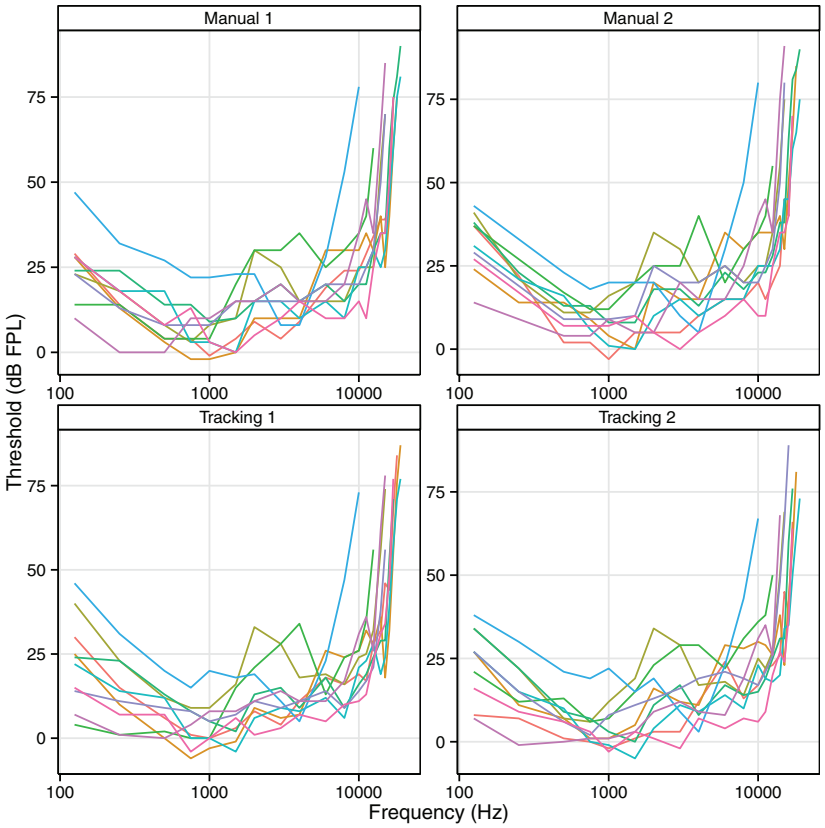

Fig. 1. Hearing thresholds from each subject measured using signals calibrated using forward pressure (dB FPL) as a function of frequency. Each panel represents data obtained using one of two methods in one of two test sessions. The titles of the panels indicate the method and trial number, with data from individual subjects are displayed using the same color consistently across panels.

frequency) Bekesy tracking method (Lee et al. 2012)?; (2) Is the time required to obtain a full set of hearing thresholds (0.125 to 20 kHz) shorter for one of the methods?

MATERIALS AND METHODS

Behavioral hearing thresholds for 21 frequencies between 0.125 and 20 kHz were recorded from one ear (left/right counterbalanced) of each of 10 participants (8 females) between the ages of 19 and 47 years (mean = 28.3; SD = 9.4) with varying degrees of hearing ability. As this was designed to be a preliminary exploration, inclusion in the study was open to those between the ages of 18 and 65 years of age with availability for testing and confirmation of a clear ear canal and healthy tympanic membrane by otoscopy before each measurement. All data collection was done in a standards-compliant sound-treated booth in accordance with the guidelines of the IRB at Northwestern University.

Each subject contributed 4 sets of data recorded during 2 sessions, 1 week apart. In each test session, behavioral hearing thresholds were obtained using identical instrumentation using two measurement protocols—the standard clinical (modified Hughson–Westlake) procedure (manual method, hereafter) and a modified Bekesy tracking method (tracking method, hereafter) as described below and detailed in Lee et al. (2012). In keeping with normal practice 5 (up) and 10 (down) dB steps were used for the manual procedure. A 2-dB step size was used for threshold determination in the tracking procedure with the listener pressing an attenuator button as long as the stimulus tone was audible and releasing the button when no longer audible. The tester determined threshold for the clinical procedure based on guidelines to accept the lowest level at which the subject responded two out of three times during an ascending trial. Threshold was algorithmically determined in the tracking procedure as the mean value of the midpoints of ascending runs. The mean value of these midpoints was continually computed and updated after the first six reversals in attenuator direction. The current mean value was accepted as the threshold once the standard error of measurement computed as the standard deviation of the midpoints divided by the square root of the sample size fell below 1 (see Lee et al. 2012 for details). Time required to obtain the entire set of thresholds (0.125 to 20 kHz) for each test procedure was recorded (start/stop time recorded on all tests and for each subject from the same clock on a computer by the same tester).

It is important to note that an identical calibration method was used for each of the four datasets. Thévenin-equivalent probe calibration was used to estimate the forward-going sound pressure incident on the tympanic membrane free of contamination from reverse-going waves after reflection at the tympanic membrane. This quantity is referred to as forward pressure and is considered an estimate of input to the tympanic membrane free of complications caused by standing waves and the variable resonances produced at varying depths of insertion (see Souza et al. 2014 for details). All data are presented in forward pressure level (FPL) to reflect the use of this calibration process.

The effect of test method and test session on hearing thresholds and test time were evaluated using a random-intercept

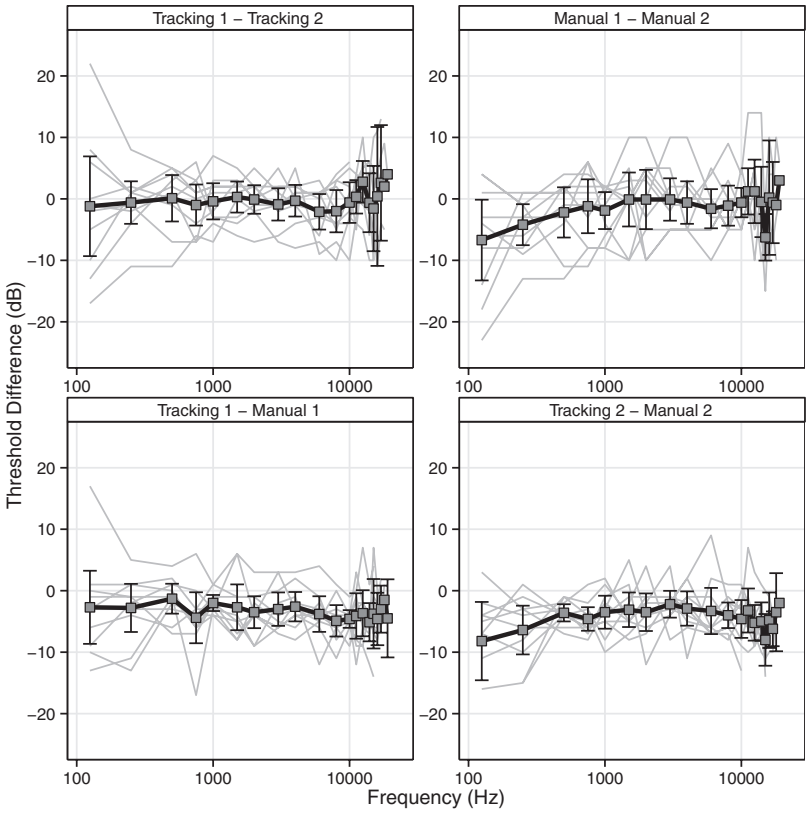

Fig. 2. Differences in hearing thresholds obtained using two test methods within one session or between two test sessions using the same method. Thin gray traces represent threshold differences from individual subjects. The symbols represent the mean difference. The error bars represent the 95% confidence interval. Each panel is labeled to mark the difference represented.

model in R (R Core Team 2014). The model included a cross-level interaction between test session and method. Differences in thresholds within subjects across test sessions and methods were accounted for in the model.

RESULTS

Similarities in threshold patterns within subjects across trials and methods are evident in the four panels of Figure 1. Data are not included in these plots when subjects did not respond at the output limit of our equipment, nominally 105 dB SPL with slight variations as a function of frequency. As anticipated this impacted the highest frequencies evaluated in subjects. The highest frequency at which a response was obtained from a given subject is also consistent across methods and sessions.

Differences in measured thresholds as a function of frequency are plotted in Figure 2 for both methods. Few differences in thresholds exceeded 10 dB (11/210 instances for manual and 8/210 for tracking) with differences most commonly observed at the lowest frequencies (<1 kHz; 7/11 instances for manual and 5/8 for tracking) across sessions as well as methods. The mean difference across sessions and methods remains between 0 and −5 dB for all comparisons with differences within individual subjects rarely exceeding 10 dB. Most pertinently, the differences between sessions are similar for the tracking and manual methods. Similarly, the differences between methods were consistent for the two sessions. Within any given session manually obtained thresholds were slightly but consistently higher (~5 dB) than those obtained using the tracking method. Thresholds obtained within sessions were not statistically significantly different across methods ($p > 0.05$). Similarly, thresholds obtained across sessions were not statistically significantly different within methods ($p > 0.05$).

The time required to obtain a complete set of thresholds for each method and trial in individual subjects is plotted in the left panel of Figure 3. Group data for each method and session are displayed in the right panel of Figure 3 in the form of box plots. Closed symbols represent the first trial with each method. The sex, age (in years), and test ear of each subject is indicated in the x axis of the left panel. With the exception of one subject (plotted as the first set in the left panel), tracking thresholds took longer to complete in all subjects within sessions. Many subjects were faster during the second trial (open symbols). The median difference in test time between the 2 sessions was approximately 2 min, irrespective of test method. Two subjects were slow to complete the tracking dataset; however, their thresholds were consistent with the group and they were not particularly older or younger than the remaining subjects. Closer examination of the tracking records for these 2 subjects revealed that both subjects had multiple prolonged descending runs well past their ultimate threshold. That is, these subjects failed to release the response button even when the signal level was well below their threshold. Once the subjects released the response button, they allowed the signal level to rise past their ultimate threshold, as expected, before starting to attenuate the signal again. These prolonged excursions delayed convergence to an acceptable value of the standard error of measurement.

As is evident in the right panel of Figure 3, the median times for the tracking method were longer than those for the manual

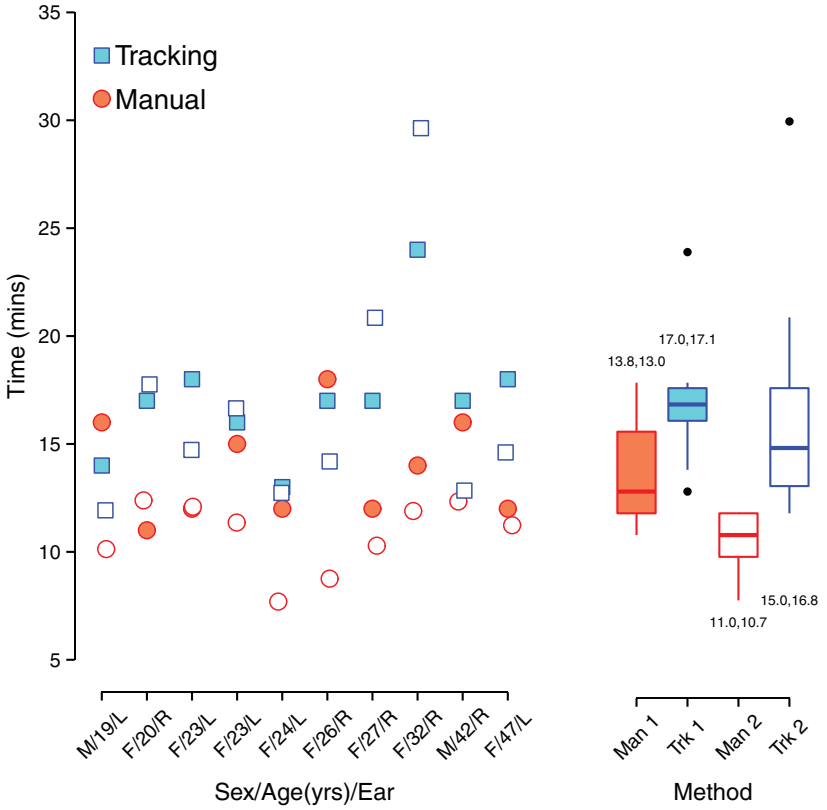

Fig. 3. Time required to obtain a full set of thresholds. Data from individual subjects for the tracking (squares) and manual (circles) methods are displayed in the left panel. Closed and open symbols are used to represent data from the first and second sessions, respectively. Subjects are represented with increasing age from the left to right of the panel. Group data are presented using box plots in the right panel. Horizontal lines mark the median, with the boxes representing the interquartile range, and the whiskers extending to 1.5 times the interquartile range. Outliers are marked with solid circles. The median and mean values for each method and session are presented near each box using the "median, mean" format.

method in each session. More remarkably, the first quartile of the range of times to completion for the tracking method was longer in duration than the third quartile of the times to completion for the manual method in each test session. The reduction in test time between sessions was not statistically significant ($p > 0.05$) for either test method.

DISCUSSION

Automated methods for establishing pure-tone hearing thresholds are desired to aide telehealth applications as well as to increase accuracy for laboratory measurements. In this report, we compared thresholds obtained using a popular psychophysical method (after Levitt 1971) with those obtained using the standard manual procedure by a trained tester. Equivalent thresholds were obtained using the two methods. However, there were differences in the time required to obtain a complete set of thresholds using the two methods.

The accuracy of different laboratory methods (e.g., 2AFC and single interval yes-no) for obtaining hearing thresholds has been compared and found to be equivalent before (reviewed by Leek et al. 2000). The accuracy of other automated methods has also been found to be comparable with audiologist-obtained thresholds (Margolis et al. 2010; 2015; Jacobs et al. 2012). Our results support these previous findings. The tracking thresholds reported here were consistently, but not statistically significantly, lower than the manual thresholds. Lower thresholds when using a tracking or other self-recording method have been consistently reported in the literature and attributed to differences in stimulus step size and response inertia (e.g., Harris 1979).

Speed is another consideration when determining the method of choice. Faster methods lead to greater subject cooperation and data volume in the laboratory. In the clinical setting, more efficient methods ultimately reduce cost without sacrificing quality. Leek et al. (2000) reported the measurement time to be 2 min 13 sec per threshold for a 2AFC method. In contrast, each threshold could be obtained in 45 sec using the single-interval yes-no procedure. This time increased to 1 min 24 sec when track length was allowed to vary to reach a stable state in the yes-no procedure. During the first trial in our dataset, the average time required to obtain a single thresholds was 37 and 48.6 sec for the manual and tracking methods, respectively. Thus, either method was faster than the times reported for the 2AFC or single-interval yes-no methods by Leek et al. Although the average speed for the tracking method was not very different from that for the manual method, 2 subjects took inordinately longer using this method. Margolis et al. (2011) have reported a similar experience with some children who took inexplicably long on their automated procedure. This highlights the need for data monitoring either by a clinician/experimenter or by the measurement algorithm to identify individuals experiencing difficulty during automated threshold estimation.

In conclusion, our results suggest that accuracy of hearing threshold estimation is not necessarily compromised by using

the standard, manual modified Hughson–Westlake method. Thresholds are obtained with slightly greater speed using the manual method as compared with the tracking method used in this report. The speed of automated threshold acquisition of course depends on the efficiency of the algorithm. In fact, specific tracking patterns were identifiable in subjects who experienced difficulty converging to a reliable threshold. This information can be leveraged to develop intelligent algorithms that detect loss of engagement or lack of attention and create an audible warning system to draw the listener back to the task. Finally, our results are not a rejection of the tracking procedure; they simply indicate that the manual method should perhaps not be dismissed off hand on account of inaccuracy. This is the only report we are aware of where the issues of calibration and measurement method have been decoupled by controlling stimulus delivery with advance calibration procedures to minimize individual variations in ear-canal geometry. Once decoupled from calibration-related uncertainties (through the use of identical calibration procedures for both methods), the manual method of measuring hearing thresholds appears to be a comparable choice. Further investigation in a larger population is warranted to explore the full clinical utility of the study findings.

ACKNOWLEDGMENTS

The authors thank Jungwha Lee for her statistical guidance. Our thanks to Amir Khan, Samir Datta, Chelsea Adams, Nicole Oliver, Jesus Echezarreta, Steven Stachowiak, Victoria Hellyer, Julia Junghwa Lee, and Jungmee Lee for their help on various aspects of this project. We are grateful to Jonathan H. Siegel for countless discussions on this topic and specific feedback on this manuscript.

This research was supported by NIH/NIDCD Grants R01DC008420, T32DC009399, and Northwestern University.

The authors have no conflicts of interest to disclose.

Address for correspondence: Sumitrajit Dhar, The Roxelyn and Richard Pepper Department of Communication Sciences and Disorders, Northwestern University, 2240 Campus Drive, Evanston, IL 60208, USA. E-mail: s-dhar@northwestern.edu

Received June 9, 2015; accepted March 3, 2016.

REFERENCES

Carhart, R., & Jerger, J. F. (1959). Preferred method for clinical determination of pure-tone thresholds. *J Speech Hear Disord*, *24*, 96–108.

Green, D. M. (1993). A maximum-likelihood method for estimating thresholds in a yes-no task. *J Acoust Soc Am*, *93*(4 pt 1), 2096–2105.

Harris, D. A. (1979). Microprocessor, self-recording and manual audiometry. *J Aud Res*, *19*, 159–166.

Hughson, W., & Westlake, H. (1944). Manual for program outline for rehabilitation of aural casualties both military and civilian. *Trans Am Acad Ophthalmol Otolaryngol*, *48*(suppl), 1–15.

Jacobs, P. G., Silaski, G., Wilmington, D., et al. (2012). Development and evaluation of a portable audiometer for high-frequency screening of hearing loss from ototoxicity in homes/clinics. *IEEE Trans Biomed Eng*, *59*, 3097–3103.

Lee, J., Dhar, S., Abel, R., et al. (2012). Behavioral hearing thresholds between 0.125 and 20 kHz using depth-compensated ear simulator calibration. *Ear Hear*, *33*, 315–329.

Leek, M. R., Dubno, J. R., He, N., et al. (2000). Experience with a yes-no single-interval maximum-likelihood procedure. *J Acoust Soc Am*, *107* (5 Pt 1), 2674–2684.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *J Acoust Soc Am*, *49*, Suppl 2:467+.

Lin, F. R., Niparko, J. K., Ferrucci, L. (2011). Hearing loss prevalence in the United States. *Arch Intern Med*, *171*, 1851–1852.

Mahomed, F., Swanepoel, d. e. W., Eikelboom, R. H., et al. (2013). Validity of automated threshold audiometry: A systematic review and meta-analysis. *Ear Hear*, *34*, 745–752.

Margolis, R. H., & Morgan, D. E. (2008). Automated pure-tone audiometry: An analysis of capacity, need, and benefit. *Am J Audiol*, *17*, 109–113.

Margolis, R. H., Frisina, R., Walton, J. P. (2011). AMTAS(®): Automated method for testing auditory sensitivity: II. Air conduction audiograms in children and adults. *Int J Audiol*, *50*, 434–439.

Margolis, R. H., Glasberg, B. R., Creeke, S., et al. (2010). AMTAS: Automated method for testing auditory sensitivity: Validation studies. *Int J Audiol*, *49*, 185–194.

Margolis, R. H., Wilson, R. H., Popelka, G. R., et al. (2015). Distribution characteristics of normal pure-tone thresholds. *Int J Audiol*, *54*, 796–805.

R Core Team. (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: URL http://www.R-project.org/.

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.

Zhou, B., & Green, D. M. (1995). Reliability of pure-tone thresholds at high frequencies. *J Acoust Soc Am*, *98*(2 Pt 1), 828–836.