PEST: Efficient Estimates on Probability Functions

Article *in* The Journal of the Acoustical Society of America · April 1967

DOI: 10.1121/1.1910407

2 authors, including:

Martin Taylor
Martin Taylor Consulting
76 PUBLICATIONS **2,445** CITATIONS

Reprinted from THE JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA, Vol. 41, No. 4 (Part 1), 782–787, April 1967

PEST: Efficient Estimates on Probability Functions*

M. M. TAYLOR

Defence Research Medical Laboratories, Toronto, Canada

C. DOUGLAS CREELMAN

University of Toronto, Toronto, Canada

An adaptive procedure for rapid and efficient psychophysical testing is described. PEST (Parameter Estimation by Sequential Testing) was designed with maximally efficient trial-by-trial sequential decisions at each stimulus level, in a sequence which tends to converge on a selected target level. An appendix introduces an approach to measuring test efficiency as applied to psychophysical testing problems.

INTRODUCTION

As an experimenter, you wish to determine the level of an independent variable (for example, sound amplitude) that leads to some predetermined probability that a related event will occur on a single discrete trial (e.g., correct judgment of which of four intervals contained a signal). You wish to make the estimate in as few trials as possible for a given precision of the final estimate.

This problem arises in a number of fields besides psychophysics. Dixon and Mood (1948) developed their classical Up–Down method in the context of explosives research. Wetherill (1963) has considered the problem in the abstract, using bioassay as the exemplary problem area. PEST (Parameter Estimation by Sequential Testing) was developed initially for psychoacoustic research, and is described here in that context.

I. FIXED AND ADAPTIVE METHODS

The classical way to find the desired level of the independent variable (L_t) is to select a number of fixed testing levels which are thought to bracket L_t, to determine experimentally an event proportion for each, and to interpolate using a function (such as the normal ogive) that is presumed to approximate the true function. Such a method may be called a "fixed" method, since all the measures to be made are determined before the experiment is begun.

Adaptive methods, of which PEST is one, do not start with a prespecified set of conditions. Rather, the level to be presented at any one trial is determined by some portion of the history of the run. Adaptive methods generally attempt to make measurements at levels near L_t. They are therefore in principle more efficient than fixed methods, which include many measurements at far removed levels. This efficiency is gained at the expense of flexibility. Adaptive procedures yield information about the target level, but not about the nature of the psychometric function at probabilities remote from the target probability. Pollack *et al.* (1966) have aptly described adaptive methods as trading serendipity for efficiency.

All the adaptive methods which, to our knowledge, have been proposed for psychoacoustic work[1] follow the same basic pattern. A sequence of trials to yield one value of L_t begins by testing at some arbitrary initial level. After some finite number of trials, a new testing level is chosen. The choice of the new level depends on the results of the trials at the initial level. Trials are performed at each of a series of testing levels, the choice of each new level depending in some way on the history of the run. When the rules for the method indicate that the sequence is finished, some calculation yields a value of L_t. A sequence can usually be completed during a single experimental session, although various adaptive methods differ in speed and accuracy.

* Defence Research Medical Laboratories Research Paper No. 622.

[1] Zwislocki *et al.* (1958), Cornsweet (1962), Campbell (1963), and Wetherill and Leavitt (1965).

II. PEST PRINCIPLES

Rules for adaptive methods can differ in four ways: when to change levels, what level to try next, when to end the run, and how to calculate L_t at the end of the run. PEST is an attempt to define efficient rules which are at the same time convenient to use. In as few trials as possible, PEST makes a decision of known power about whether the current testing level is above or below L_t; each new testing level is placed so as to obtain nearly maximum information about the location of L_t; the run is terminated only when the current estimate is as precise as the experimenter requires; and the calculated value of L_t is just the last level selected for testing.

A. When to Change Levels in PEST

A Wald (1947) sequential likelihood-ratio test determines whether the current level yields an event proportion greater or less than the target probability, P_t. This test was shown to produce a decision of any given power in as few trials as possible, and is in this sense a maximally efficient test.

While computation of the Wald test needed to give any chosen power can be rather tedious, the test itself is simple, and the selected power has been found to have only second-order effects on the efficiency[2] of the PEST technique. A Wald test to aim at any desired target probability is very easy to construct for PEST.

Starting anew with each change in testing level, the experimenter keeps a running count of the number of correct responses $N(C)$ and the total number of trials T. After each trial, the test defines permissible upper and lower bounds on $N(C)$. If $N(C)$ falls between these bounds, another trial is made at the same testing level. If $N(C)$ falls on or above the upper bound, the decision is that the current level is too high, and if $N(C)$ is on or below the lower bound, the current level is taken to be too low.

If the current testing level were exactly L_t, the expected number of correct trials $E[N(C)]=P_t\times T$, after T trials. The sequential test bounds are given by the expected number of events plus and minus a constant.

$$N_b(C)=E[N(C)]\pm W,$$

where $N_b(C)$ is the bounding number of events after T trials, and W is a constant, called the deviation limit of the sequential test.

The power of the test and the rapidity with which a PEST sequence will converge depend in opposite ways on W. Small values of W yield quick but not very powerful decisions, while large values of W give, after many trials, decisions of great power. The rapidity, but not the power, of the sequential test is affected by P_t. Values of P_t far from 0.5 yield decisions relatively slowly when the true event probability associated with the current testing level is near P_t, whatever the value of W.

For PEST, we have found $W=1$ to be useful, and this is the value we ordinarily select. With $P_t=0.75$, a decision will then ordinarily take 2–25 trials. To find a signal yielding a d' less than unity (75% correct in a two-alternative forced-choice experiment) it becomes necessary to select a larger value of W, say $W=1.5$ or 2. This eliminates a potential source of bias in the estimate of L_t, as discussed below. While runs may take longer with these larger values of W, the resulting estimates will be correspondingly more precise.

B. What Level to Try Next in PEST

The first testing level may be chosen quite casually. Ordinarily, a few trials are saved by beginning with a level near an *a priori* estimate of L_t, but in psychoacoustics it is often helpful to the observer if at the beginning of a run he is given easy discriminations. The initial testing level in such cases may be chosen to give essentially perfect discrimination without affecting the final result.

When the sequential test yields a decision about whether the current level is too high or too low, a step is made in the appropriate direction, and testing starts anew at the new level. The size of the first step does not matter much. PEST effectively does not require the experimenter to know anything about the slope of the psychometric function, except its sign. In practical psychophysical work, subjects may be disturbed by overly large changes in the difficulty of the task, so that steps are usually kept rather small. In detection studies, we never make a step of more than 4 dB no matter what the stepping rules would indicate.

Step sizes after the first are determined by the history of the run. The following rules define the PEST technique we normally follow. They differ very slightly from a set we have privately circulated and publicly described (Taylor and Creelman, 1965), but the efficiency and bias of the two sets of rules are indistinguishable in our simulations. We believe, however, that when the target level is changing over time, and PEST is used to track the change, the present rules should provide more stable tracking than the earlier set.

1. On every reversal of step direction, halve the step size.

2. The second step in a given direction, if called for, is the same size as the first.

3. The fourth and subsequent steps in a given direction are each double their predecessor (except that, as noted above, large steps may be disturbing to a human observer and an upper limit on permissible step size may be needed).

4. Whether a third successive step in a given direction is the same as or double the second depends on the se-

[2] In the sense of Appendix A. Efficiency is an inverse function of both the accuracy, or variance in the results from a test procedure and the average number of trials required by it.

quence of steps leading to the most recent reversal. If the step immediately preceding that reversal resulted from a doubling, then the third step is not doubled, while if the step leading to the most recent reversal was not the result of a doubling, then this third step is double the second.

These rules were developed partly from intuition and partly from adjustment of the rules over many hours of computer simulation. It is possible that more efficient and equally convenient rules exist, but it is unlikely that they belong to any obvious family containing the present set. It is not possible to define exactly why these rules should be so efficient, but we can give an heuristic rationale for each.

1. If two sets of trials at different levels have given opposite answers, the target level is most probably between them. Subsequent testing at a level midway between will then yield the maximum information from the next set of trials.

2. When the current level is near the target level, the sequential test is quite likely to make a wrong decision. It is better to check a previous decision immediately than possibly to waste trials checking the three-quarter point as one would do if the sequential test were infallible. Rule **2** might perhaps be modified when the step size is large, but we have not attempted simulations with rules that change during the run.

3. Once the sequential tests begin giving the same answer in spite of several moves in the same direction, it becomes likely that the target level is not now in the region being tested, so the test level is then shifted rapidly. This rule is particularly important in psychophysics, because target levels can well change during a run. It also allows PEST to escape from regions in which it might be trapped by a wrong decision or two.

4. This rule prevents "rocking" instability, a series of levels repeated over and over, which can often happen if the third step is always doubled or always not doubled. In the old set of rules, the necessary asymmetry was introduced by a rule which caused the second step to be halved sometimes, and not at other times. In each case, simulation shows that the efficiency of the estimate is increased by about 20% by the inclusion of the rule.

C. When to Stop and How to Estimate Target Level in PEST

A PEST run stops when the rules call for a step of some predetermined minimum size. The choice of stopping step size determines the final precision of the estimate, but only weakly affects its efficiency.

The estimate of L_t is that level called for by the last small step. No trials are actually run at this level. The immediate availability of the datum without further analysis is a useful feature of the method. The final level may be slightly less precise than an estimate based on some sort of averaging, but in our experience, the experimenter's life is made very much easier because he does not have to record the history of the run and because he can make immediate decisions while the subject is still in the experimental situation. We have not attempted in our simulations to find measures of efficiency for averaged estimates following PEST runs.

III. COMPUTER SIMULATION

We have simulated many versions of PEST. The set of rules given above and those earlier circulated (Taylor and Creelman, 1965) proved clearly superior to any others tried, and they have been simulated over a wide range of conditions.

In all our simulations, we have assumed that the underlying psychometric function was quasilogistic, of the form

$$P(E)=1/S+(S-1)/S(1+e^{-L}),$$

where $P(E)$ is the event probability, S a constant (initially intended as the number of alternatives in a forced-choice experiment), and L the level of the independent variable.

The initial bias of the starting point has been varied over the range ± 32 logit units.[3] Within a block of runs with the same nominal starting point bias, the actual starting point has been selected randomly from a normal distribution with mean at the bias and standard deviation ranging in different experiments from zero to ten logit units. It is meaningless to try to determine the efficiency of a psychophysical technique if the simulation assumes that the experimenter always starts a known distance away from the target, because if he knew enough to do so, he should obviously not do the experiment at all. The initial step size has been varied from one-quarter to 16 logit units, the stopping step size from 2 to 1/1024 logit units, the sequential test deviation limit W has been variously $\frac{1}{2}$, 1, $1\frac{1}{2}$, 2, and 4, the target probability has varied from 0.16 to 0.98 and the lower bound $(1/S)$ of the psychometric function has been taken as 1, $\frac{1}{8}$, $\frac{1}{4}$, and $\frac{1}{2}$. We have included conditions in which the subject "lapses" (the whole psychometric function shifts to higher values on the level axis) on a varying proportion of trials, and conditions in which the probability of correct response depends on whether the last trial was correct. There are many thousand possible combinations of these conditions, and we have selected only a few hundred for our simulations.

The complete results are too voluminous to describe here; we intend a full discussion of them elsewhere. Here we give those general descriptions of the results necessary for an appreciation of the flexibilities and limitations of PEST.

[3] A logit unit is the scale unit of L in the logit formula. A logit is approximately 0.6 S.D. on the matching normal ogive.

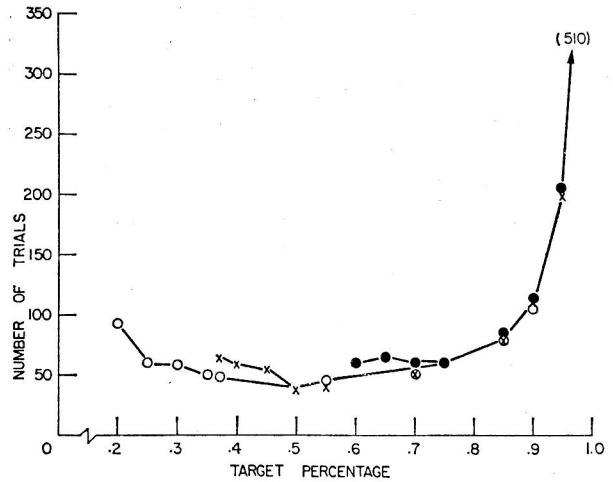

FIG. 1. Number of trials to reach a stopping rule of $\frac{1}{4}$ logit step, using forced-choice procedures. Simulation results from 400 Monte Carlo runs per point. ●: 2 AFC; ×: 4 AFC; ○: 8 AFC.

The result of most interest is that the efficiency of PEST is almost independent of conditions, with some minor exceptions, and that the estimate is unbiased, with the same minor exceptions. For general psychoacoustic work, PEST with a deviation limit of unity or greater is unbiased and highly efficient for target levels which give discriminability better than $d'=0.9$, no matter how many alternatives in the forced-choice experiment. The number of trials to reach a stopping step of $\frac{1}{4}$ logit unit, a realistic figure for actual experiments, is shown in Fig. 1 for various target probabilities and numbers of forced-choice alternatives. The expected number of trials becomes large for very high target probabilities, but with sufficient patience an experimenter can find levels which yield at least 95%. The more important result is the accuracy of the estimate. The variability of the data is shown in Fig. 2 for the same representative simulations. Here the number of alternatives becomes important, and at low target probability the variability of the measurement becomes quite large.

When the event probabilities are as independent, trial by trial, as they are in psychoacoustic work (Shipley, 1961), then no other condition shows bias in the final estimate. Slight shifts in efficiency appear here and there, mostly predictable from common-sense arguments, such as that trials are wasted by starting with small steps well off the target level. One loss of efficiency that is not immediately predictable is that which results from stopping at too large a step size. We recommend that the starting and stopping step sizes be chosen to allow at least four reductions in step size before the end of the run.

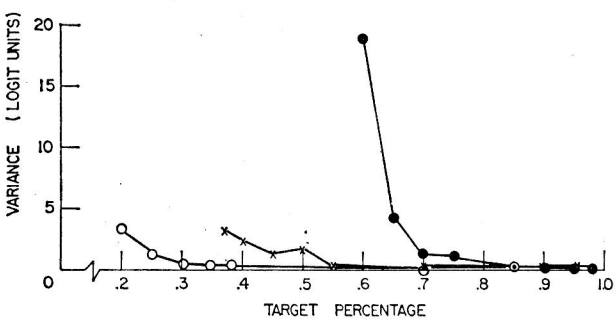

FIG. 2. Variance of results from PEST simulations, calculated around the mean level. ●: 2 AFC; ×: 4 AFC; ○: 8 AFC.

It should be particularly emphasized that we find no sign of bias in the PEST estimate when the target probability is very high. Our simulations have attempted P_t as high as 0.98 and found neither bias nor loss of efficiency.

When lapses occur the estimate of target level is slightly raised, but not very much. For example, PEST aiming at 74% in a two-alternative forced-choice experiment actually attains the level associated with 76% correct in the normal condition if the lapse probability is 0.20 and the amount of shift was 2 logit units. For most purposes one may ignore the problem of lapses.

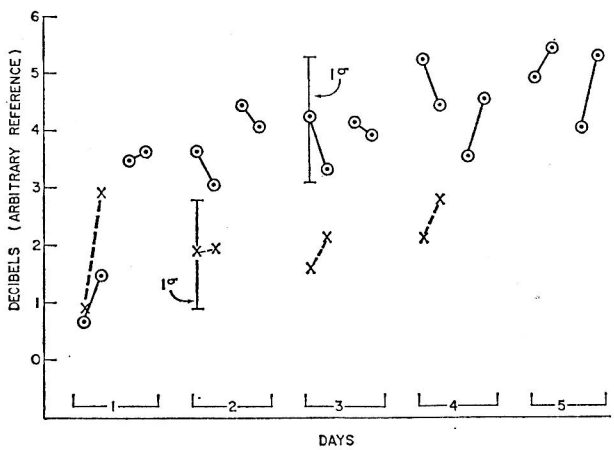

FIG. 3. PEST: results from naive observers with no prior experience, two-alternative forced-choice detection of a 1000 Hz signal in noise. Vertical bars indicate 1 standard deviation. ×: 19 students; ⊙: 14 housewives. Lines connect points from a single sitting.

IV. PEST WITH HUMAN SUBJECTS

We have used PEST in psychoacoustic work since the spring of 1965, and have found it easy to learn and convenient to use. It is usable equally with naive and with experienced subjects. When the independent variable is signal amplitude, we typically limit the step size to 4 dB and stop when a step of 0.5 dB is called for. Under these conditions, runs with a target probability of 75% or 80% in a 2AFC experiment typically take 20–80 trials, averaging about 45 trials. With experienced subjects the between-trial standard error of measurement is about 0.8 dB; with naive subjects it is about 2 dB.

Figure 3 shows results obtained with PEST that would have been difficult to obtain any other way. In separate experiments, 14 housewives and 19 university

students who had never been in a psychoacoustic experiment were given a 2 AFC task of detecting a 1000 Hz tone in white noise. Two PEST runs, totaling about 120 trials on the average, were run with only a slight break each day. The total time was 10–15 min for each pair of runs. The housewives were given two such double sessions each day with a rest between. Figure 3 shows that from the first 10 min to the second, the housewives improved an average of about $2\frac{1}{2}$ dB, while the students showed most improvement within the first 5 min. Over the second sitting and the remaining days, there was a slow improvement of at most 1 dB over all.

Note that the data in Fig. 2 represent the very early stages of learning to perform the 2 AFC task. A technique that took as much as 10 min would lose the facts that performance during the second 5 min is almost as good as on the following days and that most of the learning happens in the first 5 min.

Apart from psychoacoustic problems, we have used PEST in experiments in visual brightness discrimination, and in an experiment on visual figural aftereffects. For the aftereffect experiment, where the aim was to find a Point of Subjective Equality (PSE), the target probability was 0.5. This P_t permits steps to be made following each separate response. The variable dimension was the tilt of a line. With steps starting at 8° and stopping on the call for a $\frac{1}{4}$° step, the runs typically took only 10–20 trials and resulted in a standard error of measurement of the order of $\frac{1}{2}$°.

In sum, PEST is efficient, flexible, useful with naive or trained observers, robust, and easy to run. We find it to compare very favorably with other adaptive methods known to us.

Appendix A. Efficiency of Psychophysical Measurement

In thermodynamics, as well as in common usage, the term *efficiency* relates to the amount of work required to do a particular job, relative to the amount of work needed by some theoretical ideal device. We use "efficiency" in the same sense. The job here is the determination of the "target level" with some specified error variance, and the amount of work required is the number of trials needed to attain the desired precision of measurement. Once an ideal measuring device is specified, the numerical efficiency of a measuring technique is known.

For measurement in general, there is an arithmetic trading relation between precision and number of repeated determinations, so that it becomes possible to compare efficiency not only of techniques which yield the same variance (do the same job), but also of techniques which differ in precision. In general, when a measurement depends on N equally valuable independent determinations, the variance of the result is inversely proportional to N. For a given technique, this relation may be written: $N\sigma^2 = K$, where K is a constant. K reflects the number of determinations required to give any specified variance, and hence is a suitable index with which to compare the efficiencies of various techniques. Since K represents the amount of work in a particular measurement, we call it the "sweat factor" of the method.

In the problem for which PEST is designed, the variance in question is that of the target level estimator, while the number of determinations N is the average number of trials needed to attain an estimate. In our simulations many independent determinations of each target level were made, under constant conditions, to find an empirical value for the variance of the PEST estimator, and the mean number of trials (see Figs. 1 and 2). We thus have an empirical sweat factor for PEST under many different sets of conditions.

To determine the efficiency of PEST, or of any other technique aimed at the same problem, it is necessary to define an ideal measuring technique to be used as a yardstick. We have defined a somewhat conservative ideal device, which is not realizable in practice.

Imagine an omniscient experimenter who knows, trial by trial, the nature of the random process leading to the usual binominal variability of the probability measure. The omniscient experimenter selects, for each block of trials, that single level which will give an event proportion exactly equal to the target probability. Because the random process is in fact different from block to block, the level selected by the omniscient experimenter will also be different from block to block. After many blocks, the levels selected will be distributed, and we take the standard deviation of the distribution as the estimate for the ideal device.

Knowing the form of an assumed psychometric function, we can calculate the asymptotic variance of this ideal measurement, and hence we can compute the efficiency of PEST in the various conditions.

If the test level is fixed over several blocks of trials, the event proportion in individual blocks will have a binomial distribution that can be approximated by a

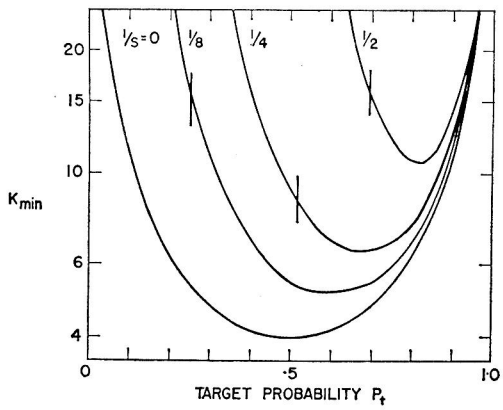

FIG. A1. Ideal sweat factor (see text) for target P_t at various event probabilities. Parameter is the lower bound of a forced-choice psychometric function.

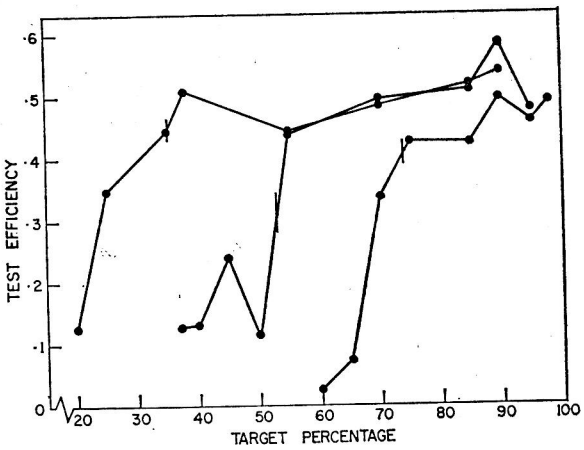

FIG. A2. Relative efficiency of PEST as a function of P_t and experimental procedure. Calculation based on Figs. 1, 2, and A1. Left curve: 8 AFC; center curve: 4 AFC; right curve: 2 AFC.

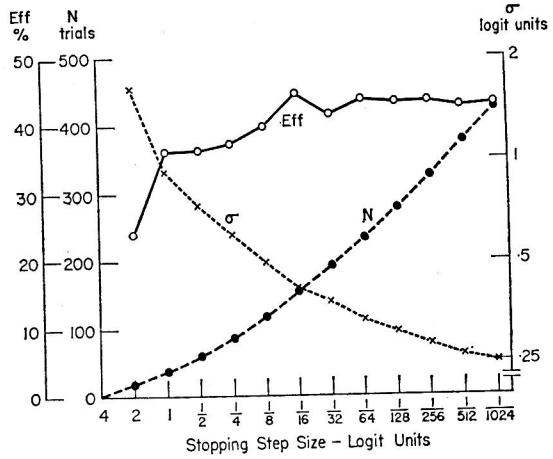

FIG. A3. Selection of various stopping rules (size of the smallest step used) affects efficiency only slightly. An experimenter can increase experimental precision at a cost in trials per experimental run. Data from Monte Carlo simulations.

normal distribution with variance $p(1-p)/N$, where N is the number of trials in a block, and p is the "true" event probability. Now consider the case when only a single block is run. If the psychometric function for probabilities in the neighborhood of p has a slope $dp/dL=V$, then it is possible to calculate the variance of the likelihood function relating levels on the psychometric function to the particular obtained probability. This variance we need in order to compute an ideal sweat factor with which to determine the numerical efficiency of PEST. The sweat factor is the product of the variance by N, and is given for the ideal by $K_{\min}=p(1-p)/V$. The ideal sweat factor is denoted by $K_{\min}$ because it is the lowest possible sweat factor for determination of the target level corresponding to the target probability p. Note that $K_{\min}$ is independent of N, to the extent that the approximations hold for small N.

Since the efficiency of a method is given by the relative amount of work necessary to do a job, it is natural to define the efficiency of a measurement technique by the ratio of the number of determinations required by the ideal to the number required by the actual technique to attain the same variance. But since the trading relation between number of trials and variance holds, the same answer may be obtained more flexibly by use of the ratio of sweat factors. Accordingly, we define phychometric test efficiency as $K_{\min}/K_{\text{empirical}}$.

Figure A1 shows the ideal sweat factor for the quasi-logistic psychometric functions used in our simulations. Note that these sweat factors are independent of the measuring technique and indicate the intrinsic difficulty of making a measurement at any particular target probability. The work required for a measurement at extreme probabilities is evident. PEST efficiency is around 40%–50%, except for low target probabilities (below the vertical bars in Fig. A1).

Figure A2 shows efficiency calculated for the data of Figs. 1 and 2. The important feature is that efficiency does not decrease at high target levels, as might be suspected on the basis of Fig. 1. The difficulty is inherent in any technique. The increasing variance at low target probabilities does, however, lead to lowered efficiency.

Figure A3 shows results of simulations where the size of the step defining the stopping rule is taken to be successively smaller. The mean number of trials to reach the stopping rule is shown along with the obtained variance and the resultant calculated efficiency. This shows the kind of choice open to the experimenter in his selection of decision rules. After the first few steps, further reduction will increase precision, directly compensated for by increased trials required for each measurement. The relative efficiency is barely affected by such choices.

REFERENCES

CAMPBELL, R. A. (1963). "Detection of a Noise Signal of Varying Duration," J. Acoust. Soc. Am. *35*, 1732–1737.

CORNSWEET, T. N. (1962). "The Staircase-Method in Psychophysics," Am. J. Psychol. *75*, 485–491.

DIXON, W. J., and MOOD, A. M. (1948). "A Method for Obtaining and Analyzing Sensitivity Data," J. Am. Statist. Assoc. *43*, 109–126.

POLLACK, I., HEADLY, P., and MASS, E. (1966). "Modest Computer-Controlled Psychoacoustical Facility," J. Acoust. Soc. Am. *39*, 1248(A).

SHIPLEY, E. F. (1961). "Dependence of Successive Judgments in Detection Tasks: Correctness of the Response," J. Acoust. Soc. Am. *33*, 1142–1143.

TAYLOR, M. M., and CREELMAN, C. D. (1965). "PEST: A Rapid Technique for Finding Arbitrary Points on a Psychometric Function" (Psychonomic Society, Chicago, Ill.).

WALD, A. (1947). *Sequential Analysis* (John Wiley & Sons, Inc., New York), esp. pp. 88–105 and 196–199.

WETHERILL, G. B. (1963). "Sequential Estimation of Quantal Response Curves," J. Roy. Statist. Soc., Ser. B, *25*, 1–48.

WETHERILL, G. B., and LEAVITT, H. (1965). "Sequential Estimation of Points on a Psychometric Function," Brit. J. Math. Stat-Psychol. *18*, 1–10.

Reprinted from THE JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA, Vol. 42, No. 5, 1097, November 1967

Erratum and Note:

PEST: Efficient Estimates on Probability Functions

[J. Acoust. Soc. Am. **41**, 782–787 (1967)]

M. M. TAYLOR AND C. D. CREELMAN
Defence Research Establishment Toronto, Downsview, Ontario, Canada

In the Appendix to the above paper (p. 787), the formula for the ideal sweat factor is incorrectly given. The correct expression is

$$K_{\min}=p(1-p)/V^2.$$

The superscript was omitted from the published paper. We wish to thank B. Lopes Cardozo for bringing this error to our attention.

A program has been written for the PDP-8 family of computers to run PEST in real-time experimental control applications. Copies of the program are available from the authors.