AN ANALYSIS OF THE CURRENT CLINICAL PROCEDURE FOR MEASURING HEARING THRESHOLDS AND RECOMMENDATIONS FOR IMPROVING THE MEASUREMENT

by

Whitney Mast

Audiology Doctoral Project submitted to the faculty of the

DEPARTMENT OF SPEECH, LANGUAGE, AND HEARING SCIENCES

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF AUDIOLOGY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2018

THE UNIVERSITY OF ARIZONA

GRADUATE COLLEGE

As members of the Audiology Doctoral Project Committee, we certify that we have read the audiology doctoral project prepared by Whitney Mast, titled An Analysis of the Standard Clinical Procedure for Measuring Hearing Thresholds and Recommendations for Improving the Measurement, and recommend that it be accepted as fulfilling the audiology doctoral project requirement for the Degree of Doctor of Audiology.

Huanping Dai Date: 4/20/18

David Velonovsky Date: 4/20/18

Linda Norrix Date: 4/20/18

Stephanie Griffin Date: 4/20/18

Final approval and acceptance of this audiology doctoral project is contingent upon the candidate's submission of the final copies of the audiology doctoral project to the Graduate College.

I hereby certify that I have read this audiology doctoral project prepared under my direction and recommend that it be accepted as fulfilling the audiology doctoral project requirement.

Date: 4/20/18

Audiology Doctoral Project Director: Huanping Dai

STATEMENT BY AUTHOR

This Audiology Doctoral Project has been submitted in partial fulfillment of the requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this Audiology Doctoral Project are allowable without special permission, provided that an accurate acknowledgement of the source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: Whitney Mast

ACKNOWLEDGEMENTS

I would like to acknowledge my mentor on this project for the past two years, Dr. Huanping Dai, for being instrumental in helping me expand my skills as a researcher and pushing me to always ask bigger questions. Additionally, I would like to thank the other members of my committee, Dr. David Velenovsky, Dr. Linda Norrix, and Dr. Stephanie Griffin, for their insightful commentary and thoughtful review on this project. Finally, I would like to thank all of my close friends and family, who have supported me through the long journey of my graduate career so faithfully.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

ABSTRACT

Measuring hearing thresholds is part of virtually all hearing evaluations. The standard clinical procedure for measuring hearing thresholds is easy to use, efficient, and its outcome is relatively reproducible (Levitt, 1971). However, the procedure has the downside that the task performance (proportion of 'yes' response, or P_{Yes}) targeted by the clinical thresholds is unspecified. Furthermore, it remains uncertain whether the procedure extracts information about threshold in an optimal way. The purpose of this paper is to address both issues. Clinical thresholds (TH^{HW}) were measured at 1 and 4 kHz for 22 normal-hearing adult listeners with 10 repetitions per listener at each frequency. For determining the target P_{Yes} for TH^{HW}, or P_{Yes}^{HW}, psychometric functions, which relate P_{Yes} to signal level, were also measured. From the measured psychometric functions, both P_{Yes}^{HW} and the threshold corresponding to $P_{Yes} = 0.5$, or $TH_{0.5}$, were derived. On average, $P_{Yes}^{HW} \approx 0.93$ and $TH_{0.5} \approx TH^{HW} - 3\ dB$. With minimal modifications to the current procedure, better estimates of thresholds were obtained either by fitting a psychometric function to the trials in the adaptive track, or by averaging the turning points (reversals) in the adaptive track. Therefore, the above simple modifications to the current clinical protocol could be utilized by clinicians for threshold estimation.

I. INTRODUCTION

Pure-tone audiometry for measuring hearing thresholds is an essential part of virtually all hearing evaluations. The standard clinical procedure for measuring hearing thresholds, the modified Hughson-Weslake (HW) protocol (Jerger and Cahart, 1959), has been adopted as an ANSI standard (ANSI, 2004) and within clinical guidelines (ASHA, 2005). Figure 1 illustrates how this procedure works via an example of the adaptive track obtained using the procedure. In this procedure, the experimenter starts an adaptive track by presenting the signal at a level well above the assumed hearing threshold of the listener (-55 dB at Trial #1). In subsequent trials, the signal level descends by 10 dB following a 'yes' response (marked by small black dots) and ascends by 5 dB if no response (marked by small red dots) is received. On an ascending run, the lowest signal level without receiving a response is referred to as a lower reversal (marked by triangles), and the first signal level receiving a 'yes' response is referred to as an upper reversal (marked by stars). As soon as two ascending runs end at the same signal level, that signal level is recorded as the threshold (marked by TH^{HW}).

This standard HW procedure has several benefits in that it is easy to use, that it takes little time—typically around 10 trials at each frequency, and that it is relatively consistent and reproducible (Mahomed et al., 2013; Song et al., 2015; Leijon, 1992). However, several important aspects about the HW procedure have remained uncertain. The first uncertainty is that, for hearing thresholds (TH^{HW}) measured using the HW procedure, the corresponding target performance is undefined. In psychophysical research, detection thresholds are commonly defined as the signal level corresponding to a specific value of proportion of correct or yes response. Thus, the threshold corresponds to a particular point on the listener's psychometric function (PF), which describes proportion of correct or yes responses as a function of

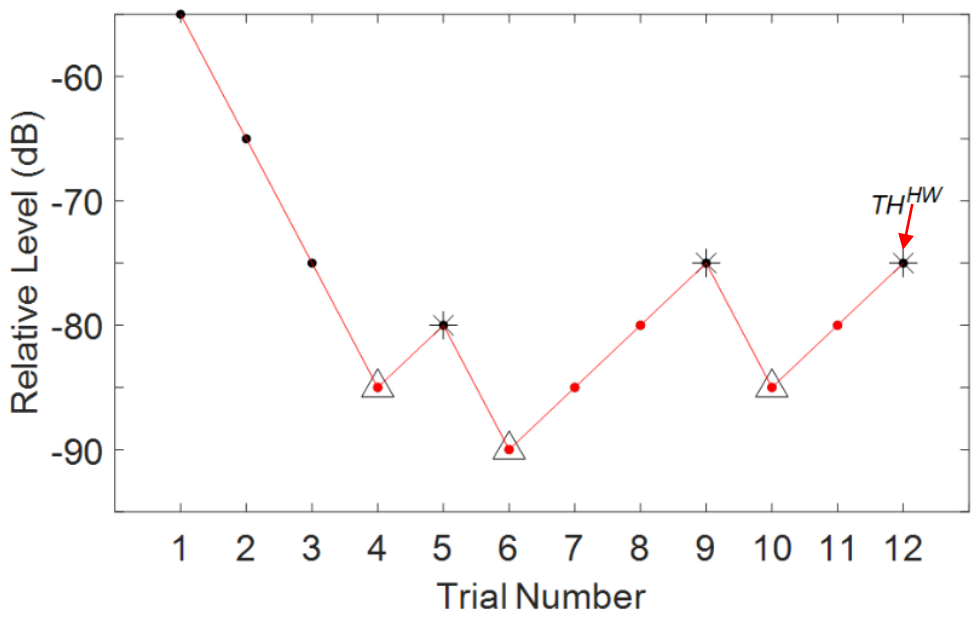

Figure 1. An example of an adaptive track of relative signal level (dB) over trials in the standard Hughson-Weslake (HW) clinical procedure for estimating hearing thresholds. Signal level is expressed relative to approximately 106-dB SPL (per spec of the Sennheiser headphone HD265). The small black dots indicate the signal level on trials where the listener's response was 'yes'; the small red dots indicate the signal level on trials where the listener either did not respond as in the HW protocol, or responded 'no' as in the experimental 'yes-no' procedure. Triangles indicate lower reversals, which mark the beginning of ascending runs; stars indicate upper reversals, which mark the end of ascending runs. The clinical threshold obtained for this particular adaptive track is labeled as TH^{HW} on trial # 12; it is the sound level $(-75\ dB)$ at which the upper reversal (stars) visited for a second time (after a visit on

signal level. In the standard HW procedure, in contrast, it is only vaguely stated that, at threshold, the proportion of yes responses, denoted as P_{Yes}, is at least 0.5. Due to the fact that the target performance is undefined for these thresholds, it is unknown whether or not these thresholds actually correspond to any particular expected value of P_{Yes}. If there is a particular expected value of P_{Yes} to which the clinical threshold corresponds, then it needs be determined. It is important to resolve this uncertainty, because the knowledge concerning the P_{Yes} value is necessary for determining the accuracy of the clinical threshold (TH^{HW}).

The second uncertainty about the standard HW procedure is if information has been extracted from the adaptive tracks in an optimal way. One reason that this is suspected not to be the case is that the step sizes (10 dB for descending runs and 5 dB for ascending runs) used in the clinical protocol are relatively large and may be too coarse to produce very precise threshold estimates. It has been reported that smaller step sizes have yielded better estimates of threshold (Marshall et al., 1996; Jerlvall & Arlinger, 1986). In the HW protocol, compounded with the large step sizes, deriving threshold based on *only* upper reversals could potentially overestimate the threshold. For these reasons, it is conceivable that both the accuracy and the consistency of threshold measurements can be improved. An important goal of this study was to find ways to better extract information from the data made available either by the current protocol or by simple modifications of the current protocol, so that a more accurate threshold measurement can be obtained. Such simple modifications would allow for easy adoption by clinicians.

The purpose of this study was to address the uncertainties described above. For this purpose, an accurate and reliable estimate of the psychometric function is needed for each listener. Therefore, extensive data from additional tests was collected to increase the total number of trials from an average of 10 in the standard clinical method to720 for obtaining each psychometric function. A psychometric function, referred to as the grand PF, or PF^{grand}, was then fitted for each listener using these data. Using this PF^{grand}, the P_{Yes} value corresponding to each estimated clinical threshold (TH^{HW}), or P_{Yes}^{HW} was determined. Furthermore, using this PF^{grand}, the best estimate of a threshold defined as the signal level corresponding to $P_{Yes} = 0.5$, or $TH_{0.5}^{Grand}$ was found. For each listener, $TH_{0.5}^{Grand}$ serves as a benchmark against which thresholds obtained using any procedure can be compared, thus the accuracy and consistency of these thresholds can be assessed.

Throughout this paper, the threshold corresponding to $P_{Yes} = 0.5$ will be referred to as $TH_{0.5}$. Defining threshold at $P_{Yes} = 0.5$ is consistent with traditional psychophysics in which thresholds are typically defined at the mid-point on the psychometric function. Furthermore, there is an apparent advantage for defining threshold at $P_{Yes} = 0.5$. It has been shown that among thresholds that are defined at different levels of target performance, the threshold defined at $P_{Yes} = 0.5$ is expected to be least variable (Green, 1993; Taylor and Creelman, 1967; Fisher, 1922). Several adaptations of the clinical protocol were explored, including a), averaging upper and lower turning points or reversals in the adaptive track, and b), fitting psychometric functions in several ways to the same trials as used in deriving the clinical threshold. For each adaptation, the accuracy and consistency of the threshold estimates was assessed.

II. METHODS

A. Listeners

Twenty one undergraduate and one graduate student at the University of Arizona participated in the study. Twenty one listeners were female and one was male, with an age range of 20-40 years. All of the listeners denied hearing difficulties at the time of the study and indicated that English was their primary language. The listeners received extra course credit for their participation. On average, listeners completed the entire task in under 2 hours, and were allowed to take short breaks as needed. This research was reviewed and approved by the Social and Behavioral Sciences division of the Human Subjects Protection Program at the University of Arizona. Informed consent was obtained for all participants.

B. Stimuli

The signals were generated digitally and played at a sampling rate of 44100 Hz with a 24-bit resolution. They were presented to each listener's right ear via Sennheiser (HD265) supra-aural headphones. The signals used were 200-ms pure tones of 1000 or 4000 Hz. Their onsets and offsets were shaped with a squared cosine and had 10-ms ramps. The signal level was specified as relative to approximately 106-dB SPL according to the specification sheet of the Sennheiser headphones (HD265). Signal level was calibrated before the start of each test. Given that the main purpose of this study was to make comparisons between different measures of thresholds, only the relative signal levels or level differences, rather than the absolute signal levels, are essential to be specified correctly.

C. Procedure

In this study, the standard HW procedure was modified to a yes-no task (see Fig.1 for an example). The yes-no task differs from the standard HW procedure in two aspects. First, whereas the HW procedure does not solicit an explicit 'no' response when the listener does not hear the signal, the yes-no task does. Second, whereas the HW procedure does not have clearly marked observation intervals thus does not specify exactly *when* the signal should be expected, the yes-no task does. In short, the listeners are subject to greater amount of uncertainty regarding to the timing of the signal in the HW procedure than in the yes-no task used in this study. These modifications were made in the current study because the goal was best served when the listeners were performing at their best sensitivity, which required minimizing any uncertainty, including that about signal timing. Because the HW procedure is more prone to timing uncertainty than the yes-no task, the above modifications to the HW procedure, and use of the yes-no task, are inevitable. Other than these modifications, the threshold-seeking procedure was kept as similar

as possible to the HW procedure, in that the signal level decreased by 10 dB after the listener provided a 'yes' response, and increased by 5 dB after a 'no' response was received.

The computer controlled the testing sequence of the signal presentation and listener responses. Listeners performed the 'yes-no' task using a computer keyboard in response to tones at different intensity levels. During testing, the participants were seated in a sound-treated room at a desk with a computer monitor and keyboard while wearing supra-aural headphones (HD265). Instructions were displayed on the monitor throughout the duration of testing. The listener was asked to press 1 on the keyboard to indicate that they heard the sound, even if it was very faint, and 0 to indicate that they did not hear any sound during the observation interval, which was indicated by the presence of a flashing figure on the computer monitor. The visual display informed the listener of the exact timing of the signal. Following each signal presentation, the computer waited for the listener's response, and upon receiving each response, executed a 500-ms pause before presenting the next signal.

Each listener completed 10 repetitions of the HW protocol thus yielding 10 threshold estimates at each frequency. To construct a psychometric function, referred to as the grand PF, or PF^{grand}, for the purpose of establishing a benchmark threshold against which obtained thresholds can be compared, additional detection tests were included for each listener. These tests used smaller step sizes (as small as 1 dB) thus yielded finer details of the psychometric function. In particular, the first additional protocol utilized an adaptive track similar to that used in the HW protocol, with the exception of having a starting step size of 8 dB and 4 dB that was then decreased to 2 dB and 1 dB. The second additional protocol used the maximum likelihood procedure to create an estimate of the listener's psychometric function based on their responses to tones at different signal levels. In total, the HW protocol and the two additional protocols

with smaller observation intervals comprised of a total of 720 trials, which were then used to obtain the PF^{grand} for each listener at each frequency.

D. Data analysis

Across listeners and for each signal frequency, probability distributions (Fig. 5) were constructed based on all estimates (220 estimates total, 10 estimates per listeners) of the clinical thresholds (TH^{HW}) relative to the benchmark threshold ($TH_{0.5}^{Grand}$). The shape of the two distributions combined across frequency reasonably resembles the shape of a normal distribution. Henceforth, the mean and standard deviation of the threshold estimates will simply be reported without further description of the distributions of the thresholds.

The procedure used for fitting the psychometric functions (PF) was similar to that described in Dai and Micheyl (2011). The trials of the adaptive tracks were sorted according to the values of signal level (dB), denoted as x, and for each value of x, P_{Yes} was computed. The resulting P_{Yes} values paired to the x values were fitted using a cumulative Gaussian probability function Φ:

$$P_{Yes} = \alpha + (1 - \alpha - \lambda)\Phi\,(x,\ \mu,\ \sigma), \qquad (1)$$

where μ is the mean and σ is the standard deviation of the underlying Gaussian probability density function (PDF); α is the false-alarm parameter indicating the offset of the lower asymptote of the PF from $P_{Yes} = 0$; λ is the inattention parameter indicating the offset of the upper asymptote of the PF from $P_{Yes} = 1$. The mean,μ, is also referred to as the *threshold* parameter, which corresponds to $P_{Yes} = 0.5$. The standard deviation, σ, describes the spread of the PF, thus is referred to as the *spread* parameter; it is inversely related to the slope of the PF. If the slope of a PF is defined as a change in z-score per dB, then σ is the inverse of the slope, i.e.,

$1/\sigma$ can be regarded as the slope parameter (Marshall & Jesteadt, 1986). All four parameters, μ, σ, α, and λ, were estimated using the maximum-likelihood method. Note that, what is referred to as $TH_{0.5}$ in the current paper is always estimated as the threshold parameter from a PF.

III. RESULTS AND DISCUSSION

A. Clinical thresholds measured using the standard HW procedure: TH^{HW}

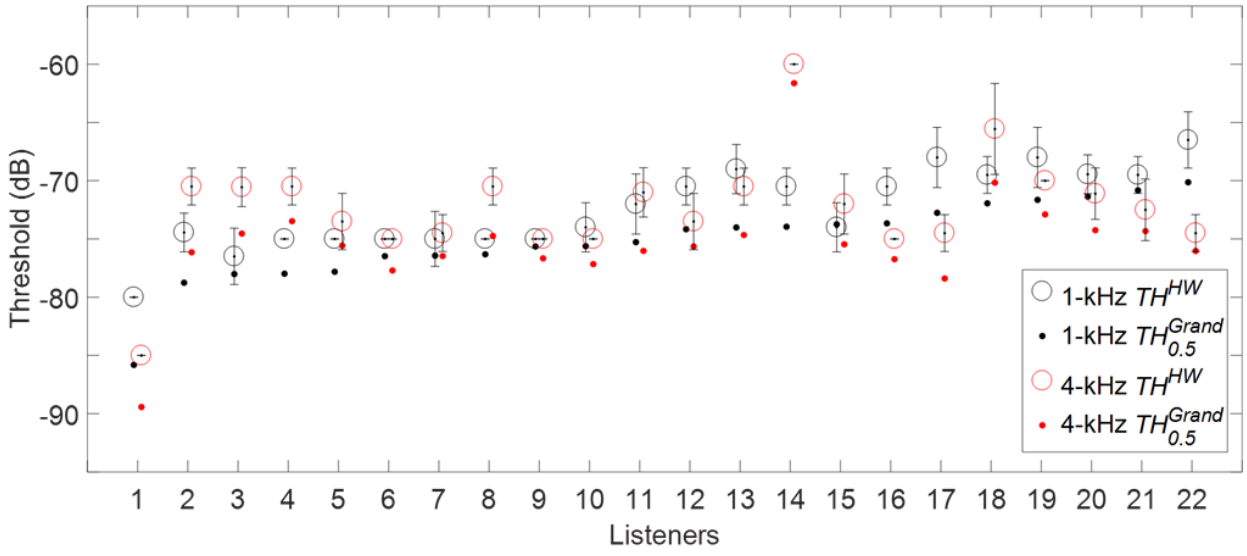

Figure 2. Thresholds of all listeners at 1 kHz (black symbols) and 4 kHz (red symbols). Each open circle represents the mean of 10 estimates of clinical thresholds (TH^{HW}) for each listener; each error bar represents the standard deviation of these estimates. Each dot represents the benchmark threshold$TH^{Grand}_{0.5}$ which was obtained via fitting the grand psychometric function based on 720 trials. The order of the listeners was based on the $TH^{Grand}_{0.5}$ values at 1 kHz (black dot) plotted from low to high.

Figure 2 shows for each listener the mean of 10 estimates of clinical thresholds (TH^{HW}, unfilled circles; the small, filled circles labeled as $TH^{Grand}_{0.5}$ will be discussed later) at two signal frequencies (1 kHz in black; 4 kHz in red). On average, each threshold estimate took 9.6 trials. The reader may get a sense about the repeatability or consistency of clinical thresholds by

looking at the size of the error bars. Each error bar represents the standard deviation of the threshold estimates. In a dozen cases there was no variability among repetitions—the threshold estimates were identical in all 10 repetitions. There is considerable individual variability in the standard deviation of threshold estimates, ranging from zero (e.g., Listener #1 at both frequencies) to about 5 dB (Listener # 18 at 4 kHz). The mean of all standard deviations across listeners and frequency was 1.68 dB. This value appeared considerably smaller than what has been previously reported. For example, the standard deviation was 3.9 dB according to a meta-analysis of existing reports by Mahomed et al. (2013), and 4.1 dB (mean of 3.4 and 4.8 dB) according to Green (1993). The standard-deviation scores in Mahomed et al. (2013) were based on differences between a pair of threshold estimates, thus it is expected to be a factor of $\sqrt{2}$ greater than if the standard deviation is based on individual threshold estimates. Converted to a measure similar to that of the present study, their estimate of standard deviation would be $3.9/\sqrt{2}$ =2.76 dB, which is still somewhat greater than the present estimate. It is possible that these differences in threshold variability are partly due to the differences in experimental procedure among studies. A standard deviation of 1.68 dB is relatively small, implying that clinical threshold estimates are relatively consistent or repeatable.

While the finding of a relatively small standard deviation of 1.68 dB implies a reasonably good consistency, this finding is not overtly meaningful without specifying the accuracy of threshold estimates, as the target performance of threshold estimates may not converge to a single expected value. To specify the accuracy and consistency for clinical thresholds, one needs to know the psychometric functions of individual listeners. To obtain a reliable estimate of PFs, data was collected over a relatively large number of trials (n=720) for each listener at each signal frequency. Given that on average only about 10 trials were used to estimate each clinical

threshold (TH^{HW}), this estimate of each PF was based on more than 70 times as many trials. These estimates of PFs, because they are based on *all* trials available, will be referred to as the grand PFs, denoted as PF^{grand}. In the next section the estimates of PF^{grand} for individual listeners are described.

A. Grand psychometric functions: PF^{grand}

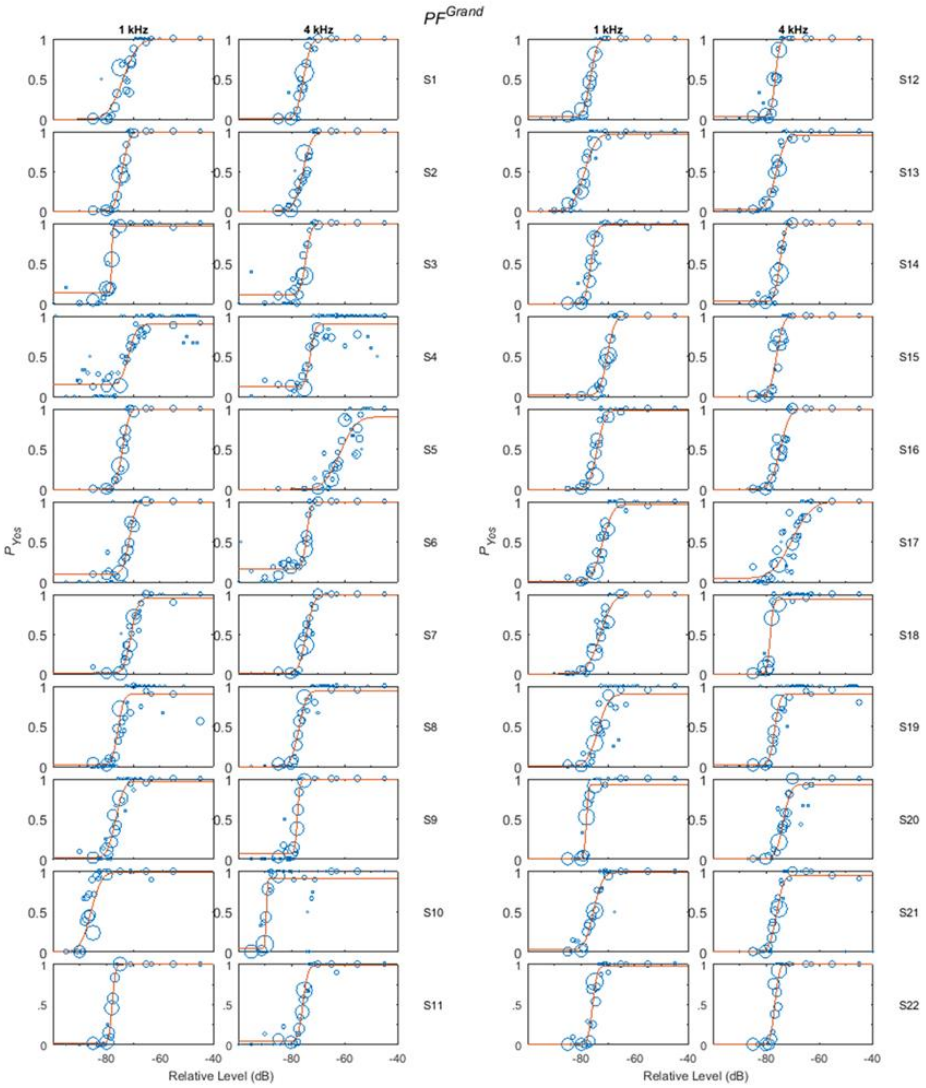

Figure 3. Raw data (blue circles) and fitted curves of the grand psychometric functions (PF^{grand})—proportion of yes responses P_{Yes} as a function of the relative signal level—for 22 listeners (rows) and 2 signal frequencies (columns). The size of each circle is proportional to the number of trials used to obtain each data point. Note that the subject number on this graph does not match the order of the listeners in Fig. 2 where the order was based on benchmark thresholds at 1 kHz.

Establishing grand psychometric functions (PF^{grand}) for each listener serves two important purposes. First, PF^{grand} allows for estimation of the expected value of P_{Yes} to which the clinical thresholds (TH^{HW}) actually correspond, or P_{Yes}^{HW}. By specifying P_{Yes}^{HW}, a major pitfall of the standard HW protocol will be overcome, which is that the task performance is unspecified. Second, PF^{grand} provides the best estimate of a benchmark threshold ($TH_{0.5}^{Grand}$) for each listener. For each listener, $TH_{0.5}^{Grand}$ is simply the threshold parameter (μ) obtained by fitting the listener's PF^{grand}. Knowing $TH_{0.5}^{Grand}$ is important because it allows for accuracy assessment of threshold estimates obtained using any procedure. Next, individual PF^{grand} are described.

Figure 3 shows the grand psychometric functions (PF^{grand}) for individual listeners (rows) and signal frequencies (columns). These PFs display some apparent variability both within listeners (across frequency) and between listeners. *Within listeners*, aside from the apparent variability, the PFs also appear to show some degree of similarity across frequency. To provide a quantitative measure of this similarity, across-frequency correlation coefficients were computed for the estimates of all four parameters. The correlation for the σ parameter was insignificant ($r = 0.07, \mathrm{P} > 0.37$). However, all three other parameters show statistically significant correlations across frequency. The correlation was strong for the false-alarm (α) parameter ($r = 0.85, \mathrm{P} < 0.0001$) and moderate for the threshold (μ) parameter ($r = 0.56, \mathrm{P} < 0.003$) and inattention ($\lambda$) parameter ($r = 0.5, \mathrm{P} < 0.01$). These significant correlations across frequency suggest that there is some consistency in a listener's behavior with regard to the three parameters. This outcome may be helpful when classifying the listeners into subgroups based on their tendencies. For example, S19, who demonstrated relatively high inattention rate at both frequencies, may be classified as an inattentive listener. Additionally, S3 and S4, who both

demonstrated relatively high false-alarm rate at both frequencies, may be classified as eager responders. Such classifications are valuable for understanding individual differences.

Furthermore, *within listeners*, for each signal frequency, correlations among the four parameters were examined to investigate any possible interactions. It was suspected that some of the parameters may interact with each other. For example, inattention could impair performance thus elevate threshold, which would yield a positive correlation between the estimates of inattention and threshold parameters. Such suspicion can be tested by examining the within-frequency correlation coefficients among the four parameters. The correlation coefficients among the parameters were computed and are shown in Table I. No correlations were statistically significant at the P=0.05 level except for the correlation between the threshold (μ) and spread (σ) parameters at 4 kHz (marked by an asterisk).

Table I. Correlation coefficients (first row: 1 kHz; second row: 4 kHz) computed among the four parameters of the PF^{grand}. The asterisk marks where the correlation is statistically significant at the P=0.05 level.

	Threshold (μ)	Spread (σ)	False-Alarm (α)	Inattention (λ)
Threshold (μ)	1 (1kHz) 1 (4kHz)	0.140 0.712*****	0.143 0.262	-0.019 -0.061
Spread (σ)		1 1	-0.325 -0.118	0.119 0.033
False-Alarm (α)			1 1	-0.125 0.163
Inattention (λ)				1 1

Between listeners, individual differences are apparent for estimates of threshold (i.e., the horizontal position of the function), spread (1/slope) of the PF, false-alarm rate (offset from the

lower asymptote, $P_{Yes} = 0$), and inattention (offset from the upper asymptote, $P_{Yes} = 1$). For instance, some listeners demonstrated relatively high rates of inattention consistently across frequency (e.g., S4, S8, S19, and S20), whereas many others did not show inattention at all. Likewise, some listeners demonstrated relatively high rates of false alarm consistently across frequency (e.g., S3, S4, and S6), whereas others showed little-to-no false alarm.

Thus, the overall pattern of results appears to be a combination of apparent between-listener individual differences and moderate to strong within-listener consistency, as indicated by the moderate cross-frequency correlations for the threshold and inattention parameters and the strong cross-frequency correlation for the false-alarm parameter. This pattern of results could be useful for classifying listeners into subgroups, as a way of describing trends in human auditory discrimination behavior. Now that the individual functions (PF^{grand}) are available as described above, these functions can be applied to estimate P_{Yes}^{HW}, or the expected value of P_{Yes} to which the clinical HW thresholds correspond, so that the target performance of the clinical protocol can be defined.

B. Estimating P_{Yes}^{HW} for clinical thresholds

Although the target performance, i.e., P_{Yes}^{HW}, for the clinical protocol was originally unspecified, the grand psychometric function (PF^{grand}) allowed for determining P_{Yes}^{HW} to which the clinical thresholds (TH^{HW}) correspond using Eq. 1. For a given value of TH^{HW}, the corresponding value of P_{Yes}^{HW} was obtained by substituting TH^{HW} into the signal-level variable, i.e., $x = TH^{HW}$, in the fitted form of the PF^{grand}. The distributions of the obtained P_{Yes}^{HW} values are shown in Fig. 4 for all estimates of TH^{HW} across listeners for 1 kHz (blue bars) and 4 kHz (green bars). The mean value of P_{Yes}^{HW} is 0.925 for 1 kHz, and 0.940 for 4 kHz, with an

overall mean of 0.932. This overall mean value is consistent with Marshall and Jesteadt (1986) who reported a mean value of 0.917. While it has often been stated that the target performance of the standard clinical protocol ranges anywhere from $P_{Yes}^{HW} = 0.5$ to 1, based on this analysis, it is recommended that the target performance of the standard clinical procedure be characterized as around P_{Yes}=0.932.

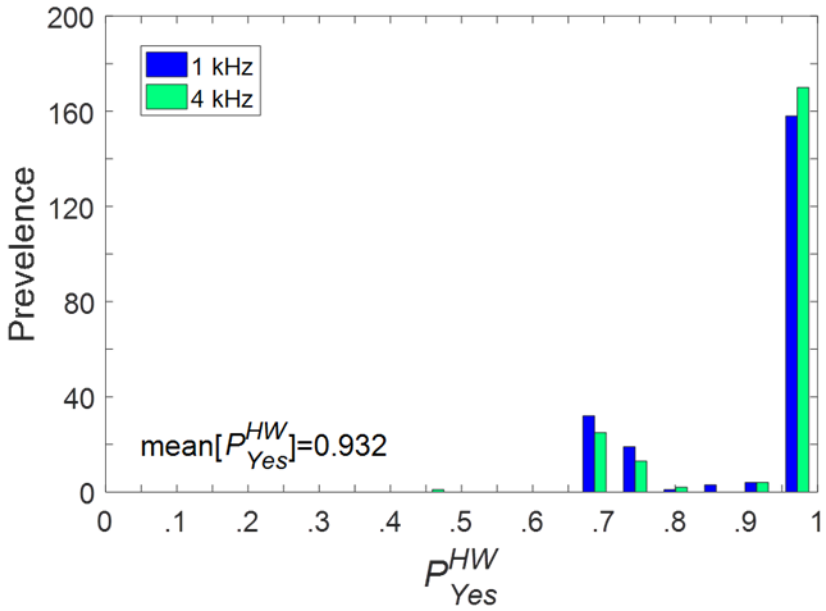

Figure 4. Distributions of P_{Yes}^{HW} corresponding to clinical thresholds for two frequencies (1 kHz in blue and 4 kHz in green). Each P_{Yes}^{HW} value was obtained based on the grand psychometric function (PF^{grand}). Each distribution is based on a total of 220 threshold estimates. The mean of P_{Yes}^{HW} was computed across listener and frequency.

C. Threshold difference: $\Delta = TH^{HW} - TH_{0.5}^{Grand}$

Given that the estimated mean target performance of the clinical protocol is relatively high ($P_{Yes}^{HW} = 0.932$), it is of interest to determine the difference between the clinical threshold (TH^{HW}) and the benchmark threshold ($TH_{0.5}^{Grand}$), i.e., $\Delta = TH^{HW} - TH_{0.5}^{Grand}$, for each listener at both signal frequencies. Figure 5 shows the distributions of Δ for all listeners at 1 kHz (blue bars) and 4 kHz (green bars). The combined distribution of Δ across frequency reasonably resembles a

normal distribution. The average Δ was 2.9 dB, with a standard deviation of 2.2 dB. These values are also shown in Fig. 6 (marked as TH^{HW}). Thus, on average, TH^{HW} is expected to be approximately 3 dB above $TH_{0.5}^{Grand}$.

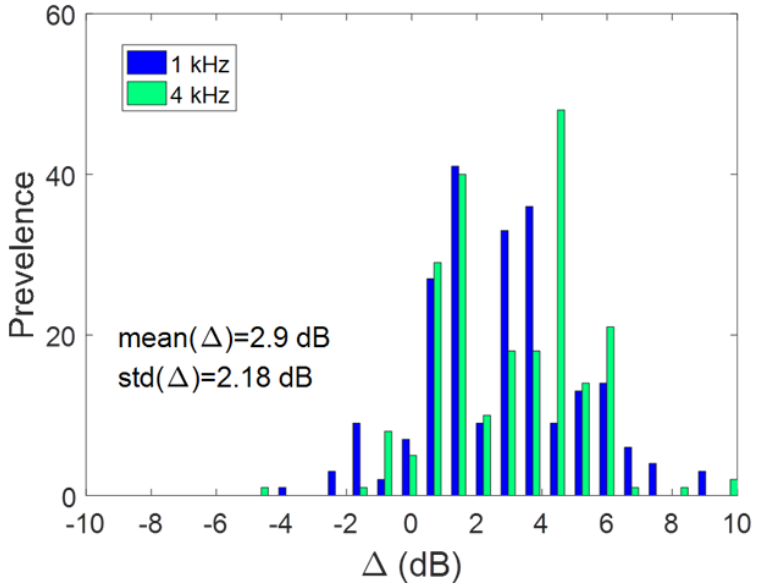

Figure 5. Distributions of threshold differences ($\Delta = TH^{HW} - TH_{0.5}^{Grand}$) at two frequencies (1 kHz in blue and 4 kHz in green). Each distribution is based on a total of 220 threshold estimates. The mean and standard deviation of Δ are computed across listener and frequency.

D. Estimating threshold based on averaging reversals: TH^{Rev} and TH^{Avg}

In this section, a simple strategy was explored in order to arrive at the threshold that would correspond to P_{Yes}=0.5. Again, the goal is to optimize the information that can be extracted from the same data available in the adaptive tracks for estimating thresholds. In psychoacoustic research, threshold has often been obtained by averaging reversals of adaptive tracks (see, e.g., Levitt, 1971). This strategy is adopted by the current approach. Briefly, in adaptive tracks, a lower reversal (see triangles in Fig. 1) is defined as the beginning level in an ascending run, and an upper reversal (see stars in Fig. 1) is the ending level of an ascending run. Here, the accuracy

and consistency of thresholds obtained by averaging *only* the signal levels that occurred on the reversals (that is, the stars and triangles on Fig. 1) contained in the adaptive tracks is examined. The threshold estimates based on averaging the reversals are denoted as TH^{Rev}. The mean (open circle) and standard deviation (error bar) of $TH^{Rev} - TH_{0.5}^{Grand}$ across listener and frequency are shown in the second column of Figure 6.

A variation of the reversal strategy for obtaining thresholds was also explored by averaging signal levels of *all* trials from the first lower reversal to the last upper reversal, which is also the end of the adaptive track. The threshold estimates obtained this way are denoted as TH^{Avg}. The mean (open circle) and standard deviation (error bar) of $TH^{Avg} - TH_{0.5}^{Grand}$ across listener and frequency are shown in the third column of Figure 6. These two strategies yielded similar outcomes: the threshold estimates are below $TH_{0.5}^{Grand}$ by about 2 dB. Therefore, threshold estimates obtained using either strategy do not match $TH_{0.5}^{Grand}$, nor does their target performance match P_{Yes}=0.5.

The suspected reason for these mismatches was the uneven step size in the ascending runs (5 dB) and descending runs (10 dB). The greater step size in the descending run may be responsible for pushing the target performance below P_{Yes}=0.5. To test this idea, a Monte-Carlo simulation was completed in which the step size was set to 5 dB in both ascending and descending runs. The results are shown as red asterisks in column 2 (TH^{Rev}) and column 3 (TH^{Avg}) of Fig. 6. On average, the simulated thresholds using both strategies but with equal step size (5 dB) were reasonably close to $TH_{0.5}^{Grand}$ (Δ =0.4 dB). The standard deviation was 1.69 dB for TH^{Rev} and 1.62 dB for TH^{Avg}, both of which are relatively small. As part of the simulation, in a control

condition, adaptive tracks using the unequal step sizes (-10 dB and +5 dB) identical to that in the standard clinical procedure were also simulated, and computed thresholds using the two strategies. The results are presented as the red triangles in columns 2 and 3 of Fig. 6. The means of the simulated thresholds using unequal step sizes are nearly identical to those obtained from the human listeners (though the standard deviations are somewhat smaller). Overall, the simulation results suggest that both strategies yield accurate threshold estimates targeting $P_{Yes} = 0.5$ *only* when using an equal step size. This outcome is worth considering for clinical adaptation.

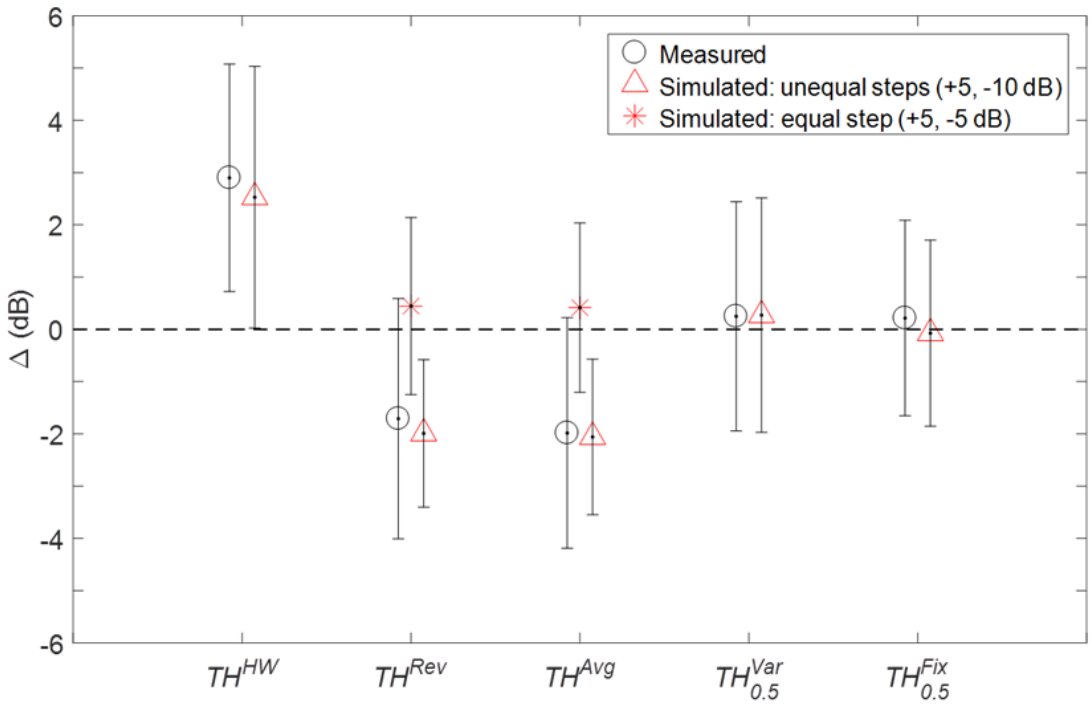

Figure 6. Measured (circles) and simulated (triangles and stars) threshold differences from the benchmark ($\Delta = TH - TH_{0.5}^{Grand}$) averaged across listener and frequency for five threshold procedures: TH^{HW} represents standard clinical procedure, TH^{Rev} the reversal-averaging procedure, TH^{Avg} the simple averaging procedure, $TH_{0.5}^{Var}$ the psychometric-function (PF) fitting procedure using variable σ, and $TH_{0.5}^{Fix}$ the psychometric-function (PF) fitting procedure using fixed σ. The error bars represent the standard deviation of Δ computed across listener and frequency for each procedure.

E.

Estimating threshold by fitting psychometric function to the trials in the adaptive tracks: $TH_{0.5}^{Var}$ and $TH_{0.5}^{Fix}$

Thresholds obtained using the standard clinical procedure correspond to a P_{Yes} value well above 0.5 ($P_{Yes}^{HW} = 0.932$). Thresholds obtained by averaging reversals in the adaptive tracks obtained using the standard clinical protocol correspond to a P_{Yes} value below 0.5. A principled way to obtain thresholds corresponding to $P_{Yes} = 0.5$ is by fitting psychometric functions. An earlier section described using PF^{Grand} to derive $TH_{0.5}^{Grand}$. However, PF^{Grand} is obtained with a relatively large number of trials, which are usually unavailable, and would be too lengthy, in a clinical setting. Therefore, fitting psychometric functions based on the relatively small number of trials used in the current protocol for obtaining the clinical thresholds was explored. Two fitting procedures were examined. First, the form of the psychometric function described by Eq. 1 as for PF^{Grand} is used, in which the spread (σ), and thus the slope, is allowed to vary as a free parameter. Second, a form of psychometric function is used in which the spread (σ) parameter is fixed and is set to the mean value estimated from all listeners ($\sigma = 2.20\ dB$). This fixed-σ fitting approach is simpler than the variable-σ fitting approach, thus is more easily adaptable in clinical settings. The steps for fitting the psychometric functions are identical to that used in fitting the PF^{Grand} thus have been described in the method section (D. Data analysis). Here, the threshold results from both fitting approaches are discussed.

The top panel of Fig. 7 shows the distributions (1 kHz in blue and 4 kHz in green) of the difference between the estimated threshold parameter using variable-σ fit, denoted as $TH_{0.5}^{Var}$, and the benchmark ($TH_{0.5}^{Grand}$) across listeners. The form of the distributions reasonably resembles a normal distribution. The mean Δ^{Var} across frequency is 0.25 dB; the standard deviation of Δ^{Var} is 2.19 dB. These values are also shown in column 4 of Fig. 6, marked as $TH_{0.5}^{Var}$. The

distributions of P_{Yes}, or the target performance corresponding to $TH_{0.5}^{Var}$ using PF^{grand} was also examined. To determine the P_{Yes} value, $TH_{0.5}^{Var}$ was substituted into the signal-level variable, i.e., $x = TH_{0.5}^{Var}$, in Eq. 1. For $TH_{0.5}^{Var}$, the overall mean value of P_{Yes} is 0.53. The small difference between $TH_{0.5}^{Var}$ and $TH_{0.5}^{Grand}$ (0.25 dB) indicates that thresholds estimated using variable- σ fitting approach are reasonably accurate.

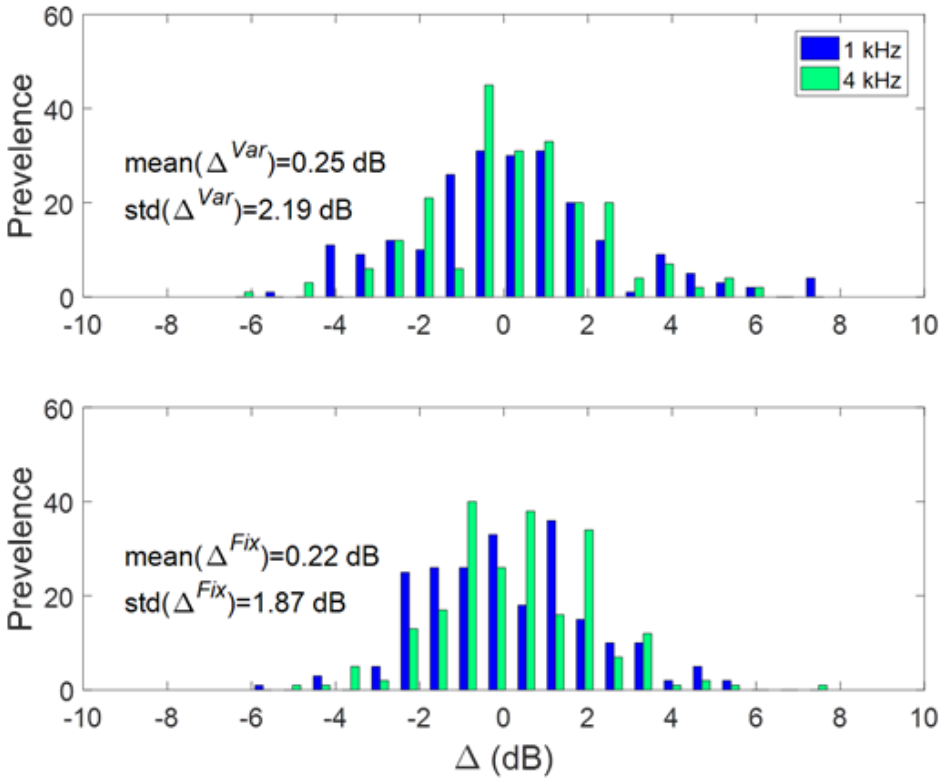

Figure 7. Distributions of threshold differences from the benchmark ($TH_{0.5}^{Grand}$) for fitting procedures using a variable σ of the psychometric function (top panel, $\Delta^{Var} = TH_{0.5}^{Var} - TH_{0.5}^{Grand}$) and a fixed σ (bottom panel, $\Delta^{Fix} = TH_{0.5}^{Fix} - TH_{0.5}^{Grand}$) at two frequencies (1 kHz in blue and 4 kHz in green). Each distribution is based on a total of 220 threshold estimates. The mean and standard deviation of Δ are computed across listener and frequency.

For fitting psychometric functions using fixed σ, the σ value was set to a constant of 2.20 dB, which is the mean σ computed across 44 grand psychometric functions (22 listeners x 2 frequencies). This mean σ is very close to the mean σ of 2.15 dB, which was computed based on the findings by Marshall & Jesteadt (1986). The bottom panel of Fig. 7 shows the distributions (1 kHz in blue and 4 kHz in green) of the difference between the estimated threshold parameter using fixed-σ fit, denoted as $TH_{0.5}^{Fix}$, and the benchmark ($TH_{0.5}^{Grand}$) across listeners. Similar to that in the top panel, the form of the distributions reasonably resembles a normal distribution. The mean Δ^{Fix} across frequency is 0.22 dB; the standard deviation of Δ^{Fix} is 1.87 dB. These values are also shown in column 5 of Fig. 6, marked as $TH_{0.5}^{Fix}$. The distributions of P_{Yes}, or the target performance corresponding to $TH_{0.5}^{Fix}$ using PF^{grand} were also exmained. To determine the P_{Yes} value, $TH_{0.5}^{Fix}$ was substituted into the signal-level variable, i.e., $x = TH_{0.5}^{Fix}$, in Eq. 1. For $TH_{0.5}^{Fix}$, the overall mean value of P_{Yes} is 0.52. Again, The small difference between $TH_{0.5}^{Fix}$ and $TH_{0.5}^{Grand}$ (0.22 dB) indicates that thresholds estimated using fixed-σ fitting approach are reasonably accurate. Given that the fixed-σ approach is simpler and that its accuracy is comparable to that of the variable-σ approach, using the fixed-σ approach is recommended. For clinicians who elect to apply this fixed-σ approach, the σ value of 2.2 dB is recommended because it is based on a large number of listeners from multiple studies.

IV. CLINICAL SIGNIFICANCE

Defining the target performance of an audiometric threshold seeking protocol relative to the psychometric function, which correlates the proportion of 'yes' responses to various signal levels, is necessary given that the target performance of the modified Hughson-Weslake protocol

is largely unknown. In general, the threshold generated from the current protocol was previously stated to correspond to the proportion of 'yes' responses, or P_{Yes} between 0.5 and 1. In this study, the target performance of the current clinical protocol, or P_{Yes}^{HW} was found to be 0.932 on average. This P_{Yes}^{HW} value is considerably higher than the classically defined threshold target performance of $P_{Yes} = 0.5$ in psychoacoustics, which corresponds to the point on the psychometric function with limited measurement variation (Green, 1993; Taylor and Creelman, 1967; Fisher, 1922). For this reason, measuring clinical thresholds at a target performance of $P_{Yes} = 0.5$ has the potential to minimize the variability of threshold estimates compared to threshold estimates at other target performance levels. As a result, extracting clinical thresholds with a target performance of $P_{Yes} = 0.5$ may minimize the normal test-retest variability for individual listeners, which has been found to be 10 dB using current the current clinical protol (Schmuziger et al., 2004)**.** Given that several auditory pathologies are in part identified based on very small changes in audiometric thresholds over time (e.g., ototoxicity, noise-induced hearing loss, tinnitus), having more reliable threshold estimates has the potential to yield more accurate and timely diagnosis and subsequent treatment plans. Due to the importance of obtaining accurate and reliable thresholds in dealing with these pathologies, defining thresholds at a target performance of $P_{Yes} = 0.5$ could be beneficial in clinical practice.

Three new adaptations of the current protocol could produce thresholds corresponding to a target performance of $P_{Yes} = 0.5$. The first adaptation is to subtract 3 dB from clinical thresholds. This simple correction value could be easily be applied to both manual and automated audiometric systems. In this study, clinical thresholds using the current protocol (TH^{HW}) were approximately 3 dB above the benchmark threshold ($TH_{0.5}^{Grand}$), defined as the

signal level corresponding to $P_{Yes} = 0.5$, derived from each listener's psychometric function. One potential limitation of this adaption is that the 3 dB correction may not be large enough for correcting HW thresholds because, as mentioned in an earlier section, actual HW thresholds are expected to be slightly higher than those obtained using the study protocol, which utilized a 'yes-no' procedure.

The second adaption is to average the trials in the adaptive track of the HW protocol and use an equal-step size of 5 dB for both ascending and descending trials. This adaptation differs from the first in that it is limited by the fact can only be easily adopted in automated audiometric systems, making it less practical for clinicians using manual audiometric systems. In this study, it was determined that thresholds targeting $P_{Yes} = 0.5$ could be obtained by averaging all signal levels at and beyond the first 'no' response in the clinical protocol. However, this accuracy was only achieved when an equal step size of 5 dB was used for both descending and ascending trials, while the current protocol utilizes an unequal step size of 10 dB for descending trials and 5 dB for ascending trials.

The third adaption is to fit a psychometric function to the few trials used in generating the HW threshold, with a fixed spread of 2.20 dB. Threshold estimates corresponding to a target performance of approximately $P_{Yes} = 0.5$ were derived by fitting a psychometric function to the few trials (n=10 on average) used in generating the clinical threshold, or TH^{HW} . The results were similar using both a spread parameter that was allowed to vary and one set to a fixed average value of 2.20 dB, however, the fixed spread approach is advantageous over the variable spread approach in that it is easier to implement in automated audiometric systems.

Although the results of this study are based on normal hearing listeners, it is expected that the results could be applied to listeners hearing impairment. One factor that may be different between normal hearing and hearing impaired listeners is the slope of the psychometric function. Arehart et al. (1990) reported slightly steeper psychometric function slopes for hearing impaired listeners as compared to listeners with normal hearing, meaning that the relationship between clinical threshold and the threshold with a target performance of $P_{Yes} = 0.5$ may be different for listeners with hearing impairment due to differences in the shape of their psychometric functions. However, this difference in slope between hearing impaired and normal hearing listeners has not been consistently observed across studies (Watson et al., 1972**;** Lecluyse & Meddis, 2009). Furthermore, even if the difference in slope exists between normal hearing and hearing impaired listeners, simulation results from this study demonstrated that when the assumed slope of the fitted function and the true slope were mismatched, the threshold was still approximately correct. This finding suggests that regardless of slope, fitting a psychometric function yields a threshold that corresponds to a target performance of $P_{Yes} = 0.5$. Finally, it would be interest to extend the current analysis to hearing impaired populations.

V. CONCLUSIONS

1. The standard clinical procedure can be characterized as that the clinical thresholds (TH^{HW}) correspond to an expected target performance of $P_{Yes}^{HW} = 0.932$. The clinical thresholds are about 3 dB above the benchmark threshold, $TH_{0.5}^{Grand}$. Given this result, a simple correction factor of -3 dB could be applied to clinical thresholds to arrive at a threshold that more closely corresponds to a target performance of $P_{Yes} = 0.5$

2. Thresholds obtained by averaging the reversals on the adaptive track used for estimating the clinical thresholds are about 2 dB below the benchmark threshold, $TH_{0.5}^{Grand}$. However, simulation results showed that, when a single step size of 5 dB is used, the thresholds obtained by averaging the reversals are within 0.4 dB of the benchmark threshold, $TH_{0.5}^{Grand}$. As changing step size is a simple modification of the standard clinical procedure, a recommendation is to use this modified procedure for obtaining accurate estimates of clinical thresholds in the future.
3. Thresholds obtained by fitting psychometric functions to the trials on the adaptive track used for estimating clinical thresholds are within 0.25 dB of the benchmark threshold, $TH_{0.5}^{Grand}$. The accuracy and consistency of the estimated thresholds are largely independent of whether the spread (σ) of the fitted psychometric functions is variable (i.e., a free parameter) or fixed at a constant of 2.20 dB. Because fitting PF with a fixed σ is simpler, an alternative recommendation is to use the fixed-σ procedure for obtaining accurate estimates of clinical thresholds using automated audiometric methods.

REFERENCES

American National Standards Institute. (2004). *Methods for manual pure-tone threshold Audiometry*, *ANSI, 3*, 21.

American Speech-Language-Hearing Association. (2005). Guidelines for manual pure-tone threshold Audiometry. Retrieved June 8, 2017, from http://www.asha.org/policy/GL2005-00014.htm.

Arehart, K. H., Burns, E. M., & Schlauch, R. S. (1990). A comparison of psychometric functions

for detection in normal-hearing and hearing-impaired listeners. *Journal of Speech, Language, and Hearing Research*, *33*(3), 433-439.

Carhart, R., & Jerger, J. (1959). Preferred method for clinical determination of pure-tone thresholds. *Journal of Speech & Hearing Disorders*, *24*(4), 330-345.

Dai, H., & Micheyl, C. (2011). Psychometric functions for pure-tone frequency discrimination. *The Journal of the Acoustical Society of America*, *130*(1), 263-272.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *222*, 309-368.

Green, D. M. (1993). A maximum likelihood method for estimating thresholds in a yes–no task. *The Journal of the Acoustical Society of America*, *93*(4), 2096-2105.

Jerlvall, L., & Arlinger, S. (1986). A comparison of 2-dB and 5-dB step size in pure-tone audiometry. *Scandinavian Audiology*, *15*(1), 51-56.

Lecluyse, W., & Meddis, R. (2009). A simple single-interval adaptive procedure for estimating thresholds in normal and impaired listeners. *The Journal of the Acoustical Society of America*, *126*(5), 2570-2579.

Leijon, A. (1992). Quantization error in clinical pure-tone audiometry. *Scandinavian audiology*, *21*(2), 103-108.

Levitt, H. C. C. H. (1971). Transformed up- down methods in psychoacoustics. *The Journal of*

the Acoustical society of America, *49*(2B), 467-477.

Mahomed, F., Swanepoel, D.W., Eikelboom, R. H., & Soer, M. (2013). Validity of automated threshold audiometry: a systematic review and meta-analysis. *Ear and hearing*, *34*(6), 745-752.

Marshall, L., & Jesteadt, W. (1986). Comparison of pure-tone audibility thresholds obtained with audiological and two-interval forced-choice procedures. *Journal of speech and hearing research*, *29*(1), 82-91.

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. *Journal of Speech, Language, and Hearing Research*, *39*(4), 687-696.

Schmuziger, N., Probst, R., & Smurzynski, J. (2004). Test-retest reliability of pure-tone thresholds from 0.5 to 16 kHz using Sennheiser HDA 200 and Etymotic Research ER-earphones. *Ear and hearing*, *25*(2), 127-132.

Song, X. D., Wallace, B. M., Gardner, J. R., Ledbetter, N. M., Weinberger, K. Q., & Barbour, D. L. (2015). Fast, continuous audiogram estimation using machine learning. *Ear and hearing*, *36*(6), e326-335.

Taylor, M., & Creelman, C. D. (1967). PEST: Efficient estimates on probability functions. *The Journal of the Acoustical Society of America*, *41*(4A), 782-787.

Watson, C. S., Franks, J. R., & Hood, D. C. (1972). Detection of tones in the absence of external

masking noise. I. Effects of signal intensity and signal frequency. *The Journal of the Acoustical Society of America*, *52*(2B), 633-643.