

Evaluation of maximum-likelihood estimators in nonintensive auditory psychophysics

KOUROSH SABERI and DAVID M. GREEN
University of Florida, Gainesville, Florida

This is a brief report on the use of maximum-likelihood (ML) estimators in auditory psychophysics. Slope parameters of psychometric functions are characterized for three nonintensive auditory tasks: forced-choice discrimination of interaural time differences (ΔITD), frequency (Δf), and duration (Δt). Using these slope estimates, the ML method is implemented and threshold estimates are obtained for the three tasks and compared with previously published data. ΔITD thresholds were additionally measured for human observers by means of two other psychophysical procedures: the constant-stimuli (CS) and the 2-down 1-up methods (Wetherill & Levitt, 1965). Standard errors were smallest for the ML method. Finally, simulations showed ML estimates to be more efficient than the CS and k -down 1-up procedures for $k = 2$ to 5. For up-down procedures, efficiency was highest for k values of 3 and 4. The entropy (Shannon, 1949) of ML estimates was the smallest of the simulated procedures, but poorer than ideal by 0.5 bits.

The maximum-likelihood (ML) method is an adaptive procedure that utilizes the maximum available statistical information, pooled across trials, in estimating an observer's threshold (Green, 1990, 1993, 1995; Gu & Green, 1994; Green, 1968; Laming & Marsh, 1988; Pavel, 1981; Pentland, 1980; Watson & Fitzhugh, 1990; Watson & Pelli, 1979, 1983). In auditory psychophysics, ML estimators have been applied to tasks in which the intensity of the signal is varied to estimate a threshold—for example, absolute or tone-in-noise detection (Green, 1990, 1993; Shelton, Picardi, & Green, 1982; Shelton & Scarrow, 1984). There is, however, a lack of information on the requirements and performance measures and capabilities of this procedure when applied to nonintensive scales. These involve tasks in which the signal does not involve a change in stimulus energy.¹

This paper describes three new results related to ML estimators. First, psychometric functions and their slope parameters for three nonintensive stimulus domains are documented. It is important to determine the slope parameter of the psychometric function on a logarithmic stimulus scale for a given stimulus dimension before implementing the ML method. Psychometric functions and slope parameters are measured for the discrimination of interaural time differences ($\Delta ITDs$), but data from the literature are used to document slope parameters for frequency (Δf) and duration (Δt) discrimination. Second, results from the first section are used to implement the ML method. Thresholds are estimated for

ΔITD , Δf , and Δt and compared with those reported in the literature. In addition, ΔITD thresholds and their variability are compared with those obtained from the same human observers but different psychophysical procedures. The latter procedures are the 2-down, 1-up adaptive method (Levitt, 1971; Wetherill & Levitt, 1965) and the method of constant stimuli (CS), both of which are commonly used in psychoacoustics.

In the third and final part of the paper, computer simulations are used to extend results from human observers and to add new simulation results to the literature on efficiency, bias, and information gain from ML estimators, the k -down 1-up procedures for values of $k = 2$ to 5, and the CS method. Although computer simulations have previously been used to compare the ML method to the 2-down 1-up and CS procedures, there are two reasons why additional simulations are warranted. First, the efficiency of the up-down method increases with the value of k (Kollmeier, Gilkey, & Sieben, 1988; Saberi & Green, 1996; Schlauch & Rose, 1990), and therefore it is helpful to make a comparison between threshold estimates from the higher, more efficient k rules and ML estimates. Second, there is disagreement on the relative efficiency of the CS method compared with the ML method and adaptive staircase procedures (McKee, Klein, & Teller, 1985; Simpson, 1988; Taylor, Forbes, & Creelman, 1983; Watson & Fitzhugh, 1990). Further simulations will contribute to a better understanding of the dynamics of each of these methods. We begin with a description of ML estimation and the relevance of the properties of the psychometric function to this procedure.

Maximum-Likelihood Estimation and the Psychometric Function²

Because likelihood values depend on binomial probabilities associated with a given stimulus level and the

¹We thank Beverly A. Wright for many helpful discussions. We also thank Z. Onsan, Q. Nguyen, and M. Fullerton for technical assistance. Portions of this article were presented at the 127th meeting of the Acoustical Society of America. This work was supported by NIH and AFOSR. Correspondence should be addressed to K. Saberi, 216-76 Division of Psychology, California Institute of Technology, Pasadena, CA 91125 (e-mail: kroshe@etho.caltech.edu).

particular model of the psychometric function, one must make an a priori decision about the psychometric model that gives rise to these probabilities. Different psychophysical tasks produce different psychometric functions, and the determination of the appropriate function is an empirical issue. Many forms of the psychometric function have been used as psychophysical models. These include the logistic (Green, 1990, 1993; Macmillan & Creelman, 1991; Madigan & Williams, 1987); Gaussian (Laming & Marsh, 1988; Saberi, 1995); the Weibull function (which is more popular in vision research; Foley & Legge, 1981; Quick, 1974; Robson & Graham, 1981; Watson, 1979; Watson & Pelli, 1983; Weibull, 1951), and the arctangent (Finney, 1971, 1978; Saberi & Green, 1996; Urban, 1910). Most of these functions are very similar in shape if the parameters are correctly adjusted.

Once an experimenter has selected a basic model of the psychometric function (e.g., logistic), two important steps must be taken. First, the functions must be transformed to a logarithmic stimulus scale; it is well known that psychometric functions defined on such a scale are parallel (Green & Luce, 1975; Green, McKey, & Licklider, 1959; Laming, 1988; Nachmias, 1981; Roufs, 1974; Watson, 1979; Watson & Pelli, 1983). The parallel nature of these functions is convenient because one observer's function usually differs from another only by its placement along the log-stimulus scale (i.e., its logarithmic mean), not by its slope.

The second step is that the experimenter must now determine this slope value for the family of psychometric functions selected in the first step. Different stimulus dimensions give rise to different slopes. In general, it is assumed that the psychophysical discriminator observes a quantity that is related to the stimulus scale, x , by a power transformation, $y \propto x^v$ (Egan, 1965; Laming, 1985, 1986, 1988). On a logarithmic scale, the slope of the psychometric function may therefore be considered to be the proportionality constant, v . If v is unity, the psychometric function has a range of about 20 dB. Laming (1986) has described a wide range of functions with $v = 1, 2, 4$, or even 8, for auditory and visual tasks that produce effective ranges of about 3–20 dB.

PSYCHOMETRIC FUNCTIONS AND SLOPE PARAMETERS FOR ΔITD , Δf , and Δt

Psychometric functions and their slope parameters are characterized here for Δf and Δt from previously published data, but for ΔITD they are measured. Some data on ΔITD psychometric function do exist (Henning, 1980; Koehnke, Colburn, & Durlach, 1986), but these functions are remeasured here for simple and complex tones using the same observers. The primary purpose was to quantify slope parameters across different observers and stimuli and to verify the parallel nature of these functions on a log- μ sec scale. In addition, because much of the comparison with other psychophysical methods re-

ported in this paper utilizes ΔITD s, we wanted a more detailed quantification of these functions.

Method for ΔITD Functions

ΔITD psychometric functions were measured for three normal-hearing subjects (within 10 dB of ANSI (1989) standard between 125 and 8000 Hz). All were experienced in ΔITD -discrimination tasks. The subjects practiced until we were confident that their performance on this task was stabilized. On each trial, 1 of approximately 20 fixed ΔITD s was randomly selected and presented in a two-interval forced-choice (2IFC) method of CS. Psychometric functions were measured for two types of stimuli: a 500-Hz pure tone and a 50-Hz, sinusoidally amplitude-modulated (SAM) tone with a carrier frequency of 3.5 kHz. These two waveforms were selected because they have been reported to produce very different thresholds; smaller values for the pure tone and larger values for the SAM carrier (Henning, 1974, 1980; Klumpp & Eady, 1956; Nuetzel & Hafer, 1976, 1981).

For the 500-Hz tone, the ΔITD values ranged from about 4 to 100 μ sec (slightly different ranges for different observers) and for the SAM tone, they ranged from 10 to 1,000 μ sec. Within a trial, the ITD in the first interval (equal to $\Delta ITD/2$) led to one ear, and in the second interval, to the other. The ear that carried the leading sound in the first interval was selected on a random basis. Subjects were instructed to determine whether the order of the perceived locations of the sounds was left then right or right then left. Feedback was provided after each trial. Each run lasted for 100 trials with unlimited practice allowed at the beginning of the run. Practice trials were ended by the subject, and usually did not exceed 5–10 trials.

Stimuli were generated on an IBM PC, presented through digital-to-analog converters (TDT-II) at a rate of 20 kHz, a lowpass filter with a cutoff at 10 kHz (Kemo VBF/24), and through Sennheiser HD-450 headphones in a sound-attenuating booth. Each stimulus was 400 msec in duration with 10-msec cosine-squared ramps and was presented at a level of 60 dB SPL. ITDs were produced by shifting the phase of the 500-Hz pure tone or the phase of the envelope of the SAM tone (no carrier delay).

ΔITD Functions

Figure 1 shows results for the 3 observers. Each function for each observer is based on 7,000 to 10,000 trials. The solid lines are logistic fits,

$$F_i(\text{ITD}) = \frac{1 - \beta}{1 + \exp(-1.7 \text{ITD}^v / \mu_i)} + 0.5\beta, \quad (1)$$

where β is an assumed inattention rate and μ may be considered the mean of the psychometric function on a log-stimulus scale (also referred to as the threshold parameter).³ It is useful, for ML procedures, to set β at 0.04 or 0.02, which produces an upper asymptote of 0.98 or 0.99, respectively. This asymptote of slightly less than

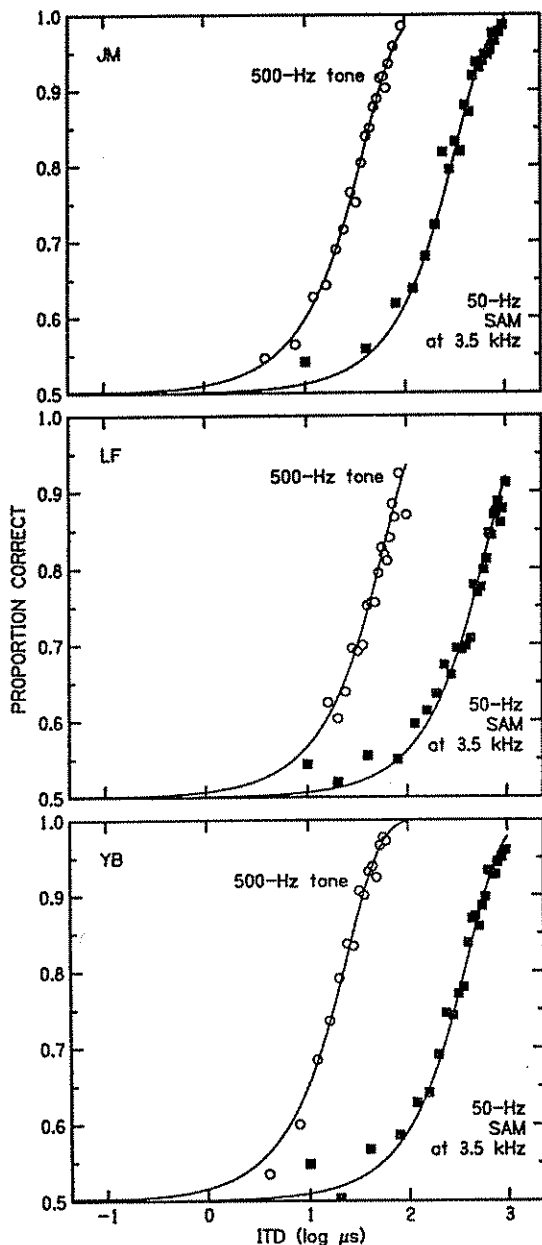


Figure 1. Psychometric functions for 3 subjects measured using the method of constant stimuli for a 500-Hz pure tone and a sinusoidally amplitude-modulated (SAM) tone with a 50-Hz modulation frequency and a 3.5-kHz carrier. The solid fits are logistic (Equation 1, $\nu = 1$).

unity is a better approximation of an observer's true function and prevents digressions of the ML search routine (Watson & Pelli, 1983).

For each observer and each function, we determined the best fitting parameters ν and μ from a MATLAB implementation of a multivariate Nelder-Meade simplex algorithm to minimize the squared deviation of each observer's data from the fit by Equation 1. For the 500-Hz tone, the

best fitting values of ν and μ for the 3 observers (J.M., L.F., and Y.B.), respectively, were $\nu = 1.02, 0.95, 1.05$, and $\mu = 45.2, 53.4, 30.5 \mu\text{sec}$. For the SAM complex, these values were $\nu = 0.94, 0.88, 0.91$, and $\mu = 238.1, 328.2$, and $247.1 \mu\text{sec}$. The values of ν are very near unity for the 500-Hz tone and slightly smaller for the SAM complex. We should note, for comparison, that for various tasks the reported values of ν have ranged from about 1 to 8 for fits with the Gaussian model (Laming, 1986).

The functions of Figure 1 are all fitted with $\nu = 1$ instead of individual values. An inspection of these functions shows that $\nu = 1$ provides a good fit to all the functions. Small deviations of observed slope from assumed slope are not likely to affect the performance of ML procedures (Emerson, 1984; Green, 1990; Madigan & Williams, 1987); a slope mismatch of a factor of 2 to 3, for example, increases the variability of threshold estimates by 20%–50%. For simplicity, we therefore used $\nu = 1$ in implementing the ML procedure.⁴

Δf and Δt Functions

We next summarize the available data from previously published work on psychometric functions for frequency and duration discrimination. These data for a variety of signals and measurement methods are shown in Figure 2. The data from these studies were transformed from d' to proportion correct and plotted as a function of log-stimulus values. The solid curves are Equation 1 with slope parameter $\nu = 1$; they are visually fitted to the data to show that they do describe the trend of the data. The different symbols represent either different observers or different signals (the asterisks represent data from the present study and are explained in the next section). The least squares estimates of ν and μ (Equation 1) were obtained individually for each symbol type in each panel. The average ν for frequency discrimination was 1.05 (with a standard deviation of 0.30 from 8 slopes) and for duration discrimination it was 1.22 (with a standard deviation of 0.26 from 8 slopes). Thus, $\nu = 1.0$ is also a good approximation for frequency- and duration-discrimination psychometric functions, and this value was used in the following ML threshold measurements.

THRESHOLD ESTIMATES FROM HUMAN OBSERVERS

Maximum-Likelihood Estimates for ΔITD , Δf , and Δt

Fifty hypothesis psychometric functions generated by Equation 1 constituted the set of hypotheses. The functions had a slope parameter $\nu = 1$ and a value of μ that increased geometrically ($\mu_i = ab^i$) where a and b are constants. The exact choice of these parameters is not critical, but 40 to 60 hypotheses in the stimulus range are usually sufficient. We first report on results for ΔITD discrimination. For ΔITDs , threshold parameters (at .707 probability of a correct response) for the 50 hypothesis psycho-

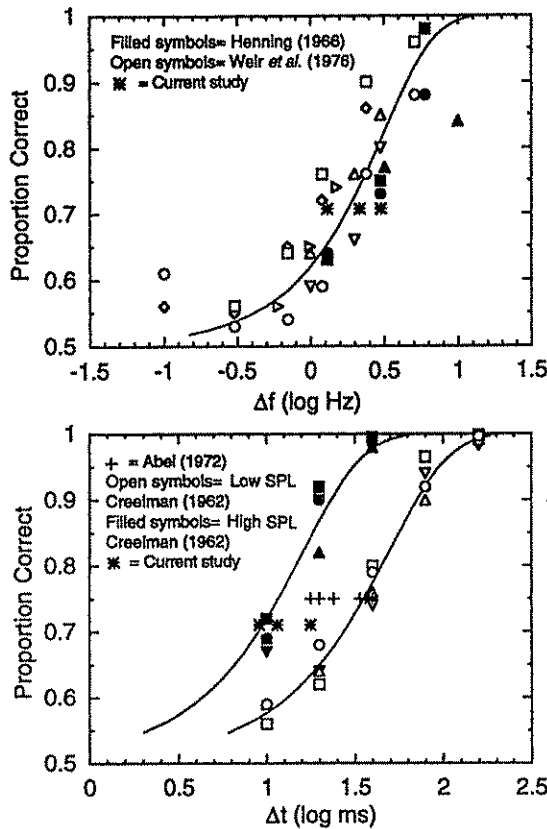


Figure 2. Top panel shows frequency-discrimination data; lower panel shows duration-discrimination data. The various symbols code either different subjects or different conditions from different studies. In the lower panel, each + symbol represents a different stimulus type used in the Abel (1972) study (different stimulus bandwidths or levels). All data were originally published in units of d' and were transformed here to proportion correct. The asterisks indicate data from the present study obtained with the maximum-likelihood method and are discussed in the section on threshold estimates from human observers. The solid curves are logistic (Equation 1, $\nu = 1$).

metric functions ranged from 0.6 to about 3.1 log μsec (re 1 μsec), which brackets the range of thresholds reported in the literature (Henning, 1974; Klumpp & Eady, 1956). The inattention parameter β was 0.02, producing an upper asymptote of 0.99.

Three normal-hearing observers (M.G., E.M., and A.N.) whose ages ranged from 19 to 24 years served as subjects. All had experience in lateralization tasks. Each observer completed 14 runs of 30 trials. The stimulus was the 400-msec, 500-Hz dichotic tone described in the previous section, and the design was 2IFC. The procedure tracked the .92 probability of a correct response (i.e., the sweetpoint of logistic functions; Green, 1990); however, to allow comparison with other techniques, threshold was defined as the .707 probability determined by interpolation on the psychometric function. Visual feedback was provided after each trial. ΔITD varied adaptively be-

tween trials according to ML rules (Green, 1990; Lamington & Marsh, 1988; Watson & Pelli, 1983).

The first set of bars in the upper panel of Figure 3 shows the average of 14 ΔITD threshold estimates for each of the 3 observers (open bars) and the mean for the 3 observers (solid bar). These thresholds are similar to those reported in the literature for a 500-Hz tone (Hershkowitz & Durlach, 1969; Klumpp & Eady, 1956; Zwislocki & Feldman, 1956). The first set of bars in the lower panel of Figure 3 shows the standard error of threshold estimates. The remaining bars are described in the next section.

Next, we report on ML threshold estimates for human observers in frequency- and duration-discrimination tasks. For frequency discrimination, observers were instructed to pick the interval with the higher pitch signal. The tone in one interval had a frequency of 3 log Hz (re 1 Hz), and the tone in the other interval had a frequency of $(3 + \Delta f)$ log Hz. The value of Δf ranged from 0.00011 to 0.01115 (total frequency range 1000.26–1026 Hz) and was selected according to ML rules. All tones were 300 msec in duration and were presented to the left ear. For duration discrimination, observers were instructed to pick the interval that carried the longer duration tone. In one in-

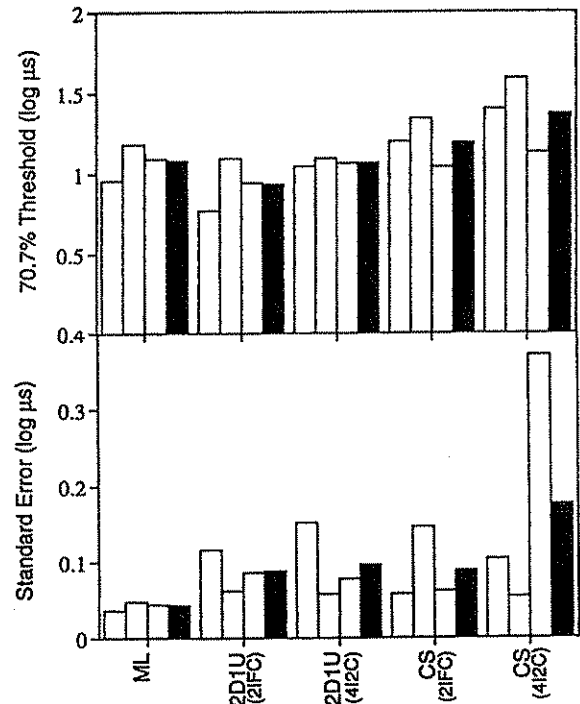


Figure 3. Top panel shows lateralization thresholds for each of 3 subjects (open bars) and their means (solid bars) using various psychophysical procedures. ML, maximum likelihood; 2D1U, 2-down 1-up; CS, constant stimuli; 2IFC, two-interval forced-choice; 4I2C, four-interval two-cue. The stimulus was a 400-msec, 500-Hz tone. The bottom panel shows the standard error of threshold estimates for the same observers and conditions.

terval, a 1-kHz tone of $t = 2.4771 \log \text{ msec}$ (re 1 msec) duration was presented, and in the other interval, the tone duration was $(2.4771 + \Delta t) \log \text{ msec}$ ($300 + \Delta t \text{ msec}$). The 50 values of Δt ranged from 0.00188 to 0.15635 log msec (1.3–130 msec) and were selected according to ML rules.

For duration discrimination, the frequency of the tone on each observation was randomized by 1% so that the use of cues based on spectral differences corresponding to different durations would be difficult. The level of the tone on each observation was uniformly randomized by 2 dB to eliminate energy-based cues for Δt thresholds less than 0.1139 log msec (Green, 1988, pp. 19–20). All stimuli were presented to the left ear. Each of the seven runs for each observer and condition consisted of 20 trials. Each observer completed both experiments in less than 15 min. These results are shown as the asterisks in Figure 2. Each asterisk represents the averaged data from 1 observer. The data fall within the range of values from the other studies (Abel, 1972; Creelman, 1962; Weir, Jesteadt, & Green, 1976). The standard errors were also quite small; for frequency discrimination, $\sigma_{\log \text{ Hz}} = 0.08, 0.12, 0.06$, and for duration discrimination, $\sigma_{\log \text{ msec}} = 0.10, 0.15, 0.09$.

Comparison With Other Psychophysical Methods (ΔITD)

ΔITDs for a 500-Hz tone were also measured for the same observers using the following psychophysical procedures: (1) the 2-down 1-up adaptive procedure in a 2IFC design; (2) the CS method in 2IFC; (3) Condition A in a 4-interval, 2-cue 2IFC; and (4) Condition B in a 4-interval, 2-cue 2IFC.

The total number of trials for each method was 420; each observer completed fourteen 30-trial runs for the ML method and seven 60-trial runs for the remaining methods (the same observers were used in all methods). For Method A (Levitt, 1971; Wetherill & Levitt, 1965), all conditions were the same as for ML runs except for the rules that determined the signal magnitude between trials. Two successive correct responses resulted in a decrease in ΔITD by a fixed log μsec stepsize, and one incorrect response resulted in an increase in ΔITD by the same stepsize.

The procedure tracks the .707 probability of a correct response, and therefore may be compared with our ML estimates. A stepsize of 0.2 log μsec was used up to the fourth reversal and 0.1 log μsec thereafter. The first four reversals were discarded, and threshold estimates were based on the average of the remaining reversals. Although we have used the average of reversals to obtain an estimate of threshold, other efficient rules of data summary could have also been used for this purpose (Schlauch & Rose, 1990; Watson & Fitzhugh, 1990). However, the present goal was to compare methods and rules that are commonly used in hearing research.

For the 2IFC CS method (B), after pilot tests, four values of ΔITD were selected (0.7, 1.0, 1.3, and 1.6 log μsec) to cover a reasonable range of the observers' psychometric functions. On each trial, all four values of ΔITD had equal a priori probabilities of being used. At the end of the seventh run, the proportion of correct responses for each ΔITD value was pooled across the seven runs and transformed into z scores; from a weighted least squares fit (Finney, 1971), threshold corresponding to .707 probability of a correct response was then determined. For these fits, the slope parameter ν was taken to be unity, consistent with results from the section on psychometric functions.

The two other methods (C and D) were identical to A and B, except that instead of the standard 2IFC, a 4-interval 2-cue 2IFC design was used. This design has previously been used to measure ΔITD thresholds with up-down procedures (Trahiotis & Bernstein, 1990); the cuing intervals are presumed to facilitate discrimination without affecting the forced-choice statistics. Method C was similar to A except for the following. Four intervals were used in which three of the intervals had a zero ΔITD (diotic), and one interval carried the entire ΔITD to be detected (instead of $\Delta \text{ITD}/2$). The starting stimulus level was 650 μsec . The signal (nonzero ITD) was either in the second or third interval. In effect, the first and fourth intervals served as cues and the observers were to pick the interval that differed from the other three, knowing that this was either Interval 2 or 3. Method D was identical to B except that it was placed in the context of the 4-interval 2-cue design.

Results are plotted in the remaining bars of Figure 3. Thresholds are nearly the same for all procedures, though there is a slight tendency for the CS method to produce a higher value. A t test between estimates from the CS methods and estimates from the other methods was significant [$t(13) = 3.13, p < .05$]. The cuing paradigm did not seem to help the discrimination task and generally increased the variability of estimates, in addition to increasing the time per trial. The ML method produced the smallest standard error compared with the remaining methods [$t(13) = 1.88, p < .05$]. We also compared the ML standard errors with only the 2-down 1-up, 2IFC (Condition A) and again found a statistically significant difference [$t(4) = 2.83, p < .025$].

SIMULATIONS

Previous simulations have been used to compare the ML method with the CS method (Watson & Fitzhugh, 1990), the PEST method (parameter estimation by sequential testing; Hall, 1981; Pentland, 1980), the 2-down 1-up adaptive procedure (Hall, 1981; Watson & Fitzhugh, 1990), and in forced-choice compared with *yes-no* tasks (Green, 1993; King-Smith, Grigsby, Vingrys, Benes, & Supowit, 1994; Madigan & Williams, 1987). Other fea-

tures of the ML method that have been examined with simulations include effects of slope mismatches between the assumed and true psychometric function (Emerson, 1984; Green, 1990; Madigan & Williams, 1987), effects of momentary lapses in attention on performance (Green, 1990, 1995; Hall, 1981; Madigan & Williams, 1987), of stimulus-placement policy (Green, 1990; King-Smith et al., 1994), and of small-sample statistics on bias and efficiency (Watson & Fitzhugh, 1990). For a recent review of various adaptive methods, see Treutwein (1995).

Two points from previous simulation are especially relevant to the present study. First, ML estimators seem to provide threshold values that are less variable than those obtained from the 2-down 1-up procedure (Watson & Fitzhugh, 1990). Second, there is disagreement in the literature as to the relative efficiency of ML and CS methods. It is suggested, on the one hand, that the CS method is more efficient than the ML (Simpson, 1988) or adaptive staircase procedures (McKee et al., 1985). Watson and Fitzhugh (1990), on the other hand, have disputed this suggestion by noting the realistic effects of experimenter uncertainty about threshold and showing that with such uncertainty included in simulations, ML procedures are in fact more efficient than the CS method. To further examine the relative efficiencies of the CS and ML methods, we used about half as much experimenter uncertainty in the following simulations as the Watson and Fitzhugh study in favor of the CS method. This reduced uncertainty in stimulus placement should reduce the variability of threshold estimates when measured with the CS method.

In comparing the ML method with other procedures, we chose methods and rules that are commonly used in psychoacoustic research. These are the k -down 1-up and the CS methods (the same methods as used with human subjects). In simulating the up-down procedure, we elected to use the conventional average of levels at reversal points to estimate threshold because of the universal usage of this rule. As noted, more efficient rules of data summary are available and should be considered by researchers. For example, one may make ML estimates at the end of the run from the track history (Schlauch & Rose, 1990; Watson & Fitzhugh, 1990), which has been shown to be more efficient and less biased than averaging of reversals. Nonetheless, most researchers use the averaging rule, and because our goal was to make comparisons with currently established psychophysical methods, the averaging rule was used in the following simulations.

Efficiency and the Ideal Sweat Factor for Various Procedures

Taylor and Creelman (1967) have described a very useful measure that allows comparison among techniques that track different probabilities or that use different rules and numbers of trials. They have defined the empirical sweat factor (S_{emp}) of a procedure as the product of the variance of estimates and number of trials; that is, $S_{emp} = n\sigma^2$. They have defined the ideal sweat factor as the binomial variance divided by the squared slope of the psychometric function; that is, $S_{ideal} = pq/F^2$. The efficiency of a psychophysical procedure is $\eta = S_{ideal}/S_{emp}$.

We simulated runs for the ML method, the CS method, and the k -down 1-up procedure for $k = 2$ to 5. The simulated observer had a .707 probability threshold of 2.0 log μ sec and the number of trials (n) was 60, except for $k = 4$ and 5, for which it was 80 to ensure a sufficient number of reversals. For each condition, 1,000 runs were simulated. At the end of each simulation, the standard deviation, $\sigma_{\log \mu sec}$, of log-transformed thresholds was calculated. For the CS method, four values of ΔITD (1.75, 2.0, 2.25, and 2.5 log μ sec [4-dB stepsize]) were used to cover the effective range of the observer's psychometric function (these values bracket probabilities between .62 and .94 for the simulated observer). The $\sigma_{\log \mu sec}$ of estimates for this procedure was based on thresholds determined from a weighted least squares fit (Finney, 1971) at the end of each run with the slope parameter ν fixed at unity. If the obtained threshold was greater than twice the stepsize above the .99 or below .51 probability on the observer's psychometric function, threshold was taken as that limit. This rule was adopted because the CS method occasionally generates data that are insufficient to bound threshold (Watson & Fitzhugh, 1990).

Although the stepsize between the four selected stimulus levels in the CS method was 4 dB (0.2 log μ sec), we also simulated other values (2, 8, and 12 dB) and found little difference in variability of estimates as a function of stepsizes. Unlike adaptive methods, the experimenter's uncertainty about threshold may have significant effects on the performance of the CS method. Watson and Fitzhugh (1990) simulated this uncertainty by maintaining a constant stepsize between selected stimuli while, between runs, the mean of the stimulus levels was a Gaussian random variable with a zero expected value and a standard deviation that was approximately half the range of the psychometric function. Their psychometric functions were quite steep (14 dB between .51 and .99 probabilities; with a 6-dB uncertainty), whereas our psychometric func-

Table 1
Measures of Sweat Factor and Efficiency

	CS	$k = 2$	$k = 3$	$k = 4$	$k = 5$	ML
Sweat factor	4.32	2.91	1.38	1.10	2.10	0.77
Efficiency	0.28	0.42	0.47	0.47	0.23	0.57

Note—CS, constant stimuli; k , k -down 1-up; ML, maximum likelihood.

tions are approximately 20 dB between .55 and .95 and 40 dB between .51 and .99. We assumed about half as much uncertainty as Watson and Fitzhugh, in favor of the CS method; a standard deviation of 10 dB, which is about one fourth the range of the psychometric function.

None of the methods showed any significant bias in threshold estimation. Sweat factors and efficiencies are shown in Table 1. The CS method produced the largest sweat factor and was the least efficient. The ML method was the most efficient. As for the up-down method, $k = 3$ and 4 were more efficient than $k = 2$ and 5. This latter observation for $k = 5$ is surprising because S_{ideal} for this case is quite small (i.e., S_{emp} was large). It seems that the poor efficiency for $k = 5$ may be related to the number of trials ($n = 80$); when k is high, the procedure requires a large n for efficient and unbiased tracking. When we increased n to 140, the sweat factors became nearly the same for $k = 3, 4,$ and 5 and in fact smallest for $k = 5$ (1.00, 0.90, and 0.86, respectively). However, most experimenters prefer $n < 100$, and in such a case, $k = 5$ would probably not be a very practical alternative. Kollmeier et al. (1988) and Schlauch and Rose (1990) also reported better efficiency for $k = 3$ than for 2.

Entropy and Information Gain

The information gained (or lost) from a system can be estimated from its entropy before and after a specified process (Shannon, 1949). The entropy of a system is a measure of its disorganization. For a Gaussian process, the entropy in bits (Shannon, 1949, p. 56) is

$$H = \log_2(\sigma \sqrt{2\pi e}) \tag{2}$$

The usefulness of this measure for psychophysical theory is that it characterizes a procedure and knowledge about the distribution of thresholds in the strict definition of information (entropy, H), gain or loss of that information (H_{Δ}), and rate of gain or loss (dH/dn). Watson and Fitzhugh (1990) have used this concept for evaluating the performance of psychophysical procedures and have shown that the Quest ML procedure (Watson & Pelli, 1983) is about 1.5 bits more informative than the CS method by the time n reaches 64. Watson and Fitzhugh assumed that the distribution of threshold estimates is approximately normal, and to the extent that this assumption holds, Equation 2 is a good estimate of the entropy of each method.

We define, in addition, H_{min} as the minimum attainable entropy by assuming that the well-known expression $\sigma_x = \sqrt{pq/F'}$ (Finney, 1971; Robbins & Monroe, 1951; Taylor, 1971; Wetherill, 1966) is approximately Gaussian distributed for $n \geq 20$.⁵ Figure 4 shows the empirically measured entropy H_e for the various procedures and H_{min} for $p = .92$. Because the minimum entropy in Figure 4 is a function of the tracking probability, H_{min} will be different for the various procedures. However, since this function is determined from the optimum tracking probability (Green, 1990), it is a lower bound on H_{min} .

Clearly, for all n , the most information about threshold is acquired from the ML method, which itself is approximately 0.5 bit less informative than ideal. By 60 tri-

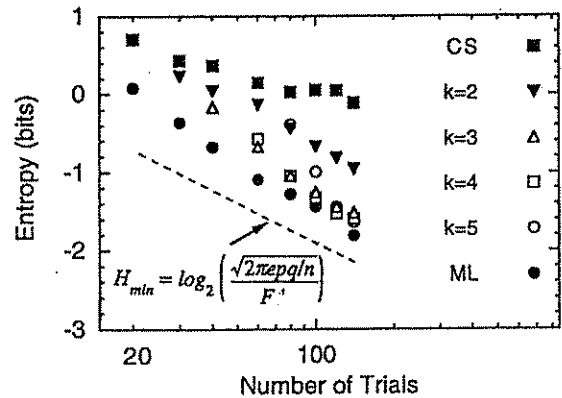


Figure 4. Entropy of different procedures determined from simulations. CS, constant stimuli; k , k-down 1-up procedure; ML, maximum likelihood. The dashed line (H_{min}) is minimum attainable entropy; p and $q = 1 - p$ are binomial probabilities, n is number of trials per run, and F' is the slope of the logistic psychometric function.

als, ML estimates are 1.3 bits more informative than estimates from the CS method. This value is very near the 1.5-bit difference (between QUEST and CS) reported by Watson and Fitzhugh (1990) for $n = 64$. The up-down method with $k = 2$ provides relatively little information about threshold, but it provides more than the CS. An interesting feature of the up-down method is that if n is small, higher k s are in some cases less informative than smaller k s; for example, compare the open square and triangle for $n = 60$, or compare open circle ($k = 5$) with other symbols when $n = 80$ or 100. On the other hand, the rate of gain in information, dH/dn , is much greater for $k = 5$, and by the time $n \geq 120$, the $k = 5$ rule provides as much information about threshold as $k = 3$ and 4. However, the need for such a large n , as we have noted, would make $k = 5$ a less desirable choice.

Finally, it should be clear that the lower the efficiency (η) of a procedure, the greater is the loss of threshold information from the use of that method relative to an ideal psychophysical procedure. This loss, which is the difference between the empirically measured and minimum entropies,

$$\text{information loss (bits)} = H_{\Delta} = H_e - H_{min},$$

may be related to Taylor and Creelman's (1967) measure of efficiency. Substituting the appropriate definitions and simplifying, the loss of information (in bits) from using a nonoptimum method is

$$H_{\Delta} = -0.5 \log_2 \eta \tag{3}$$

For Taylor and Creelman's efficiency measure bound by zero and unity, H_{Δ} is bound by $+\infty$ and zero, respectively.

SUMMARY AND CONCLUSION

ML methods may be applied to nonintensive auditory stimulus domains as long as the slope parameters (v in Equation 1) from psychometric functions are calculated

on a logarithmic stimulus scale; for the three stimulus domains examined, $v = 1$ is a reasonable value. Discrimination thresholds measured by the ML method with only 20 to 30 trials per run verified the accuracy of this method for such tasks by producing thresholds similar to those reported in the literature.

Data from the same human observers show that the ML method produces threshold estimates similar to those from the 2-down 1-up and CS methods. The standard error of threshold estimates were smallest for the ML method, followed by the up-down procedure. These results with human observers support the computer simulation results of Watson and Fitzhugh (1990), who also found more efficient estimates for the ML method than for the 2-down 1-up and CS methods. Cued variants of the up-down and CS methods, which have been suggested to produce more stable estimates, were not very efficient.

Efficiency of ML estimates were compared in simulations with k -down 1-up rules for higher and more efficient $k > 2$ rules (Kollmeier et al., 1988; Saberi & Green, 1996; Schlauch & Rose, 1990). Results showed a standard deviation for ML estimates that was smaller than all the up-down rules for both small and large numbers of trials per run (20 to 140); of the up-down rules, $k = 3$ and 4 were more efficient than $k = 2$ and 5, and all procedures were more efficient than the CS method.

In summary, we therefore recommend against the use of the CS method unless the experimental apparatus and setup prevent adaptively changing the stimulus level. If k -down 1-up procedures are to be used, it is best to use k values of 3 or 4. The $k = 2$ case is a popular rule; however, it is less efficient than the higher $k = 3$ and 4 rules. We also recommend avoiding the $k = 5$ rule; in spite of the lower ideal sweat factor associated with its higher tracking probability (87.1%), simulations show that the procedure is inefficient unless large numbers of trials are used on each run (e.g., 120).

As an alternative to the $k = 3$ and 4 rules, the ML method may be used to measure thresholds for nonintensive scales. Data from both human observers and simulations show that this method produces more efficient results with fewer trials. The ML method does, of course, require more restrictive assumptions than the k -down 1-up rule; however, these assumptions may be verified if the psychometric function has been characterized for that task. In addition to higher efficiency, the rapid and unbiased threshold estimation (with as few as 20 trials⁶) is another useful feature of the ML method. If a test population requires rapid measurements, such as in clinical tests (Laming & Marsh, 1988), or when large groups are to be tested, the ML method allows this additional feature.

REFERENCES

- ABEL, S. M. (1972). Duration discrimination of noise and tone bursts. *Journal of the Acoustical Society of America*, *51*, 1219-1223.
- AMERICAN NATIONAL STANDARDS INSTITUTE (1989). *Specification for audiometers*. New York: Author.
- CARR, C. E., & KONISHI, M. (1990). A circuit for detection of interaural time differences in the brain stem of the barn owl. *Journal of Neuroscience*, *10*, 3227-3246.
- CREELMAN, C. D. (1962). Human discrimination of auditory duration. *Journal of the Acoustical Society of America*, *34*, 582-593.
- EGAN, J. P. (1965). Masking-level differences as a function of interaural disparities in intensity of signal and of noise. *Journal of the Acoustical Society of America*, *38*, 1043-1049.
- EMERSON, P. L. (1984). Observations on a maximum likelihood method of sequential threshold estimation and a simplified approximation. *Perception & Psychophysics*, *36*, 199-203.
- EVANS, E. F. (1978). Place and time coding of frequency in the peripheral auditory system: Some physiological pros and cons. *Audiology*, *17*, 369-420.
- FINNEY, D. J. (1971). *Probit analysis* (3rd ed.). Cambridge: Cambridge University Press.
- FINNEY, D. J. (1978). *Statistical method in biological assay* (3rd ed.). London: Charles Griffin.
- FOLEY, J. M., & LEGGE, G. E. (1981). Contrast detection and near-threshold discrimination in human vision. *Vision Research*, *21*, 1041-1053.
- GREEN, D. M. (1988). *Profile analysis*. New York: Oxford University Press.
- GREEN, D. M. (1990). Stimulus selection policy in adaptive psychophysical procedures. *Journal of the Acoustical Society of America*, *87*, 2662-2674.
- GREEN, D. M. (1993). A maximum-likelihood method for estimating thresholds in a yes-no task. *Journal of the Acoustical Society of America*, *93*, 2096-2105.
- GREEN, D. M. (1995). Maximum-likelihood procedures and the inattentive observer. *Journal of the Acoustical Society of America*, *97*, 3749-3760.
- GREEN, D. M., & LUCE, R. D. (1975). Parallel psychometric functions from a set of independent detectors. *Psychological Review*, *82*, 483-486.
- GREEN, D. M., MCKEY, M. J., & LICKLIDER, J. C. R. (1959). Detection of a pulsed sinusoid in noise as a function of frequency. *Journal of the Acoustical Society of America*, *31*, 1446-1452.
- GU, X., & GREEN, D. M. (1994). Further studies of a maximum-likelihood yes-no procedure. *Journal of the Acoustical Society of America*, *96*, 93-101.
- HALL, J. L. (1968). Maximum-likelihood sequential procedure for estimation of psychometric functions. *Journal of the Acoustical Society of America*, *44*, 370.
- HALL, J. L. (1981). Hybrid adaptive procedure for estimation of psychometric functions. *Journal of the Acoustical Society of America*, *69*, 1763-1769.
- HENNING, G. B. (1966). Frequency discrimination of random amplitude tones. *Journal of the Acoustical Society of America*, *39*, 336-339.
- HENNING, G. B. (1974). Detectability of interaural delay in high-frequency complex waveforms. *Journal of the Acoustical Society of America*, *55*, 84-90.
- HENNING, G. B. (1980). Some observations on the lateralization of complex waveforms. *Journal of the Acoustical Society of America*, *68*, 446-454.
- HERSHKOWITZ, R. M., & DURLACH, N. I. (1969). Interaural time and amplitude jnds for a 500-Hz tone. *Journal of the Acoustical Society of America*, *46*, 1464-1467.
- JEFFRESS, L. A. (1948). A place theory of sound localization. *Journal of Comparative Physiological Psychology*, *41*, 35-39.
- KING-SMITH, P. E., GRIGSBY, S. S., VINGRYS, A. J., BENES, S. C., & SUPOWIT, A. (1994). Efficient and unbiased modifications of the QUEST threshold method: Theory, simulations, experimental evaluation, and practical implementation. *Vision Research*, *34*, 885-912.
- KLUMPP, R. G., & EADY, E. R. (1956). Some measurements of interaural time difference thresholds. *Journal of the Acoustical Society of America*, *28*, 859-860.
- KNUDSEN, E. I., & KONISHI, M. (1978). A neural map of auditory space in the owl. *Science*, *200*, 795-797.
- KOEHNKE, J., COLBURN, H. S., & DURLACH, N. I. (1986). Performance in several binaural-interaction experiments. *Journal of the Acoustical Society of America*, *79*, 1558-1562.

- KOLLMEIER, B., GILKEY, R. H., & SIEBEN, U. K. (1988). Adaptive staircase techniques in psychoacoustics: A comparison of human data and a mathematical model. *Journal of the Acoustical Society of America*, **83**, 1852-1862.
- LAMING, D. (1985). Some principles of sensory analysis. *Psychological Review*, **92**, 462-485.
- LAMING, D. (1986). *Sensory analysis*. London: Academic Press.
- LAMING, D. (1988). Précis of sensory analysis. *Behavioral & Brain Sciences*, **11**, 275-339.
- LAMING, D., & MARSH, D. (1988). Some performance tests of QUEST on measurements of vibrotactile thresholds. *Perception & Psychophysics*, **44**, 99-107.
- LEVITT, H. (1971). Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America*, **49**, 467-477.
- LICKLIDER, J. C. R. (1959). Three auditory theories. In S. Koch (Ed.), *Psychology: A study of a science* (Vol. 1, pp. 41-144). New York: McGraw-Hill.
- MACMILLAN, N. A., & CREELMAN, C. D. (1991). *Detection theory: A user's guide*. Cambridge: Cambridge University Press.
- MADIGAN, R., & WILLIAMS, D. (1987). Maximum-likelihood psychometric procedures in two-alternative forced-choice: Evaluation and recommendations. *Perception & Psychophysics*, **42**, 240-249.
- McKEE, S. P., KLEIN, S. A., & TELLER, D. Y. (1985). Statistical properties of forced-choice psychometric functions: Implications of probit analysis. *Perception & Psychophysics*, **37**, 286-298.
- NACHMIAS, J. (1981). On the psychometric function for contrast detection. *Vision Research*, **21**, 215-223.
- NUETZEL, J. M., & HAFTER, E. R. (1976). Lateralization of complex waveforms: Effects of fine structure, amplitude and duration. *Journal of the Acoustical Society of America*, **60**, 1339-1346.
- NUETZEL, J. M., & HAFTER, E. R. (1981). Lateralization of complex waveforms: Spectral effects. *Journal of the Acoustical Society of America*, **69**, 1112-1118.
- PAVEL, M. (1981). A new adaptive method for forced-choice experiments. *Journal of the Optical Society of America*, **71**, 215-223.
- PENTLAND, A. (1980). Maximum likelihood estimation: The best PEST. *Perception & Psychophysics*, **28**, 377-379.
- PITMAN, J. (1993). *Probability*. New York: Springer-Verlag.
- QUICK, R. F. (1974). A vector-magnitude model of contrast detection. *Kybernetik*, **16**, 65-67.
- ROBBINS, H., & MONRO, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, **22**, 400-407.
- ROBSON, J. G., & GRAHAM, N. (1981). Probability summation and regional variation in contrast sensitivity across the visual field. *Vision Research*, **21**, 409-418.
- ROUFS, J. A. (1974). Dynamic properties of vision—VI. Stochastic threshold fluctuations and their effect on flash-to-flicker sensitivity ratio. *Vision Research*, **14**, 871-888.
- SABERI, K. (1995). Some considerations on the use of adaptive methods for estimating interaural-delay thresholds. *Journal of the Acoustical Society of America*, **98**, 1803-1806.
- SABERI, K., & GREEN, D. M. (1996). Adaptive psychophysical procedures and imbalance in the psychometric function. *Journal of the Acoustical Society of America*, **100**, 528-536.
- SACHS, M. B., & KIANG, N. Y.-S. (1974). Rate versus level functions for auditory nerve fibers in cats: Tone-burst stimuli. *Journal of the Acoustical Society of America*, **56**, 1835-1847.
- SCHLAUCH, R. S., & ROSE, R. M. (1990). Two-, three-, and four-interval forced-choice staircase procedures: Estimator bias and efficiency. *Journal of the Acoustical Society of America*, **88**, 732-740.
- SHANNON, C. E. (1949). The mathematical theory of communication. In C. E. Shannon & W. Weaver (Eds.), *The mathematical theory of communication* (pp. 3-91). Urbana: University of Illinois Press.
- SHELTON, B. R., PICARDI, M. C., & GREEN, D. M. (1982). Comparison of three adaptive psychophysical procedures. *Journal of the Acoustical Society of America*, **71**, 1527-1533.
- SHELTON, B. R., & SCARROW, I. (1984). Two-alternative versus three-alternative procedures for threshold estimation. *Perception & Psychophysics*, **35**, 385-392.
- SIMPSON, W. A. (1988). The method of constant stimuli is efficient. *Perception & Psychophysics*, **44**, 433-436.
- SMITH, R. L. (1988). Encoding of sound intensity by auditory neurons. In G. M. Edelman, W. E. Gall, & W. M. Cowan (Eds.), *Auditory function: Neurobiological bases of hearing* (pp. 243-274). New York: Wiley.
- STEVENS, S. S., & GALANTER, E. H. (1957). Ratio scales and category scales for a dozen perceptual continua. *Journal of Experimental Psychology*, **54**, 377-411.
- TAYLOR, M. M. (1971). On the efficiency of psychophysical measurement. *Journal of the Acoustical Society of America*, **49**, 505-508.
- TAYLOR, M. M., & CREELMAN, C. D. (1967). PEST: Efficient estimates on probability functions. *Journal of the Acoustical Society of America*, **41**, 782-787.
- TAYLOR, M. M., FORBES, S. M., & CREELMAN, C. D. (1983). PEST reduces bias in forced choice psychophysics. *Journal of the Acoustical Society of America*, **74**, 1367-1374.
- TRAHIOTIS, C., & BERNSTEIN, L. R. (1990). Detectability of interaural delays over select spectral regions: Effects of flanking noise. *Journal of the Acoustical Society of America*, **87**, 810-813.
- TREUTWEIN, B. (1995). Adaptive psychophysical procedures. *Vision Research*, **35**, 2503-2522.
- URBAN, F. M. (1910). Die psychophysischen Massmethoden als Grundlagentheorie empirischer Messungen [Psychophysical methods as a basis of empirical measurement]. *Archiv für die Gesamte Psychologie*, **16**, 168-227.
- WATSON, A. B. (1979). Probability summation over time. *Vision Research*, **19**, 515-522.
- WATSON, A. B., & FITZHUGH, A. (1990). The method of constant stimuli is inefficient. *Perception & Psychophysics*, **47**, 87-91.
- WATSON, A. B., & PELLI, D. G. (1979). The QUEST staircase procedure. *Applied Vision Association Newsletter*, **14**, 6-7.
- WATSON, A. B., & PELLI, D. G. (1983). QUEST: A Bayesian adaptive psychometric method. *Perception & Psychophysics*, **33**, 113-120.
- WEIBULL, W. A. (1951). A statistical distribution function of wide applicability. *Journal of Applied Mechanics*, **18**, 292-297.
- WEIR, C. C., JESTEADT, W., & GREEN, D. M. (1976). A comparison of method-of-adjustment and forced-choice procedures in frequency discrimination. *Perception & Psychophysics*, **19**, 75-79.
- WETHERILL, G. B. (1966). *Sequential methods in statistics*. London: Methuen.
- WETHERILL, G. B., & LEVITT, H. (1965). Sequential estimation of points on a psychometric function. *British Journal of Mathematical & Statistical Psychology*, **18**, 1-10.
- ZWISLOCKI, J., & FELDMAN, R. S. (1956). Just noticeable differences in dichotic phase. *Journal of the Acoustical Society of America*, **28**, 860-864.

NOTES

1. Intensive and nonintensive stimulus cues involve fundamentally different neural processes (Evans, 1978; Jeffress, 1948; Licklider, 1959; Smith, 1988; Stevens & Galanter, 1957). Intensity is assumed to be coded by changes in neural discharge rate (i.e., rate-intensity functions; Sachs & Kiang, 1974), whereas nonintensive stimuli are in many cases differentiated in value by excitation of different neural populations (e.g., interaural delays; Carr & Konishi, 1990; Knudsen & Konishi, 1978). It is important to exercise care in extending results from one to the other class of auditory measurement.

2. Full descriptions of the maximum-likelihood procedure and implementation rules are given in Watson and Pelli (1983), Laming and Marsh (1988), and Green (1990).

3. The 1.7 constant in Equation 1 adjusts the logistic function in such a way that it produces probability ranges nearly equal to that of the Gaussian and Weibull functions, given the same threshold parameter μ . The slope parameter ν represents an internal nonlinear transformation of the stimulus scale and can generally be considered independent of the psychometric model used to fit the data. Increasing the value of ν increases the slope of the log psychometric function similarly for the logistic (Equation 1); Gaussian, $\phi(x/\mu)$; and Weibull, $W(x) = 1 - 0.5 \exp(-x^\nu/\mu)$, functions.

4. The measurement of psychometric functions and their slopes may be affected by the measurement technique (Taylor, Forbes, & Creelman, 1983). It has been suggested that the presence of a serial correlation between responses on successive trials of a CS method produces more variable or different estimates than do adaptive techniques. The slope estimates of the present study were extremely stable across subjects; nonetheless, one should be aware of possible effects of measurement method on characterization of slope parameters.

5. How closely does $P \sim \text{binomial}(n \approx 20, 0.707)$ approximate the Gaussian? If we denote $N(a \text{ to } b)$ as the normal approximation with continuity correction to a binomial probability $P(a \text{ to } b)$, then the *worst* error over all integers a and b with $0 \leq a \leq b \leq n$ is

$$W(n, p) = \max_{0 \leq a \leq b \leq n} |P(a \text{ to } b) - N(a \text{ to } b)| \approx \frac{|1 - 2p|}{10\sqrt{npq}},$$

and the largest error for $n \geq 20$ and $p = .707$ is 2% (Pitman, 1993, p. 103).

6. In simulations not reported here, we have verified the results of Watson and Fitzhugh (1990) that the ML method produces unbiased threshold estimates (within 0.5 dB) with as few as 20 trials.

(Manuscript received August 25, 1995;
revision accepted for publication August 7, 1996.)