Perception & Psychophysics
1992, 51 (3), 247-256

Estimation of psychometric functions from adaptive tracking procedures

MARJORIE R. LEEK
Army Audiology and Speech Center, Walter Reed Army Medical Center, Washington, D.C.

and

THOMAS E. HANNA and LYNNE MARSHALL
*Naval Submarine Medical Research Laboratory, Naval Submarine Base New London
Groton, Connecticut*

Because adaptive tracking procedures are designed to avoid stimulus levels far from a target threshold value, the psychometric function constructed from the trial-by-trial data in the track may be accurate near the target level but a poor reflection of performance at levels far removed from the target. A series of computer simulations was undertaken to assess the reliability and accuracy of psychometric functions generated from data collected in up-down adaptive tracking procedures. Estimates of psychometric function slopes were obtained from trial-by-trial data in simulated adaptive tracks and compared with the true characteristics of the functions used to generate the tracks. Simulations were carried out for three psychophysical procedures and two target performance levels, with tracks generated by psychometric functions with three different slopes. The functions reconstructed from the tracking data were, for the most part, accurate reflections of the true generating functions when at least 200 trials were included in the tracks. However, for 50- and 100-trial tracks, slope estimates were biased high for all simulated experimental conditions. Correction factors for slope estimates from these tracks are presented. There was no difference in the accuracy and reliability of slope estimation due to target level for the adaptive track, and only minor differences due to psychophysical procedure. It is recommended that, if both threshold and slope of psychometric functions are to be estimated from the trial-by-trial tracking data, at least 100 trials should be included in the tracks, and a three- or four-alternative forced-choice procedure should be used. However, good estimates can also be obtained using the two-alternative forced-choice procedure or less than 100 trials if appropriate corrections for bias are applied.

Adaptive testing procedures have become popular in psychophysical experiments over the past 20 years due to their efficiency and speed. In these procedures, the level of a stimulus on each experimental trial is determined by performance on previous trials. Such methods are characterized by their ability to converge rapidly on a given level of performance and to concentrate experimental trials in the vicinity of the final measurement of interest. Little experimental time and subject energy is expended on trials placed far from the point of interest on the psychometric function.

The trade for this high efficiency, however, is the loss of information about the underlying function that defines the subject's responses to a wide range of stimuli. Since few estimates of performance are obtained at levels removed from the threshold, the function may be well defined near that point, with considerably less precision at the extremes of the function. Although in many studies this price is easily paid (for example, when only the threshold at a specified point on the function is desired), there are instances when more complete descriptions of performance are desirable. This is most notably true when new phenomena are under investigation, when performance across stimulus levels cannot be estimated adequately on the basis of one performance level and a review of the pertinent literature. In such cases, adaptive methods may not be the procedure of choice, and speed of experimentation may have to be sacrificed to allow a more complete investigation of the entire psychometric function.

Some experimenters have ignored this problem and have generated psychometric functions based on the listener's performance on levels determined by the adaptive track. However, it has not been shown that reliable and unbiased estimates of psychometric function slope can be obtained from a post hoc analysis of trial-by-trial responses. While considerable attention has been devoted to evaluating the properties of threshold estimates from adaptive tracks, lit-

This work was supported by the American Society for Engineering Education and Naval Medical Research and Development Command, Navy Department, Research Work Unit No. 65856N-M0100.001-1021. It was undertaken while the first author was an American Society for Engineering Education Summer Faculty Fellow at the Naval Submarine Medical Research Laboratory, Groton, CT. The views expressed in this article are those of the authors and do not reflect the official policy or position of the Department of the Navy, Department of the Army, Department of Defense, or the U. S. Government. Correspondence should be addressed to M. R. Leek, Army Audiology and Speech Center, Walter Reed Army Medical Center, Washington, DC 20307-5001.

tle effort has been expended on assessing the quality of slope estimates. Levitt (1971), in his much-cited work on adaptive tracking techniques, discussed the optimal choice of signal levels for estimating the slope of a psychometric function. He suggested that placement of trials one standard deviation on either side of the mean could simultaneously provide reasonable estimates of both mean and slope, but he did not evaluate the accuracy of these slope estimates. Hall (1981) did provide information on the accuracy and reliability of threshold and slope estimates from post hoc fits to data obtained with a simulated four-alternative forced-choice PEST procedure. His results indicated that slope estimates calculated from those data are biased and only moderately reliable.

This paper reports the accuracy, precision, and efficiency of estimates of psychometric function slope from simulated adaptive threshold tracks under a variety of experimental conditions. These simulations demonstrate that not only is it reasonable to use the trial-by-trial data for slope estimation but that there are specific selections of experimental variables, such as psychometric procedure, track length, and step size, that will enhance those post hoc estimates. Further precision in slope estimation may be obtained by corrections for bias outlined here on the basis of the simulations.

Method

Several sets of simulations were performed with varying combinations of experimental variables as described below. For each condition, 1,000 independent simulated adaptive threshold tracks were generated on the basis of a known psychometric function. The form of the psychometric function used to generate the tracks and subsequently fit to the trial-by-trial data was

$$d' = m(E_s)^k, \tag{1}$$

where m is a measure of sensitivity (taken as 1 here), E_s represents signal energy, and k is the slope of the function. This function was shown by Egan, Lindner, and McFadden (1969) to be a good representation of human psychophysical performance. On each "trial," the current level of the adaptive track (E_s) and the input slope (k) were used to obtain a value for d'. This value was then transformed into a proportion correct using a calculation described by Elliott (1964) based on chance performance for the appropriate n-alternative forced-choice procedure (see also Green & Dai, 1991). This predicted performance level (the probability of detection, given the current signal level and psychometric function) was compared with a random number from 0 to 1. A random number less than the current probability of detection produced a correct response; otherwise an incorrect response was scored. The movement of the adaptive track was driven by the series of correct and incorrect responses.

The starting level for the "signal" on each adaptive track was selected randomly between 15 and 20 dB above the known threshold level. Changes in signal level were made in 5-dB steps until two reversals in the track had occurred. These trials were not used in the calculation of track length or for fitting psychometric functions, but they were treated as familiarization trials. This procedure allowed each track to start at a similar point relative to threshold, but not always at the same level. After two track reversals, the step size was changed to 2 dB, and signal level was adjusted according to the selected adaptive rule for that experimental condition until the specified number of trials was completed.

For each simulated track, the trial-by-trial data were organized as number correct and incorrect for each stimulus level visited by the track. A maximum likelihood procedure was used to derive the parameters producing the best fit of the assumed psychometric function (Equation 1) to the data. The maximum likelihood fits assumed binomial variability, that is, the likelihood, L, for a particular set of responses is given by

$$L = \prod_{i=1}^{N} P(s_i)^{k_i}[1-P(s_i)]^{n_i-k_i}, \tag{2}$$

where N signal levels were used, n_i signals were presented at level s_i, k_i correct responses were given at level s_i, and $P(s_i)$ is the probability of a correct response at level s_i for a given psychometric function. The likelihood was maximized by empirically searching the parameter space defining the possible psychometric function given in Equation 1.

Estimates of signal level needed for the target level of performance (71% or 79%) were calculated from the best-fitting psychometric functions. In addition, as is commonly done for up-down adaptive tracks, thresholds were estimated by averaging the upper and lower levels of each ascending run after the first two reversals (Levitt, 1971). Summary statistics of these three estimates (one slope estimate and two threshold estimates) were computed for the 1,000 simulations for each condition.

Each set of 1,000 adaptive track simulations was performed with a different combination of the following experimental variables:

Psychophysical procedure. Three psychophysical procedures were examined: two-alternative forced choice (2AFC), three-alternative forced choice (3AFC), and four-alternative forced choice (4AFC).

Target level of adaptive track. Either 71% or 79% correct level of performance was estimated by using a two-down, one-up rule or a three-down, one-up rule. That is, two (or three) consecutive correct responses led to a decrease of the signal level, and a single incorrect response led to an increase of the signal level.

Track length. Track lengths of 50, 100, 200, 300, 400, and 600 trials were investigated.

Slope of the underlying psychometric function. Psychometric functions with three different slopes were used to generate the adaptive tracks. The functions were derived from Equation 1 using slope values (k) of 0.5, 1.0, and 2.0 in units of 10 logd'/dB.

Results

The problem of degenerate psychometric functions. For the shorter track lengths, some of the simulations produced estimates of psychometric functions with an infinitely steep slope. This occurred when the maximum likelihood fitting procedure attempted to fit data reflecting chance performance at one signal level and perfect performance at the next higher level visited by the adaptive track. When the number of trials in the track was small and, therefore, the number of visits to a particular signal level was minimal, there was a higher probability that this situation would occur by chance alone than when there were sufficient trials in the track to sample adequately at each level visited. A similar problem has been described previously by O'Regan and Humbert (1989) for small sample sizes.

In the implementation of the fitting procedure, these indeterminate-slope conditions resulted in a very large slope estimate (the maximum allowed by the procedure). Summary statistics of the 1,000 simulations for experimental conditions that included a significant number of these instances, then, would be inappropriately biased and not truly reflective of the accuracy of the psychometric function reconstruction. However, because the infinite slopes *were* a characteristic of some of the simulated conditions, comparisons across conditions must include this factor in some respect.

The solution to this dilemma selected here was to ignore individual simulations producing indeterminate slope values but increase the number of simulations so that 1,000 valid estimates were obtained. This reasoning was based on what an experimenter might do if faced with this problem—that is, collect more data. The 1,000 simulations for each condition were analyzed to obtain measures of central tendency and variability. A third measure reflecting the efficiency of each experimental condition (the *sweat factor*, described below) included a factor representing the extra simulations necessary to obtain the set of valid estimates.

Accuracy of slope estimates. Figure 1 shows the geometric mean psychometric function slopes estimated from adaptive tracks generated by true underlying functions with slopes of 0.5, 1.0, and 2.0. Track length is shown on the abscissa. The left panel shows results for a target level of 71% correct; the right panel displays slopes for a target of 79% correct. The dotted, dashed, and solid horizontal lines indicate the underlying slopes of 0.5, 1.0, and 2.0, respectively. The psychophysical procedures are indicated by different symbols.

Slope estimates stabilized with longer track lengths, converging on the underlying slope value in all conditions. Little improvement in slope estimates was seen for tracks longer than 200 trials. For track lengths less than 200 trials, the estimated slopes were high, relative to asymptotic values. For tracks longer than 200 trials, the estimated slope was an accurate reflection of the true slope for all three generating slopes.

Small differences among procedures appeared primarily at the shallower underlying slope values. The most accurate slope estimates were consistently provided by the 4AFC procedure. The 2AFC procedure produced the poorest slope estimates at the the lower input slope levels. Essentially, no difference could be attributed to choice of target level for any of the procedures, indicating that this experimental decision had little effect on accuracy of slope estimates.

All conditions resulted in positively biased slope estimates at track lengths of 50 to 200 trials. Figure 2 demonstrates the nature of the bias for the 2AFC procedure at both target values. The dotted line on each panel shows the predicted results if the true underlying slope were accurately reflected in the slopes obtained from the post hoc fits to the trial-by-trial data. The solid lines represent the linear regression (on logarithmic coordinates) of the measured slope values on the true slopes. Note that for 100- and 200-trial tracks, the bias is represented primarily by a simple offset from the diagonal, but the 50-trial tracks produced slope biases dependent on the true slope.

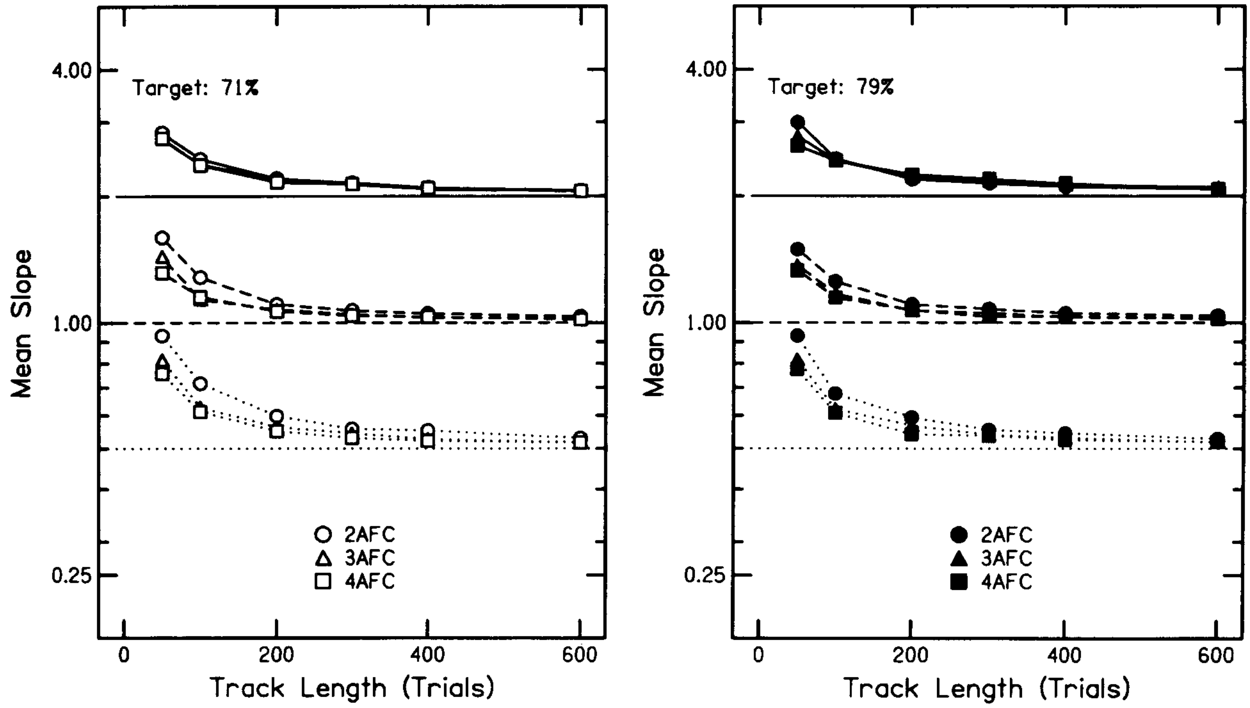

Figure 1. Geometric mean slope estimates from reconstructed functions as a function of length of adaptive track. The left panel and open symbols show results for a target level of 71% correct; the right panel and solid symbols are for a target level of 79% correct. The horizontal lines at slopes of 0.5, 1.0, and 2.0 indicate the underlying slopes of the functions that generated the tracks, and the results associated with each slope are shown in the same type of line. The parameter in each panel is the psychophysical procedure.

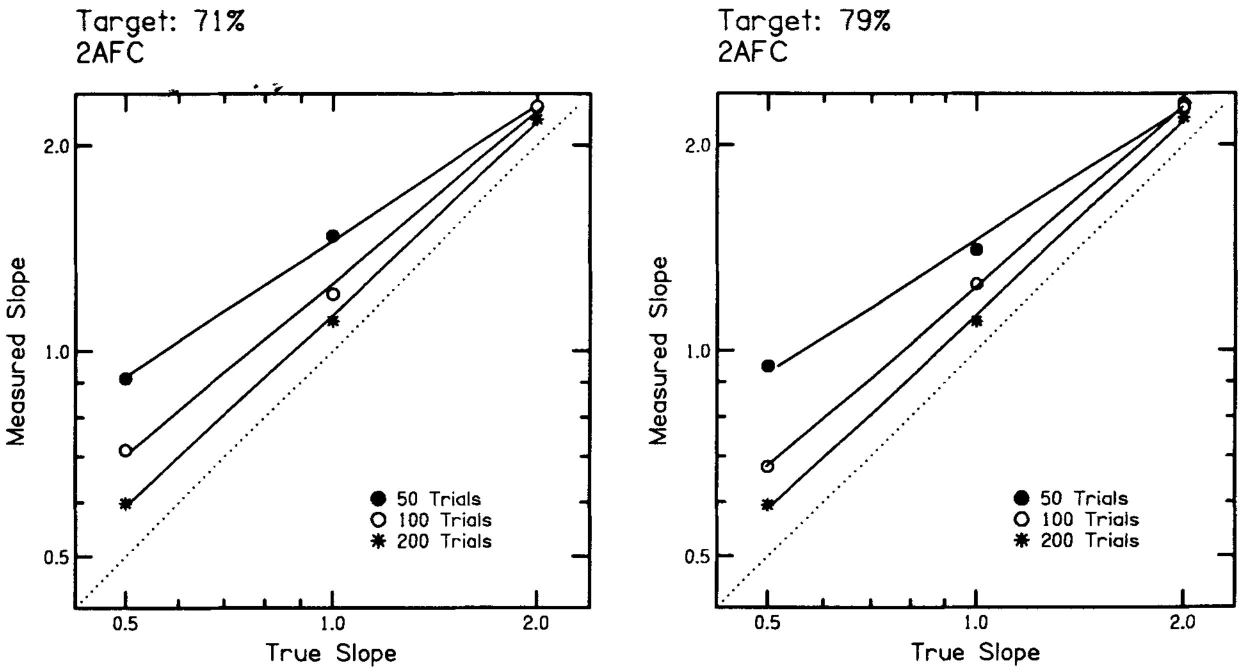

Figure 2. Geometric mean slopes measured in the simulations for three track lengths using a 2AFC procedure plotted as a function of the true underlying slope, at two target levels. The dotted line indicates results if no bias were present. The solid lines are linear regression lines.

This estimation bias may be reduced by applying correction factors to measured slope values based on the relationships shown in Figure 2. A reversal of the regression produces an equation of the form

$$k_T = a \cdot k_M^b, \tag{3}$$

where k_T is the true slope value, k_M is the measured slope, and a and b are fitting constants. Table 1 indicates the fitting constants calculated for each condition to correct for slope estimation bias. Measured slopes may be corrected by applying Equation 3 using the fitting constants shown in Table 1. If some estimation bias can be tolerated, a more convenient correction is to assume $b = 1.0$, and simply multiply the obtained slope by the appropriate value of a. This simpler correction is most appropriate for tracks of 100 or 200 trials, in which the bias changes minimally with slope, and when k_M is near 1.0.

Table 1
Fitting Constants for Bias Correction of Slope Estimates From Adaptive Tracking Procedures

Number of Trials	Target = 71% *a*	Target = 71% *b*	Target = 79% *a*	Target = 79% *b*
		2AFC		
50	.57	1.52	.56	1.56
100	.76	1.19	.78	1.14
200	.88	1.07	.88	1.06
		3AFC		
50	.67	1.20	.66	1.33
100	.84	1.06	.83	1.06
200	.91	1.02	.91	1.02
		4AFC		
50	.73	1.14	.70	1.27
100	.85	1.05	.85	1.06
200	.92	1.02	.92	1.00

Note—For any measured slope, k_M, the true slope, k_T, may be estimated using the equation: $k_T = a \cdot k_M^b$. 2AFC, 3AFC, and 4AFC = two-, three-, and four-alternative forced choice, respectively.

The exponential functions shown in Figure 2 indicate a compressive relationship between measured and true slope, particularly for the shortest track length. To estimate the limits of the compression, additional simulations were performed using the 71%-2AFC-50-trial-track condition and input slopes of 0.5 to 3.0 in steps of 0.1. Figure 3 illustrates the nature of the compressive relationship. Here the measured slope values are shown as a function of the true slopes, on linear coordinates. The solid line represents an exponential fit to these points. This figure suggests a ceiling effect in terms of slope estimates from the maximum likelihood fits. When any of a wide range of psychometric functions with steep slopes underlies performance, similar slope values will be estimated. An upper limit will artificially reduce the variability of slope estimates. Thus, for short adaptive tracks when the psychometric function is indeed steep, even though estimated slopes appear to demonstrate reasonably good reliability, it is still difficult to distinguish differences in true underlying slope. Figure 3 represents a "worst case," in that longer adaptive tracks demonstrate considerably less bias in slope estimation. However, it would be prudent to doubt very steep measured slopes whenever small sample sizes (short track lengths) are used.

Reliability of slope estimates. Figure 4 displays the variability in estimated slopes as a function of track length

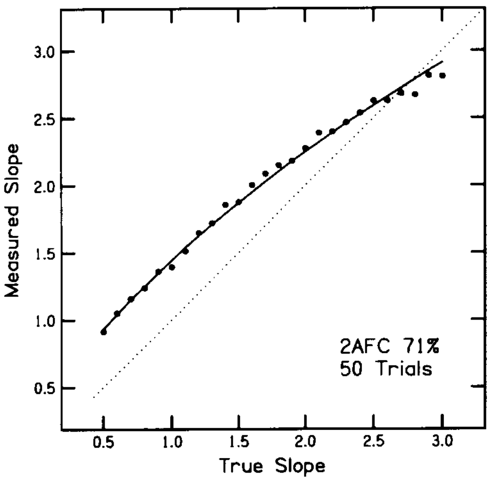

Figure 3. Measured slope as a function of true underlying slope for 50 trial tracks in the 2AFC–71% target condition. Data points are geometric mean slopes, and the solid line represents an exponential fit to the data. The dotted line represents the predicted results if no bias were present.

in terms of percent deviation from the geometric means. This corresponds to the standard deviation of the logarithms of the values used to calculate the means. The panels on this figure correspond to the three underlying slope values, with the psychometric procedure and target level as the parameters in each panel. Variability decreases systematically with longer track lengths. The major influence on variability is due to the track length, with only small differences resulting from different procedures or target levels. In general, slope estimates from tracks generated with 3AFC and 4AFC procedures show slightly better reliability than those from tracks generated with the 2AFC procedure, and the best estimates are produced when the true slope is 1.0.

Efficiency of slope estimation. The cost of the increased precision of measurement with increased track length shown in Figure 4 is the extra experimental time to produce longer tracks. While one would expect to trade longer track length for a decreased measurement error, the relative benefit of this trade may be evaluated by calculating a *sweat factor*, as defined by Taylor and Creel-

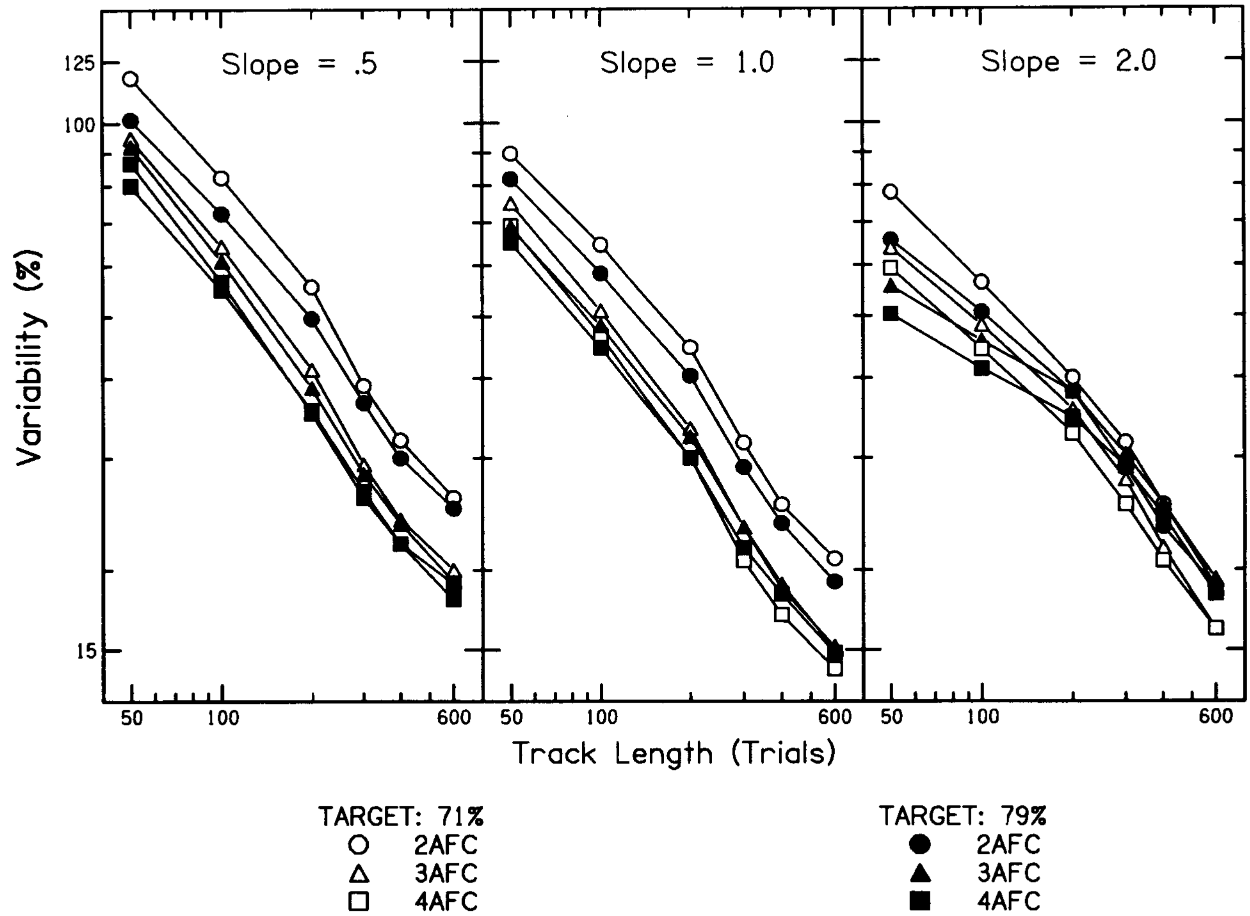

Figure 4. Relative variability of slope estimates for three true slope values. The parameters in each panel are the psychometric procedure and target level. Variability is shown as percent of the geometric mean for each condition, corresponding to the standard deviation of the logarithms of the individual slope estimates.

man (1967). This is a measure of the efficiency of a psychophysical procedure based on the variance of threshold estimates resulting from a specified number of trials. The sweat factor, K, is defined as

$$K = N \cdot \sigma^2, \tag{4}$$

where N is the number of trials in the track, and σ^2 is the variance of estimates produced. The smaller the value of K, the greater the efficiency of measurement.

The sweat factor is designed for comparisons of the amount of experimental effort necessary for a particular level of measurement reliability. To be a valid reflection of effort for the various experimental conditions simulated here, two modifications to the original sweat-factor parameters were made.

First, N (number of trials) was increased where appropriate to represent the extra "experiments" (i.e., simulations) required due to the degenerate slope estimates produced by some of the shorter track lengths as described earlier. Recall that simulations producing indeterminate slope values were redone so that 1,000 valid slope estimates were collected for each set of experimental parameters. For each condition that included extra simulations, the track length was multiplied by a factor representing the proportion of the simulations that were repeated. For example, if 50 simulations resulted in degenerate slope values, then a total of 1,050 simulations were carried out for that condition rather than 1,000, producing a measure of variability used in the calculation of sweat factor. Calculation of a sweat factor for this condition used a value of N equal to the number of trials in the track multiplied by 1,050/1,000, to reflect the comparable number of trials required.

The second variable in the sweat-factor calculation, the variance, was computed using the standard deviation of the logarithms of the slope estimates. However, these values were adjusted to take into account the estimation bias that was found for some conditions. When a bias correction is applied to estimates, the standard deviation of those estimates can change. For example, if all estimates are multiplied by a constant, the standard deviation will increase by that constant. When Equation 3 is applied to the slope estimates, the logarithms of the slope estimates are actually multiplied by b. Therefore, we multiplied the obtained standard deviations by b (shown in Table 1) to compute the sweat factors. This correction was, of course, only necessary for the shorter track lengths ($b > 1.0$), reflecting those conditions where the amount of bias in slope estimates was a function of the true slope (a simple additive relationship between the logarithms of the true and measured slope would not require this adjustment). Although this correction is only approximate, it gives a truer assessment of the relative efficiency of the various conditions. Intuitively, this adjustment reflects the fact that a small standard deviation is not necessarily indicative of a good measure if the predictions fall within a compressed range.

Figure 5 presents sweat factors as a function of track length. The top three panels show results for a target level of 71%; the bottom three panels show results for a 79% target level. Each panel shows calculations from one of the three underlying slope values, and the parameter in each panel is psychometric procedure.

For all conditions except the 79% target at the steepest input slope (lower right panel), the smallest sweat factors are associated with the 4AFC procedure and the largest are associated with the 2AFC procedure. Thus, in general, the 3AFC and 4AFC conditions provide the most efficient measurements, with essentially no effect of targeted level on the adaptive track. The 2AFC procedure consistently produces the least efficient measurement. However, it should be noted that, in actual practice, if stimulus presentations are sequential, the 3AFC and 4AFC may not result in the best use of experimental time, since there is an increase in the time to present each trial with those conditions. In terms simply of number of presentations per trial, the 3AFC procedure requires 50% more than does the 2AFC. In fact, Figure 5 suggests that the sweat factor for 3AFC is approximately two thirds that for 2AFC, perhaps reflecting a near-even trade between precision and number of presentations.

In most conditions, sweat factors decrease as track length increases up to 100 or 200 trials, with little change in efficiency for longer blocks of trials. In large measure, the greater sweat factors for short track lengths are conditioned by the compressive relationship between the measured and predicted slopes and the associated adjustments to the variances.

In general, the sweat factors are smallest for the moderate input slope value, with larger values for slopes of 0.5 and 2.0. This suggests that a slope of 1.0 is "well tuned" for efficient measurement of psychometric functions with the 2-dB step size used here. In describing the best placement of trials in measurement of the slope of psychometric functions, Levitt (1971) noted that the variability in slope measurements of a given psychometric function increased when trials were placed either too far from or too near to the midpoint of the function, and the optimal placement was related to the slope of the function. Thus, in an adaptive procedure, if the step size is too small for a given slope, too limited a region about the mean of the function will be sampled. If the step size is too large, trials will be spaced too broadly to provide a good estimate of the function. The best sweat factors shown in Figure 5 for the input slope of 1.0 are probably an empirical reflection of the fortunate combination of that slope with the step size used here.

Summary of threshold estimates. The ability of adaptive tracking procedures to estimate threshold has been evaluated in several previous articles, so only a brief summary of threshold estimates produced by these simulations is presented here.

At the 71% target level, thresholds estimated from the reconstructed psychometric functions were within 1 dB of the true threshold on the generating function for all conditions. The more traditional threshold measurement taken from adaptive procedures (the mean of the turnarounds) was also unbiased, except in the 2AFC procedure. These

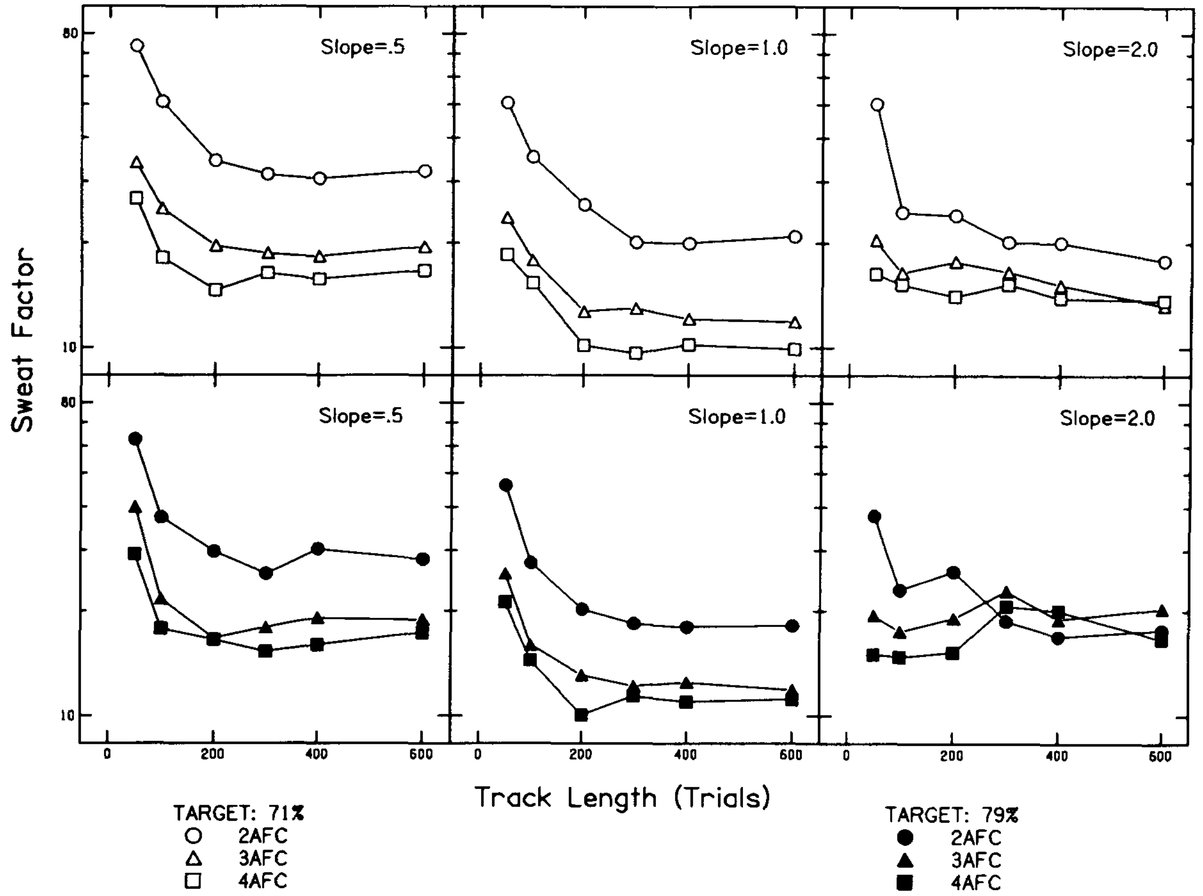

Figure 5. Sweat factors associated with slope estimates as a function of track length. The three columns of panels represent the three slopes. Target levels are shown in the two rows. The parameter in each panel is the psychophysical procedure.

values were always 2–3 dB lower than the true threshold, even with the longest track. Schlauch and Rose (1990) reported similar findings for this combination of 2AFC tracking at 71% correct.

The pattern of results for the higher target level was somewhat more complicated. The two steeper slope conditions (1.0 and 2.0) produced threshold estimates within 1 dB of true for both the reconstructed function method and the tracking method for all conditions. However, for the shallowest slope value, the 3AFC and 4AFC procedures were biased high by 1–2 dB for the tracking estimates for all track lengths and by the same amount with the reconstructed function method at the two shortest track lengths.

Discussion

Track length and step size. These simulations have revealed two experimental factors that are critical in the estimation of psychometric functions from data generated by adaptive tracking procedures. The first of these is sample size (track length), and the second is the selection of step size as reflected in the findings for the various input slope values.

At track lengths less than 200 trials, all conditions produced estimates biased high. Although this was most acute for the 2AFC procedure, some bias was seen for all conditions at short track lengths. Shelton, Picardi, and Green (1982), using a tracking target of 71% correct, suggested a minimum track length of about 50 trials for threshold estimation from adaptive tracking procedures. Longer tracks may be necessary to target 79% correct. It appears, however, that 50-trial tracks should be adequate if threshold measurement is all that is required; however, if slope estimates are also to be made, longer adaptive tracks, at least 100 trials, should be used.

The bias in slope estimates was effectively eliminated and the variability was reduced maximally when tracks were longer than 100 trials in all conditions. A judicious choice of psychophysical procedure and track target level would allow accurate and reliable slope estimates with as little as 100-trial tracks (i.e., 3AFC or 4AFC tracking at 71%).

In some cases, however, shorter tracks may be mandated by other factors in the experimental situation. The corrections suggested here will reduce the bias in slope estimation and should be valid for all except very steep underlying slopes (e.g., slopes greater than about 2.5). Fortunately, for most psychophysical measurement, slopes fall within the range of reasonable estimation and therefore may be estimated accurately with the application of these correction factors.

The limitations in slope estimation when the actual slope is steep may be partially compensated by the selection of a smaller step size. A steep underlying slope will produce the same adaptive track with a small step size as will a shallower underlying slope measured with a larger step size in the track. If a subject with an underlying slope of 2.0 is tested with a 1-dB step, the degree of bias and variability should be similar to those reported here for an underlying slope of 1.0 with a 2-dB step.

The simulations reported here necessarily included some assumptions as to form of the underlying function and method of fitting psychometric functions that may have aggravated any inherent problems in slope estimation. The maximum likelihood fitting procedure was bounded, in that a search region was specified. However, the upper boundary was never reached in any of the simulations except when an infinitely steep slope was indicated. That is, whenever an individual simulation resulted in a slope estimate reflecting the limits of the search space for the fitting procedure, the function was defined by a chance-to-perfect performance jump within one step, so that the function as defined was indeed infinitely steep. Further verification that the maximum slopes did not truly reflect a defined slope was sought by doubling the upper limit of the search space. Extending the limit did not result in more usable slope estimates.

O'Regan and Humbert (1989) reported similar problems when attempting to fit psychometric functions to small data sets, using both a maximum likelihood method and a probit fitting procedure. Therefore, this problem is not limited to data collected adaptively, nor is it limited to the maximum likelihood fitting method.

In practice, the problem of indeterminate slopes may be avoided by either increasing the number of trials in the track or by decreasing the step size.

Selection of psychophysical procedure. Several authors reporting simulated data have described the statistical properties of 2AFC as less than optimal, with biased threshold estimates and large variability (Kershaw, 1985; McKee, Klein, & Teller, 1985; Rose, Teller, & Rendleman, 1970). Shelton and Scarrow (1984), reporting real rather than simulated data, found only small differences in threshold measurements for a 2AFC compared with a 3AFC procedure, but reported greater variability associated with the 2AFC. Hall (1983) also favored a 3AFC procedure, arguing that the lower chance probability produces more stable thresholds and a faster convergence on threshold value. In comparing human data with the results of a proposed model of adaptive threshold measurements, Kollmeier, Gilkey, and Sieben (1988) were more equivocal in their recommendation of a 3AFC over a 2AFC procedure. Their model predicted more efficient measurement with 3AFC, but estimates of threshold measurement efficiency were inconsistent across human observers.

In general, the differences among procedures in these simulations were not large, but whenever there was a difference due to experimental conditions, the 2AFC was always poorest. Especially for short tracks, the findings as to accuracy, reliability, and efficiency of slope estimates presented here parallel the earlier work on threshold estimation. The 2AFC procedure produced less accurate slope estimates, with greater variability and with little improvement in efficiency from longer track lengths.

Although the results of these simulations as well as the reports of other authors would support the use of 3AFC or 4AFC procedures rather than the commonly used 2AFC method, many experimenters are reluctant to make the change. The most significant objections to the 3AFC and 4AFC procedures are the increased experimental time due to longer trial durations (when stimuli are presented sequentially) and the nontrivial issue of response bias in the multialternative procedures.

The first difficulty has been addressed recently by Schlauch and Rose (1990). These authors reported the usual finding of greater variability in threshold estimates from 2AFC procedures, relative to 3AFC and 4AFC procedures. However, they analyzed the efficiency of the procedures taking into account the differing amounts of experimental time required. The savings in trial duration for the 2AFC over the multi-interval procedures was not sufficient to compensate for the excess variability in measurement.

The second objection, a possible response bias in 3AFC or 4AFC procedures, has not been addressed adequately. Johnson, Watson, and Kelly (1984) reported significant differences in performance on the individual intervals of a 3AFC task. Performance tended to be best in the third interval and worst in the first interval. In an appendix to their article, these authors reported some evidence indicating that sensitivity to signals is not different in each interval but, instead, the effect of interval resulted from more central factors, such as attention and memory. This work calls into question the commonly accepted assumption of equal performance on all intervals of a procedure as long as the probabilities of signal presentation are equal across intervals, and this requires clarification through further research.

Selection of target level. There was little difference in the quality of slope estimation due to the choice of 71% or 79% correct as the target level for the adaptive track. Threshold estimates also were not strongly affected by choice of target level when the reconstructed function was used to extract a threshold value; however, when means of the turnarounds in the track were used, the 2AFC procedure coupled with the 71% target underestimated thresholds.

The results of these simulations suggest that the choice of target level is inconsequential to the estimation of slopes. However, to produce accurate and stable estimates of both threshold and psychometric function slopes, the higher level should be selected if a 2AFC procedure is to be used. (See also Green, 1990, for a discussion of the best placement of stimulus trials.)

Fitting the psychometric function. Psychometric functions with an underlying shape corresponding to a power function were fit to data generated by these simulated adaptive tracks using a maximum likelihood procedure. An earlier version of this work used probit analysis (Finney, 1971) to generate the tracks and to fit the functions (Leek, Hanna, & Marshall, 1988). This fitting procedure assumes a cumulative normal function, rather than a power function, and scales the dynamic range to correspond to the psychometric procedure employed (e.g., 50%-100% for a 2AFC procedure). Results of that study indicated slopes strongly biased for the 2AFC-71% conditions, with little improvement from longer track lengths. Probit analysis has also been used recently by O'Regan and Humbert (1989), Schlauch and Rose (1990), and Arehart, Burns, and Schlauch (1990) in assessing both threshold and slope estimation from various psychometric procedures, resulting in similar cautions against the use of the popular 2AFC-71% experimental conditions. However, results reported here using the maximum likelihood procedure and the power form of the underlying psychometric function indicated much less penalty associated with the use of those conditions, suggesting that, at least for slope estimation, other factors (e.g., track length and step size) take on more importance than either target level or psychometric procedure.

This discrepancy may be attributed to the statistical behavior of the probit procedure in fitting transformed functions with a range not extending from 0% to 100% correct. In a true 0%-100% function, data around the lower boundary have small error. However, in a forced-choice procedure, the lower boundary is associated with relatively large variability. The transformation for the probit fit changes the lower boundary to zero, but makes no transformation of the variability associated with that point. Since the probit analysis weights are also influenced by the variability at each level, this transformation may artificially alter the properties of the forced-choice data. Moreover, this form of bias due to the fitting procedure would be greatest for the 2AFC procedure because the transformed data would have the greatest variability near the lower boundary. The fitting problem would also be greater for the 71% condition because there would be more trials near the lower boundary. The maximum likelihood procedure makes no demands for rescaling the function and, therefore, does not vary the characteristics of the fit depending on chance levels, possibly resulting in greater stability of estimates.

Conclusions

Adaptive tracking procedures have been developed to quickly and efficiently obtain accurate performance measures at a targeted point on a psychometric function, with secondary concern for other characteristics of the function. Alternatively, accurate slope estimates require the placement of trials at more than one point on the function. Levitt (1971) suggested placing trials at plus or minus one standard deviation from threshold for estimating slope. Since a threshold value is seldom known before an experiment, slope measurement has traditionally involved the placement of trials across the dynamic range of the psychometric function, using a method of constant stimuli. These simulations have shown that, with a thoughtful choice of experimental procedures and sufficient trials in the track, accurate and reliable estimates of slope of the function can be obtained from the same adaptive track used to measure threshold. The best slope estimates resulted from either a 3AFC or a 4AFC procedure, with 100 or more trials in the track. However, good estimates can also be obtained using the 2AFC procedure or less than 100 trials if appropriate corrections for bias are applied. Moreover, because results from shallower underlying slopes can be interpreted as smaller step sizes, a judicious choice of step size in light of expected steepness of the underlying function could improve the reliability and efficiency of slope estimates.

REFERENCES

Arehart, K. H., Burns, E. M., & Schlauch, R. S. (1990). A comparison of psychometric functions for detection in normal-hearing and hearing-impaired listeners. *Journal of Speech & Hearing Research*, **33**, 433-439.

Egan, J. P., Lindner, W. A., & McFadden, D. (1969). Masking-level differences and the form of the psychometric function. *Perception & Psychophysics*, **6**, 209-215.

Elliott, P. B. (1964). Tables of *d'*. In J. A. Swets (Ed.), *Signal detection and recognition by human observers* (pp. 651-684). New York: Wiley.

Finney, D. J. (1971). *Probit analysis*. Cambridge: Cambridge University Press.

Green, D. M. (1990). Stimulus selection in adaptive psychophysical procedures. *Journal of the Acoustical Society of America*, **87**, 2662-2674.

Green, D. M., & Dai, H. (1991). Probability of being correct with 1 of *M* orthogonal signals. *Perception & Psychophysics*, **49**, 100-101.

Hall, J. L. (1981). Hybrid adaptive procedure for estimation of psychometric functions. *Journal of the Acoustical Society of America*, **69**, 1763-1769.

Hall, J. L. (1983). A procedure for detecting variability of psychophysical thresholds. *Journal of the Acoustical Society of America*, **73**, 663-667.

Johnson, D. M., Watson, C. S., & Kelly, W. J. (1984). Performance differences among the intervals in forced-choice tasks. *Perception & Psychophysics*, **35**, 553-557.

Kershaw, C. D. (1985). Statistical properties of staircase estimates from two interval forced choice experiments. *British Journal of Mathematical & Statistical Psychology*, **38**, 35-43.

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.

Leek, M. R., Hanna, T. E., & Marshall, L. (1988). *Psychometric function reconstruction from adaptive tracking procedures* (Report No. 1095). Groton, CT: Naval Submarine Medical Research Laboratory.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America*, **49**, 467-477.

McKee, S. P., Klein, S. A., & Teller, D. A. (1985). Statistical properties of forced-choice psychometric functions: Implications of probit analysis. *Perception & Psychophysics*, **37**, 286-298.

O'Regan, J. K., & Humbert, R. (1989). Estimating psychometric functions in forced-choice situations: Significant biases found in threshold and slope estimations when small samples are used. *Perception & Psychophysics*, **46**, 434-442.

Rose, R. M., Teller, D. Y., & Rendleman, P. (1970). Statistical properties of staircase estimates. *Perception & Psychophysics*, **8**, 199-204.

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.

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.

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

(Manuscript received August 7, 1989;
revision accepted for publication October 8, 1991.)