Adaptive psychophysical procedures and imbalance in the psychometric function

Kourosh Saberi[a] and David M. Green
Psychoacoustics Laboratory, Department of Psychology, University of Florida, Gainesville, Florida 32611

(Received 16 December 1994; revised 8 November 1995; accepted 15 February 1996)

One class of adaptive psychophysical procedures was studied, using simulated and human observers. These procedures are those which require an increase in stimulus intensity after an incorrect response, and a decrease after k successive correct responses. This paper analyzes how step size and the value of k affect the mean and standard deviation of threshold estimates based on a k-down 1-up adaptive procedure. Computer simulations are used to study the bias in threshold estimates, which are most evident when larger step size and small values of k are used. The adaptive procedure can be characterized by a function called the imbalance of the track, the relative probability of adjusting the stimulus either up or down at equal stimulus distances from the equilibrium point. These imbalance functions can be used to understand the threshold biases obtained in the computer simulations. The computer simulations also show that the average number of reversals obtained per trial is dependent on different values of k, but are largely independent of step size. The standard error of the threshold estimates, however, varies systematically with step size, but are nearly independent of k. Finally, we compare the stability of threshold estimates for human listeners using two very different sets of parameters: a very large step size (approximately half the range of the psychometric function) with $k=4$, and the conventional $k=3$ with an initial 4-dB and a final 2-dB step size.

PACS numbers: 43.66.Yw, 43.66.Dc, 43.66.Cb [WJ]

INTRODUCTION

The adaptive up–down procedure is by far the most popular procedure used to estimate discrimination or detection performance in auditory psychophysics (Wetherill and Levitt, 1965; Levitt, 1971). It is frequently used in a two-alternative forced-choice task in which the listener is discriminating between two stimuli. The physical difference between the two stimuli is decreased if the listener is correct for some number of trials and is increased if the listener is incorrect. This is the so-called k-down 1-up rule. Part of its popularity stems from its simplicity. The signal level presented on the next trial is based on the listener's performance over at most the past k trials. The use of $k=2$ or $k=3$ is particularly ubiquitous because the probabilities of being correct at the equilibrium point are 0.707 and 0.794, which are both near the classical definition of threshold in a two-alternative task, namely, 0.75.

Some important properties of the up–down procedure have previously been considered in studies using computer simulations or psychophysical measurements of human performance. These studies have demonstrated that a greater number of intervals on a single trial (Hall, 1983; Shelton and Scarrow, 1984; Green *et al.*, 1989; Kollmeier *et al.*, 1988; Schlauch and Rose, 1990) or increasing k from 2 to 3 increases the reliability of threshold estimates (Kollmeier *et al.*, 1988; Schlauch and Rose, 1990; Leek *et al.*, 1992). The latter result is partly due to the reduced binomial variance at higher probabilities as k increases. Another factor is that an increase in the number of observation intervals increases the slope of the psychometric function which in turn reduces statistical uncertainty along the stimulus scale. The slope of the *transformed* psychometric function (Levitt, 1971), as will be described shortly, also increases as k increases, further contributing to a lower variability of threshold estimates (Wetherill, 1966; Fisher, 1922; Taylor and Creelman, 1967; Taylor, 1971; Rose *et al.*, 1970; Green, 1990, 1993). Increases in either k or the number of intervals, of course, extends the time required for obtaining a fixed number of samples (i.e., reversals). When efficiency measures are evaluated (e.g., the sweat factor of Taylor and Creelman, 1967), higher values of k and greater numbers of intervals remain the more efficient alternatives (Kollmeier *et al.*, 1988; Schlauch and Rose, 1990).

In addition to its extensive use in psychophysical measurement, the up–down procedure has proven important to psychophysical theory in a variety of ways. The up–down procedure has been used: (1) to reconstruct psychometric functions (Leek *et al.*, 1988, 1992), (2) to estimate possible instability in psychometric functions from *post hoc* analysis of the stimulus tracks (Leek *et al.*, 1991), (3) to analyze small-sample statistics (Edwards and Wakefield, 1988), and (4) in comparison to maximum-likelihood and PEST techniques (Shelton *et al.*, 1982).

Although many features of this procedure have been studied, current usage is often dictated as much by custom and habit as by any consideration of optimizing the efficiency of the procedure. In this paper, we concentrate on one parameter that has not been previously discussed, what we call the imbalance of the adaptive procedure. This imbalance reflects unequal up and down probability forces at stimulus

[a]Present address: Research Laboratory of Electronics, 36-765, Massachusetts Institute of Technology, Cambridge, MA 02139.

values away from equilibrium, and may, in some instances, introduce a bias in threshold measurement. This discussion leads naturally to a consideration of step size, the amount the stimulus difference is changed when it is adjusted, up or down. Computer simulations of different adaptive procedures and different step sizes lead us to a definition of the relative yield of the procedure, that is, average number of turnarounds produced by a fixed number of trials. Consideration of these factors leads us to explore the use of larger step sizes than are currently in vogue and a larger number for k.

Finally, we estimated human listeners' thresholds in two detection tasks, the absolute threshold for a sinusoidal signal and the detection of a sinusoid in noise using two quite different adaptive procedures. One was a conventional 3-down 1-up procedure using first a 4-dB step size that was reduced to 2 dB after four reversals. The threshold estimate was based on a run of 60 trials. The second was a 4-down 1-up procedure using a step size that was constant throughout the run and approximately half the width of the psychometric function (10-dB step size for the sinusoidal increment, and 5-dB step size for the tone in noise). For this procedure, these threshold estimates were obtained by interleaving three adaptive tracks, each using a total of 20 trials.

I. UP–DOWN DYNAMICS

In most discrimination tasks, we assume that the observer's choice can be described as a stationary process governed by a psychometric function, $F(x)$, that relates the probability of a correct response, F, to the stimulus magnitude, x. It is usually assumed that $F(x)$ is monotonic increasing. In the case of up–down rules, a stimulus value of considerable importance is the stimulus value at which the probability of moving up is equal to the probability of moving down—what is called the equilibrium point. The observer's response pattern in the up–down process will cause the stimulus to oscillate about the equilibrium point, so that it is also treated as the threshold value for the stimulus in these tasks. It is easy to derive the equilibrium point for a k-down 1-up rule. Equilibrium occurs when the probability of moving the stimulus down, $F^k(x)$ is equal to the complimentary probability of moving the stimulus up,

$$F^k(x)=1-F^k(x)\Leftrightarrow F(x)=0.5^{1/k}. \tag{1}$$

The equilibrium probabilities are 0.7071 for $k=2$, 0.7937 for $k=3$, 0.8409 for $k=4$, and 0.8706 for $k=5$. Levitt called $F^k(x)$ the transformed psychometric function. Note that on the transformed function the equilibrium probability is 0.5 for any k.

It is convenient to denote the value of the stimulus that satisfies the equilibrium condition as x_0, also called the threshold stimulus value. If we increase or decrease the stimulus to some other value, then the probabilities of moving up or down are unequal, and the ratio of these two probabilities is the relative likelihood of moving the stimulus in one direction or another. A related concept is the relative probability of moving up or down at two stimulus values an equal distance above and below the equilibrium point. It would be desirable to have these probabilities nearly equal to one another but operating in opposite directions, so that a drift away from the equilibrium point in either direction will invoke equal and opposite tendencies to return to it. Clearly, the ratio of such forces will be nearly the same in the region of the equilibrium stimulus, because at that point the two probabilities are exactly equal. But what of more remote stimulus values?

We may quantify an imbalance ratio as the probability of moving down over the probability of moving up for two stimuli located the same stimulus distance from the equilibrium point but in opposite directions:

$$\mathrm{Im}(\Delta)=\frac{F^k(x_0+\Delta)}{1-F^k(x_0-\Delta)}. \tag{2}$$

The numerator of the ratio is the probability of moving the stimulus down, given a stimulus Δ above the equilibrium point. The denominator of the ratio is the probability of moving the stimulus up, given a stimulus Δ below the equilibrium point.

The imbalance ratio at the equilibrium point, Im(0), will be exactly unity. At other stimulus values, Im(x) will generally not be equal to unity. If Im is greater than unity, it indicates that the downward force is larger than the upward force; if Im is smaller than unity, the reverse is true. Unless the ratio is nearly one, then the stimulus track will have a tendency to spend more time on one or the other side of the equilibrium point. Threshold estimates based on average values of the stimulus or average values of some features of the track, such as the turnaround values, will provide a biased estimate of the true threshold.

We next display the imbalance ratio, Im(Δ), for a number of different psychometric functions. Six different psychometric functions were considered. They were chosen to represent a number of reasonable psychometric functions ranging from variants of the Gaussian [(3a) and (3c)] and logistic [(3b) and (3d)] to the arctangent [(3e)] and the Weibull [(3f)]. The parameters of these functions were chosen to produce a two-alternative forced-choice psychometric function that has a range of about 20 dB. The first psychometric function has a very simple form and the parameters of the other functions were adjusted to mimic the first. The equations for each are listed below. Different theorists have proposed that the basic stimulus variable is either a change in stimulus energy or the logarithm of that quantity, a decibel-like quantity (Laming, 1986; Grantham and Yost, 1982). The psychometric function produced by these two assumptions have very different properties. We consider both stimulus level in decibels (SL), Eqs. (3c) and (3d), or the amplitude of the stimulus, x, Eqs. (3a), (3b), (3e), and (3f) where $\mathrm{SL}=20\log_{10}(x)$:

$$F(x)=\int_{-\infty}^{x/\sqrt{2}}\phi(z)dz,\quad x>0; \tag{3a}$$

$$F(x)=\frac{1}{1+\exp(-1.2x)},\quad x>0; \tag{3b}$$

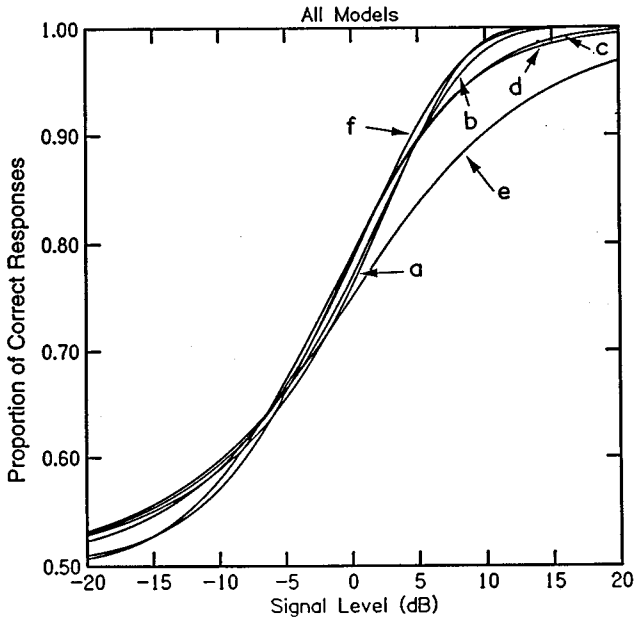

FIG. 1. Six different psychometric functions corresponding to those described in Eqs. (3a)–(3f).

$$F(x)=\frac{1}{2}+\frac{1}{2}\int_{-\infty}^{0.12\,\mathrm{SL}+0.2}\phi(z)\,dz,\quad \mathrm{SL}=20\log(x); \tag{3c}$$

$$F(x)=\frac{1}{2}+\frac{1}{2}\,\frac{1}{1+\exp(-0.21\ \mathrm{SL}-0.3)},\quad \mathrm{SL}=20\log(x); \tag{3d}$$

$$F(x)=\frac{1}{2}+\frac{1}{2}\,\frac{\arctan(x)}{\pi/2},\quad x>0; \tag{3e}$$

$$F(x)=1-\frac{1}{2}\exp\left[-\left(\frac{x}{1.157}\right)^{1.25}\right],\quad x>0. \tag{3f}$$

Figure 1 illustrates these six psychometric functions. The means of the functions have been adjusted to position them near each other along the stimulus axis. Except for the arctangent [model (3e)], these functions also produce nearly the same probability of being correct for all stimulus values. The functions are so similar that it would be difficult to determine experimentally any difference between them. Despite this similarity, the imbalance ratios, Im(Δ), shown in Fig. 2 are quite different for the different psychometric functions. We have computed the imbalance ratios for $k=2-6$ and for $\Delta\leqslant 10$ dB, a stimulus change of roughly half the range of the psychometric function. The imbalance ratio for models (3a)–(3c) is always greater than unity for any k and all stimulus changes. Models (3d)–(3f) show a somewhat different pattern, being nearly unity for all stimulus changes when k is equal to or larger than four. Models (3c)–(3f) show greater imbalance for $k=2$ than for any other rule for all Δ, while for models (3a) and (3b) this is true only for the larger stimulus changes.

One implication of these figures is that any psychometric function will show a bias in the estimate of threshold if a large step size and small value of k are used. The bias will be to estimate a threshold somewhat below x_0, because the imbalance ratio is greater than unity, that is, the downward force is larger than the upward force, except at the equilibrium point. We now turn to computer simulations of the k-down 1-up adaptive rule for these different psychometric functions.

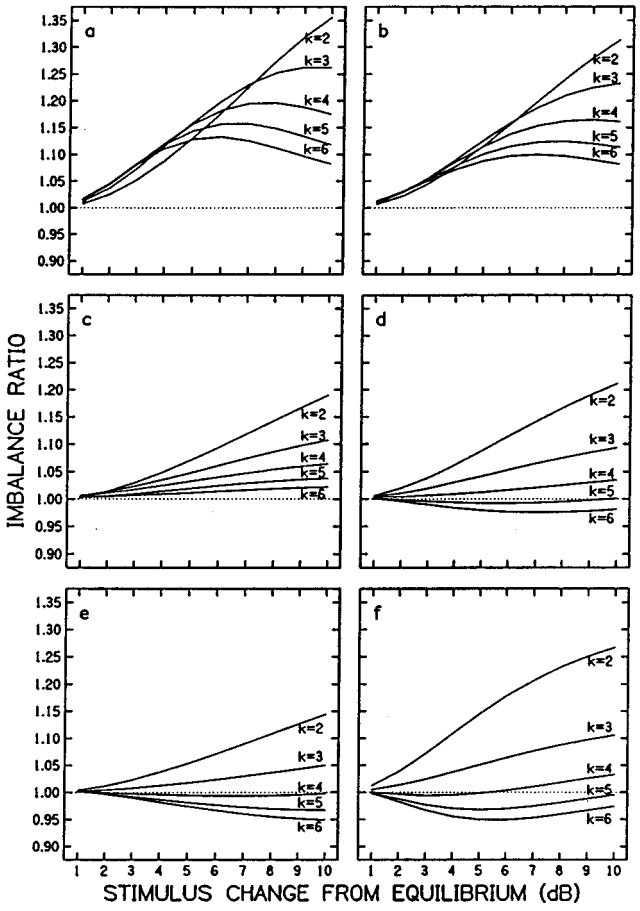

FIG. 2. The imbalance ratio (Im) shown for the six psychometric functions of Fig. 1. An imbalance ratio greater than unity means that the downward force is greater than the upward force for equal distances (Δ) above and below the equilibrium probability. In such a case, the adaptive track will be biased low.

II. COMPUTER SIMULATIONS

Monte Carlo methods were used to simulate the performance of two observers having different psychometric functions. One of the simulations used the first [Gaussian, Eq. (3a)] and the other, the fourth [logistic, Eq. (3d)] psychometric functions. We selected these two psychometric functions because their imbalance ratios were markedly different (Fig. 2). Simulations were conducted using the k-down 1-up rule for k between 2 and 5. For each condition and psychometric function, 1000 thresholds were measured. From each set of simulations, the mean and standard deviation of the threshold estimates were computed as well as the mean number of reversals.

The two major parameters of the simulations, besides the k of the up–down rule, were (1) the number of trials per run and (2) the step size used in the adaptive procedure. The number of trials per run was 10, 20, 30, 40, or 50. Step sizes of 2, 5, 10, 15, and 20 dB were examined. The step size remained constant throughout the run. The threshold was defined as the signal level corresponding to the equilibrium probability for each k of the up–down rule. The starting value for the stimulus was randomly chosen from a uniform distribution with a range of plus and minus two times the step size centered at the threshold value. Thus, for the step size of 2 dB, the first stimulus value was chosen at random from a range of 8 dB, whereas for the step size of 20 dB, the initial stimulus value was chosen from a range of 80 dB. In terms of step size, all starting points were dispersed about the threshold value by an equal distance. The estimated value of threshold was based on the average of the stimulus values on *all* reversals. If a run had no reversals, it was discarded (except when considering the yield of the procedure as described below in Sec. II A 2). Virtually no runs were discarded for any value of k or step size when there were more than ten trials. When the number of trials was 10, the approximate percentages of discarded runs for $k=2-5$ were 0.5, 4, 12, 30, and 35, respectively.

A. Results

1. Threshold estimates

Figures 3–6 plot the results of these simulations for different step sizes (see left panel, second from top) and for different values of k. The left columns show the results for an observer using the first psychometric function [Eq. (3a)]. The right columns show the results for an observer using the fourth psychometric function [Eq. (3d)]. The top panels are the mean threshold estimates corresponding to equilibrium probability (e.g., 0.707 for $k=2$), which the procedure is tracking. The solid lines are the stimulus values at that equilibrium probability. For the lower values of k (Figs. 3 and 4) and for the larger step sizes, there is a noticeable bias in threshold estimation for both psychometric functions. This is in agreement with results from Schlauch and Rose (1990) and Edwards and Wakefield (1988) who have also reported biased estimates from simulations, particularly for a large step size and $k=2$. As the value of k increases (Figs. 5 and 6), the mean threshold estimates are closer to true threshold values. The bias is nearly always in one direction; the procedure underestimates threshold, and is in accord with the imbalance ratio shown in Fig. 2. When the downward probability force is greater than the upward force ($\mathrm{Im}>1$), the track will spend more time below the equilibrium probability than above it. This imbalance is a property of the adaptive procedure and not related to the number of trials used in the procedure. Thus, in the simulations, we expect the bias to be practically independent of the number of trials used to estimate the threshold value, as it is. The bias in threshold estimates is substantially reduced with increasing k. It is between 1–2 dB for $k>3$ for the Gaussian model and nearly zero for the logistic model.

Another way to minimize the bias in threshold estimates is to reduce the step size. This keeps the track within a narrow region about the equilibrium point. As Fig. 2 shows, the imbalance ratio is never greater than about 15% in a region within 5 dB of the equilibrium point for any of the six different psychometric functions. Schlauch and Rose (1990) have also shown that biases in thresholds may be reduced if thresholds are estimated from a psychometric function fitted

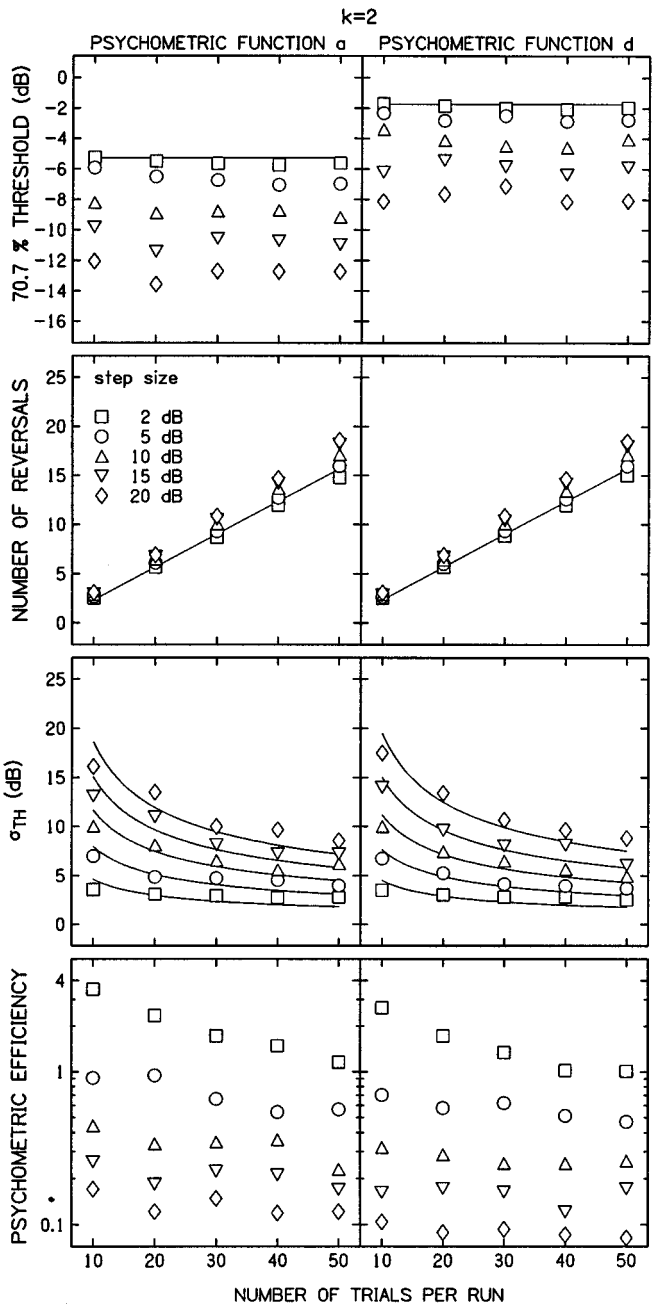

FIG. 3. Results from simulations for the 2-down 1-up rule ($k=2$). The left panels show results for a simulated observer who performs according to psychometric function *a* (Gaussian) and the right panels for psychometric function *d* (logistic). Top panels are threshold estimates based on 1000 simulations per point. The solid lines show stimulus values at equilibrium probability (true threshold). The parameter is step size. The second row of panels shows the average number of reversals as a function of the number of trials in a run. The solid line is Eq. (4). The third row of panels shows the standard deviation of threshold estimates. The solid lines are based on the equation, $a/\sqrt{N_{\mathrm{revs}}}$, where a is determined from a least-squares fit to the data for various step sizes and N_{revs} is determined from Eq. (4). The bottom panels are the psychometric efficiencies determined from Eq. (5).

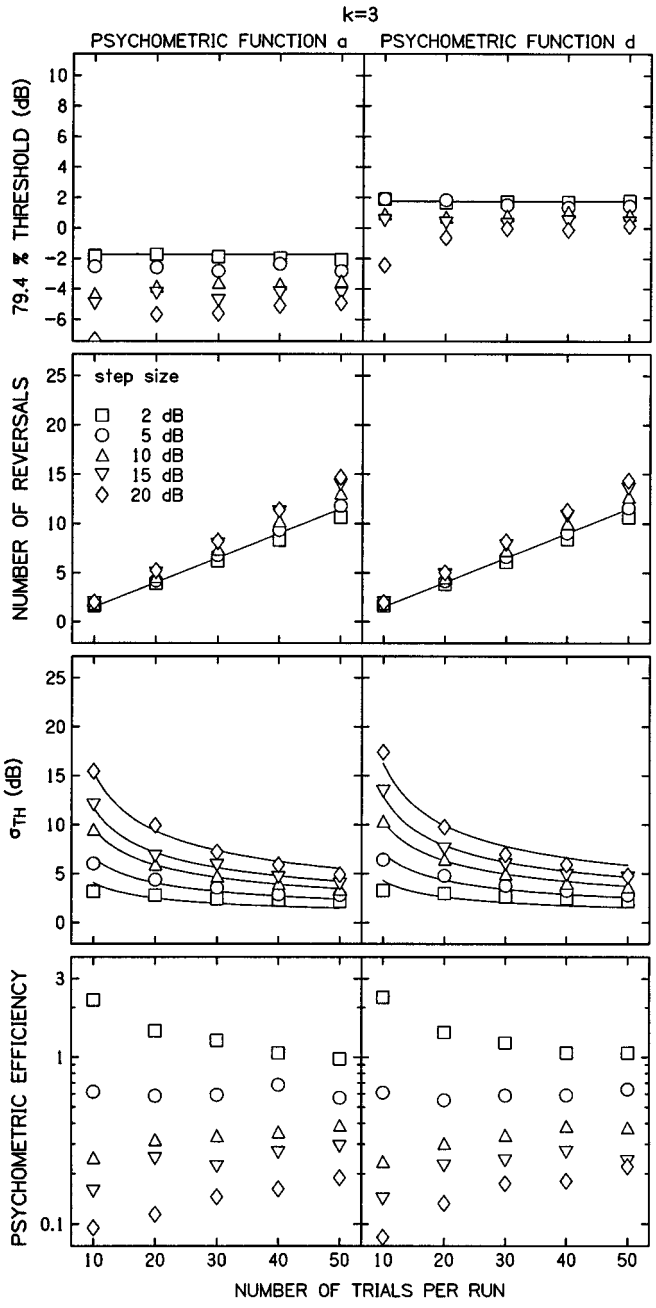

FIG. 4. Same as Fig. 3 for $k=3$.

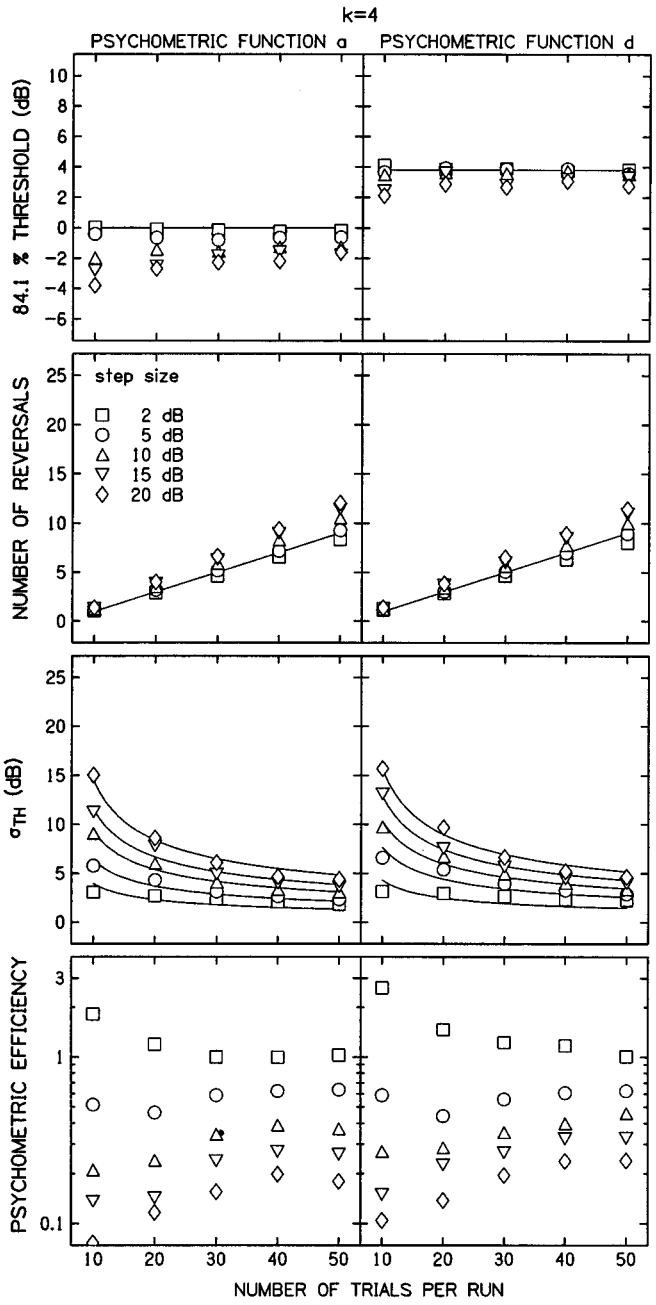

FIG. 5. Same as Fig. 3 for $k=4$.

to the trial history, instead of determining threshold from averaging the reversals.

2. Number of reversals

The second row of panels in Figs. 3–6 show the relation between the number of reversals and the number of trials used to obtain each threshold estimate. To obtain an accurate estimate of the number of reversals, no runs were discarded, and therefore the average number of reversals per condition is based on all 1000 runs for that condition. The solid line is a linear function relating the number of trials and the average number of reversals. The slope of the linear function changes systematically with k, but it is largely independent of the form of the psychometric function, that is, the left and right panels for a given value of k are nearly the same.

The value of the intercept can be rationalized on the basis of the following argument. It takes some fixed number of trials to produce the first reversal. It is clear that this number is at least $k+1$ (a sequence of k correct followed by one incorrect response, or one incorrect followed by k correct). Thus, if we subtract $k+1$ trials from the total number of trials, then we might expect for any value of k, the same starting point for the average number of reversals.

The slope of the line relating the number of effective trials to the number of turnarounds is nicely approximated by the value $1/(k+1)$. This slope value is the number of reversals produced per trial. It is what we might call the yield of

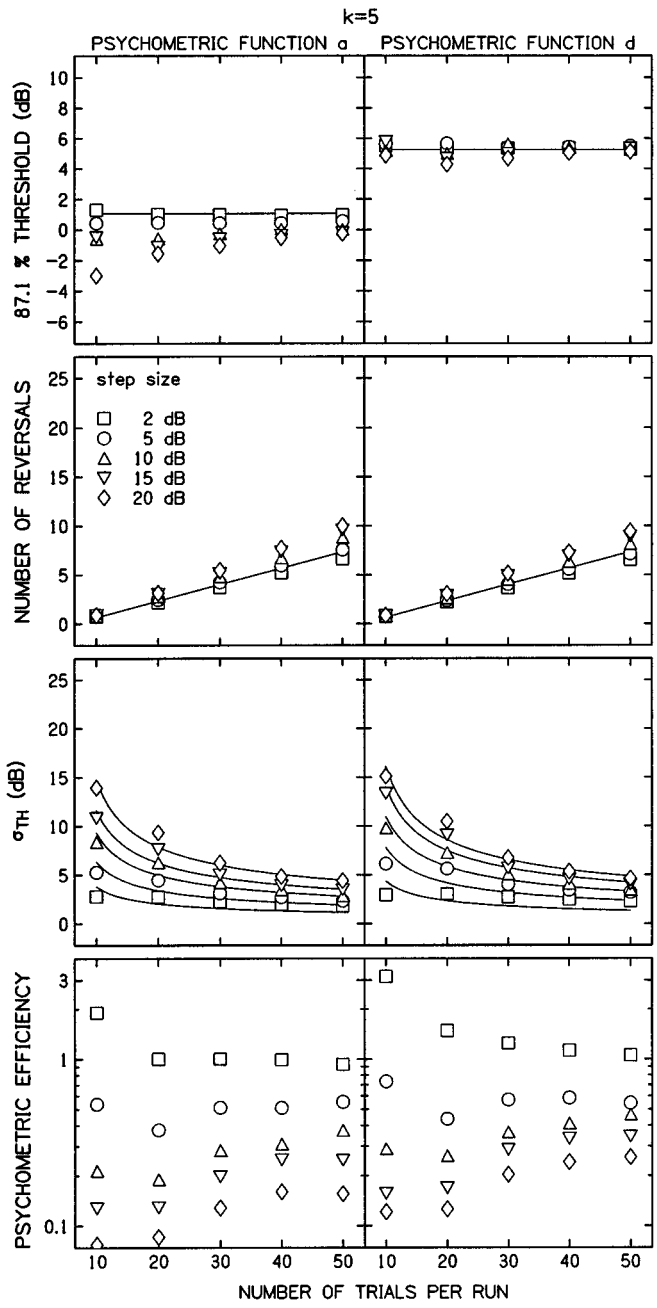

FIG. 6. Same as Fig. 3 for $k=5$.

that adaptive procedure, because it is the fraction of trials that produces a turnaround, on average.

A good approximation of our simulation results is therefore

$$N_{\text{revs}}=\frac{1}{k+1}[N_{\text{trials}}-(k+1)]=\frac{1}{k+1}N_{\text{trials}}-1. \qquad (4)$$

The solid lines in the second rows of Figs. 3–6 are calculated from Eq. (4). This simple equation provides a reasonably good approximation to the data of the simulations.

An interesting concept is the maximum possible yield for a k-down 1-up adaptive procedure. Such a track would consist entirely of alternate increases and decreases in stimulus level. Because a decrease requires k correct responses while an increase requires one incorrect response, on average two reversals arise from $(k+1)$ trials [an optimum slope of $2/(k+1)$]. Thus, our empirical slope value of $1/(k+1)$ is roughly half the maximum possible yield.

A better approximation would include the step size as a parameter, because it is evident from the data that the larger step sizes produce, on average, slightly more reversals. The least-squares slope of a linear function relating the number of reversals to the number of trials is a quantity that describes the empirical yield of an adaptive procedure. The ratio of the empirical yield to the maximum possible yield expresses the relative reversal yield of the procedure and was computed for the different up–down rules and averaged over the different step sizes. The relative reversal yields are 0.51 ($k=2$), 0.52 ($k=3$), 0.51 ($k=4$), and 0.53 ($k=5$). Thus, all the different rules produce remarkably similar values. There is a small but consistent trend for the relative reversal yield to increase with step size. For each step size, the relative reversal yields are, averaging over k, 0.52 (ss=2 dB), 0.53 (ss=5 dB), 0.53 (ss=10 dB), 0.55 (ss=15 dB), and 0.58 (ss=20 dB).

3. Variability of threshold estimates

Finally, we examine what is generally considered the critical feature of any procedure that estimates threshold, the variability of repeated measures. The third rows of panels in Figs. 3–6 show the standard deviation of the threshold estimates in decibels plotted against the number of trials, n, with the various step sizes coded by different symbols. Five lines are shown in each panel; all show the standard deviation decreasing inversely with the square root of the number of trials. Because the distribution of values from which thresholds were estimated was the distribution of reversal values, N_{revs}, Eq. (4) was used to determine N_{revs} from the number of trials, and the fitted lines were calculated using N_{revs} ($a/\sqrt{N_{\text{revs}}}$, where a is a constant). The five lines are based on different values of the parameter a, obtained individually from a least-squares fit to the data for each step size.

This equation describes the general trend of the data quite well. There are, however, some systematic departures. For the smaller k values, it is evident that the point at $n=10$ is generally below the fitted line. Thus, the measured standard deviations are lower than predicted for small n, but follow the square-root law for larger n. This may be explained by considering two different distributions, what Kollmeier *et al.* (1988) refer to as the starting and limiting distributions. The starting distribution is the distribution of stimulus levels at the initial stages of the track, while the limiting distribution is the distribution of stimulus levels near the end of a track. Often, near the beginning of a track, the variability in stimulus levels is smaller. This is partially due to the usually restricted range from which the starting stimulus value is selected, particularly if this selection is guided by *a priori* information about the general region of threshold. To verify this reasoning, we ran simulations for $k=2$ in which the starting distribution was defined as the distribution of stimulus values on the first incorrect response of a run. We defined the limiting distribution as the distribution of values on the very last trial of the run. Results showed that the

limiting distribution always had a larger standard deviation than the starting distribution. This difference was greater for the larger step sizes, a maximum difference of 4 dB, and smaller for the smaller step size, a maximum of 1 dB, although, relative to step size, the difference was proportionately greater for the smaller step size; 50% for 2-dB and 20% for 20-dB step size. Thus, the small departure from the square-root law is consistent with a consideration of the properties of the starting and limiting distributions of the stimulus track.

One other surprising aspect of these results is the similarity of the standard deviations for all values of k. The data for the third rows of panels are nearly the same for $k=2$ to $k=5$. As we have just seen, the yield of an adaptive procedure (slopes in the second panels) decreases as k increases. Thus, the same number of trials produces considerably more reversals when $k=2$ as opposed to $k=5$. The threshold estimate is based on an average of these reversal levels. While increases in k decrease the number of values used to calculate the average, the variability of the estimates is essentially independent of k. How is this possible?

The answer to this question is complex and depends on a number of factors, but the principle, and apparently controlling, ingredient is simple. The slope of the transformed psychometric function increases as k increases. A steeper slope for the psychometric function makes the point of transition or threshold easier to estimate, which is to say that the variability of the threshold estimate decreases as the slope increases. This increase in slope is enough to offset the decrease in the number of reversals (or equivalently, the decrease in the number of independent trials), and the resulting variability in threshold estimates is nearly independent of k, at least for changes in k from 2–5.

Finally, we compare the efficiency of the procedure for the four values of k and different step sizes. The ideal minimum variability for a single trial is simply the ratio of the binomial variance divided by the squared slope of the psychometric function, $p(1-p)/F'^2$. The ratio of that factor, called its ''ideal sweat factor'' by Taylor and Creelman (1967) to the empirically observed variance based on n trials, its ''empirical sweat factor,'' is a measure of the psychometric efficiency of a psychophysical procedure,

$$\text{psychometric efficiency} = \frac{\dfrac{p(1-p)}{n(dF/dx)^2}}{\text{empirical variance}}. \qquad (5)$$

The attractive feature of this measure is that it allows comparison across various procedures and number of trials.

In the bottom panels of Figs. 3–6, we plot these psychometric efficiencies on a logarithmic scale, to present the values for all step sizes in the same panel. There are two features of these calculations which are immediately noticeable.

First, for the smallest step size, the psychometric efficiency is often greater than unity, which means that the procedure is performing better than ideal. As previously discussed, this better-than-ideal performance may be attributed to the smaller variance associated with the starting distribution compared to the limiting distribution.

Second, the psychometric efficiency depends on n, the number of trials used in the threshold estimate. The efficiency will only be constant across various n if the $\sqrt{n}$ law holds. When $k=2$ and step size is larger than 10 dB, or when $k>2$ and the step size is 5 dB, the efficiency is in fact constant. However, for other cases, one obtains an approximate monotonic dependence of efficiency on the number of trials. For small step sizes, the psychometric efficiency decrease monotonically because the starting distribution is too narrow. For large step sizes and k values, the effect reverses and one obtains a monotonically increasing efficiency. This latter trend may be due to a larger variance of the starting distribution compared to the limiting distribution when step size is large and k has a high value.

III. EXPERIMENTS WITH HUMAN LISTENERS

The preceding simulations indicate that despite changes in step size from 2 to 20 dB, the average estimated thresholds are nearly the same if values of k larger than 3 are used. In experiments with real observers, step sizes larger than 5 dB are rarely used and then only when the final step size is much smaller. Would procedures using large step sizes produce threshold estimates even close to those produced by more conventional procedures? The following experiment was designed to probe that question. The purpose was not to compare extensively the two procedures, but to determine if a large threshold bias would result if a large step size was employed with a large value of k. In one condition a standard 3-down 1-up procedure was employed, with an initial step size of 4 dB and a final step size of 2 dB. In the second condition, a $k=4$ method was employed with a much larger step size, 10 dB for tone in quiet and 5 dB for tone-in-noise detection.

A. Experimental procedure

Thresholds were estimated in two forced-choice detection tasks using two different adaptive procedures. In one task, what we call the absolute-detection task, the signal to be detected was a 1000-Hz tone of 200-ms duration presented in quiet. In the second task, what we call the tone-in-noise task, the same tone was presented in a broadband noise of 25-dB spectrum level. The experimental details are similar to that used by Gu and Green (1994). The auditory signals were generated on a microcomputer using a 50 000-Hz sampling rate and played out over 16-bit digital-to-analog converters. The same three undergraduate listeners participated in all the experiments. Their hearing was within 10 dB of normal as determined from a Bekesy audiometric test, and they were paid for their participation in the experiments.

B. Experimental conditions

1. Conventional up–down procedure

One adaptive procedure used to estimate the thresholds in these two detection tasks was the conventional 3-down 1-up procedure. The initial step size was 4 dB, but after four reversals the step size was reduced to 2 dB. A threshold estimate was based on 60 trials, calculated as the average of successive pairs of reversals once the smaller step size had

TABLE I. The threshold values obtained with each procedure.

Listener	Conventional proc. ss=4 dB→2 dB	Large-step proc. ss=10 dB	Difference (ratio of SE)
(a) threshold estimates obtained for absolute threshold task for each listener [standard error of estimate (SE)].			
1	2.8	2.8	0
	(0.5)	(0.4)	(1.2)
2	−0.4	1.0	−1.4
	(0.3)	(0.5)	(0.6)
3	1.7	2.4	−0.7
	(0.4)	(0.4)	(1.0)
Listener	Conventional proc. ss=4 dB→2 dB	Large-step proc. ss=10 dB	Difference (ratio of SE)
(b) Threshold estimates obtained for tone-in-noise task for each listener [standard error of estimate (SE)].			
1	13.5	15.0	−1.5
	(0.3)	(0.3)	(1.0)
2	11.4	12.6	−1.2
	(0.2)	(0.3)	(0.7)
3	11.7	12.5	−0.8
	(0.3)	(0.3)	(1.0)

been reached. Approximately 6–10 reversals contributed to the threshold estimate. Thresholds were measured in six blocks of five runs each. On the basis of 30 runs, we computed the average threshold value and the standard error of that estimate. The first of the five 60-trial runs, within a block, started with a signal level about 15 dB above threshold, but the starting value for each consecutive run was the threshold estimate of the first run. Thus, only the first threshold estimate in a block required a number of trials to reach a near-threshold value.

2. Large-step procedure

As the second procedure, we used a 4-down 1-up adaptive rule with a much larger step size, which was held constant throughout the adaptive run. The step size was 5 dB for the tone-in-noise task and 10 dB for the absolute threshold experiments. Only 20 trials per run were used to estimate threshold in this adaptive procedure, and the threshold value was computed as the average of the stimulus value at reversal points, including the first reversal. In order to equate the time spent in the listening booth, three 20-trial runs were interleaved. The successive trials were selected in strict alternation from the three tracks. Thirty such 60-trial runs, in blocks of five, were administered to provide an average threshold estimate and a standard error. As with the conventional procedure, the first of the five 60-trial runs in a block started with a signal level about 15 dB above threshold, but the starting value for each consecutive run was the threshold estimate of the first run.

In sum, for different detection tasks, the average and standard error of the threshold estimates were computed, based on a total of 1800 trials per subject and condition. In the conventional procedure, thirty 60-trial runs contributed to that estimate. In the large-step procedure, ninety 20-trial runs contributed to that estimate.

C. Results

Table I gives the threshold values obtained with each procedure and for each task as well as the difference in estimates between the two adaptive procedures. As can be seen, the two procedures produce very similar threshold estimates; the difference is less than 1 dB for the absolute threshold task and less than 2 dB for the tone-in-noise task (fourth column). We would expect some small difference in the threshold estimates because the two procedures track different probability values, 0.794 for $k=3$ and 0.841 for $k=4$. These differences lead us to expect about a 1-dB difference in the threshold value for the tone in noise task and about a 2-dB difference in the absolute threshold task. The measured thresholds are in the expected direction and approximately the expected size for the tone-in-noise task.

The standard errors for the two procedures are also similar in size, and are always less than 0.5 dB. The ratio of the two standard errors for each subject and condition is shown in the parentheses in the last column. Thus, despite the difference in the step size (10 vs 2 dB for the absolute threshold estimates, 5 vs 2 dB for the tone-in-noise estimate), there was no consistent difference in the standard error of the threshold estimate.

We conclude that this experiment demonstrates that larger step sizes can be used in an adaptive psychophysical procedure. For an equal number of trials, the standard error of the threshold estimate is approximately the same as that obtained with more conventional procedures, and slightly better than those estimated from simulations. The smaller variability of estimates for human observers is partly related to choosing the starting level, after the first run, at the first estimated threshold. The nearly identical standard errors for large and conventional steps obtained for human observers may be related to the larger number of runs in the former case. While the total number of trials was 1800, the 90-run estimate from the large-step procedure produces more independent estimates of threshold than the 30-run conventional procedure, hence, offsetting the larger variance associated with a larger step size.

IV. CONCLUSION

We introduced in this paper the concept of an imbalance of the up–down procedure. This imbalance is related to a difference in the up and down probability forces which drive adaptive tracking. The imbalance approaches zero as the stimulus value approaches equilibrium threshold. Because of this imbalance, it is important that the experimenter be aware of the effects of step size on threshold estimation. Our simulations show that larger step sizes result in larger biases consistent with the value of the imbalance ratio. On the other hand, this bias may be reduced if the experimenter uses a larger value of k (4 or 5).

The use of a large step size in adaptive psychophysical procedures may be appealing in certain experimental situations. One major advantage of a larger step size is that the stimulus level reaches the threshold value quickly. Because this is the case, the experimenter needs little *a priori* information about the approximate threshold level. A larger step size also minimizes concern about the starting stimulus value causing a bias in the threshold estimate. Another by-product of a large step size is that it opens the possibility of using fewer trials per threshold estimate (as we have done in our

tests of human observers). Fewer trials per estimate allows one to interleave a number of adaptive tracks which will reduce intertrial correlation and produce more independent samples of threshold, given a fixed number of total trials in the experiment.

Our observers generally report that larger step sizes made the detection task easier. The larger change in the stimulus value makes the signal characteristics more obvious. This feature, together with the ability to use fewer trials, may make the use of a larger step size appealing for studies involving animals or children. It may also prove useful in clinical settings where elaborate training on the nature of the signal is less practical and where the testing time must, of necessity, be short.

Finally, in spite of these advantages, we do not, as a rule, recommend the use of large step sizes because it increases the variability of threshold estimation. However, if anyone for any reason is compelled to use larger step sizes, then it would be judicious to choose a high k value (4 or 5) to reduce bias in these estimated thresholds.

V. SUMMARY

Computer simulations of k-down 1-up adaptive procedures, show the following:

(1) Larger step sizes produce biases in threshold estimates. These biases are generally underestimates of the true threshold and can be as large as 8 dB for $k=2$, but are reduced to less than 2 dB for $k=4$ and 5. The size and direction of these biases can be understood by considering the imbalance ratio inherent in up–down adaptive procedures. The imbalance bias is a property of the psychometric function and does not diminish with increasing number of trials per run.
(2) The yield, the average number of reversals per trial, of a k-down 1-up procedure is approximately $1/(k+1)$.
(3) Despite the changes in yield, the standard deviation of repeated threshold measurements is very similar for different values of k. Larger step size produces monotonic increases in the standard deviation for any number of trials.

Measurements of human listeners' thresholds using two different adaptive procedures, show that a procedure using a larger step size (10 dB) can produce standard error of threshold estimates comparable in size to those produced by more conventional small-step-size (4–2 dB) procedures, given the same total number of trials.

ACKNOWLEDGMENTS

This work was supported by NIH and AFOSR. We thank the three formal reviewers, Dr. C. D. Creelman, Dr. C. Kaernbach, and Dr. B. Kollmeier for helpful discussions. We also thank Elizabeth A. Strickland and Huanping Dai for commenting on an earlier draft of this paper, and Z. A. Onsan, Q. T. Nguyen, and Mary Fullerton for technical assistance.

Edwards, B. W., and Wakefield, G. H. (**1988**). ''Small-sample analysis of Levitt's psychophysical procedure,'' J. Acoust. Soc. Am. Suppl. 1 **83**, S17.
Fisher, R. A. (**1922**). ''On the mathematical foundations of theoretical statistics,'' Philos. Trans. R. Soc. London Ser. A **222**, 309–368.
Grantham, D. W., and Yost, W. A. (**1982**). ''Measures of intensity discrimination,'' J. Acoust. Soc. Am. **72**, 406–410.
Green, D. M. (**1990**). ''Stimulus selection in adaptive psychophysical procedures,'' J. Acoust. Soc. Am. **87**, 2662–2674.
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., Richards, V. M., and Forrest, T. G. (**1989**). ''Stimulus step size and heterogeneous stimulus conditions in adaptive psychophysics,'' J. Acoust. Soc. Am. **86**, 629–636.
Gu, X., and Green, D. M. (**1994**). ''Further studies of a maximum-likelihood yes–no procedure,'' J. Acoust. Soc. Am. **96**, 93–101.
Hall, J. L. (**1983**). ''A procedure for detecting variability of psychophysical thresholds,'' J. Acoust. Soc. Am. **73**, 663–667.
Kollmeier, B., Gilkey, R. H., and Sieben, U. K. (**1988**). ''Adaptive staircase techniques in psychoacoustics: A comparison of human data and a mathematical model,'' J. Acoust. Soc. Am. **83**, 1852–1862.
Laming, D. (**1986**). *Sensory Analysis* (Academic, London).
Leek, M. R., Hanna, T. E., and Marshall, L. (**1988**). ''Psychometric function reconstruction from adaptive tracking procedures'' (Report No. 1095), Groton, CT: Naval Submarine Medical Research Laboratory.
Leek, M. R., Hanna, T. E., and Marshall, L. (**1991**). ''An interleaved tracking procedure to monitor unstable psychometric functions,'' J. Acoust. Soc. Am. **90**, 1385–1397.
Leek, M. R., Hanna, T. E., and Marshall, L. (**1992**). ''Estimation of psychometric functions from adaptive tracking procedures,'' Percept. Psychophys. **51**, 247–256.
Levitt, H. L. (**1971**). ''Transformed up–down methods in psychophysics,'' J. Acoust. Soc. Am. **49**, 467–477.
Rose, R. M., Teller, D. Y., and Rendleman, P. (**1970**). ''Statistical properties of staircase estimates,'' Percept. Psychophys. **8**, 199–204.
Schlauch, R. S., and Rose, R. M. (**1990**). ''Two-, three-, and four-interval forced-choice staircase procedures: Estimator bias and efficiency,'' J. Acoust. Soc. Am. **88**, 732–740.
Shelton, B. R., Picardi, M. C., and Green, D. M. (**1982**). ''Comparison of three adaptive psychophysical procedures,'' J. Acoust. Soc. Am. **71**, 1527–1533.
Shelton, B. R., and Scarrow, I. (**1984**). ''Two-alternative versus three-alternative procedures for threshold estimation,'' Percept. Psychophys. **35**, 385–392.
Taylor, M. M. (**1971**). ''On the efficiency of psychophysical measurement,'' J. Acoust. Soc. Am. **49**, 505–508.
Taylor, M. M., and Creelman, C. D. (**1967**). ''PEST: Efficient estimates on probability functions,'' J. Acoust. Soc. Am. **51**, 782–787.
Wetherill, G. B. (**1966**). *Sequential Methods in Statistics* (Methuen, London).
Wetherill G. B., and Levitt, H. (**1965**). ''Sequential estimation of points on a psychometric function,'' Br. J. Math. Stat. Psych. **18**, 1–10.