Perception & Psychophysics
1990, 47 (1), 87-91

The method of constant stimuli is inefficient

ANDREW B. WATSON
NASA Ames Research Center, Moffett Field, California

and

ANDREW FITZHUGH
Hewlett-Packard Laboratories, Palo Alto, California

Simpson (1988) has argued that the method of constant stimuli is as efficient as adaptive methods of threshold estimation, and has supported this claim with simulations. We show that Simpson's simulations are not a reasonable model of the experimental process, and that more plausible simulations confirm that adaptive methods are much more efficient than the method of constant stimuli.

A common problem in psychophysics is to estimate the signal strength required by the observer to accomplish some perceptual task. A general approach is to estimate, from performance on repeated trials, the probability of success at each of a number of strength levels. In a two-alternative forced-choice experiment, the probability of success typically rises from 50% to 100% as strength increases. From this psychometric function, one can estimate the *threshold* strength—that is, the strength yielding a preselected probability of detection.

There are two general approaches to the testing process: adaptive and nonadaptive. In the former, the distribution of trials at various strengths is dependent on the outcome of previous trials. In the latter, the distribution is specified in advance. The nonadaptive approach is often called "the method of constant stimuli" (MCS). A major concern for both approaches is efficiency: The fewer the trials required to reach a particular standard deviation of the estimate, the better. Trials remote from the threshold provide little information, and they reduce overall efficiency. Adaptive procedures have been argued to be more efficient than nonadaptive ones, because they permit the outcome of previous trials to be used to place future trials at efficient testing locations (Cornsweet, 1962; Emerson, 1986; Hall, 1981; Levitt, 1971; Lieberman & Pentland, 1982; Taylor, 1971; Taylor & Creelman, 1967; Watson & Pelli, 1979, 1983; Wetherill & Levitt, 1965).

Arguing against this conventional wisdom, Simpson (1988) has recently asserted that the method of constant stimuli is as efficient as adaptive methods of data collection. To support this claim, he offers simulations of MCS and an adaptive method (Lieberman & Pentland, 1982).

Simulations of psychometric procedures are meaningful only if they are a reasonable model of the "real life" testing situation. In particular, the knowledge assigned to the simulated experimenter must be plausible. The purpose of this paper is to show that, under reasonable assumptions about the experimenter's knowledge, adaptive methods are much more efficient than MCS.

The plan of the discussion is as follows: First, we present new simulations of MCS and three adaptive methods, in order to illustrate that—at least for the conditions simulated—adaptive methods are much more efficient than MCS. Second, we note the flaw in Simpson's simulations that led him to the mistaken conclusion that MCS is efficient.

Correspondence may be addressed to Andrew B. Watson, Vision Group, NASA Ames Research Center, Moffett Field, CA 94035.

SIMULATIONS

Data Format

In considering the quality of a psychometric procedure, we are concerned with both accuracy and bias, and with how they both depend on the number of trials collected. This information is provided by plots of the standard deviation and mean of the distribution of estimates as a function of number of trials. In the present study, each complete simulation consisted of a number of runs of a given procedure under particular conditions. Each run consisted of a number of blocks, each containing a certain number of trials. In each run, after each block, a threshold was estimated. From these data, we obtained the mean and standard deviation of the threshold estimate as a function of the number of trials. For MCS, we used 500 runs with 10 blocks, each of 10 trials. For the other methods, we used 1,000 runs of 16 blocks, each of 4 trials.

Simulated Observer

The observer was simulated with a Weibull psychometric function,

$$P(x) = \mathrm{Min}\left\{1-\delta,\ 1-(1-\gamma)\exp\left[-\left(\frac{\epsilon x}{\alpha}\right)^{\beta}\right]\right\}, \quad (1)$$

where $P(x)$ is the probability of a correct answer at strength x. The parameters of this equation, and their default values, are α (threshold) = 1 (0 dB), β (slope) = 3.5, γ (guess rate) = 0.5, δ (finger error rate) = 0.01, ϵ (ideal test point) = 1.189 (1.5 dB). These parameters are dis-

cussed elsewhere (Watson & Pelli, 1983). This function has been shown to be a good representation of human performance in many situations (Nachmias, 1981; Watson, 1979). The Weibull function is very similar to the logistic function used by Simpson.

Experimenter Knowledge

We assume that before data collection begins, the experimenter knows the location of threshold to within some error. We assume that this error is normally distributed (on a logarithmic strength axis), with a mean of zero and a standard deviation of 6 dB. The mean of zero says that, on the average, the experimenter has no bias for guessing threshold to be above or below the true value. Starting points (guesses) for each experiment are selected from this normal distribution.

Procedures Simulated

QUEST: Each trial is placed at the mode of the current posterior density for threshold (Watson & Pelli, 1979, 1983). The procedure is initialized with a Gaussian prior density whose mean is the initial guess. The representative adaptive method used by Simpson (Lieberman & Pentland, 1982) is essentially identical to QUEST.

UDTR-ML: the "transformed up-down" method (Levitt, 1971; Wetherill & Levitt, 1965). This widely used method increases strength by one step after one error, and decreases it by one step after two correct responses. We used a step size of 1 dB. The first trial was placed at the initial guess. The ML indicates that thresholds are estimated by the maximum-likelihood method.

UDTR-AVG: the same as UDTR-ML, except that thresholds are estimated by taking the average of the last four reversals. This estimation method is also in wide use.

MCS: The method of constant stimuli as described by Simpson (1988). An equal number of trials is placed at each of five levels, equally spaced on a logarithmic strength axis. The step between levels (*grain*) was varied in different simulations, between 1 and 6 dB. Before each experiment, the center level was set to the initial guess of threshold.

Threshold Estimation

All thresholds (except UDTR-AVG) were estimated by fitting Equation 1 to data, using a maximum-likelihood method (Watson, 1979). All parameters except α were fixed at their default values. MCS not infrequently generates data that are insufficient to bound the threshold estimate. This may occur, for example, when all the trials are correct. In these unbound cases, maximum likelihood may occur at plus or minus infinity. Our procedure to deal with these cases is as follows: We note that the psychometric function equals 99% at about $12/\beta$ dB above threshold, and 51% at about $36/\beta$ dB below. Thus, if threshold is outside the range guess $-$ 2 grain $-$ $12/\beta$, guess + 2 grain + $36/\beta$, the data are essentially useless. Accordingly, we evaluate the likelihood over this interval, and if the maximum is at either bound, that bound is taken as the estimate of threshold.

Simulations of MCS were performed in Mathematica (Wolfram, 1988). All other methods were simulated by programs written in C on a UNIX workstation.

RESULTS

Method of Constant Stimuli with Various Grains

The method of constant stimuli has only two important parameters that the experimenter can set. The first is the number of sample points, which Simpson set at a typical value of five, and which we shall not investigate further. The second is the step in strength between sample points, which we call the grain. Figure 1 shows the effect of grain upon the performance of MCS. For a small grain (1 dB), the standard deviation declines slowly, indicating poor performance. The best performance is obtained with a grain of about 4 dB.

The poor performance of MCS with a small grain is a consequence of threshold's often lying outside the testing interval. Reasonable performance can only be obtained by setting the grain large enough to insure against this possibility. But this large grain also ensures that some of the testing points provide no useful data, so that a large percentage of the trials are wasted. For example, with a slope (β) of 3.5, the psychometric function goes from 51% to 99% in the space of about 14 dB. With a grain of 4, then, and a resulting testing range of 16 dB, we are guaranteed that at least one testing point will be entirely outside the useful testing range. This is the best case; ordinarily, more than one point will fall outside this range. This is the essential problem with MCS.

Figure 2 shows the bias for MCS as a function of trial number for 6 values of grain. Note that the scale on this figure is enlarged relative to that of Figure 1. Small grains appear to lead to a positive bias, and large grain values lead to a negative bias; but for a reasonable num-

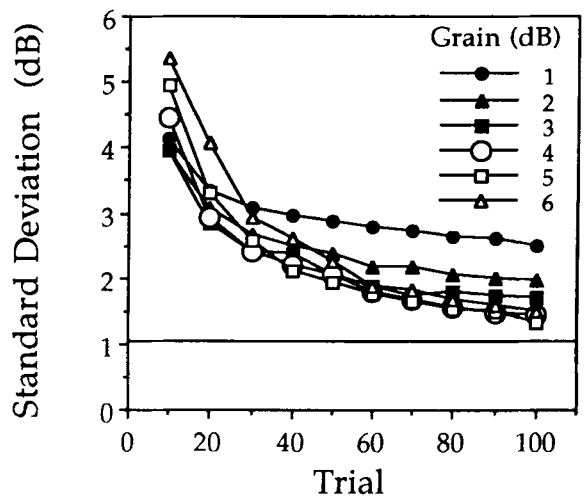

Figure 1. Standard deviation of the threshold estimate as a function of trial number for the method of constant stimuli at six values of the grain parameter. The initial standard deviation was 6 dB. A line is drawn at 1 dB as a visual guide.

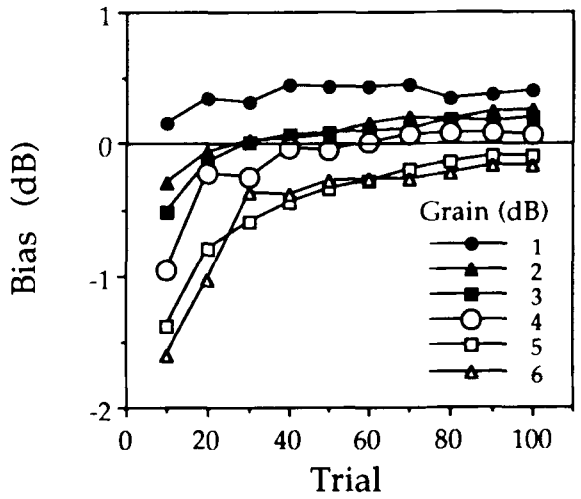

Figure 2. Bias (mean) of the threshold estimate as a function of trial number for the method of constant stimuli at six values of the grain parameter. The initial standard deviation was 6 dB.

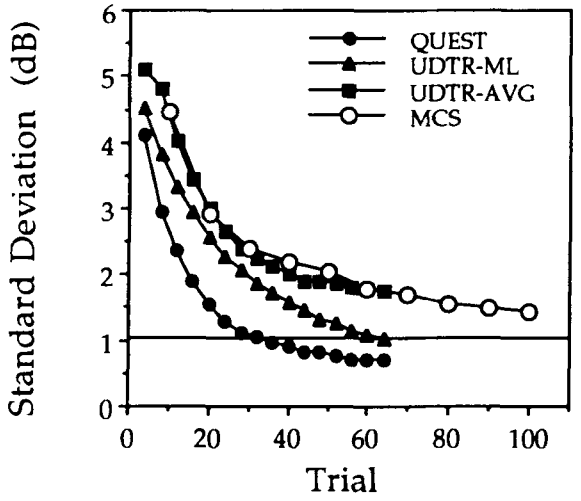

Figure 3. Standard deviation of the threshold estimate as a function of trial number for three adaptive methods and the method of constant stimuli with grain = 4. The initial standard deviation was 6 dB.

ber of trials, all biases are modest. A grain of 4, shown above to yield the lowest standard deviation, also gives the least bias.

Method of Constant Stimuli versus Adaptive Methods

Figure 3 compares the standard deviation of the most efficient version of MCS (grain = 4) with those of various adaptive methods. The results for MCS are comparable to those for UDTR-AVG, but considerably poorer than those obtained with the other adaptive methods tested.

Figure 4 shows the biases of the best MCS and the adaptive procedures. All procedures lead to modest bias, so that bias is clearly not a basis upon which to select a procedure.

Efficiency

The relative performances of psychometric procedures are best measured in terms of efficiency. The most natural definition of relative efficiency is the ratio of numbers of trials required by each procedure to achieve a given standard deviation. For example, we may ask: For a given number of MCS trials, how many QUEST trials are needed to yield an equivalent standard deviation? The ratio of these two numbers of trials is easily visualized as the horizontal distance between the two curves in Figure 5, which plots standard deviation versus log trials for the two procedures. One particular distance, corresponding to an efficiency of 28% at 60 MCS trials, is shown by

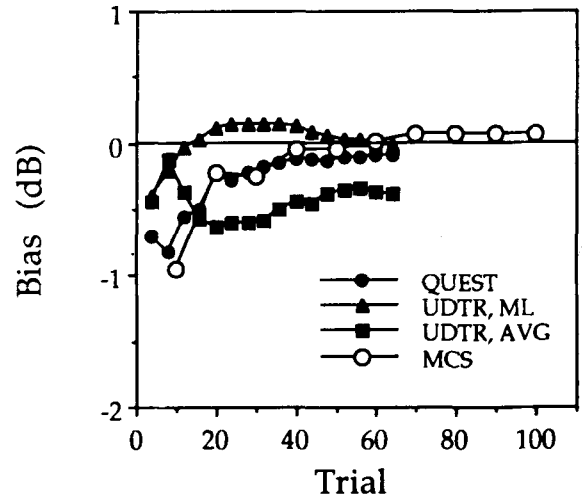

Figure 4. Bias (mean) of the threshold estimate as a function of trial number for three adaptive methods and the method of constant stimuli with grain = 4. The initial standard deviation was 6 dB.

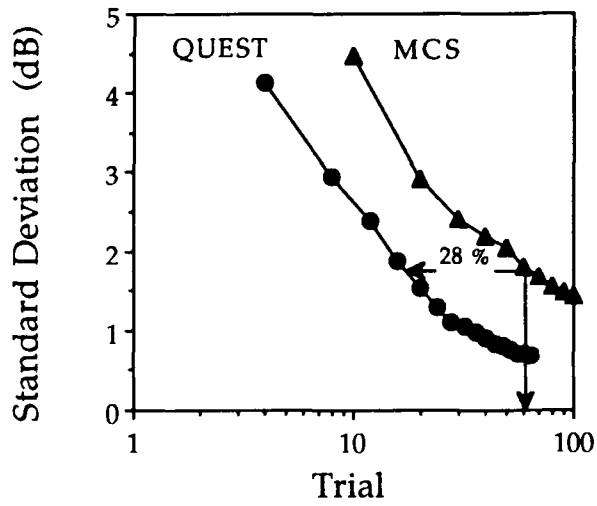

Figure 5. Standard deviation versus log trials for QUEST and MCS. Relative efficiency is given by the horizontal distance between the curves. The arrows indicate that at 60 trials, MCS is 28% as efficient as QUEST.

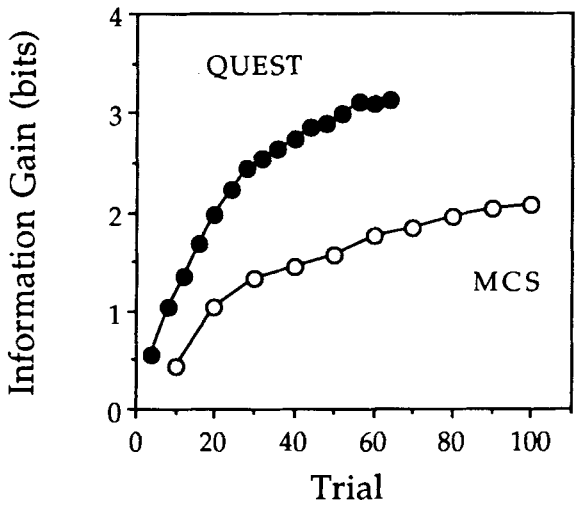

Figure 6. Information gain for QUEST and the method of constant stimuli. MCS used a grain of 4 dB. Initial standard deviation was 6 dB.

arrows in the figure. The efficiency of MCS, relative to QUEST, ranges between 20% and 40%. In short, MCS wastes about 60%-80% of the trials. This agrees with our earlier observation that several of the five MCS testing points collect no data of any value. The same analysis applied to UDTR-ML shows that it is about half as efficient as QUEST.

Information Gain

Another measure of the performance of a psychometric procedure is the information gain after a certain number of trials (Pelli, 1987). This is equal to the difference in the entropy of the estimate distribution before and after the trials. For a normal distribution with standard deviation σ, the entropy is (Shannon & Weaver, 1949)

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

Assuming that each distribution is approximately normal, we may use this to compute entropy after a number of trials.[1] If the initial standard deviation is 6 dB, then the starting entropy is 4.632 bits, and the difference between this number and the entropy after a given number of trials is the information gain. This quantity is pictured for MCS and QUEST in Figure 6. It is evident that the adaptive procedure acquires information much more rapidly than does the method of constant stimuli. After 64 trials, QUEST is ahead of MCS by almost 1.5 bits.

DISCUSSION

Efficiency of Adaptive and Nonadaptive Procedures

We have not attempted an exhaustive survey of adaptive and nonadaptive methods, nor have we extensively varied the parameters of the procedures we have used, or the parameters of the simulated observer. Instead, we have simulated a few, widely used procedures, and have set the parameters at what we believe are reasonable values. With these caveats, adaptive methods, particularly efficient adaptive methods such as QUEST, prove to be much more efficient than the method of constant stimuli. In particular, QUEST is about 2.5 to 5 times as efficient as the method of constant stimuli, depending on the number of trials. We also find that QUEST is about twice as efficient as UDTR-ML.

Simpson's Simulations

Why do Simpson's (1988) simulations show MCS to be at least reasonably efficient?—in short, because his method of simulating the variability of threshold ensures that threshold always lies within the testing interval. Upper and lower bounds are selected from uniform distributions between 0 and R, and 0 and $-R$, respectively, where R is a fixed value. Thus the testing interval always includes true threshold (0). As we have noted, in real life, MCS is inefficient largely because threshold may lie outside the testing interval, unless that interval is made extravagantly large. Thus Simpson's simulations were arranged in such a way as to avoid the conditions under which MCS suffers. This would be reasonable if these conditions could also be avoided in real life, but they cannot.

The upper and lower bounds of the testing interval selected in this way may be converted into an interval midpoint and width: width = upper − lower; midpoint = (lower + upper)/2. The midpoint is equivalent to the initial guess in our simulations, while the width is equal to five times our grain. Note that we fixed the grain within a simulation, whereas Simpson varied it randomly from experiment to experiment. The joint distribution of midpoint and width in Simpson's simulations is shown in Figure 7. This illustrates that the two quantities are not independent, and, in particular, that narrow widths can only occur when the midpoint is close to threshold. This contrasts with our simulations, in which width was fixed and midpoints were selected from a normal distribution of fixed standard deviation.

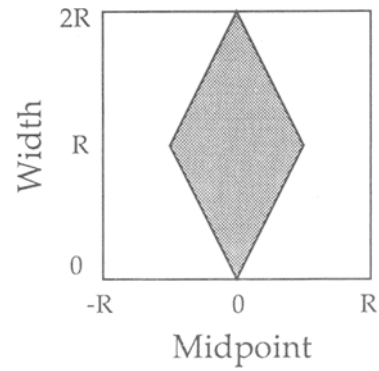

Figure 7. Joint distribution of the midpoint and width of the testing interval in Simpson's simulations. The width is dependent on the midpoint, and small widths are used only when the midpoint is near zero.

In Simpson's simulations, the distribution of the midpoint is a triangle function with a width of R and a height of $2/R$. The standard deviation is therefore $R/(2\sqrt{6})$ (e.g., $R = 5$, $SD = 1.02$ logit units).[2] So far, this is not too different from our normal distribution with a fixed standard deviation. We then use a fixed width in the presence of this fixed variability. Simpson, in contrast, selects a width based on the distance of the guess from the true threshold. But in real life, the experimenter does not have this information.

The issue is: What prior information do we assume on the part of the experimenter? In our simulations, we assume that the shape of the psychometric function is known and that threshold is known to within some standard deviation. Simpson assumes, in addition, that we know the distance between guess and threshold. But, of course, if we had this information, there would be no need to run the experiment.

Simpson quotes the statement from McKee, Klein, and Teller (1985) that variabilities of estimates from adaptive methods "can never be less than those from the method of constant stimuli selected for the optimal deployment of trials." We must presume that "optimal deployment" means "optimal based on the true location of threshold." Of course, if this location is known, there is no need to run the experiment. In the real world, threshold is never known exactly, even after the experiment has been completed. Thus a more accurate statement would be that *the method of constant stimuli can never be as efficient as a properly designed adaptive method.*

REFERENCES

CORNSWEET, T. N. (1962). The staircase-method in psychophysics. *American Journal of Psychology*, **75**, 485-491.

EMERSON, P. L. (1986). Observations on maximum-likelihood and Bayesian methods of forced-choice sequential threshold estimation. *Perception & Psychophysics*, **39**, 151-153.

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

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

LIEBERMAN, H. R., & PENTLAND, A. P. (1982). Microcomputer-based estimation of psychophysical thresholds: The best PEST. *Behavior Research Methods & Instrumentation*, **14**, 21-25.

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.

PELLI, D. G. (1987). The ideal psychometric procedure. *Investigative Ophthalmology & Visual Science*, **28**(3, Supplement), 366.

SHANNON, C. E., & WEAVER, W. (1949). *The mathematical theory of communication*. Urbana: University of Illinois Press.

SIMPSON, W. A. (1988). The method of constant stimuli is efficient. *Perception & Psychophysics*, **44**, 433-436.

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.

WATSON, A. B. (1979). Probability summation over time. *Vision Research*, **19**, 515-522.

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.

WETHERILL, G. B., & LEVITT, H. (1965). Sequential estimation of points on a psychometric function. *British Journal of Mathematical & Statistical Psychology*, **18**, 1-10.

WOLFRAM, S. (1988). *Mathematica: A system for doing mathematics by computer*. New York: Addison-Wesley.

NOTES

1. Since for a given σ, the normal is the distribution with the largest entropy, the normality assumption will always overestimate entropy (Shannon & Weaver, 1949). This error, for the distributions encountered here, is small, and it is largest for the least efficient procedures and conditions (e.g., small numbers of trials).

2. For a Weibull function with $\beta = 3.5$, and a logistic with slope = 1, one "logit unit" (the strength unit of the logistic) equals about 1.74 dB. Thus Simpson's midpoints had standard deviations of 1.8, 3.6, and 7.1 dB.

(Manuscript received March 13, 1989;
revision accepted for publication August 7, 1989.)