A simple single-interval adaptive procedure for estimating thresholds in normal and impaired listeners

Wendy Lecluyse and Ray Meddis
Department of Psychology, University of Essex, Wivenhoe Park, Colchester, Essex CO4 3SQ, United Kingdom

(Received 8 March 2009; revised 23 June 2009; accepted 1 September 2009)

This report presents a single-interval adaptive procedure for measuring thresholds in untrained normal and impaired listeners. The accuracy of the procedure is evaluated using Monte Carlo methods and human data allowing a method to be proposed for deciding in advance the number of trials required to achieve a specified level of accuracy. The number of trials depends on the slope of the psychometric function. The slope of the psychometric function is evaluated in normal and impaired listeners, and is found to give a useful guide to the required number of trials. The single-interval up/down procedure is subsequently compared with two other popular traditional methods (two-interval forced-choice, two-down/one-up and maximum-likelihood procedures) and is shown to yield similar thresholds and be more efficient.
 [DOI: 10.1121/1.3238248]

PACS number(s): 43.66.Yw [MW] Pages: 2570–2579

I. INTRODUCTION

Researchers who need to make many threshold measurements in both normal and impaired participant groups are confronted with the problem of finding a fast, participant-friendly, and reliable measurement procedure. They must choose between standard clinical methods based on single-interval "yes/no" procedures and those used in psychoacoustic research laboratories based on a multiple-interval forced-choice approach. Clinical methods, such as the modified Hughson–Westlake procedure (Carhart and Jerger, 1959), have been optimized for speed and patient acceptability while laboratory methods aim for greater accuracy and theoretical rigor. The former are simple to administer, require little patient training, can be easily automated, and involve only a small number of trials.

Unfortunately, standard clinical procedures are not acceptable to most of the laboratory-based scientific community because they are believed to overestimate thresholds and fail to accommodate differences in response bias (Marshall and Jesteadt, 1986). This is a problem for the clinical researcher who needs to obtain thresholds that are meaningful within a wider research context where the choice of method is overwhelmingly a multiple-interval, forced-choice approach. Standard laboratory procedures, on the other hand, are more complicated, often require considerable training, and typically need many trials. Opting for the apparently more rigorous laboratory procedure comes at a very high price. The large number of trials (often around 50–60) is a major disincentive. Also, when using these procedures, the patient must choose one of two temporal windows where one is occupied by a stimulus and the other is empty. Patients who may be elderly or have a lower educational level may experience considerable difficulty with this method. The problem is particularly pressing when the stimulus is below threshold and neither window is a straightforward choice. Here participants are required to guess. This is not an intuitively obvious method for measuring anything and can weaken participant confidence in the procedure. For this group yes/no procedures might be more suitable.

This report, therefore, addresses two issues concerning the use of adaptive, single-interval (yes/no) methods for measuring absolute thresholds for tones in quiet. Do they give different results from multiple-interval forced-choice methods and are they more or less efficient? There is already a substantial literature on the statistical aspects of single-interval methods (e.g., Brownlee *et al.*, 1953; Choi, 1990; Cornsweet, 1962; Dixon and Mood, 1948; Levitt, 1971; von Bekesy, 1947) that focuses on appropriate procedures for finding the mean of an underlying psychometric function. However, the subsequent widespread adoption of a signal-detection approach to the nature of threshold (Green and Swets, 1966; Swets, 1964) encouraged the adoption of "objective techniques" such as multiple-interval, forced-choice where the listener's response could be classified as "right" or "wrong." The signal-detection approach seeks to decompose the psychometric function into two components, sensitivity and criterion, claiming that sensitivity is the relevant measure when assessing threshold.

Later, Green (1993) sought to renew interest in the more subjective yes/no approach by emphasizing how efficient it could be. Leek *et al.* (2000), more recently, highlighted the benefits of using this approach, particularly in a clinical context. However, even if it could be shown that adaptive single-interval methods are more efficient, the suspicion remains that subject caution may contaminate the measurements and render them valueless. Kaernbach (1990) previously sought to reconcile the two approaches by introducing trials when no stimulus was presented in a single-interval procedure. This permitted a full assessment of hit rates and false-alarm rates as required by signal-detection theory. This procedure included a sophisticated method for the selection of stimulus levels that promoted substantial efficiencies. Nevertheless, these "subjective" methods remain minority approaches.

The debate has been strongly influenced by Marshall and Jesteadt's (1986) report showing that standard clinical methods for assessing absolute threshold gave overestimates when compared with a two-interval forced-choice (2IFC) methodology. This is often construed as a vindication of the "objective" approach. However, this is a misreading of their results. Marshall and Jesteadt (1986) showed, in the same report, that single-interval methods *when appropriately applied* can yield thresholds similar to those obtained using the 2IFC approach. They attributed the overestimates obtained using the standard clinical methodology to simple, easily remedied, procedural problems.

Marshall and Jesteadt (1986) identified various procedural deficiencies in the clinical approach. The first problem is purely statistical and concerns the clinical practice of choosing the "lowest stimulus level that is reliably heard by the patient." This automatically biases the threshold estimate to be above the 50%-point of the psychometric function by an amount that depends on the step size. A large step size such as the 5-dB step size used in their clinical procedure exaggerates the effect compared to the smaller 2-dB step size used in their comparison 2IFC-procedure. The second factor is psychological and concerns the timing of the presentation of the test stimulus. For 2IFC, the stimulus timing is precisely locked to a visual cue while the clinical procedure has variable timing and no visual cue. When Marshall and Jesteadt (1986) equated these factors using a computer-controlled yes/no procedure, the difference between the estimated thresholds using the two procedures was much smaller. The investigations reported below used precisely timed stimuli with an audible cue and small step sizes in a one-down, one-up procedure in an attempt to minimize the problems identified in their study.

Green (1993) recommended a new method for specifying the sequence of stimulus levels presented to the listener. He suggested that stimulus levels should always be presented at the estimated "sweet point" (most informative level) of the underlying psychometric function, which, in the absence of guessing, would be its 50%-point. This estimate would be updated after each trial by fitting the accumulated data to the best-fit logistic function using a maximum-likelihood (ML) procedure. He found that this procedure was "moderately efficient." This project used Green's (1993) procedure as its starting point. However, it was found that it did not always produce rapid convergence on the true mean and could produce misleading estimates. These flawed estimates had been noticed before (Green, 1995; Leek *et al.*, 2000) but had been identified as secondary consequences of listener lapses of attention. Computer simulations described below, however, show that these errors are intrinsic to Green's (1993) method. As a consequence, it was necessary to evaluate the efficiency of the simpler one-up, one-down rule.

Leek *et al.* (2000) used Green's (1993) method in an extensive study using both normal and clinical populations. Their results indicate that the method is generally acceptable to untrained listeners and gave reliable threshold estimates based on only 24 trials that were comparable with 2IFC-methods. They used catch trials to monitor the false-alarm rate of their listeners although they found that false-alarms were rare (around 5%). The use of catch trials is an import feature of the procedure to be described below where listeners are constrained to keep false-alarms to a minimum and trials containing false-alarms are rejected. While Green (1993) offered a "correction for guessing," an estimation of the guessing rate is possible only after more catch trials than are feasible in practice. This problem will be minimized by keeping rates as close to zero as possible.

The small number of trials used in the study of Leek *et al.* (2000) raises the question of how to estimate the number of trials necessary to achieve a desired precision of threshold estimation. Different studies have different requirements, and it should be possible to adjust the number of trials to take this into account. Computer simulations of the single-interval up/down (SIUD)-procedure using Monte Carlo methods will be used below to give insight into this issue. A simple formula for specifying the required number of trials will be derived on the assumption that the underlying psychometric function takes the form of a logistic curve with a known slope. The results indicate that a surprisingly small number of trials will be necessary in many situations, particularly for hearing impaired listeners with steep psychometric functions.

II. THE SIUD-PROCEDURE

A. Procedure

The SIUD-procedure for measuring absolute thresholds is based on a simple yes/no task. A single stimulus is presented to the participant who responds "yes" or "no" according to whether or not the stimulus was heard. The participant responds by means of a button box linked to a visual display. Part of this display is made invisible when a stimulus is presented thus marking the observation interval. The level of the stimulus is changed from trial to trial using a one-down, one-up adaptive procedure. If the participant says yes, the stimulus level is decreased by a fixed amount. If he says no, the level is increased by the same amount.

The run starts with an initial phase where the stimulus level is set at supra-threshold level (generating a guaranteed yes-response) and is adjusted using a large step size until the first no-response. This initial step size is typically set at 10 dB. The start level is different in each run and randomly located in a range ± 5 dB relative to the nominal start value.

After the first no-response, the stimulus level is set to the mid-point between the previous two levels, and a small step size, say, 2 dB, is used from this point on. The run then continues for a fixed number of trials counting from the trial immediately before the first no-response ("trial 1" in Fig. 1). An illustration of a typical threshold track is shown in Fig. 1.

The choice of a 2-dB small step size was guided by computer simulations (not shown) comparing various step sizes. The 2-dB step gave the lowest variance of the threshold estimates on a range of psychometric slopes (k varied between 0.25 and 1) and trial numbers. In other words, 2-dB steps provide satisfactory reliability in long or short threshold runs. Also, a 2-dB step is commonly used in adaptive

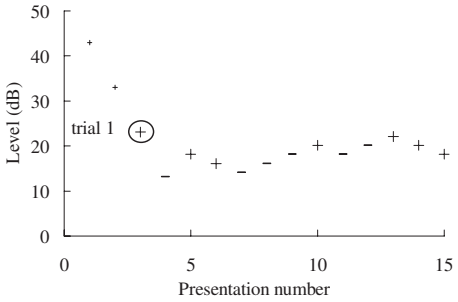

FIG. 1. Illustration of a threshold run. Plus-signs represent the yes-responses, whereas minus-signs represent the no-responses. A large step size of 10 dB is used until the first reversal. After the first reversal the stimulus level is set to the mid-point between the two previous levels and a small step size of 2 dB is used from this point onwards. The trial count starts from the presentation before the first reversal (circled). The responses preceding this point are not included in the threshold estimates (small markers).

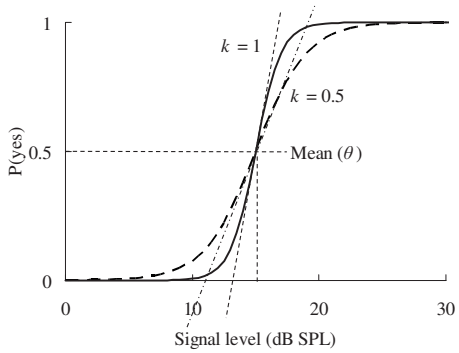

FIG. 2. Psychometric functions with two different slope parameters, $k=1$ (thick continuous line) and $k=0.5$ (thick dashed line). The threshold θ is defined by the mean of the function where the proportion of yes-responses is 0.5.

procedures. Dixon and Mood (1948) suggested a step size close to the standard deviation of the underlying psychometric function.

B. Catch trials

Catch trials are trials where no stimulus is presented and the participant is expected to say no. Catch trials are primarily intended to identify situations where the participant is either not attending or adopting some strategy that is inconsistent with the aims of the investigation. A catch trial is always presented on the second trial in a run to provide a reminder of how "no-stimulus" sounds. 20% of successive trials are catch trials, presented at random without constraint. If the participant is "caught out," the run is stopped and restarted; possibly after resting the participant and giving further instructions. Participants are encouraged not to guess but to report hearing a tone only when they are confident that they have heard it. The restart process following the rare false-alarms acts as an additional incentive for patients to make only confident judgments.

C. Threshold estimation

The threshold is estimated at the end of the run. All stimulus levels from trial 1 (defined above and see Fig. 1) onwards are included in the estimate of the threshold. These are indicated by large markers in Fig. 1. Earlier trials are discarded (small markers in Fig. 1). The threshold can be estimated by using the mean of all these levels (Dixon and Mood, 1948). Alternatively, Cornsweet (1962) suggested that the median level could be used, as this will reduce the effect of any extreme values.

However, in what follows, we are interested in estimating the *accuracy* of our threshold estimate as a function of the number of trials. In this case, it is expedient to estimate the threshold after each trial. We do this by assuming that the participant's decisions close to threshold are approximated by an underlying psychometric function of the form

$$p(L) = 1/(1 + \exp(-k(L - \theta))), \quad (1)$$

where $p(L)$ is the proportion of yes-responses, L is the level of the stimulus [decibel sound pressure level (SPL)], k is a slope parameter, and θ (decibel SPL) is the threshold to be estimated. The threshold θ is the level of the stimulus at which the proportion of yes-responses is 0.5. A psychometric function as described in Eq. (1) is fitted to the responses using a least-squares, best-fit procedure, with θ and k as free parameters.

Figure 2 illustrates this function for two values of slope ($k=0.5$ and $k=1.0$). The value of k typically is close to 0.5 for normal hearing (Green, 1993, using data from Watson *et al.*, 1972). However, Arehart *et al.* (1990) and Carlyon *et al.* (1990) using a d' statistic showed that the slope of the psychometric function can be steeper than normal for patients with a moderate hearing impairment. The slope of the function influences the variability of the threshold estimate. When the slope is steep (continuous line, $k=1.0$), the transition (across level) from yes to no occurs over a narrower range of levels, and fewer trials will be required to estimate the threshold with a given degree of accuracy.

III. EVALUATION I: COMPUTER SIMULATIONS

The accuracy of a threshold estimate improves as the length of a threshold run is increased. The trade-off between accuracy and speed needs to be considered when setting up a measurement protocol, and a compromise is always required. It would therefore be helpful to be able to predict the number of trials needed to obtain a given level of accuracy. In this section "Monte Carlo" computer simulations will be used to assess the improvement in accuracy associated with increasing the number of trials when using the SIUD-procedure. It will be shown that the variability of the estimates can be approximated by a simple mathematical formula and that this formula can be used to specify the number of trials that will be needed to achieve a given level of accuracy.

A. Method

The listener's response was simulated by assuming that it is determined by the psychometric function in Eq. (1). The threshold parameter θ was fixed at 15 dB SPL and the slope value k fixed at 0.5 for the first simulations and at 1.0 for a second evaluation. For each trial, the stimulus level L was used in Eq. (1) to compute the probability p that a yes-response will occur. A uniformly distributed, random number

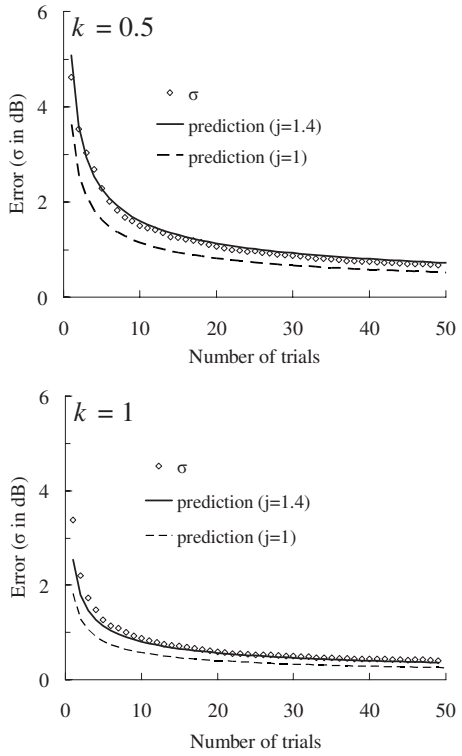

FIG. 3. Standard deviation, σ, of threshold estimates, θ, as a function of the number of trials in a Monte Carlo computer simulation. Top panel: standard deviations (open diamonds) are based on 1000 threshold estimates assuming a psychometric slope k of 0.5. The predictive functions of σ are generated using Eq. (2) with a standard slope value $k=0.5$ and an adjustment factor, $j=1$ (dashed line) or $j=1.4$ (continuous line). Bottom panel: same as top panel but assuming a psychometric slope k of 1. The prediction is generated using Eq. (2) with $k=1$ and $j=1$ (dashed line) or $j=1.4$(continuous line).

between 0 and 1 was then generated to determine the response. If the random number was less than $p(L)$, a yes-response was judged to have occurred on that trial; otherwise, a no-response was assumed. Catch trials were not included in the computer simulations.

The SIUD-procedure was followed exactly as described in Sec. II. The initial starting level was randomly set in a range ± 5 dB relative to a nominal start value of 40 dB SPL. This value was chosen to be above the psychometric function asymptote, guaranteeing only yes-responses on the first presentation. The simulation consisted of 1000 runs. Each run continued for 50 trials. This generated 1000 threshold estimates updated at each of the 50 trial times.

B. Results

The accuracy of the estimates was assessed in terms of the unbiased standard deviations of the thresholds estimated about the true value of θ (i.e., 15 dB SPL). Standard deviations were calculated after each trial across the 1000 runs. The individual data points (open diamonds) in Fig. 3 show how the standard deviation, σ, of the threshold estimate, θ, decreases (i.e., accuracy improves) as the run progresses.

When assuming a slope of 0.5, the accuracy is better than 2 dB after only 10 trials, and after 30 trials, the accuracy is better than 1 dB. The second set of simulations, using a slope parameter $k=1$, shows standard deviations that are lower compared to the standard deviations for a slope of 0.5, and an accuracy of 1 dB is found after less than ten trials. Threshold estimates were approximately normally distributed, and accuracy can be defined to mean that 68.3% of the possible values of the true mean lie within $\pm 1\sigma$ of the true threshold value. No bias in the threshold estimates was observed.

C. Predicting the accuracy of threshold estimates

The formula given in Eq. (1) to describe the hypothesized psychometric function is the logistic function, the cumulative distribution function of the logistic distribution (Hastings and Peacock, 1975). We can therefore use the standard equation for the variance of the logistic probability density function ($\pi^2/3k^2$) to provide an approximate estimate of the reliability of our threshold measurements

$$\sigma = \frac{j\pi}{k\sqrt{3}\sqrt{n}}, \tag{2}$$

where σ is the standard deviation of the threshold estimates, k is the slope parameter of the psychometric function, n is the number of trials in a single threshold run, and j is an adjustment factor to improve the approximation. The fit to the data when $j=1$ is shown in Fig. 3 as a dashed line. It has the correct shape but it underestimates the error.

This underestimation is a consequence of the fact that the stimulus levels are not statistically independent. In our case, each presentation level is related to the previous level by the up/down rule. Some correction is, therefore, required. It is difficult to find a correction factor based on an analytical solution (see Choi, 1990, for a fuller explanation), but a numerical approach based on the Monte Carlo simulations suggests a correction factor j of 1.4. This is illustrated as the continuous line in Fig. 3 for psychometric slopes $k=0.5$ and $k=1.0$. The rms errors of the fit are 0.10 and 0.15 dB, respectively.

D. How many trials?

The number of trials, n, needed to acquire a certain level of accuracy can be calculated by rearranging Eq. (2) as follows:

$$n = \frac{6.4}{k^2\sigma^2}, \tag{3}$$

where σ is the required level of accuracy and a correction factor $j=1.4$ is assumed. Note that the number of trials, n, does not include catch trials or trials in the initial stage (see Fig. 1).

This application of Eq. (2) is considered an important addition to the SIUD-procedure since it allows researchers to predict the number of trials required for a given level of accuracy of the threshold estimates.

In the standard case ($k=0.5$), Eq. (3) reduces to $n=26/\sigma^2$. We can see that in this standard situation a required accuracy (in σ) of 1 dB would indicate the use of 26 trials while an accuracy of 2 dB would indicate a requirement of only 7 trials per threshold run. When the psychometric slope k is 1.0 (for example, for some participants with impaired

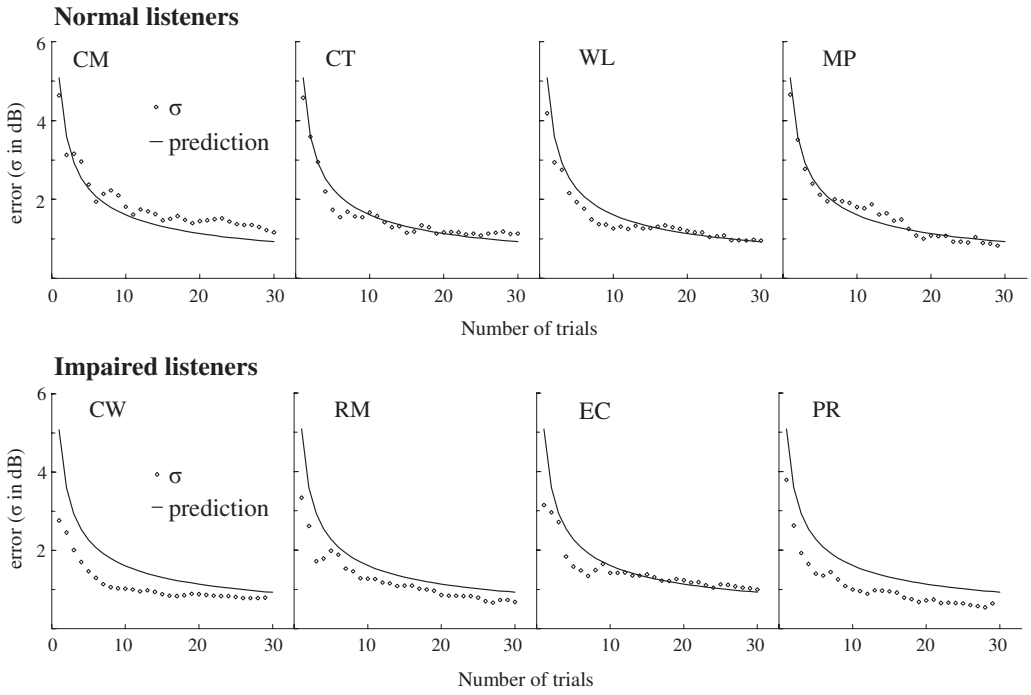

FIG. 4. Standard deviation, σ, of threshold estimates as a function of the number of trials, for four normal listeners (top row) and four impaired listeners (bottom row). Standard deviations (open diamonds) are based on 20 threshold estimates. Continuous lines represent the prediction of the σ values using Eq. (2) with a standard slope value $k=0.5$ and a j-value of 1.4.

hearing), $n=6.4/\sigma^2$ and an accuracy of 1 dB can be achieved with only seven trials. This required number of trials is four times less than for normal listeners.

IV. EVALUTION II: HUMAN LISTENERS

The question remains as to whether these computer simulations of a mathematical abstraction do indeed represent what happens when a human listener is seated in a booth making the same kind of decisions. The predictive value of Eq. (2) was therefore evaluated using human data.

A. Method

Absolute thresholds for a pure tone were measured in four normal and four impaired listeners using the SIUD-procedure described above. For the normal listeners, the stimulus was generally a 2-kHz, 100-ms tone. For one normal listener (MP) the tone frequency was 1 kHz (as a result of an operator error). The participants were aged between 21 and 32 years and have normal audiograms.

The impaired listeners were tested using a 1-kHz, 100-ms stimulus. They were aged between 58 and 76 years. They all had raised thresholds over a large range of frequencies. Participant RM has normal thresholds at the test frequency, but he is considered an impaired listener since he has a sloping loss from 2 kHz onwards. Each participant was tested for 20 threshold runs. Each run consisted of 30 trials. The data were collected in a single session with a 2 min break after every third run. The step size was set at 2 dB and the initial step size was set at 10 dB. Thresholds were estimated after each trial. The adaptive procedure was the same as for the numerical simulations with the addition of 20% catch trials that are not included in the analyses below.

Listeners were seated in a sound-proof booth and stimuli were presented through circumaural headphones (Sennheiser HD600) linked directly to the computer sound card (Audiophile 2496, 24-bit, 96 000-Hz sampling rate). Responses (yes/no) were made using a button box. A monitor in front of the participant showed a display of the button box. While the stimulus was presented displayed button symbols disappeared. Immediately after stimulus presentation the buttons reappeared on the screen signaling that a response was required.

A cue tone at the same frequency and with the same duration as the stimulus tone but 10 dB more intense preceded the stimulus tone by 0.5 s. The cue/stimulus pair was initiated under computer control 0.5 s after the listener's previous response. A raised cosine ramp of 4 ms was applied to both cue and target sounds. When a catch trial occurred, only the cue was presented at the level appropriate if the trial were not a catch trial.

B. Results

The accuracy of the estimates was assessed in terms of the standard deviations, σ, of the threshold estimates θ, after each trial across the 20 runs. These are shown in Fig. 4 for the normal listeners (top row) and for the impaired listeners (bottom row).

The continuous lines in Fig. 4 (top row) show the predicted standard deviations [Eq. (2)] for the normal listeners when assuming a slope parameter $k=0.5$ and applying the correction factor $j=1.4$ (see above). The predictive function fits the normal data with an average rms error across listeners of 0.24 dB.

The standard deviations found in the impaired group

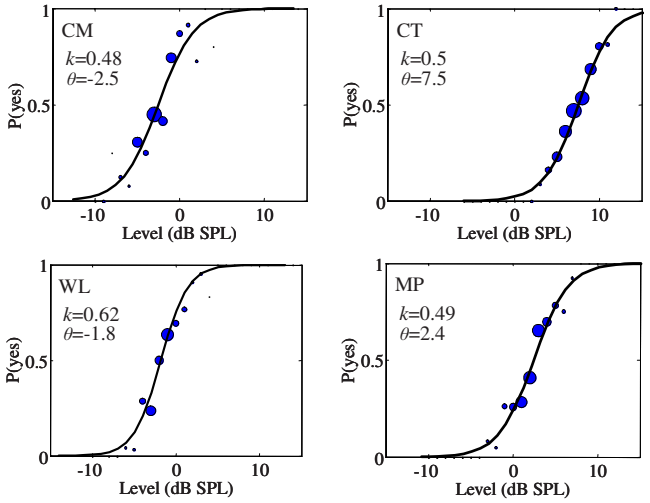

FIG. 5. (Color online) Psychometric function for 4 normal listeners based on 600 yes-/no-responses. The size of the circles represents the number of responses contributing to that point of the psychometric function. The full line represents the best-fit logistic function to the responses. The inset shows the slope parameter k and threshold θ (dB SPL) associated with the best fit. The stimulus was a 100-ms tone with tone frequency 2 kHz for listeners CM, CT, and WL, and 1 kHz for listener MP.

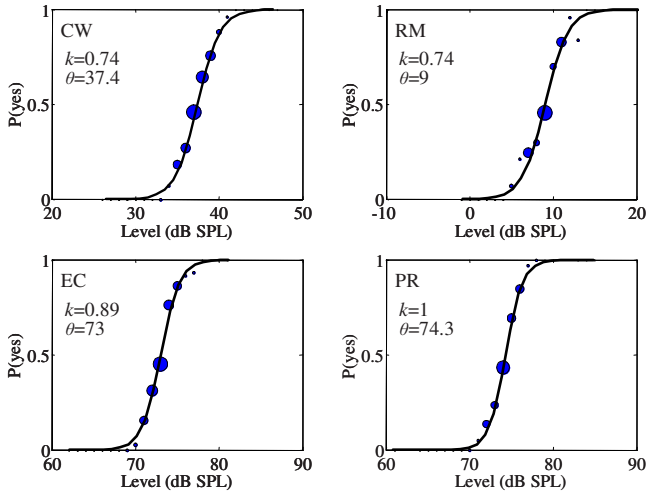

FIG. 6. (Color online) Psychometric functions for four impaired listeners (see Fig. 5 for more details). The stimulus was a 100-ms tone with tone frequency 1 kHz.

were lower than the normal group (Fig. 4, bottom row). The predictive function used for the normal listeners (assuming a slope parameter $k=0.5$ and a correction factor j of 1.4) was also fitted to the data of the impaired listeners. In almost all cases this results in a conservative prediction of the error of the threshold estimates, and this is consistent with the possibility that the impaired listeners have steeper psychometric functions ($k>0.5$).

C. Conclusion

Equation (2) offers a general guide to the accuracy of the SIUD-method as a function of number of trials in a threshold run, particularly if the slope parameter of the psychometric function k is known. If the slope is not known, then a slope parameter $k=0.5$ (and correction factor $j=1.4$) gives a useful, if sometimes conservative, estimate. Equation (3) can be used to decide how many trials are needed to achieve a required level of accuracy.

V. PSYCHOMETRIC SLOPES OF NORMAL AND IMPAIRED LISTENERS

The computer simulations described in Sec. III assumed that the psychometric slope k was either 0.5 or 1. To check these assumptions, the behavioral data collected in the previous experiment (Sec. IV) were reanalyzed to establish appropriate values for the slope.

A. Method

The 600 yes-/no-responses at various signal levels, obtained when measuring the thresholds described in the previous experiment (Sec. IV), were used to generate a psychometric function for each normal and impaired listener (Figs. 5 and 6). Responses were aggregated into bins of 1-dB width, and the proportion of yes-responses in each bin was calculated. The relative size of the circles is used to indicate the relative number of responses at each stimulus level. The best-fit function [Eq. (1)] is shown as the continuous line through the data points. The k-value and threshold θ, associated with this best fit, are shown in the insets.

B. Results

Figure 5 shows the psychometric functions of the four normal listeners. The slope estimates k range from 0.48 for listener CM to 0.62 for listener WL. The psychometric functions for the four impaired listeners are shown in Fig. 6. Their slopes (0.74, 0.74, 0.89, and 1.00) were considerably steeper. The observation of steeper slopes is in line with other studies (Arehart *et al.*, 1990; Carlyon *et al.*, 1990).

VI. COMPARISON WITH OTHER PROCEDURES

The use of the SIUD-method can only be recommended if it is as accurate as other methods currently used in research laboratories. The SIUD-procedure was therefore compared with two other procedures in common use: (1) 2IFC, two-down/one-up method described by Levitt (1971) and (2) the single-interval, ML-method of Green (1993).

A. Computer simulations

1. Method

Monte Carlo simulations were made in the same manner as described above. Numerical simulations assumed an underlying psychometric function given in Eq. (1) where the true threshold θ was fixed at 15 dB SPL, and the slope of the psychometric function k was fixed at 0.5. Threshold estimates were simulated over 500 runs for all three procedures. The step size was always 2 dB except for the initial step size (10 dB).

In all conditions, the initial starting level was randomly set in a range ± 5 dB relative to a nominal start value of 40 dB SPL. Catch trials were not used in the SIUD-condition. A threshold estimate was computed at the end of each run and the standard deviation computed over the 500 runs.

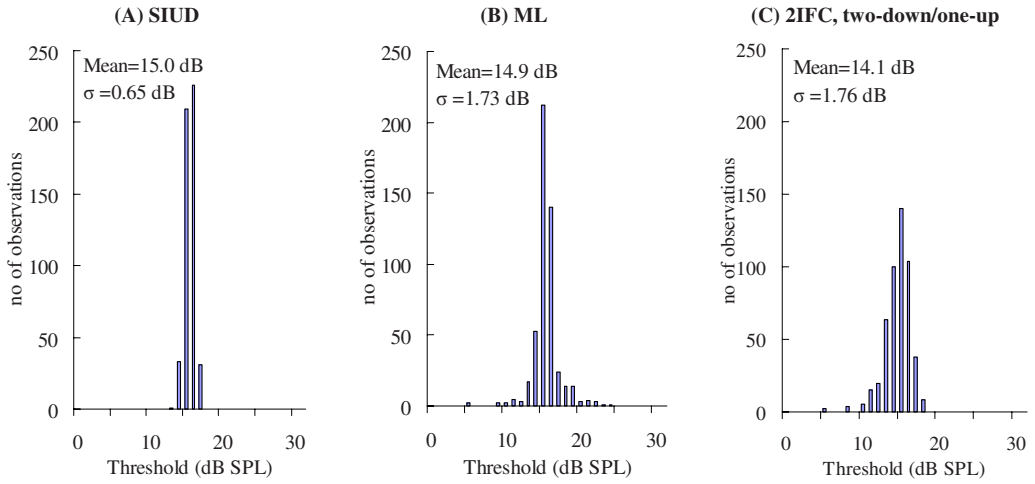

FIG. 7. (Color online) Histogram for threshold estimates using (a) SIUD, (b) ML, and (c) 2IFC, two-down/one-up. 500 threshold estimates were obtained per condition. Threshold runs were terminated after 50 trials for the ML- and SIUD-condition (not including the initial trials) and after 14 reversals in the 2IFC-condition. The inset shows the mean threshold estimate and the standard deviations (σ) for each condition.

(a) *2IFC.* The thresholds for the 2IFC-procedure were measured using the two-down/one-up adaptive procedure described by Levitt (1971). The signal level was initially adjusted in 10-dB steps. After two reversals, the step size was reduced to 2 dB. Each run was terminated after 14 reversals. The threshold was computed by averaging the levels of the last 12 reversals. The 2IFC-condition was the first to be simulated. It was found that the average number of trials required for 14 reversals was 50. The other two conditions were then simulated using this number of trials so that a fair comparison of accuracy could be made.

(b) *ML.* The ML-procedure followed as closely as possible the protocol described by Green (1993). The initial step size (10 dB) was used to adjust the stimulus level up to the first reversal. After that, the ML-procedure was used to set the new stimulus level after each trial. The best-fit psychometric function [see Eq. (1)] was obtained using the ML-method described by Green (1993) on the basis of all observations up to that point. The 50%-point of the function was then used to determine the level of the next stimulus to be presented. Each run continued for 50 trials (not including the initial trials). The threshold estimate for the run was taken to be the 50%-point of the final best-fit psychometric function. The false-alarm rate was fixed at zero.

(c) *SIUD.* The SIUD-procedure was exactly as described above using 50 trials per threshold run.

2. Results

The distribution of the threshold estimates across 500 runs for all three procedures (SIUD, ML, and 2IFC) is given in Fig. 7. The mean threshold estimates are 15.0, 14.9, and 14.1 dB SPL for SIUD, ML, and 2IFC, respectively. The mean standard deviations σ are 0.65, 1.73, and 1.76 dB for SIUD, ML, and 2IFC, respectively.

The threshold estimate for the 2IFC simulation (14.1 dB SPL) is below the true threshold (15 dB). This is partly because the 2IFC-procedure estimates the 70.7%-point of a psychometric function where the minimum hit rate is 50%. The equivalent point on the true psychometric function (ranging from 0% to 100% correct) is 41.4%, i.e., an underestimate of the true mean (see Fig. 8). For a slope k of 0.5, an adjustment of +0.7 dB is necessary to establish the level at the 50%-point of the underlying (yes/no) psychometric function. This adjustment was previously suggested by Leek *et al.* (2000). The new threshold estimate of 14.8 dB SPL is closer to, but still an underestimate of, the true threshold.

Our main concern here, however, is *reliability* as represented by the standard deviation of the threshold estimates over many runs. The spread of threshold estimates in the SIUD-condition is considerably less than the spread in the ML-condition or the 2IFC-condition.

The lower accuracy of the ML-procedure is, at least partly, explained by a number of extreme threshold estimates both above and below the true threshold. These can be seen as unexpected outliers in the distribution of threshold estimates in Fig. 7(B). To investigate the matter further, a limited set of tracks of the threshold estimates is considered

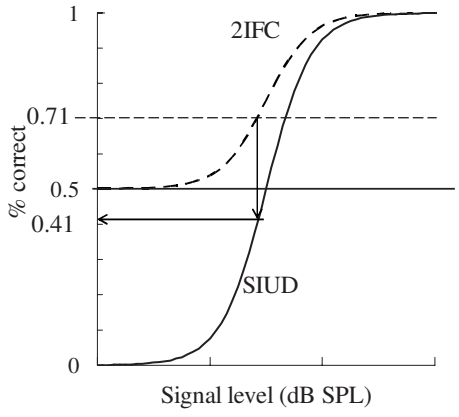

FIG. 8. Schematic representation of the psychometric function for a 2IFC-condition (dashed line) ranging from 50% to 100% correct responses and the psychometric function generated in a single-interval procedure (continuous line) ranging from 0% to 100% correct responses. The 70.7%-point on the 2IFC-psychometric function corresponds to the 41.4%-point on the SIUD-psychometric function.

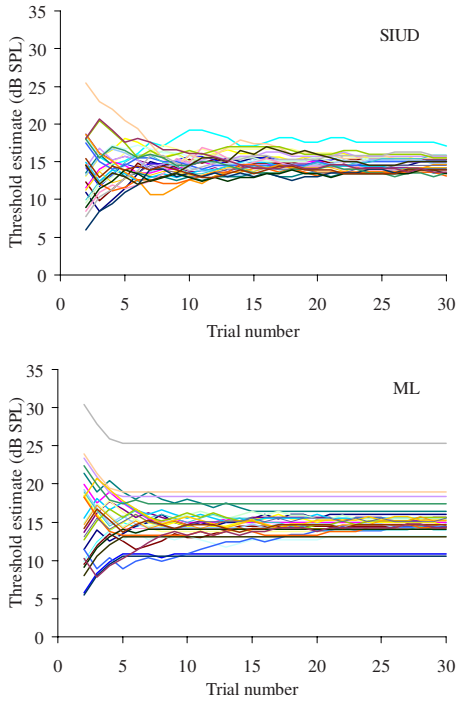

FIG. 9. (Color online) Threshold estimate as a function of trial number in a SIUD-procedure (top panel) and a ML-procedure (bottom panel) obtained using Monte Carlo simulations. Threshold tracks associated with 30 threshold runs are pictured for each condition. Each run consists of 30 trials.

individually. These are shown in Fig. 9. After each observation, the ML-procedure estimates the 50%-point of the psychometric function. The bottom panel of Fig. 9 shows that these estimates normally converge quickly on a value close to the true mean (15 dB) of the function. However, some tracks settle quickly and permanently on a *false estimate*. These rogue estimates inflate the overall standard deviation of the threshold estimation procedure.

The top panel of Fig. 9 shows equivalent tracks for the SIUD-condition. In this case the tracks all appear to be converging on the true threshold. This agrees with Fig. 7(A) where no outliers in the distribution are present.

The relatively widespread of estimates in the 2IFC-procedure may also be partly explained by outliers. For example, the distribution of estimates in Fig. 7(C) shows two very low estimates. Moreover, the distribution of estimates is asymmetric with the long tail toward values below the mean. This is almost certainly attributable to random adjustments in level when the stimulus is below threshold and the patient is required to guess. During this guessing phase, the patient may guess correctly and the stimulus level reduces even further below threshold. At this time there is a chance that a random walk be initiated with peaks and troughs below threshold levels. In this case the estimated threshold will be an underestimate of the true threshold. A presentation of the individual threshold tracks in the 2IFC-condition is not included in Fig. 9 since the nature of this procedure does not allow for threshold estimates to be made after each trial.

B. Human data

As a final reassurance concerning the reliability of the SIUD-procedure, we compared the thresholds measured using both SIUD and 2IFC of a group of student volunteers with no prior experience of audiometric methods.

TABLE I. Individual thresholds (in dB SPL) and standard deviations, σ, for the SIUD and 2IFC-condition. In the 2IFC-condition the adjusted thresholds are between parentheses (targeting the 50%-point on a psychometric function ranging from 0% to 100% correct). The bottom row shows the average thresholds and average standard deviations across listeners for each condition.

	SIUD		2IFC	
Listener	Threshold	σ	Threshold	σ
S1	16.6	2.0	16.3 (17.0)	4.6
S2	8.2	1.9	6.4 (7.1)	1.7
S3	10.9	1.8	10.0 (10.7)	1.9
S4	−2.4	2.3	−1.1 (−0.4)	2.2
S5	11.7	1.9	5.0 (5.7)	6.7
S6	5.4	1.3	3.1 (3.8)	3.1
S7	9.0	2.2	8.4 (9.1)	1.6
S8	8.9	1.9	9.2 (9.9)	4.3
S9	5.2	0.7	4.2 (4.9)	4.4
Average	8.2	1.8	6.9 (7.6)	3.4

1. Method

Nine listeners were used with audiometric thresholds within the normal range. The procedures used are exactly as described above. Five threshold estimations were made using both the SIUD and 2IFC-procedure. The SIUD-condition used only ten trials per run. The 2IFC used eight reversals and the thresholds were estimated as the mean of the last six reversal levels. This required an average of 32 trials per run.

2. Results

Table I shows the average threshold and associated standard deviation for each condition for each individual listener. The average threshold across all listeners for the SIUD-method was 8.2 dB SPL. The average for the 2IFC-thresholds was 6.9 dB SPL. After applying the 50%-adjustment suggested above (Sec. VI A) the mean 2IFC-threshold (across listeners) is 7.6 dB SPL. These adjustments are shown between parentheses in Table I alongside the initial 2IFC-thresholds.

The average thresholds per listener are similar for both conditions. Although the average threshold across listeners is slightly higher for the SIUD-condition compared to the 2IFC-condition, this was the case for only five out of nine listeners. This suggests that there is no consistent pattern for SIUD-threshold to be higher than 2IFC-thresholds. The standard deviation of the threshold estimates, however, is similar or substantially higher for the 2IFC-condition (average $\sigma = 3.4$) compared to the SIUD-condition (average $\sigma = 1.8$). These findings are consistent with the Monte Carlo simulations presented above (Sec. VI A). The average standard deviation across listeners in the SIUD-condition is half the standard deviation in the 2IFC-condition despite the fact that almost three times fewer trials were used in the SIUD-condition.

The average number of catch trials presented per threshold run in the SIUD-condition was 3.6. A caught-out incident occurs when a listener reports to have heard the stimulus when no stimulus was presented. The rate of caught-out incidents had an overall average of 2.1% of the catch trials. Six of the nine participants had a zero caught-out rate. The average in the remaining three listeners (S2, S4, and S9) was 6.9%.

C. Conclusion

The SIUD-method produces a narrower spread of threshold estimates than either the ML- or the 2IFC-method over runs of 50 trials, and shows no obvious tendency to overestimate the threshold.

VII. DISCUSSION

In summary, the SIUD-procedure is recommended as a fast and reliable threshold procedure to estimate absolute threshold in both normal and impaired groups of listeners. Numerical simulations suggest that it is substantially more efficient than either Green's (1993) ML-method or the conventional 2IFC-procedure. They also showed that the number of trials needed to estimate threshold using the SIUD-method can be specified approximately using a simple formula based on the required accuracy and the steepness of the psychometric slope. Psychometric slopes were found to be steeper for hearing impaired listeners and, as a consequence, fewer trials are required for this group.

The new results reported here extend Leek *et al.*'s (2000) and Green's (1993) studies of single-interval methods by allowing the number of trials to be varied according to the accuracy requirements of the study. For example, if an accuracy of ± 2 dB is adequate then only seven trials are needed resulting in a considerable saving in testing time over other laboratory practices. Participants with a hearing impairment will often have steeper psychometric slopes than normal hearers. In this case even fewer trials will be needed. A fourfold reduction in the required number of trials applies if the slope is as steep as 1.0. In our experience with impaired hearers, ten trials give thresholds that are satisfactory for a repeatable clinical description of the impairment.

The single-interval method has been discussed solely in the context of absolute thresholds. Clearly, it could also be used in the context of a wide range of threshold measurements. It must be stressed, however, that the estimate of the number of required trials must be based on knowledge of the slope of the underlying psychometric function. If suprathreshold levels are used, compression may apply and the slope will be more shallow (Schairer *et al.*, 2008). This implies that more trials will be required.

Green's (1993) ML-method was found to be subject to occasional false estimates that arise from time to time as the result of a premature convergence on an inappropriate threshold value. These erroneous values make a proper comparison of efficiency inadmissible. They were noted by Leek *et al.* (2000) as well as by Green (1995) where they were attributed to secondary consequences of "attentional lapses." However, the numerical simulations showed that they are intrinsic to the procedure itself. Moreover, the simulations also show that these false estimates do not improve if the number of trials is increased (Fig. 9, bottom panel). In Green's (1993) procedure each successive stimulus level is set at the current best estimate of the threshold. Unfortunately, this is self-reinforcing and a false estimate can quickly become permanent across an indefinite number of trials. In contrast, the SIUD-method is not subject to the same problem and will always eventually converge on the true threshold.

Further simulations and experimental observations showed that single-interval methods gave similar threshold estimates to the 2IFC approach. This result replicates the findings of Leek *et al.* (2000) as well as Marshall and Jesteadt (1986). While the latter study was primarily aimed at comparing 2IFC with standard clinical practice based on the ANSI (1997) threshold search strategy, they did also include a single-interval comparison condition using the method of constant stimuli. Their single-interval procedure gave similar results to 2IFC.

Adherents of the signal-detection theory of the nature of absolute threshold will be puzzled by our finding that SIUD- and 2IFC-thresholds obtained using human listeners did not show any systematic differences. Of course, the absence of any measured effect does not prove that none exists; one might be observed in different testing circumstances where listeners choose to apply extra caution to their judgments. However, our listeners were asked to be very cautious and no difference was seen. The matter clearly invites further investigation. If response bias is a matter of concern, however, Kaernbach's (1990) single-interval adjustment-matrix (SIAM) procedure uses single-interval methodology while taking listener's criterion into account.

Gu and Green (1994) recommended that catch trials be used to estimate the listener's "guessing rate." However, we abandoned this approach because it was impossible to obtain an accurate estimate of guessing based on only a small number of catch trials. We were, however, reassured by the estimates of Leek *et al.* (2000) who found very low guessing rates. We encouraged our listeners to be conservative in their judgments. We defended this to them on the grounds that it was impossible to make useful measurements if listeners reported yes when no stimulus had been presented. In the event, our listeners gave very few "false-alarms." When they did occur, the run was restarted and this further discouraged guessing. Catch trials are, however, a useful guide to the attentional state of the listener and rest periods can be arranged if they begin to occur during a measurement session. Moreover, catch trials offer a regular reminder to the listener of what a no-stimulus condition sounds like. This may add to listener's confidence later when a stimulus is presented just above threshold.

Marshall and Jesteadt (1986) drew attention to the importance of the visual cue normally given in 2IFC-procedures to help the listener pay attention at the right time. This effect had previously been studied by Watson and Nichols (1976) who found a 2-dB improvement in threshold when an appropriate cue was given. The procedure followed in this study involved giving an audible cue 0.5 s before the

test stimulus. The cue has the same frequency and the same duration as the test stimulus and may well have minimized errors due to temporal uncertainty.

VIII. CONCLUSIONS

We recommend that the SIUD-method with catch trials be used for studies where it is necessary to limit the number of trials as much as possible. The procedure is simple to administer and requires little participant training. The estimated thresholds are comparable in value and are less variable than commonly used ML- and 2IFC-method. They also yield a known degree of accuracy for a given number of trials. The use of an audible cue similar to, but preceding, the test stimulus by a fixed time interval is also recommended as an aid to better threshold estimation.

ACKNOWLEDGMENTS

The authors would like to thank Christine Tan and Soumini Menon for assisting in data collection. During the course of this study, the authors benefited from helpful discussions with Marjorie Leek concerning procedural issues and Graham Upton concerning statistical issues. Helpful comments on a preliminary draft manuscript were also received from Graham Upton, Enrique Lopez-Poveda, and Brian Walden. They would also like to thank the two anonymous reviewers who made insightful comments on a previous version of this article.

ANSI S3.21-1978 (**1997**). American National Standard Method for Manual Pure-Tone Threshold Audiometry (American National Standards Institute, New York).

Arehart, K. H., Burns, E. M., and Schlauch, R. S. (**1990**). "A comparison of psychometric functions for detection in normal-hearing and hearing-impaired listeners," J. Speech Hear. Res. **33**, 433–439.

Brownlee, K. A., Hodges, J. L., and Rosenblatt, M. (**1953**). "The up-and-down method with small samples," J. Am. Stat. Assoc. **48**, 262–277.

Carhart, R., and Jerger, J. F. (**1959**). "Preferred method for clinical determination of pure-tone thresholds," J. Speech Hear. Disord. **24**, 330–345.

Carlyon, R. P., Buus, S., and Florentine, M. (**1990**). "Temporal integration of trains of tone pulses by normal and by cochlearly impaired listeners," J. Acoust. Soc. Am. **87**, 260–268.

Choi, S. C. (**1990**). "Interval estimation of the LD50 based on an up-and-down experiment," Biometrics **46**, 485–492.

Cornsweet, T. N. (**1962**). "The staircase-method in psychophysics," Am. J. Psychol. **75**, 485–491.

Dixon, W. J., and Mood, A. M. (**1948**). "A method for obtaining and analyzing sensitivity data," J. Am. Stat. Assoc. **43**, 109–126.

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

Green, D. M. (**1995**). "Maximum-likelihood procedures and the inattentive observer," J. Acoust. Soc. Am. **97**, 3749–3760.

Green, D. M., and Swets, J. A. (**1966**). *Signal Detection Theory and Psychophysics* (Wiley, New York).

Gu, X., and Green, D. M. (**1994**). "Further studies of a maximum-likelihood yes-no procedure," J. Acoust. Soc. Am. **96**, 93–101.

Hastings, N. A. J., and Peacock, J. B. (**1975**). *Statistical Distributions: A Handbook for Students and Practitioners* (Butterworth, London).

Kaernbach, C. (**1990**). "A single-interval adjustment-matrix (SIAM) procedure for unbiased adaptive testing," J. Acoust. Soc. Am. **88**, 2645–2655.

Leek, M. R., Dubno, J. R., He, N., and Ahlstrom, J. B. (**2000**). "Experience with a yes-no single-interval maximum-likelihood procedure," J. Acoust. Soc. Am. **107**, 2674–2684.

Levitt, H. (**1971**). "Transformed up-down procedures in psychoacoustics," J. Acoust. Soc. Am. **49**, 467–477.

Marshall, L., and Jesteadt, W. (**1986**). "Comparison of pure-tone audibility thresholds obtained with audiological and two-interval forced-choice procedures," J. Speech Hear. Res. **29**, 82–91.

Schairer, K., Messersmith, J., and Jesteadt, W. (**2008**). "Use of psychometric-function slopes for forward-masked tones to investigate cochlear nonlinearity," J. Acoust. Soc. Am. **124**, 2196–2215.

Swets, J. A. (**1964**). *Signal Detection and Recognition by Human Observers* (Wiley, New York).

von Bekesy, G. (**1947**). "A new audiometer," Acta Oto-Laryngol. **35**, 411–422.

Watson, C. S., Franks, J. R., and Hood, D. C. (**1972**). "Detection of tones in the absence of external masking noise. I. Effects of signal intensity and signal frequency," J. Acoust. Soc. Am. **52**, 633–643.

Watson, C. S., and Nichols, T. L. (**1976**). "Detectability of auditory signals presented without defined observation intervals," J. Acoust. Soc. Am. **59**, 655–668.