Pergamon

Vision Res. Vol. 34, No. 7, pp. 885–912, 1994

0042-6989/94 $6.00 + 0.00

Efficient and Unbiased Modifications of the QUEST Threshold Method: Theory, Simulations, Experimental Evaluation and Practical Implementation

P. EWEN KING-SMITH,* SCOTT S. GRIGSBY,* ALGIS J. VINGRYS,* SUSAN C. BENES,† AARON SUPOWIT‡

Received 6 January 1993; in revised form 23 August 1993

QUEST [Watson and Pelli, *Perception and Psychophysics, 13*, 113–120 (1983)] is an efficient method of measuring thresholds which is based on three steps: (1) Specification of prior knowledge and assumptions, including an initial probability density function (p.d.f.) of threshold (i.e. relative probability of different thresholds in the population). (2) A method for choosing the stimulus intensity of any trial. (3) A method for choosing the final threshold estimate. QUEST introduced a Bayesian framework for combining prior knowledge with the results of previous trials to calculate a current p.d.f.; this is then used to implement Steps 2 and 3. While maintaining this Bayesian approach, this paper evaluates whether modifications of the QUEST method (particularly Step 2, but also Steps 1 and 3) can lead to greater precision and reduced bias. Four variations of the QUEST method (differing in Step 2) were evaluated by computer simulations. In addition to the standard method of setting the stimulus intensity to the *mode* of the current p.d.f. of threshold, the alternatives of using the *mean* and the *median* were evaluated. In the fourth variation—the Minimum Variance Method—the next stimulus intensity is chosen to minimize the expected variance at the end of the next trial. An exact enumeration technique with up to 20 trials was used for both yes–no and two-alternative forced-choice (2AFC) experiments. In all cases, using the mean (here called ZEST) provided better precision than using the median which in turn was better than using the mode. The Minimum Variance Method provided slightly better precision than ZEST. The usual threshold criterion—based on the "ideal sweat factor"—may not provide optimum precision; efficiency can generally be improved by optimizing the threshold criterion. We therefore recommend either using ZEST with the optimum threshold criterion or the more complex Minimum Variance Method. A distinction is made between "measurement bias", which is derived from the mean of repeated threshold estimates for a single real threshold, and "interpretation bias", which is derived from the mean of real thresholds yielding a single threshold estimate. If their assumptions are correct, the current methods have no interpretation bias, but they do have measurement bias. Interpretation bias caused by errors in the assumptions used by ZEST is evaluated. The precisions and merits of yes–no and 2AFC techniques are compared. Practical implementation of the ZEST method is described in the Appendix, with emphasis on the flexibility of the current methods in circumventing experimental problems, and on enhancements to allow for variations in the slope of the psychometric function, drifts in threshold, and correlation between thresholds for different stimuli.

Threshold methods Forced-choice method Yes–no method Efficiency Bias

INTRODUCTION

Visual threshold measurements have provided much of our knowledge about normal visual function and also provide important information about visual function in infants and patients. Typically, time limitations restrict infant and clinical measurements to fewer trials than would be used in a study of normal visual processes. For any subject population, there is a need for threshold techniques which will provide accurate and unbiased information in a given number of trials. In this paper, the performance of a number of threshold techniques is compared by using computer simulations; a preliminary report of these simulations has been presented

*College of Optometry, The Ohio State University, 338 West Tenth Avenue, Columbus, OH 43210-1240, U.S.A.
†Department of Ophthalmology, The Ohio State University, Columbus, OH 43210, U.S.A.
‡Academic Computing Services, The Ohio State University, Columbus, OH 43210, U.S.A.

(King-Smith, Grigsby, Vingrys, Benes & Supowit, 1991). In addition, implementation and experimental evaluation of an efficient threshold technique are described.

This paper is mainly concerned with threshold measurements which are *independent* in the sense that the testing strategy and threshold calculations for any one stimulus (e.g. one spatial frequency of a contrast sensitivity function) are independent of the results for any other stimulus in the same experiment. Greater efficiency can be obtained in special circumstances where there is a known correlation between thresholds for different stimuli—e.g. in automated perimetry (Johnson & Shapiro, 1990; see also Appendix). This paper is also limited to threshold methods based on a series of discrete responses to stimulus presentations; for a more general discussion of psychophysical methods, see Pelli and Farell (1994).

Adaptive threshold methods

Threshold methods may be "adaptive" or "non-adaptive". In adaptive threshold methods, the intensity used on any trial depends on the subject's responses to previous trials whereas in non-adaptive methods, the stimulus intensities are predetermined and independent of the subject's responses (the most common example being the "method of constant stimuli"—McKee, Klein & Teller, 1985). When there is considerable initial uncertainty about the threshold value, adaptive methods are thought to be more efficient than non-adaptive methods, for the following reason. Test intensities which are close to threshold are generally more informative than those which are far from threshold (Taylor, 1971). For example, in a "yes–no" experiment, a measured response probability of 100% indicates only that threshold is considerably below the stimulus intensity (but not, say, whether it is 1 rather than 2 log units lower); however, a measured response probability of say, 50%, indicates that threshold is relatively close to the stimulus intensity. Adaptive methods are designed to present most stimulus intensities close to threshold, and so they are typically more efficient than non-adaptive methods (Watson & Fitzhugh, 1990).

Many different adaptive threshold methods have been developed. In a simple "staircase" method (Cornsweet, 1962), the stimulus intensity is reduced or increased by a (typically) fixed step after correct or incorrect responses respectively. Modified decision rules (e.g. step changes of intensity after two correct responses or one incorrect response) have been developed by Wetherill and Levitt (1965) and Levitt (1971); additionally, the step size can be reduced systematically as a function of the trial number, i,—e.g. as c/i (where c is a constant, Robbins & Munro, 1951) or as $c/2^i$ (MOBS, Modified Binary Search, Tyrrell & Owens, 1988; Johnson & Shapiro, 1989).

Other adaptive methods are based on blocks of trials. In PEST (Parameter Estimation by Sequential Testing—Taylor & Creelman, 1967; Findlay, 1978; Hall, 1981) a block of trials is presented at a fixed intensity; the block is terminated when the number of correct responses deviates significantly from that expected from the threshold-criterion probability. A new block is then presented at a lower or higher intensity, depending on whether the observed probability was, respectively, above or below the criterion. In APE (Adaptive Probit Estimation, Watt & Andrews, 1981), each block has a fixed number of trials and uses four intensities; after each block (except the first), a new set of four intensities is derived from previous responses.

An adaptive threshold method which offers the potential of high efficiency is the "maximum likelihood" method (Hall, 1968; Pentland, 1980; Klein 1981; Watson & Pelli, 1983; Green, 1993). After each trial, the currently most likely value of threshold is determined, and this intensity is used for the next stimulus intensity; the final threshold estimate is the most likely value of threshold after the last trial.

The QUEST method

The most popular maximum likelihood method is probably the QUEST method of Watson and Pelli (1983). QUEST is based on the following assumptions:

(1) The psychometric function has the same shape under all conditions, when expressed as a function of log intensity.

(2) The subject's threshold does not vary from trial to trial.

(3) Individual trials are statistically independent.

The first trial of this method for a yes–no experiment is illustrated in Fig. 1. The experimenter's prior knowledge about the probability of different threshold values is represented in Fig. 1(A). This is known as the initial probability density function (p.d.f.) and denoted $q_0(T)$, where T is threshold in log units; $q_0(T)\delta T$ is the probability that the log threshold is in a small range δT at T. Watson and Pelli (1983) assumed a Gaussian initial p.d.f., but the p.d.f. assumed here (a modified hyperbolic secant) is skewed and is based on a histogram of thresholds measured in our laboratory (see below). In Fig. 1(A), this p.d.f. has been normalized so that the total area under the curve is unity, i.e.

$$\int q_0(T)\mathrm{d}T = 1 \qquad (1)$$

although this is not necessary for normal application of the QUEST method, it is useful for computer simulations.

In the standard QUEST method, the first stimulus intensity is chosen to correspond to the mode (maximum likelihood), x_1, of this initial p.d.f.—the vertical line in Fig. 1(A). If the subject responds "yes" to the first trial, it indicates that the threshold is probably below the stimulus intensity and so the new p.d.f. of threshold should be shifted towards lower intensities; similarly, a "no" response should shift the p.d.f. to higher intensities.

Watson and Pelli (1983) have shown how Bayes' theorem may be applied to calculate the new p.d.f. of threshold. Suppose that $p(r, x, T)$ is the probability of

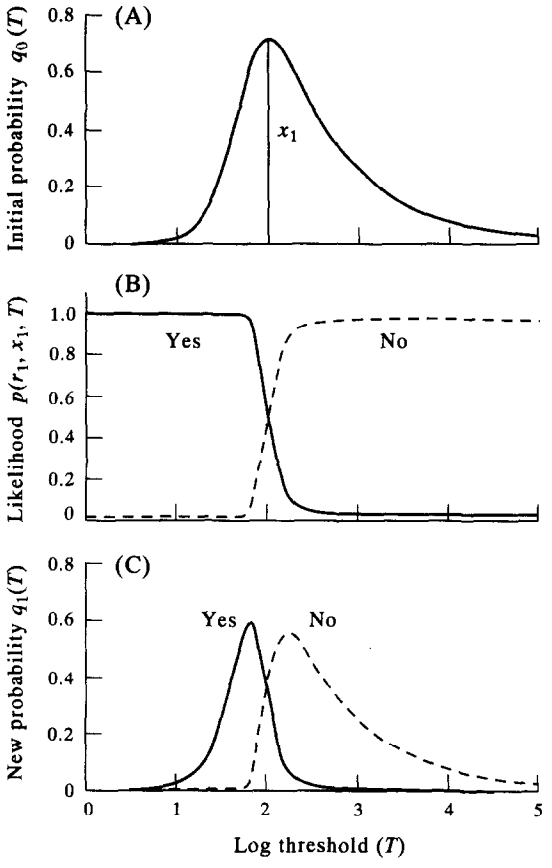

FIGURE 1. An illustration of the calculations associated with the first trial of the QUEST method using a yes–no method. (A) The assumed initial probability density function (p.d.f.), $q_0(T)$, of log threshold, T. The vertical line is drawn at the maximum-likelihood value of log threshold—i.e. the mode of the p.d.f.; this would be used as the log intensity, x_1, of the first stimulus in the standard QUEST method. (B) Likelihood functions for "yes" and "no" responses; for example, the "yes" likelihood function, $p(1, x_1, T)$ is the probability that the subject will respond "yes" to the log stimulus intensity, x_1, as a function of the subject's log threshold, T. This likelihood function is a left-to-right mirror image of the psychometric function (probability of seeing curve). (C) P.d.fs., $q_1(T)$, of log threshold after "yes" or "no" responses to the first trial. By Bayes's Theorem, these p.d.fs. are the product of the initial p.d.f. (A) with the corresponding likelihood function (B) [equation (2)].

a response, r ($r = 1$ for "yes", $r = 0$ for "no") for a subject with log threshold, T, to a stimulus of log intensity, x; the functional form of $p(r, x, T)$ is discussed later [equation (9)]. For $r = 1$, this is simply the psychometric function (probability of seeing as a function of log intensity, x) for a subject with log threshold, T. Consider the first trial of a threshold measurement [log intensity x_1, Fig. 1(A)]. Then the probability that the log threshold is T *and* that the subject will give response r_1 is simply the product of the probabilities of these two events, i.e.

$$q_1(T) = p(r_1, x_1, T) q_0(T). \tag{2}$$

Thus $q_1(T)$ is a measure of the probability that the log threshold is T, given both the prior knowledge, $q_0(T)$, and the subject's response, r_1.

The multiplication in equation (2) is represented in Fig. 1. For the stimulus intensity, x_1, the solid and dashed curves in Fig. 1(B) give the probability $p(r_1, x_1, T)$ of yes ($r_1 = 1$) and no ($r_1 = 0$) responses respectively; these are functions of T and are called "likelihood functions" (cf. the psychometric function is a function of x). Corresponding solid and dashed curves in Fig. 1(C) give the product of the functions in Fig. 1(A) and (B) and hence [from equation (2)] the two possible p.d.fs. of threshold, $q_1(T)$, after the first trial. For a relatively broad p.d.f., such as that in Fig. 1(A), this multiplication effectively "cuts off" part of the p.d.f.; e.g. for a "yes" response, the high threshold end of the p.d.f. is cut off [Fig. 1(C)].

In theory, $q_1(T)$ could be normalized by multiplying by a constant so that the area under this function is unity as in equation (1); $q_1(T)\delta T$ would then be the probability that log threshold now lies in a range δT at T. In practice it is not necessary to do this normalization to implement the QUEST method. For the simulations of this paper, it is actually advantageous to use equation (2) *without* normalization; given the assumptions made by the QUEST method, the probability that the subject's first response is r_1 can then be calculated as follows. $q_1(T)\delta T$ is the probability that the log threshold is in a small range δT at T *and* that the first response is r_1. The overall probability, $P(r_1)$, of response r_1 is then simply obtained by integration i.e.

$$P(r_1) = \int q_1(T)\, dT. \tag{3}$$

It should be noted that this probability, which depends on the distribution of thresholds in the population and the intensity of the first trial, is *not* equivalent to the probability of seeing which is defined as threshold.

The function $p(r, x, T)$ is potentially complicated (being a function of two continuous variables, x and T, and one binary variable, r) but two simplifications can be made. First, for any x and T, the sum of the probabilities of "yes" and "no" responses must be unity so that

$$p(0, x, T) = 1 - p(1, x, T). \tag{4}$$

This inverse relationship between "no" and "yes" probabilities is seen in Fig. 1(B).

The second simplification is that, from assumption 1 (see above), the shape of a psychometric function, when plotted as a function of log intensity, x, is independent of log threshold; log threshold determines the position of the psychometric function along the log intensity axis, but does not affect its shape. This is known as Crozier's law (Crozier, 1936; Blackwell, 1963; le Grand, 1968; Nachmias, 1981). Thus, for any value of log threshold, T', the psychometric function is given by the equation

$$p(1, x, T') = \Psi(x - T') \tag{5}$$

where $\Psi(X)$ is a canonical (standard) form of the psychometric function [e.g. equation (9)] and $X = (x - T')$ is the difference between log stimulus intensity and log threshold.

A similar equation applies to likelihood functions (probability vs log threshold, T); for any log intensity, x', the "yes" likelihood function will be given by

$$p(1, x', T) = \Psi(x' - T) \tag{6a}$$

and from equation (4), the "no" likelihood function will be

$$p(0, x', T) = 1 - \Psi(x' - T). \quad (6b)$$

Thus likelihood functions for different values of log intensity, x', all have the same shape and may be derived from a standard shape by sliding them along the log threshold axis by a distance x'. By comparing equation (6a) (function of T) with equation (5) (function of x), it is seen that the shape of the "yes" likelihood function [equation (6a)] is simply a left-to-right mirror image of the psychometric function [equation (5)]; the "yes" likelihood function in Fig. 1(B) illustrates this mirror reversal of a typical psychometric function. From equations (6a) and (6b), the "yes" and "no" likelihood functions for any log intensity, x', are readily derived from the standard form of the psychometric function, $\Psi(X)$, which need be calculated only once and then can be stored as a table in computer memory (Watson & Pelli, 1983).

The process of Fig. 1 can now be repeated as many times as required. For example, in the standard QUEST method, if the subject responded "yes" to the first trial, the next intensity, x_2, would be set to the mode of the "yes" p.d.f. in Fig. 1(C); after the second response, r_2, a new p.d.f. would be derived by a multiplication like that in equation (2). In general, the p.d.f. after trial i is given by

$$q_i(T) = p(r_i, x_i, T) q_{i-1}(T). \quad (7)$$

The preceding analysis applies equally well to forced-choice experiments; in that case, $r_i = 1$ and $r_i = 0$ would correspond to correct and incorrect responses respectively.

In Watson and Pelli's (1983) implementation, calculations were performed on the *logarithms* of probabilities and likelihoods [so that the multiplication of equation (7) became an addition which was considerably faster to calculate with the slow laboratory computers then available]. We prefer to use *unmodified* probabilities because this facilitates calculation of the mean and median of the current p.d.f. of threshold. It also facilitates calculation of the variance and range of this p.d.f.; this is valuable for estimating the precision of the final threshold estimate, and also may be used for a termination rule [i.e. the experiment may be terminated when the variance or range of the p.d.f. falls below a predetermined value—this method of termination makes fewer assumptions than the χ^2 method of Watson and Pelli (1983) and the variance estimator of Laming and Marsh (1988)].

The initial probability density function (p.d.f.)

Watson and Pelli (1983) assumed that the initial p.d.f. was a Gaussian function of threshold whose mean and standard deviation could be estimated from the experimenter's experience. However, when equipment is being used in a fairly consistent manner, it is possible to fine tune the QUEST method by analyzing threshold data collected in these conditions over a certain time period.

Figure 2(A) shows a histogram of 18,944 thresholds measured on an oscilloscope display system using a yes–no method; the number of thresholds in 0.05 log unit ranges of contrast threshold are plotted as a function of the logarithm of contrast threshold (in %). Data are from both normal subjects and patients, and include measurements of both contrast sensitivity functions and flicker modulation sensitivity (de Lange curves); conditions were similar to those used by Grigsby, Vingrys, Benes and King-Smith (1991). It may be noted that some contrast threshold values were estimated to be over 100% (log threshold greater than 2) and so are greater than contrasts that are physically possible. This ability of the QUEST method to estimate thresholds beyond the physical range of the equipment is discussed in the Appendix.

The dashed line in Fig. 2(A) is a least squares fit to the histogram using a modified hyperbolic secant of the form

$$q_0(T) = A/[B e^{-C(T-t)} + C e^{B(T-t)}] \quad (8)$$

where T is log threshold, A determines the overall height of the fitted curve, B and C determine the slopes of the fall-off at high and low thresholds respectively, and t corresponds the most probable log threshold value. Values of B and C were 1.22 and 5.07 respectively indicating a much shallower fall-off in probability at high thresholds than at low. This function (shifted, for convenience, to have a peak at $T = t = 2$) is the one illustrated in Fig. 1(A) and is used in most of the current simulations.

False positive rate for the experiments of Fig. 2(A) was 942 in 29,940 blank trials, i.e. 3.15%. The histogram of Fig. 2(A) was determined using only those threshold measurements in which the subject responded at least once; the subject did not respond to any of the stimuli on 3.87% of attempted threshold measurements, and these very high thresholds can be considered to correspond to the high-threshold "tail" of the dashed curve in Fig. 2(A).

The generality of equation (8) is supported by a similar analysis, in Fig. 2(B), of 70,247 color-mixture thresholds measured on a color video display. A yes–no technique was again used, both normal subjects and patients were included, and conditions were similar to those of Grigsby *et al.* (1991). [Unlike the sinusoidal stimuli used in Fig. 2(A) for which the maximum possible physical contrast was 100%, these stimuli were circular test spots whose contrast or Weber fraction could be greater than 100% for incremental test spots; thus, the histogram extends to higher threshold contrasts than those of Fig. 2(A).] In this case, the slope parameters, B and C, are 1.80 and 5.10, and thus the slope at high thresholds is somewhat steeper than in Fig. 2(A). False positive rate was 1305 in 58,778 blank trials, i.e. 2.22%. As in Fig. 2(A), the histogram does not include attempted threshold measurements where the subject did not respond to any stimulus—these accounted for 1.85% of the total threshold measurements.

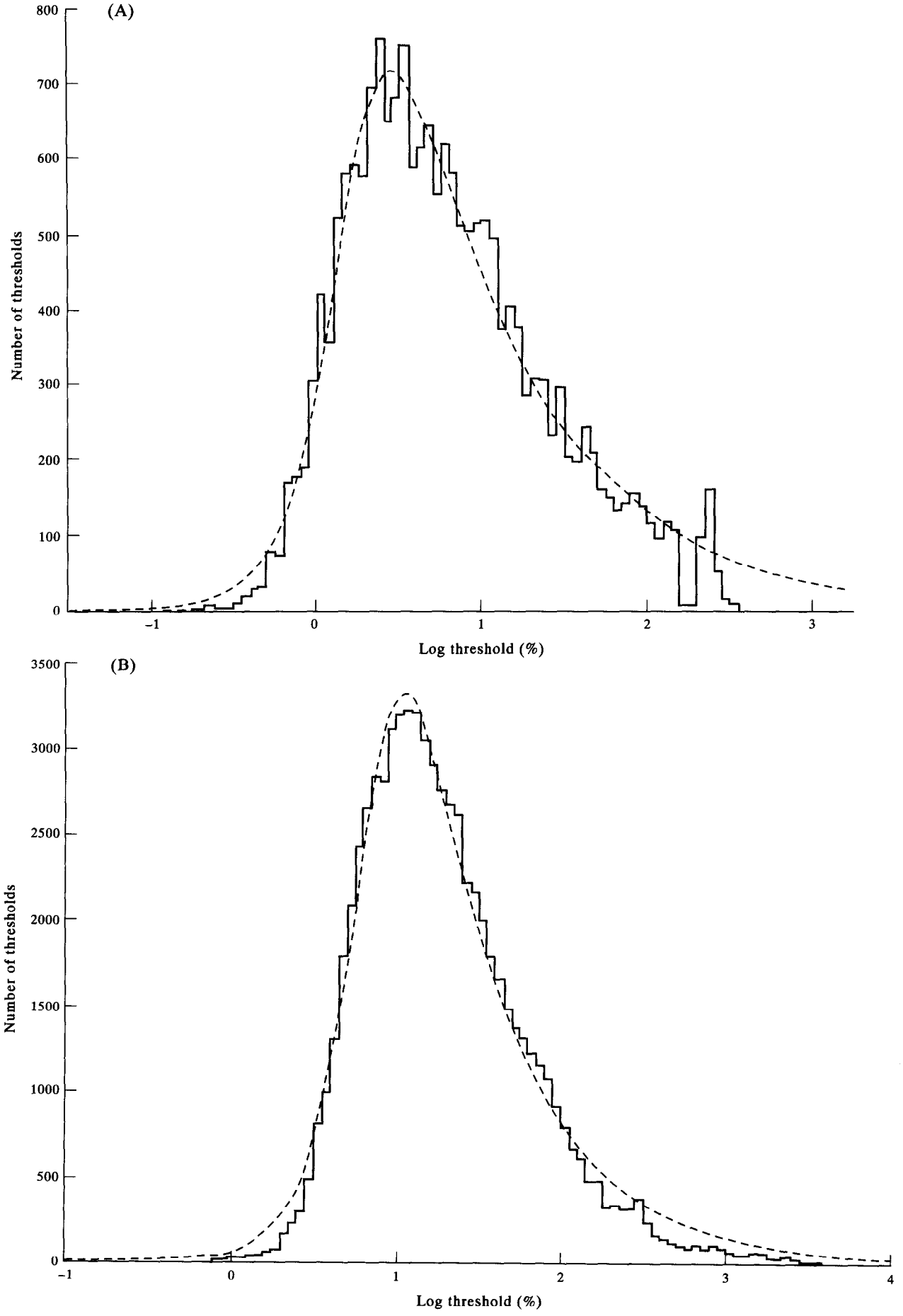

FIGURE 2. Histograms of log threshold values determined experimentally. The dashed lines are least squares fits to these histograms using equation (8). (A) Histogram for 18,944 thresholds measured on an oscilloscope display system. (B) Histogram for 70,247 thresholds measured on a color video display system.

The initial p.d.f. used for most of the current simulations is based on Fig. 2(A) (oscilloscope display), and would thus be suitable for measuring, say, contrast sensitivity functions. It should be emphasized that this initial p.d.f. is not necessarily suitable for all types of visual threshold measurements—e.g. a broader initial p.d.f. would be more suitable if very high or low threshold values were more probable.

The psychometric function

The psychometric function (and hence the likelihood function) used here is a Weibull (1951) distribution similar to that used by Watson and Pelli (1983). It is given by

$$p(1, x, T) = \Psi(x - T)$$
$$= 1 - \delta - (1 - \gamma - \delta)\exp[-10^{\beta(x - T + \epsilon)}] \quad (9)$$

where γ is the false positive rate (e.g. 0.5 for two alternative forced choice, 2AFC), δ is the false negative rate (e.g. caused by blinks or mental lapses), β determines the slope of the psychometric function and ϵ determines the threshold criterion—i.e. the probability of seeing which is defined as threshold.

For yes–no simulations, the false positive rate, γ was set to 0.03 which is similar to experimental values determined above. For both yes–no and 2AFC simulations, we used the same value of false negative rate, $\delta = 0.01$, as Watson and Pelli (1983); however, our method of incorporating false negatives is slightly different from theirs, as it allows for the possibility of false negatives occurring for all stimulus strengths (rather than just for strong stimuli), but the difference between the two psychometric functions is small (less than 0.01 change in probability). In our experiments, and in most of our simulations, the slope factor, β was 3.5, the same as Watson and Pelli's. In our yes–no experiments (Appendix), the threshold criterion, ϵ was set to zero, corresponding to a threshold probability of 0.637; this and other values of ϵ were used in the current simulations.

The ideal sweat factor

Watson and Pelli (1983) advocated a value of ϵ in equation (9) corresponding to the "ideal" sweat factor (Taylor, 1971; Green, 1990). The "sweat factor" is a measure of the amount of effort required to obtain a certain accuracy of threshold estimate, for a sequence of trials at a *fixed* intensity, and is given by

$$K(X) = \Psi(X)[1 - \Psi(X)]/[\mathrm{d}\Psi(X)/\mathrm{d}X]^2 \quad (10)$$

where $\Psi(X)$ is the standard form of the psychometric function [equation (9)] and X is the difference between log intensity and log threshold (i.e. $x - T$). The ideal sweat factor is the minimum value of $K(X)$; it is possible to arrange that the ideal sweat factor occurs at threshold intensity (i.e. $x = T$ or $X = 0$) by suitable choice of ϵ in equation (9)—this ideal value will be called ϵ_{id}. The variance of a threshold estimate based on M trials (where M is large) at a fixed intensity equals $K(X)/M$ (Taylor, 1971); thus if $\epsilon = \epsilon_{id}$, and intensity is set to a fixed value very close to threshold, this should yield the minimum variance in the threshold estimate from such a sequence of trials. It was for this reason that Watson and Pelli (1983) advocated using the ideal sweat factor; however, it is not clear that ϵ_{id} is the optimum value of ϵ for adaptive methods based on a limited number of trials and this is investigated in the present study.

For a yes–no experiment [$\beta = 3.5$, $\gamma = 0.03$ and $\delta = 0.01$ in equation (9)], $\epsilon_{id} = 0.052$ log units and the corresponding probability of seeing at threshold is 0.780; for a 2AFC experiment ($\gamma = 0.5$), $\epsilon_{id} = 0.063$ log units and the probability of a correct response at threshold is 0.897. Slight differences from the values of Watson and Pelli (1983) are due to the slight difference in the current psychometric function [equation (9)].

Variations on the QUEST method

For asymmetrical p.d.fs. such as those in Fig. 1(A, C), the mode, mean and median will generally differ from each other; for example, for the initial p.d.f. in Fig. 1(A), the mode, mean and median are 2.0, 2.45 and 2.25 respectively [when rounded to the step size of 0.05 log units used by Watson and Pelli (1983) and here]. Previous studies (King-Smith, 1984; Emerson, 1986) have demonstrated that greater efficiency and less measurement bias may be obtained when the next log intensity, x_i, is set to the *mean* of the current p.d.f. rather than the *mode* which is used in the standard QUEST method (Watson & Pelli, 1983). In the current simulations, further comparisons are made between all three measures; mode, mean and median. Methods based on these three strategies will be called "mode-QUEST", "mean-QUEST" and "median-QUEST".

The minimum variance method and the ideal psychometric procedure

The Minimum Variance Method (King-Smith, 1984) is another modification of the QUEST procedure. In this method, the log stimulus intensity, x_i, for the next stimulus is chosen so as to minimize the expected variance, $\hat{V}$, at the end of the next trial, where expected variance is defined by

$$\hat{V} = P_y V_y + P_n V_n. \quad (11)$$

P_y and P_n are the probabilities of "yes" and "no" responses, and V_y and V_n are variances of the corresponding p.d.fs. For example, on the first trial, one possible stimulus intensity would be the mode of the initial p.d.f. which is illustrated in Fig. 1. In this case, P_y and P_n are given by equation (3) and V_y and V_n can be calculated from the corresponding p.d.fs. in Fig. 1(C) using the standard statistical formula

$$V = \int (T - m)^2 q(T)\,\mathrm{d}T \Big/ \int q(T)\,\mathrm{d}T \quad (12)$$

where m is the mean of the p.d.f. $q(T)$, i.e.

$$m = \int Tq(T)\,\mathrm{d}T \Big/ \int q(T)\,\mathrm{d}T. \quad (13)$$

Equation (11) is evaluated for a range of different stimulus intensities and the stimulus intensity is chosen which minimizes $\hat{V}$. For the initial p.d.f. of Fig. 1(A), [and for $\epsilon = 0$ in equation (9)], the minimum expected variance is for $x_1 = 2.65$, i.e. somewhat greater than the mode, mean and median of this p.d.f. which are given above.

This method minimizes the expected variance at the end of the next trial and so "looks *one* trial ahead". A variation of this method is to "look *two* trials ahead" by trying a range of log intensities, x_i, for the next trial and, for each possible outcome ("yes" or "no") a range of log intensities, x_{i+1}, for the following trial; the value of x_i is chosen which minimizes (together with the corresponding best choices for x_{i+1}) the expected variance at the end of the next two trials [i.e. from an equation like equation (11) but with four terms corresponding to the four possible response pairs to two trials—"yes–yes", "yes–no", "no–yes" and "no–no"]. The log intensity for the following trial, x_{i+1}, is *not* determined from these calculations made *before* trial i, but is determined by repeating the whole calculation process *after* trial i. For the final trial, there is no reason to try to minimize the expected variance after the next two trials, so the simpler "one-trial look-ahead" method is used.

In the Ideal Psychometric Procedure (Pelli, 1987), the method is extended to many steps ahead—e.g. to the end of the experiment. The computation time for a look-ahead of M trials grows as about L^M where L is the number of expected variances [equation (11)] computed per look-ahead trial. For this reason, the current simulations have been limited to one and two trial look-ahead.

The final threshold estimate

In our simulations, each experimental run was terminated after a predetermined number of trials, N. For such an experiment, there are 2^N possible sequences of responses, because the subject may respond "yes" or "no" to any of the trials. Any sequence may be uniquely specified by a "sequence number", j, $(1 \leqslant j \leqslant 2^N)$ defined by

$$j = r_1 2^{N-1} + r_2 2^{N-2} \ldots + r_i 2^{N-i} \ldots + r_N + 1. \quad (14)$$

Our method of calculating the final threshold estimate differs in two ways from that of Watson and Pelli (1983). First, Watson and Pelli derived a likelihood function of log threshold, $l_j(T)$, by dividing the final p.d.f. for sequence j, $q_{Nj}(T)$, by the initial p.d.f., $q_0(T)$; they used this likelihood function to calculate threshold [from the mode of $l_j(T)$]. In this way, they eliminated all assumptions about the initial p.d.f. from their final estimate of threshold. At large positive and negative values of log threshold, T, the likelihood function tends to constant, non-zero, values, which are typically small and different for positive and negative extremes; this means that, although the mode (peak) can be calculated, the mean, median and variance are not calculable, at least in the case where the whole function $-\infty < T < \infty$, is considered. It is, of course, possible to calculate the mean, median and variance if the range of the likelihood function is restricted e.g. $T_1 < T < T_2$; however, these values will then depend on the rather arbitrary choice of T_1 and T_2.

These problems can be avoided if the final p.d.f., $q_{Nj}(T)$, is used instead of the likelihood function, because this p.d.f. tends rapidly (exponentially) to zero for large positive and negative values of T and so it has calculable mean, median and variance (even when an infinite range of T is considered). In our experiments, where the initial p.d.f. has been determined from a histogram of log thresholds (Fig. 2) it is reasonable to include this information in the final threshold estimate—i.e. to use the final p.d.f., rather than to divide it by the initial p.d.f. to derive a likelihood function. Our simulations will show that estimated thresholds are typically little affected by any reasonable choice of initial p.d.f.

A second difference from Watson and Pelli (1983) is that the final threshold estimate, E_j, was taken as the *mean*, rather than the mode, of the final p.d.f.; thus

$$E_j = \int T q_{Nj}(T)\,\mathrm{d}T \Big/ \int q_{Nj}(T)\,\mathrm{d}T. \quad (15)$$

The mean was used for two reasons. The first advantage of using the mean is that it minimizes the variance (mean square error) of the final threshold estimate; for example, in equation (12), the minimum value of variance, V, is given by setting m equal to the mean of the p.d.f. $q(T)$ [equation (13); Hays, 1988, p. 177]. Because we evaluated different threshold methods by calculating an overall weighted variance for the threshold estimates (see Methods), this provides the optimum performance for all techniques and so provides a fair comparison. If the mode of the final p.d.f. had been used (cf. Watson & Pelli, 1983), the overall variance would have been higher. A second reason for using the mean, rather than the mode, of the final p.d.f. is to eliminate bias—see below.

An illustrative example

To help the reader envisage the threshold methods and simulations of this paper, Fig. 3 is a perspective plot which represents the final p.d.fs., $q_{Nj}(T)$, which can be obtained after $N = 3$ trials of the QUEST method. Each final p.d.f. has been plotted at its corresponding value of log estimated threshold, E_j, and corresponds to one of the 8 $(=2^3)$ possible response sequences in three trials; for example, the leftmost p.d.f., which has the highest estimated threshold, corresponds to a sequence of three "no" responses. The final threshold estimate, E_j, equals the *mean* of the final p.d.f., given by each thick solid vertical line; this is emphasized by the diagonal line, $(E_j = T)$, which passes through the means of the p.d.fs. For comparison, each dashed vertical line gives the *mode* of the p.d.f.; if Fig. 3 is now considered to represent an intermediate stage of a longer threshold determination $(N > 3)$, it is seen that setting the next intensity equal to the mode (the standard QUEST procedure) would differ considerably from setting it to the mean. The thin lines at $T = 2$ and $T = 4$ are discussed below.

In addition to the mean, two other parameters of the final p.d.fs. are important for the current simulations. First, if the assumptions of the method are correct, the

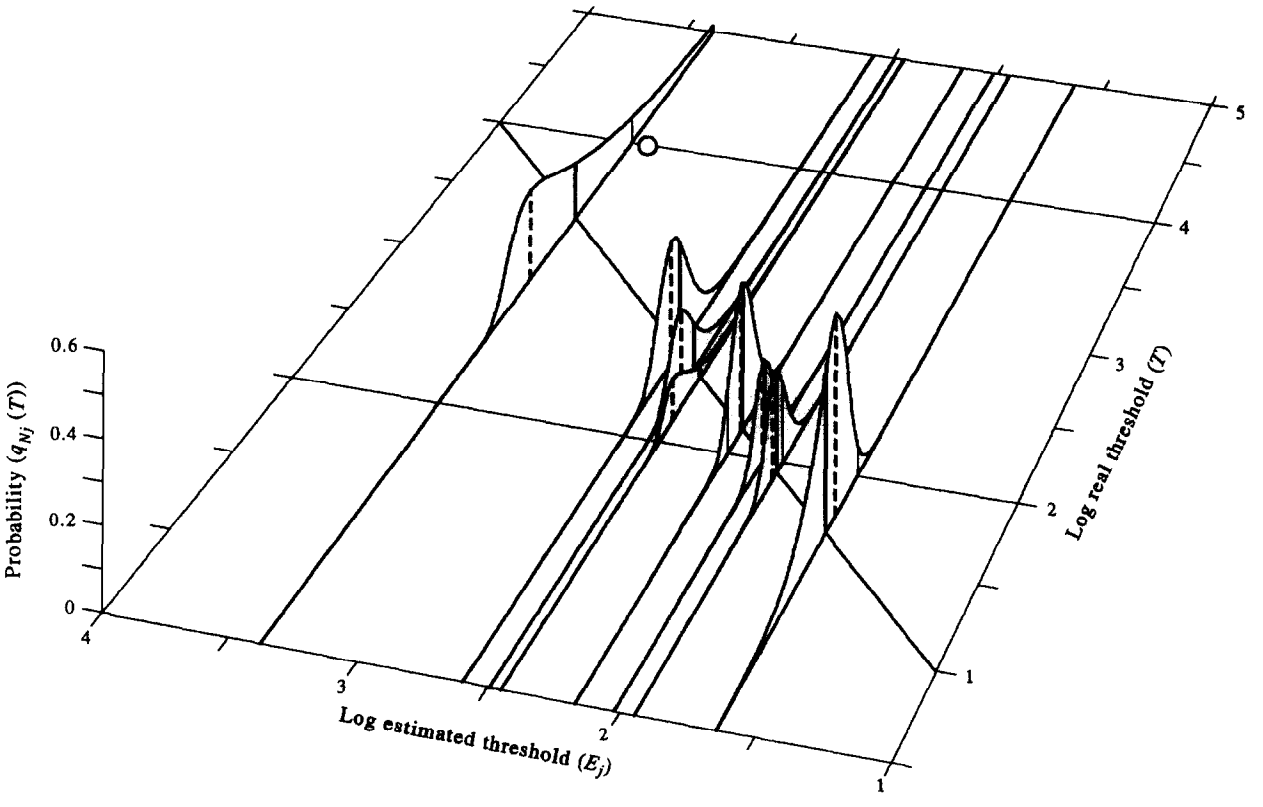

FIGURE 3. A perspective plot of final p.d.fs., $q_{Nj}(T)$, for a simulation of the yes–no QUEST procedure [i.e. next intensity set to the mode of the current p.d.f. and using the "ideal sweat factor", i.e. $\epsilon = \epsilon_{id} = 0.052$ log units in equation (9)], with $N = 3$ trials. There are eight final p.d.fs. corresponding to the 2^3 possible sequences of responses. Final p.d.fs. are plotted at the corresponding values of log estimated threshold, E_j, which for our simulations are the *means* of the final p.d.fs.—thick solid vertical lines; this derivation of estimated threshold from the mean of the final p.d.f. is illustrated by the diagonal line ($E_j = T$). The dashed vertical lines give the *modes* of the final p.d.fs.—for a longer threshold run ($N > 3$), these would be used for the next intensity in the standard QUEST method. Thin lines at $T = 2$ and $T = 4$ and circle at $T = 4$ help to illustrate measurement bias (see text for details). For best perspective view, hold vertically and view from above. In this and subsequent figures, the following conditions are used unless otherwise indicated. The initial p.d.f. is that shown in Fig. 1(A), and is the same shape as the curve fitted to the histogram of oscilloscope thresholds in Fig. 2(A). The psychometric function is given by equation (9) with $\beta = 3.5$, $\gamma = 0.03$ for yes–no and 0.5 for 2AFC simulations, and $\delta = 0.01$.

area under each p.d.f. gives the probability of the corresponding response sequence [cf. equation (3)]. Second, the *variance* of a final p.d.f. gives the variance of the threshold estimate.

Measurement and interpretation bias

An advantage of using the mean, rather than the mode, of the final p.d.f., $q_{Nj}(T)$, as the threshold estimate is that the mean is an unbiased estimate of threshold, in the sense that, if all assumptions are correct, this estimate will be correct "on average". For a given sequence of "yes" and "no" responses which yields a log threshold estimate, E_j, the final p.d.f. gives the relative probabilities of different values of log real threshold, T, i.e.

$$p(T|E_j) = q_{Nj}(T) \Big/ \int q_{Nj}(T)\,dT \qquad (16)$$

where $p(T|E_j)$ means the probability of T given E_j. [The denominator in equation (16) ensures that $\int p(T|E_j)\,dT$ is unity]. The expected value of log real threshold (i.e. the weighted average of all possible values of log real threshold) is thus given by

$$\hat{T}_j = \int T p(T|E_j)\,dT = \int T q_{Nj}(T)\,dT \Big/ \int q_{Nj}(T)\,dT = E_j \qquad (17)$$

[cf. Hays, 1988, p. 164 and equation (15)]. Thus the log threshold estimate, E_j, equals the expected threshold associated with the observed sequence of yes and no responses. For example, in Fig. 3, the expected values of log threshold, $\hat{T}_j$, are given by the thick solid vertical lines; these values lie on the diagonal line $E_j = T$ and are thus equal to the log threshold estimate, E_j.

To understand the meaning of this lack of bias, it is important to distinguish between two types of bias which we call "measurement bias" and "interpretation bias" and which the following diagram helps to explain:

$$\text{Real threshold} \underset{\text{Interpretation}}{\overset{\text{Measurement}}{\rightleftarrows}} \text{Estimated threshold.}$$

The term "measurement" is used here to include both the collection of data (e.g. "yes" or "no" responses at certain intensities) and the calculation of an estimated threshold from these data. "Measurement bias" could be

demonstrated by performing (or simulating) many repeat threshold measurements of a single real threshold; this bias would then be the difference between the mean of these log threshold estimates and the log real threshold. This type of bias has been evaluated in many threshold simulations and experiments (Emerson, 1986; Green, 1990, 1993; Green, Richards & Forrest, 1989; Hall, 1981; Kollmeier, Gilkey & Sieben, 1988; Laming & Marsh, 1988; Leek, Hanna & Marshall, 1992; McKee *et al.*, 1985; O'Regan & Humbert, 1989; Schlauch & Rose, 1990; Shelton, Picardi & Green, 1982; Swanson & Birch, 1992).

The term "interpretation" is used here as the inverse of "measurement"—i.e. given a single estimated value of threshold (corresponding, say, to a particular sequence of "yes" and "no" responses), one may ask what values of real threshold could have given rise to this threshold estimate. More specifically, what are the relative probabilities of different real thresholds which could have given rise to this threshold estimate, and what is the weighted average of these real thresholds? As noted above, these relative probabilities are given by the final p.d.f. [equation (16)] in our versions of the QUEST procedure. Thus the average value of log real threshold which could have caused this threshold estimate is equal to the mean of the final p.d.f., which is, in fact, the log threshold estimate [equation (17)]; our measurements are therefore free of interpretation bias.

Perhaps surprisingly, a method which is free of interpretation bias is typically not free of measurement bias; in other words, if many measurements are performed on a subject with log real threshold, T, the expected value of log estimated threshold, $\hat{E}$, generally differs from T, i.e.

$$\hat{E} \neq T \tag{18}$$

[cf. equation (17)]. Figure 3 helps to illustrate this inequality. $\hat{E}$ is a weighted mean of the eight possible log threshold estimates, i.e.

$$\hat{E} = \sum_{j=1}^{2^N} E_j p(E_j|T). \tag{19}$$

Bayes's theorem may be used to evaluate $P(E_j|T)$ yielding

$$P(E_j|T) = q_{Nj}(T)/q_0(T). \tag{20}$$

In Fig. 3, $P(E_j|T)$ is therefore proportional to the heights of the p.d.f.s at T, e.g. the thin vertical lines which are shown for $T = 2$. $\hat{E}$ is then the center of gravity or weighted mean of the eight possible log threshold estimates, where the weights are given by the heights of the vertical lines. For $T = 2$, the calculated value [from equations (19) and (20)] of $\hat{E}$ is 2.01 and so is close to the log real threshold. It can readily be shown that measurement bias is inevitable for some values of T, e.g. $T = 4$, represented by a thin line in Fig. 3. The highest log threshold estimate (leftmost p.d.f.) is 3.37, and $\hat{E}$ cannot be greater than this, implying a large measurement bias. In fact, the weighted mean of log threshold estimates is $\hat{E} = 3.30$, (circle, Fig. 3) implying a measurement bias, $\hat{E} - T$, of -0.70; this bias is the distance (along the line $T = 4$) of the circle from the diagonal line, $E_j = T$.

It may be concluded that the current methods have measurement bias [equation (18)] even though they have no interpretation bias [equation (17)]. An example of measurement bias is given in Fig. 11, whereas interpretation bias caused by wrong assumptions is illustrated in Figs 14 and 15.

An alternative way of thinking about measurement bias (Pelli, personal communication) is that it is due to violation of an assumption of the QUEST and related methods. The procedure for assessing measurement bias fixes threshold to a particular value, violating the assumption that the successive threshold measurements are for random samples from the initial p.d.f. Pelli notes that the evaluation of *measurement* bias in the numerous papers quoted above may be inappropriate for methods such as the current ones, which are free from interpretation (but not measurement) bias.

METHODS

Simulations were performed using Fortran on a Cray YMP supercomputer. Experiments of up to 20 trials were simulated, using yes–no and two alternative forced choice (2AFC) conditions. (Although 20 trials is fewer than is typically used in a 2AFC experiment, even for infant and clinical studies, the results are valuable in demonstrating the performance of different methods in the early trials, where there is the greatest difference between different techniques; also, trends in the data may indicate performance differences for a greater number of trials.) The intensity of the next stimulus was determined by using the mode, mean or median of the current p.d.f., or by using the Minimum Variance Method. The effects of changing the threshold criterion [ϵ in equation (9)], the slope of the psychometric function [β in equation (9)] and the initial p.d.f. were also studied.

An exact enumeration technique was used (McKee *et al.*, 1985) as follows. For any of the 2^N sequences, j [equation (14)], the probability, P_j, of that sequence can be calculated, as well as the variance, V_j, of the final log threshold estimate. For example, Fig. 1 illustrates the two possible sequences ("yes" and "no") for the very simple case of an experiment with only one trial ($N = 1$); the probability, P_j, of say, the "yes" sequence is given by equation (3), the area under the "yes" p.d.f. in Fig. 1(C). In general

$$P_j = \int q_{Nj}(T)\,dT \tag{21}$$

where $q_{Nj}(T)$ is the final p.d.f. [cf. equation (3)]. The variance of the log threshold estimate will be the variance of this p.d.f., i.e.

$$V_j = \int (T - E_j)^2 q_{Nj}(T)\,dT \Big/ \int q_{Nj}(T)\,dT \tag{22}$$

where E_j is the threshold estimate for sequence j [equation (15)]. These calculations can be performed for

all 2^N sequences in an experimental run of N trials (e.g. Fig. 3). An overall weighted variance is then given by

$$V_N = \sum_{j=1}^{2^N} P_j V_j. \quad (23)$$

An overall standard deviation is the square root of this variance; for a given condition, this was calculated as a function of N, the number of trials per run. For completeness, our plots show the standard deviation for $N = 0$, which is simply the standard deviation of the initial p.d.f. [0.745 log units for the p.d.f. of Fig. 1(A)]. Calculations were performed over a range of 5 log units of threshold (as in Fig. 1), using a step size of 0.05 log units as in Watson and Pelli (1983). For calculating the next intensity, the mean, mode or median of the current p.d.f. was rounded to the nearest 0.05 log unit step; however, the final estimate of threshold [mean of the final p.d.f., equation (15)] was not rounded in this way.

For the Minimum Variance Method (with one-trial look-ahead), the stimulus intensity is chosen which minimizes the expected variance at the end of the next trial [equation (11)]. In practice, the mean (rounded to the nearest 0.05 log units) of the current p.d.f. was used as a starting point and the expected variance from equation (11) was calculated for this intensity and for the two neighboring intensities (i.e. at ± 0.05 log units). If the central intensity value provided the minimum expected variance, this was the intensity used for the next trial; otherwise, the sequence of intensities was extended by 0.05 log units from the intensity which yielded the lowest expected variance, and this was repeated, as necessary, until a (local) minimum of expected variance was obtained. This local-minimum method was also used for two-trial look ahead. In some simulations, a more exhaustive search for a (global) minimum was performed (by considering all intensities within ± 1 log unit from the mean of the current p.d.f.); the two methods typically agreed to within 0.2% of overall standard deviation, so the results reported here are for the first (local minimum) method, which requires considerably less computation.

Practical implementation of mean-QUEST (ZEST) is described in the Appendix.

RESULTS

Simulations using the "ideal sweat factor"

For the first simulations, the threshold criterion parameter, ϵ in equation (9), was set to ϵ_{id} corresponding to the "ideal sweat factor" (see Introduction; Taylor, 1971; Watson & Pelli, 1983). Figure 4(A) is a plot of overall standard deviation [from equation (23)] as a function of number of trials for simulated yes–no experiments; to improve the separation of data for different conditions, the *logarithm* of the standard deviation has been plotted. As noted earlier, the standard deviation for zero trials is simply the standard deviation of the initial p.d.f. [Fig. 1(A)] and so is the same for all methods.

It is seen that the standard deviations for *mean*-QUEST (stimulus intensity set to the mean of the current p.d.f.) are lower than for *median*-QUEST, which in turn

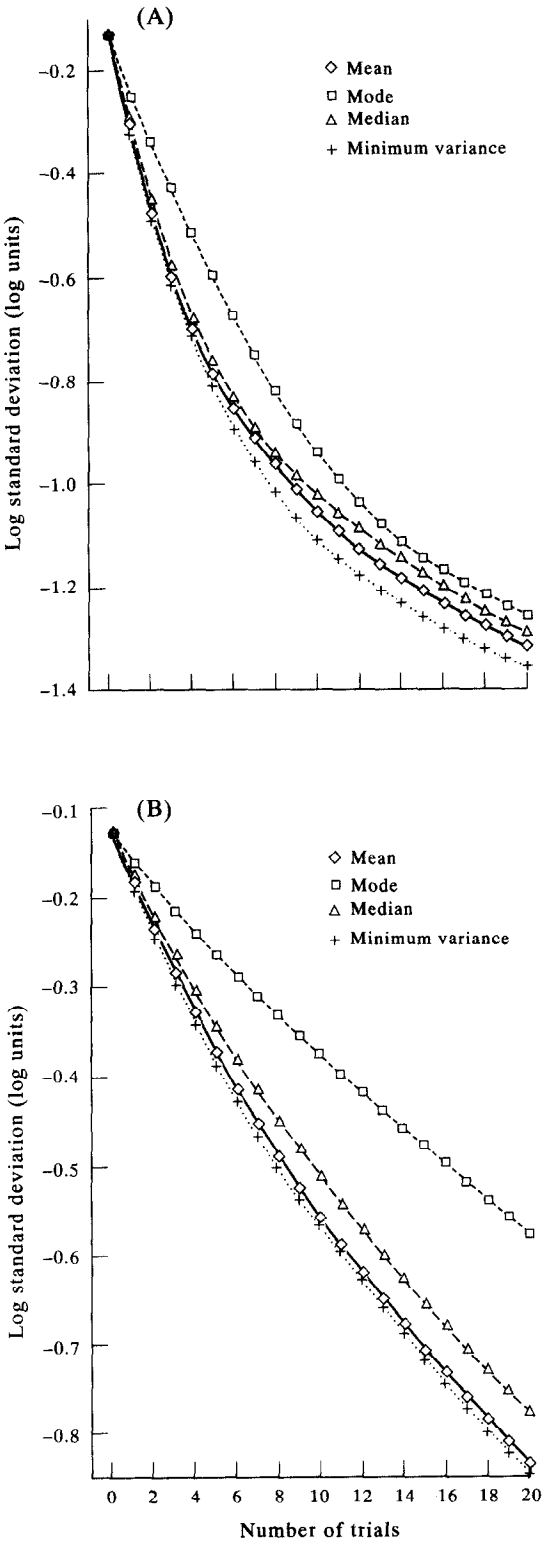

FIGURE 4. Overall standard deviation [the square root of the overall weighted variance from equation (23)] plotted as a function of the number of trials. The simulations were performed using the ideal sweat factor. Diamond, squares and triangles correspond to variations of the QUEST method where the next log intensity is set respectively to the mean, mode and median of the current p.d.f. of log threshold (mean-, mode- and median-QUEST). Using the mode corresponds to the standard version of the QUEST method, whereas the name ZEST is proposed for using the mean. The crosses correspond to the Minimum Variance Method (see text). The values plotted for zero trials correspond to the standard deviation of the initial p.d.f. (A) Yes–no method, $\epsilon_{id} = 0.052$ log units. (B) 2AFC method, $\epsilon_{id} = 0.063$ log units. See caption of Fig. 3 for other details.

are lower than for *mode*-QUEST. For example, after eight trials the overall standard deviations are 0.1101, 0.1136 and 0.1522 log units for mean-, median- and mode-QUEST respectively. {For the standard QUEST procedure (Watson & Pelli, 1983) of using the mode for the next stimulus intensity and setting the final threshold estimate equal to the mode of a likelihood function derived by dividing the final p.d.f. by the initial p.d.f., the overall error [derived by replacing E_j in equation (22) with this final threshold estimate] would have been 0.2481—i.e. over twice the value for mean-QUEST.} On account of the greater rapidity of achieving a certain accuracy when setting the stimulus intensity to the *mean* of the current p.d.f., we propose that this variation of the QUEST method be named ZEST—Zippy Estimation by Sequential Testing.

Corresponding results for 2AFC simulations are given in Fig. 4(B). The same ordering of accuracies is seen, with mean-QUEST (ZEST) giving the lowest overall standard deviation, followed by median-QUEST and mode-QUEST. In this case, the advantage of using the mean rather than the mode is considerably greater; accuracy from 20 trials using the mode is poorer than that from only 11 trials using the mean, so the standard condition uses over nine extra trials at this stage. Judging from the trends of the data in Fig. 4(B), this deficit would probably be greater for longer experiments.

The optimum value of threshold-criterion

The ideal sweat factor provides the greatest threshold accuracy, for a number of trials at a *fixed* intensity; however, in QUEST and other threshold methods where a considerable range of intensities are used, it is not clear that the "ideal" sweat factor is optimum. Therefore, threshold simulations were performed, each with a different value of ϵ, and typical results for yes–no simulations with eight trials are given in Fig. 5. Overall weighted standard deviation (left scale) has been plotted as a function of ϵ (epsilon), for mean-, mode- and median-QUEST; the corresponding probability of seeing at threshold is given by the crosses (right scale). The value of ϵ_{id} (0.052 log units) is shown by the vertical dashed line. It is seen that, for this limited number of trials, ϵ_{id} does not give the lowest standard deviation; for the mode, minimum standard deviation is given by a larger value of ϵ, whereas for the mean, a smaller ϵ yields the best performance.

In Fig. 6(A), the overall standard deviation of yes–no simulations is plotted against number of trials when the optimum ϵ (to the nearest 0.005 log units, plotted in lower panel) is used for each condition and for each number of trials; the horizontal line in the lower panel corresponds to ϵ_{id}. Standard deviations are generally slightly lower than for the ideal sweat factor (Fig. 4), but the ordering of accuracies is the same; the mean (ZEST) does better than the median which, in turn, does better than the mode. It should be noted that the optimum ϵ using the mean is close to zero from about 8 to 16 trials; this is the value of ϵ which we use in practice (Appendix). Results for 2AFC simulations are shown in Fig. 6(B); again, using the mean (ZEST) gives the lowest error, followed by the median and then the mode. For increasing number of trials, the optimum value of ϵ tends towards ϵ_{id}, for both yes–no [Fig. 6(A)] and 2AFC [Fig. 6(B)] simulations, but there is still a considerable discrepancy between the optimum ϵ and ϵ_{id} for a run of 16 trials.

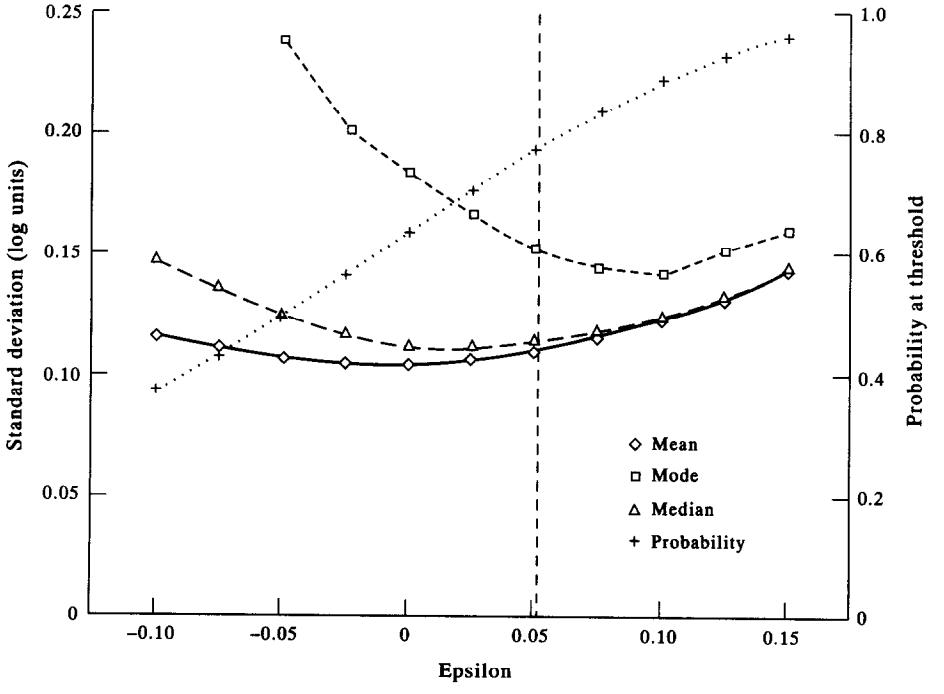

FIGURE 5. Overall standard deviation (left scale) for yes–no simulations using eight trials, plotted as a function of the threshold criterion factor, ϵ [equation (9)], for mean- (diamonds), mode- (squares) and median-QUEST (triangles). The crosses give the probability of seeing at threshold (right scale) and the vertical dashed line corresponds to the ideal sweat factor, $\epsilon = \epsilon_{id} = 0.052$ log units. See caption of Fig. 3 for details.

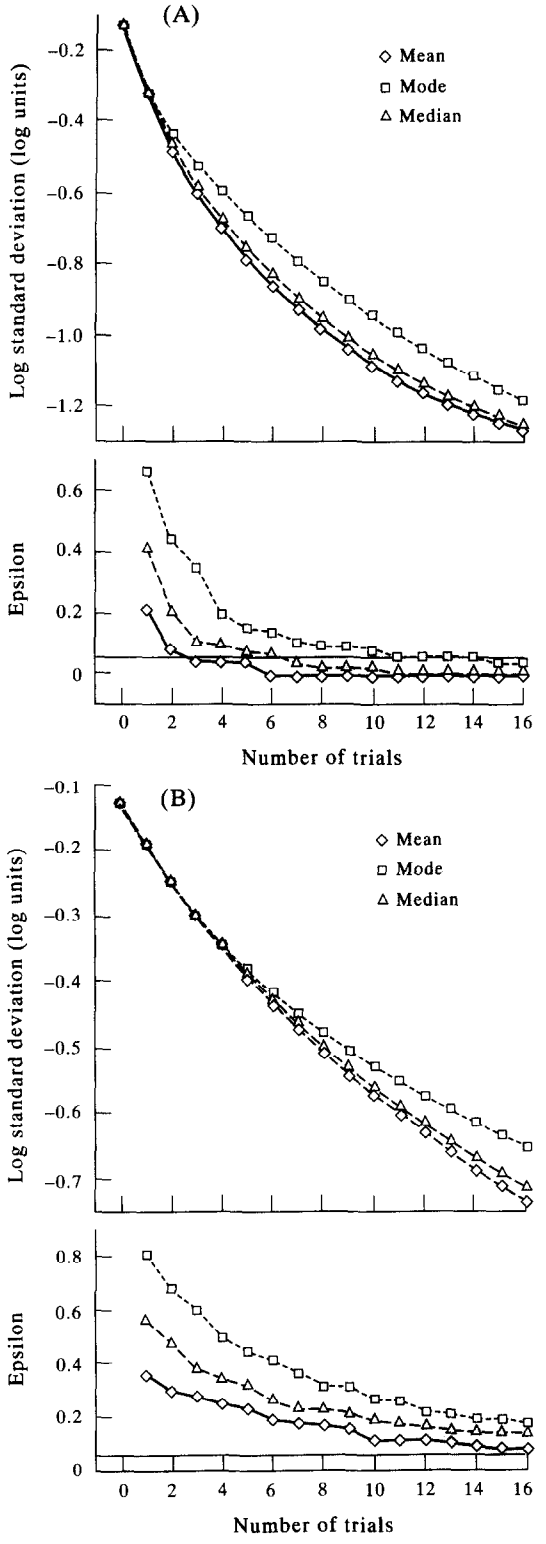

FIGURE 6. Overall standard deviation plotted as a function of the number of trials. The simulations were performed using the value of the threshold-criterion factor ϵ in equation (9) which yields the lowest standard deviation (cf. Fig. 5), for each condition (mean-, mode- or median-QUEST) and for each number of trials; this optimum ϵ is plotted in the lower panels. See caption of Fig. 3 for details. (A) Yes–no method. (B) 2AFC method.

Comparison with the Minimum Variance Method

As might be expected, the Minimum Variance Method (using one-trial look-ahead) is the most accurate of the four methods in Fig. 4. For the yes–no simulations of Fig. 4(A), the accuracy for 20 trials using ZEST may be obtained by only about 18 trials using the Minimum Variance Method; for 2AFC simulations [Fig. 4(B)], the advantage of the Minimum Variance Method compared to ZEST is relatively small—less than one trial after 20 trials.

The results shown in Fig. 4 for the Minimum Variance Method were for *one*-trial look-ahead. Our simulations indicate that *two*-trial look-ahead (as described in the Introduction) yields a relatively small advantage compared to one-trial look-ahead—0.03 trials after 16 trials for yes–no simulations and 0.4 trials for 2AFC. The Minimum Variance Method is theoretically independent of ϵ (because it should always choose the intensity which minimizes variance, regardless of the value of ϵ) and this was confirmed by simulations.

A summary comparison of some of the preceding data is shown in Fig. 7(A) (yes–no) and (B) (2AFC). Diamonds and circles correspond to using ZEST with the ideal sweat factor and optimum ϵ respectively; crosses are for the Minimum Variance Method with two-trial look-ahead. It is seen that, for yes–no experiments [Fig. 7(A)], there is some advantage (about two trials in 16) to using the optimum ϵ rather than the ideal sweat factor; the advantage is smaller for 2AFC [Fig. 7(B)]. For yes–no simulations [Fig. 7(A)], the Minimum Variance Method is slightly better (about one trial in 16) than using ZEST with the optimum ϵ. For 2AFC simulations [Fig. 7(B)], the advantage of the Minimum Variance Method is less; to our surprise, for about eight trials, it was actually slightly less accurate than ZEST using optimum ϵ—this is presumably because the Minimum Variance Method, which is optimal in the short term (two trials ahead), is not always the better strategy in the medium term.

Comparison of yes–no and 2AFC simulations

Comparison of the yes–no simulations in Figs 4(A) and 7(A) with the 2AFC simulations in Figs 4(B) and 7(B) illustrates the considerably higher accuracy of yes–no experiments [note the difference in vertical scales between Fig. 4(A) and (B)]. For example, after 20 trials using the Minimum Variance Method (one trial look-ahead), the overall weighted standard deviations for yes–no and 2AFC simulations are 0.0445 and 0.1428 log units respectively [cf. Fig. 4(A, B)]; the standard deviation for 2AFC simulations is therefore 3.21 times higher. The accuracy obtained from 20 trials for 2AFC is poorer than that from six trials for the yes–no simulation (Fig. 4).

Selection of stimulus intensity

The threshold methods described in this paper differ only in the selection of the stimulus intensity used for any trial. Figure 8 illustrates some differences between the selection of stimulus intensity for mode-QUEST (A), mean-QUEST (ZEST, B) and the Minimum Variance Method (C). For each of the three methods, the initial p.d.f., $q_0(T)$, is shown at the top and the corresponding

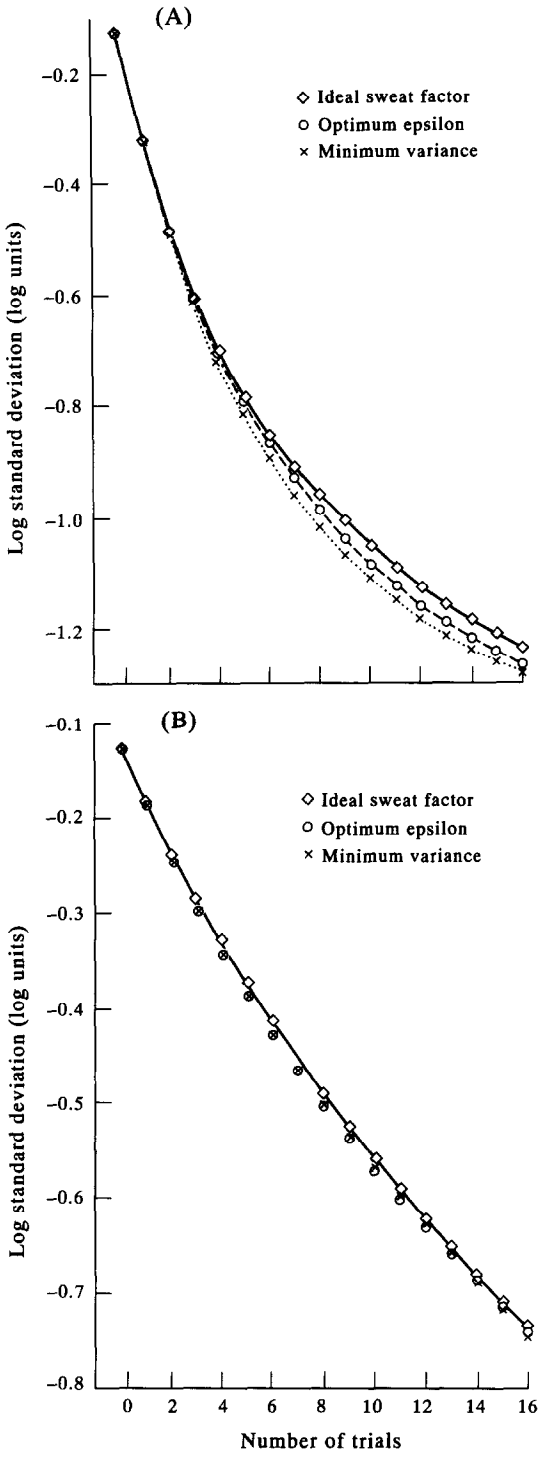

FIGURE 7. Summary of overall standard deviation plotted as a function of the number of trials. Two versions of ZEST (mean-QUEST) are shown; diamonds correspond to using the ideal sweat factor and circles correspond to using the optimum value of ϵ for a given number of trials (cf. Figs 5 and 6). Crosses correspond to the Minimum Variance Method with two-trial look ahead. See caption of Fig. 3 for other details. (A) Yes–no method. (B) 2AFC method.

initial stimulus intensity, x_1, is given by a vertical line. Assuming that the subject responds "no" to both the first and second trials ($r_1 = r_2 = 0$), the lower two plots for each method show $q_1(T)$ and $q_2(T)$, and corresponding stimulus intensities, x_2 and x_3 [see Fig. 1 and equation (7) for the derivation of these p.d.fs.].

It is seen that, after a "no" response, the stimulus intensity increases rather little for mode-QUEST (A) whereas it increases more for ZEST (B) and even more for the Minimum Variance Method (C). The relatively small increase in intensity for mode-QUEST (A) is because the stimulus intensity chosen is close to the steep end of the skewed p.d.f.; therefore the QUEST multiplication [equation (7)] does not "cut off" much of the p.d.f. By comparison, the intensities selected by ZEST (B) and the Minimum Variance Method (C) are further from the steep end of the p.d.f. and therefore the QUEST multiplication cuts off more of the p.d.f. Results for median-QUEST (not shown) were intermediate between mode-QUEST (A) and ZEST (B).

Figure 9 further illustrates the selection of stimulus intensity by mode-QUEST in comparison to the Minimum Variance Method. Log stimulus intensity is plotted as a function of trial number for the first five trials. "No" responses cause upward "steps" (i.e. increases of stimulus intensity at the next trial, cf. Fig. 8) whereas "yes" responses cause downward steps; circles and crosses are for response sequences with initial "no" and "yes" responses respectively. Results for mean- and median-QUEST are not shown but were intermediate between the two plots of Fig. 9.

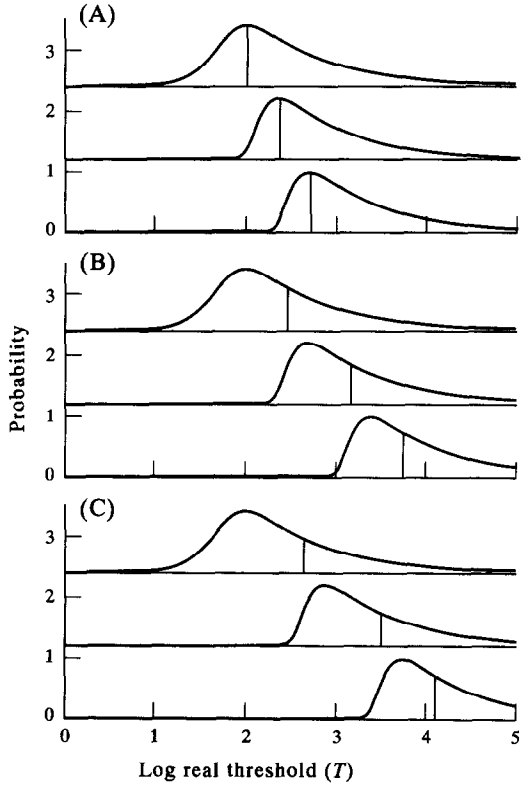

FIGURE 8. P.d.fs, $q_0(T)$, $q_1(T)$ and $q_2(T)$, before first three trials for mode-QUEST (A), ZEST (B) and the Minimum Variance Method (C); p.d.fs. have been scaled to a maximum of unity and shifted vertically for clarity. The subject responded "no" to the first two trials. Vertical lines give choice of stimulus intensities for the first three trials. A yes–no procedure was simulated, with $\epsilon = 0.09$ for mode-QUEST, $\epsilon = -0.01$ for ZEST (these values give the lowest error in an eight trial experiment) and $\epsilon = 0$ for the Minimum Variance Method (one-trial look-ahead). Other details are in caption to Fig. 3.

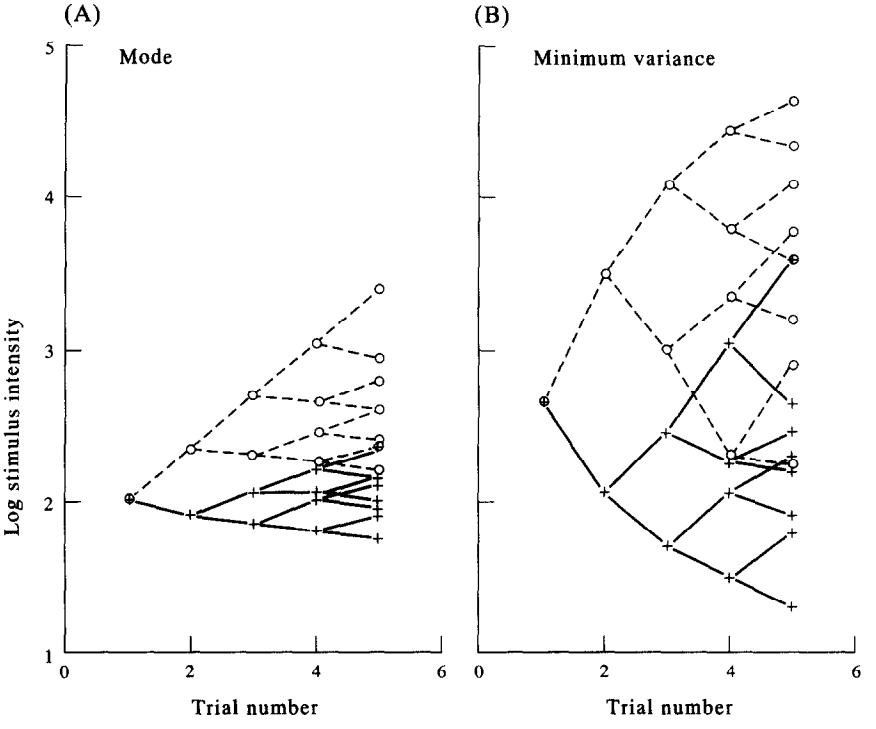

FIGURE 9. Stimulus intensities for the first five trials using mode-QUEST (left) and the Minimum Variance Method (right). Connecting lines of positive and negative slope correspond to "yes" and "no" responses, respectively; crosses and circles represent sequences of responses beginning with a "yes" or a "no" response respectively. Other details are in caption to Fig. 8.

As in Fig. 8, the initial steps in stimulus intensity for the Minimum Variance Method are much greater than for mode-QUEST. This should enable the Minimum Variance Method to obtain better threshold estimates when the real threshold is relatively high or low. The Minimum Variance Method also seems quicker in correcting for possible errors; for example, when an initial "yes" response is followed by three "no" responses, the intensity on the fifth trial (uppermost cross) is relatively high compared to that in mode-QUEST, as if the Minimum Variance Method had taken more account of the possibility that the initial "yes" response was a false positive.

Error as a function of estimated threshold

In a yes–no experiment with eight trials, there are $2^8 = 256$ possible sequences of responses and hence 256 possible threshold estimates. In Fig. 10, the standard deviation [from equation (22)] of each of these estimates is plotted as a function to the estimated threshold [equation (15)], for both mode-QUEST and the Minimum Variance Method. The area of each circle is proportional to the probability of the threshold estimate [equation (21)]. The leftmost and rightmost points for each method in Fig. 10 correspond respectively to sequences of eight "yes" and eight "no" responses. The grouping of points for mode-QUEST corresponds to different numbers of "yes" responses in the sequences; for example, the three near-vertical columns at about 4, 3.2 and 2.5 log units correspond respectively to 1, 2 and 3 "yes" responses in the eight trials.

The overall standard errors [equation (23)] are 0.1413 and 0.0965 log units for mode-QUEST and the Minimum Variance Method respectively. Corresponding plots for mean- and median-QUEST are intermediate in appearance between the extremes shown in Fig. 10.

Two major differences between mode-QUEST and the Minimum Variance Method can be seen. First, sequences yielding high standard deviations (e.g. over 0.4 log units) are more common for mode-QUEST than for the Minimum Variance Method; (these are often sequences which may contain false positive or false negative responses, such as one "yes" response followed by seven "no" responses). Secondly, the Minimum Variance Method gives a broader range of threshold estimates (about 4.2 log units compared to 3.3 log units for mode-QUEST); therefore more accurate estimates of high and low thresholds can be made.

As noted in the Introduction, if the assumptions made are correct, these methods are free of "interpretation bias", i.e. the threshold estimate equals the expected mean of the possible real thresholds which could have given rise to that sequence of responses.

Error as a function of real threshold

When error is considered as a function of real, rather than estimated threshold, two differences should be noted. First, real threshold is a continuous variable whereas, for a measurement with N trials, there are only 2^N discrete values of estimated threshold (e.g. Fig. 3). Second, in addition to random measurement error, bias must also be considered (see discussion of measurement bias in the Introduction).

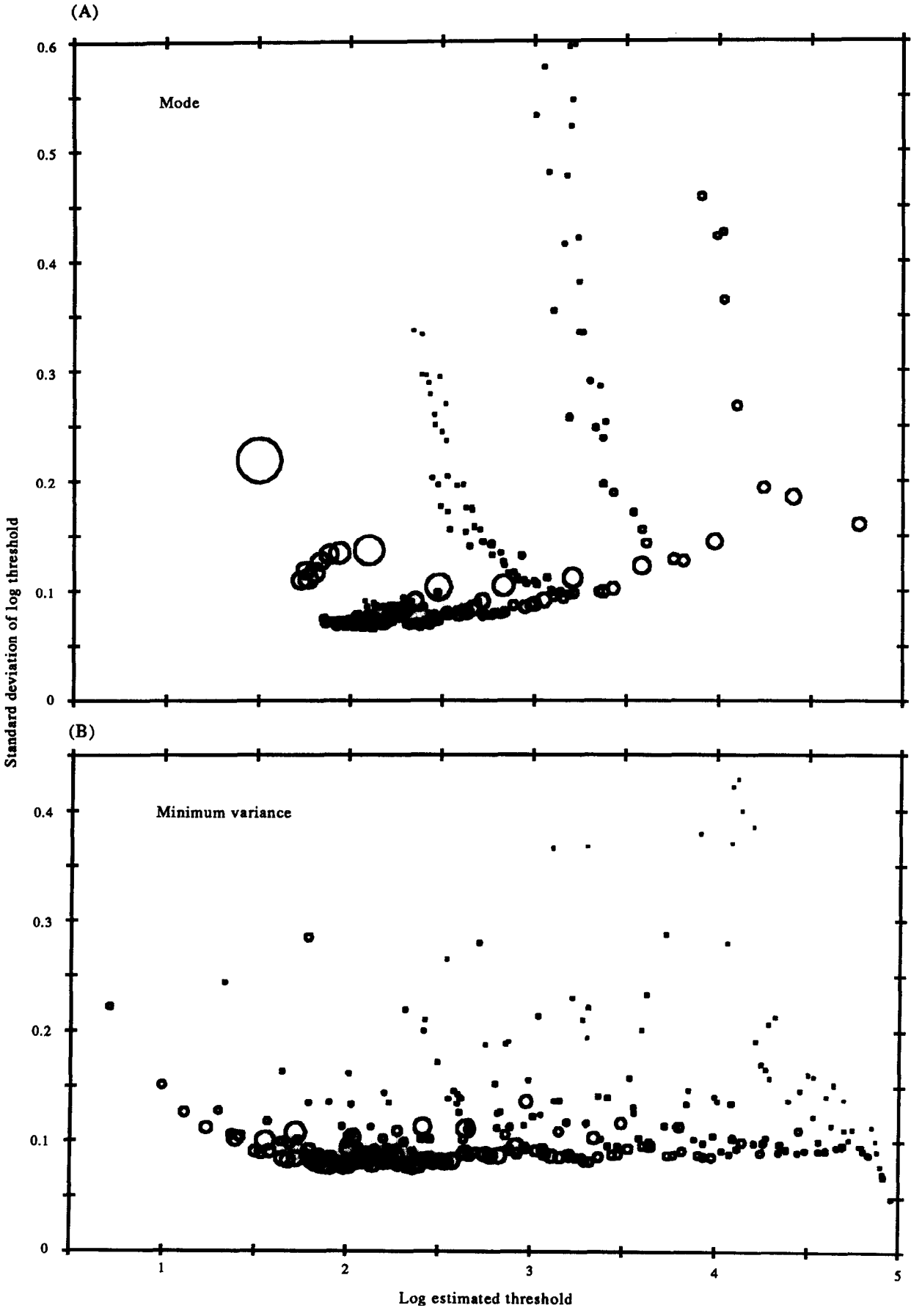

FIGURE 10. Characteristics of the 256 possible threshold estimates in a yes–no experiment with eight trials, using mode-QUEST (upper) and the Minimum Variance Method (lower). The standard deviation of each log threshold estimate is plotted as a function of the log estimated threshold; the area of each circle indicates the probability of the sequence of responses which yielded that threshold estimate. Other details are in caption to Fig. 8.

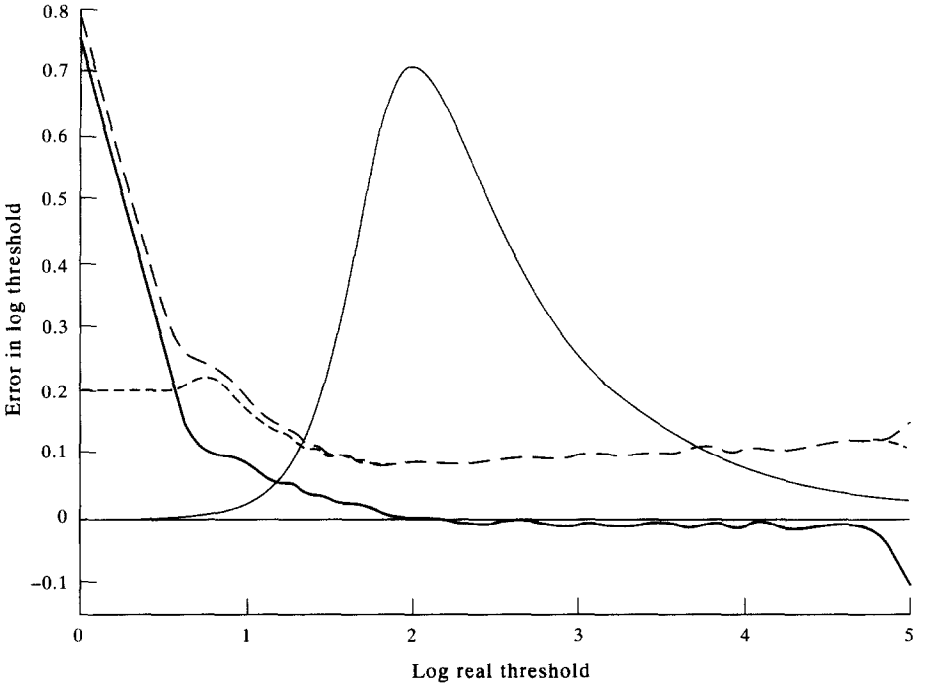

FIGURE 11. Errors in the log threshold plotted as a function of log *real* threshold; a yes–no, Minimum Variance Method was simulated using eight trials. The thick solid line gives "measurement bias"; the short-dashed line gives random measurement error (standard deviation of the threshold estimates about their mean value); the long-dashed line gives total error (r.m.s. deviation of threshold estimates about the log real threshold). The initial p.d.f. is given by the thin solid curve. See text for details.

For any log real threshold, T, measurement bias, b is given by

$$b = \hat{E} - T \tag{24}$$

where $\hat{E}$, the expected value of the log threshold estimate, is given by equation (19). This bias is plotted as the thick solid line in Fig. 11, which shows various types of error associated with the Minimum Variance Method (conditions as in Figs 9 and 10). The short-dashed line corresponds to random measurement error—the standard deviation, σ, of the log estimated thresholds about their expected value, $\hat{E}$, given by

$$\sigma^2 = \sum_{j=1}^{2^N} (E_j - \hat{E})^2 P(E_j|T). \tag{25}$$

$P(E_j|T)$ can be derived from equation (20). The long-dashed line gives the total r.m.s. error, s, of the threshold estimates relative to the real threshold value, T, given by

$$s^2 = \sum_{j=1}^{2^N} (E_j - T)^2 P(E_j|T) = \sigma^2 + b^2. \tag{26}$$

The thin solid curve is the initial p.d.f. of log real threshold [from Fig. 1(A)].

It is seen that there is a broad range of log real threshold where measurement bias (thick solid line) is relatively small compared to overall error (long dashed line). Estimated thresholds tend to be too high when the real threshold is improbably low, and too low when the real threshold is improbably high. Measurement bias is relatively severe (0.76 log units) for the lowest log real threshold, $T = 0$, but it may be noted that this value of real threshold is very improbable (thin solid curve).

Figure 12 gives a comparison of total error, s [equation (26)], as a function of real threshold for mode-QUEST (solid line), median-QUEST (short dashes), ZEST (medium dashes) and the Minimum Variance Method (long dashes); a yes–no procedure with eight trials was again simulated and optimum values of ϵ were used, as in Fig. 6(A). For the middle of the range of log real thresholds (about 1.5–3.5) all four methods give similar errors. However, for relatively high and low thresholds, the Minimum Variance Method gives the least total error, followed by ZEST, median- and mode-QUEST. This superiority of the Minimum Variance Method compared to mode-QUEST is consistent with the results of Figs 9 and 10.

Effect of initial p.d.f.

The preceding simulations were based on the initial p.d.f. of threshold shown in Fig. 1(A) which was derived from Fig. 2(A), the threshold histogram for an oscilloscope display. Figure 13 shows the effect of changing the initial p.d.f. to one derived from the threshold histogram for a color video display [Fig. 2(B)]; yes–no ZEST simulations were used with $\epsilon = 0$. It is seen that the results are not greatly affected by the choice of initial p.d.f.; the standard deviation of the initial p.d.f. for the color video display is somewhat lower than for the oscilloscope display, and this leads to a slight advantage (less than one trial) throughout the range of trials studied (up to 20 trials). The similarity of the two sets of results in Fig. 13 indicate that similar data would probably be obtained from any reasonable initial p.d.f.

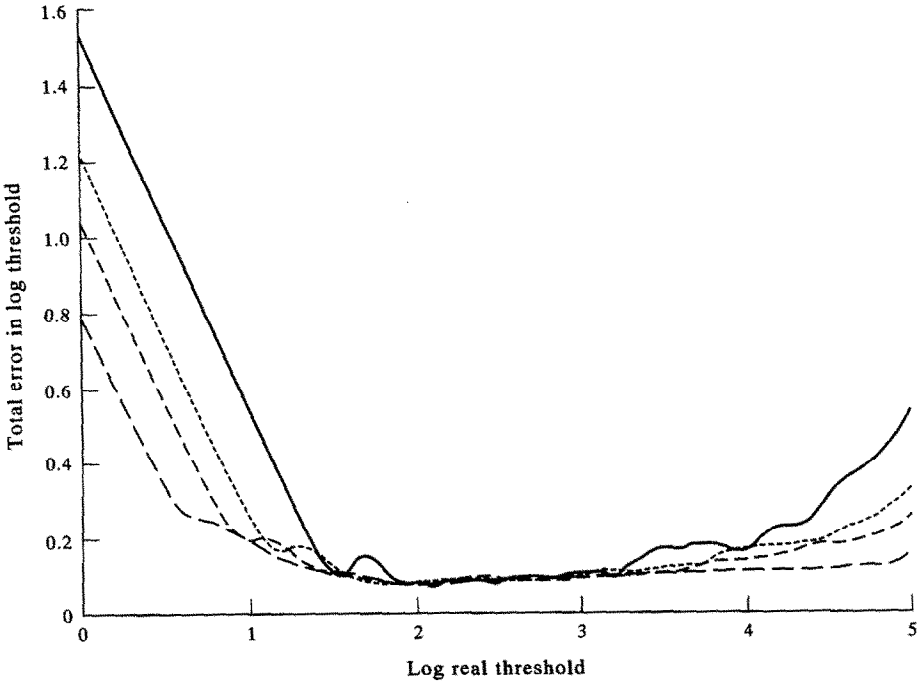

FIGURE 12. Total error (cf. the long dashes in Fig. 11) plotted as a function of log real threshold for yes–no simulations using eight trials. The Minimum Variance Method, mean-QUEST (ZEST), median-QUEST and mode-QUEST are represented by the long-dashed, medium-dashed, short-dashed and solid lines respectively. Optimum values of ϵ [from Fig. 6(A)] were used for mean-, median- and mode-QUEST. Other details are in caption to Fig. 8.

Figure 14(A) shows the effect of an error in the lateral position of the initial p.d.f. Simulations were performed in which the *assumed* initial p.d.f. was the same as in Figs 3–12 [i.e. that in Fig. 1(A)] but the *real* initial p.d.f. was shifted by 1 log unit to higher thresholds [i.e. the peak was now at 3 log units, rather than 2 log units as in Fig. 1(A)]; in the simulations, separate "assumed" and "real" p.d.fs. were calculated based on the corresponding initial p.d.fs. The *assumed* p.d.fs. were used for deriving stimulus intensities and for the final estimate of log threshold, E_j; however, to calculate the overall weighted error from equation (23), the mean square error was calculated by using the *real* final p.d.f., $q_{Nj}(T)$, in equation (22). It is seen from Fig. 14(A), that this rather severe error in the lateral position of the initial p.d.f. causes a rather small degradation of performance—a given level of accuracy typically requires less than one extra trial.

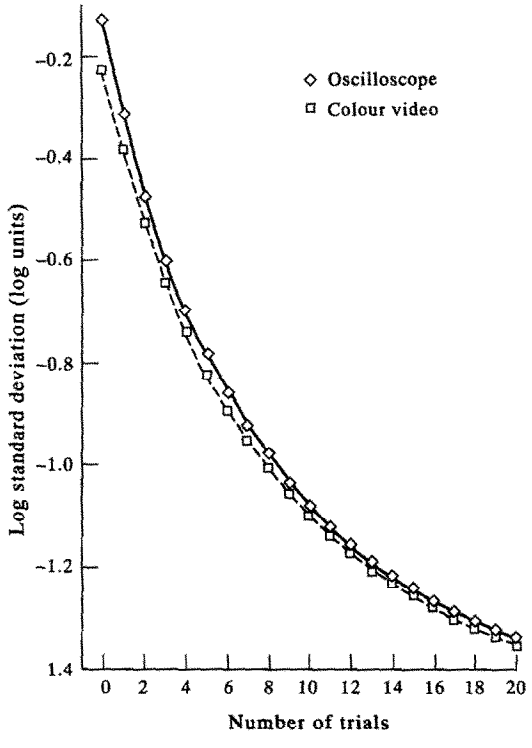

FIGURE 13. Effect of the initial p.d.f. on overall standard deviation. A yes–no, ZEST method was used with $\epsilon = 0$. Diamonds correspond to the initial p.d.f. of Fig. 1(A), which is based on the threshold histogram for an oscilloscope display in Fig. 2(A); this is the initial p.d.f. used in Figs 3–12. Squares correspond to an initial p.d.f. based on the threshold histogram for a color video display in Fig. 2(B). Other details are in caption to Fig. 3.

The error plotted as the circles in Fig. 14(A) includes two components; random error corresponds to the variance of the real final p.d.f. whereas interpretation bias, as defined in the introduction, is the difference between the mean of the "real" final p.d.f., $\hat{T}_j$, (an unbiased estimate of threshold) and the actual estimate of log threshold, E_j, (mean of the "assumed" final p.d.f.). This interpretation bias is plotted as a function of the actual log threshold estimate in Fig. 14(B) which illustrates the 256 different sequences of responses in an experiment with eight trials; the probability of each sequence [equation (21) using the real final p.d.f.] is indicated by the area of the corresponding circle (as in Fig. 10). For most sequences, there is little interpretation bias; the few sequences with relatively large biases of over 0.1 log units are represented by relatively small circles and so have relatively low probability (about 0.1% or less).

The weighted average of the interpretation bias for eight trials [Fig. 14(B)] is 0.0120 log units which is relatively small compared with the overall weighted error

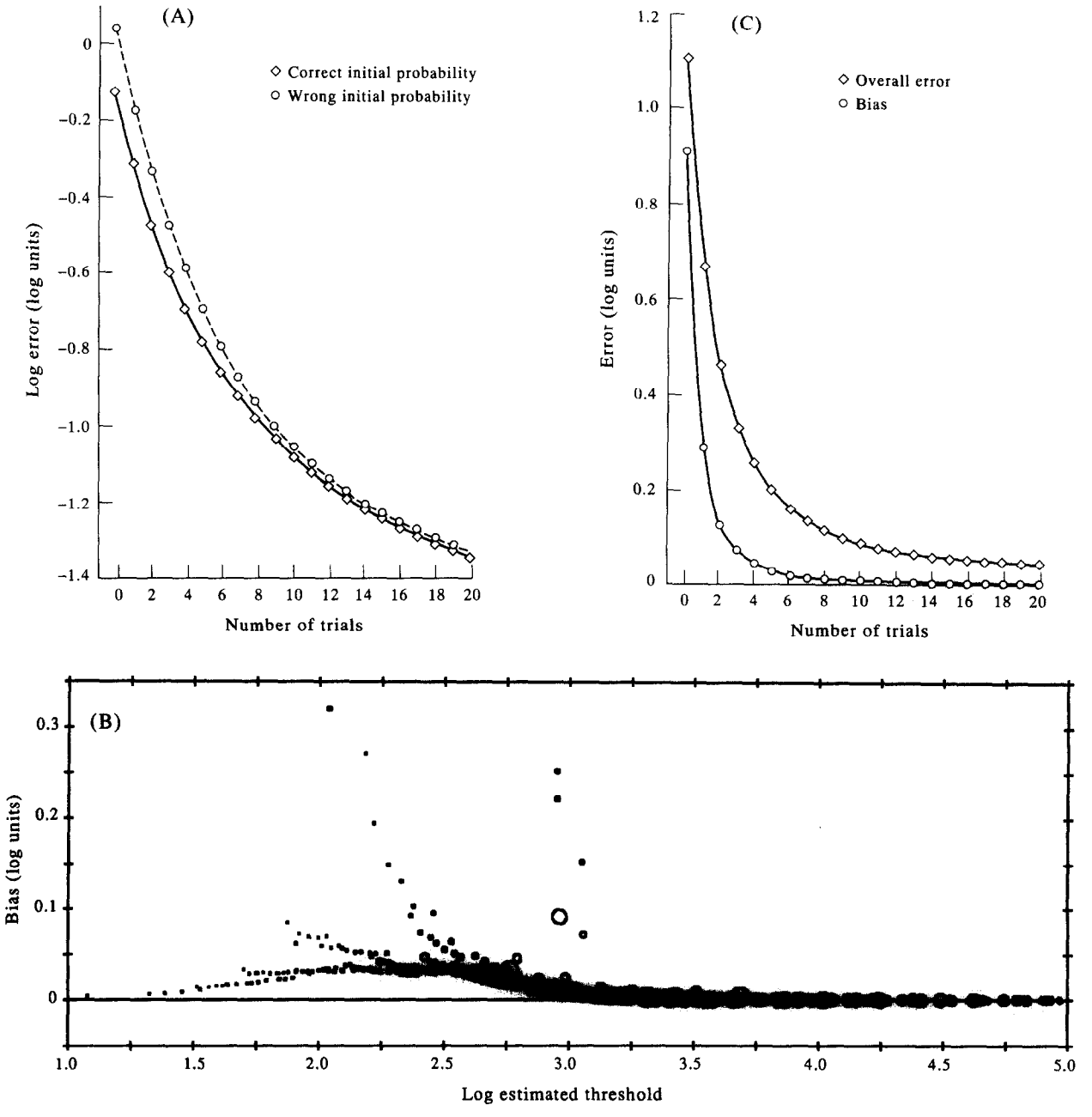

FIGURE 14. Effects of an error in the initial p.d.f. (A) Overall r.m.s. (root mean square) error plotted as a function of the number of trials for a yes–no ZEST method ($\epsilon = 0$). Diamonds correspond to the standard initial p.d.f. [Fig. 1(A)]. Circles correspond to a *real* initial p.d.f. which is shifted to higher thresholds by 1 log unit; however, the *assumed* initial p.d.f. is the standard function in Fig. 1(A). In this case, the r.m.s. error is determined from the mean-square error using the *real* final p.d.f. and the estimated threshold is calculated from the *assumed* initial p.d.f. (see text). Other details are in caption to Fig. 3. (B) A scatter plot of "interpretation" bias as a function of log estimated threshold for all 256 possible sequences of responses using eight trials. Interpretation bias is the difference between the mean of the "real" final p.d.f. (an unbiased estimate of log threshold) and the actual estimate of log threshold (mean of the "assumed" p.d.f.). The relative probability of each sequence is given by the area of the corresponding circle (as in Fig. 10). (C) Diamonds are the overall r.m.s. error replotted from the circles in (A) on a linear (rather than log) scale. Circles give the average interpretation bias [cf. (B)].

[circles, Fig. 14(A)] of 0.1157 log units. Figure 14(C) shows how this average bias varies as a function of number of trials and compares this with the overall error. The bias is relatively small after the first few trials.

The estimated standard deviation of threshold is also relatively unaffected by this error in the initial p.d.f.; for example, after eight trials, the calculated overall weighted standard deviation was 0.113 log units, whereas the r.m.s. deviation of real thresholds about the estimated threshold values was 0.116 log units.

Effect of slope of the psychometric function

Figure 15(A) indicates the effect of halving the real slope of the psychometric function from the standard value, $\beta = 3.5$, used in all previous simulations, to $\beta = 1.75$ (yes–no ZEST method, $\epsilon = 0$). Diamonds show

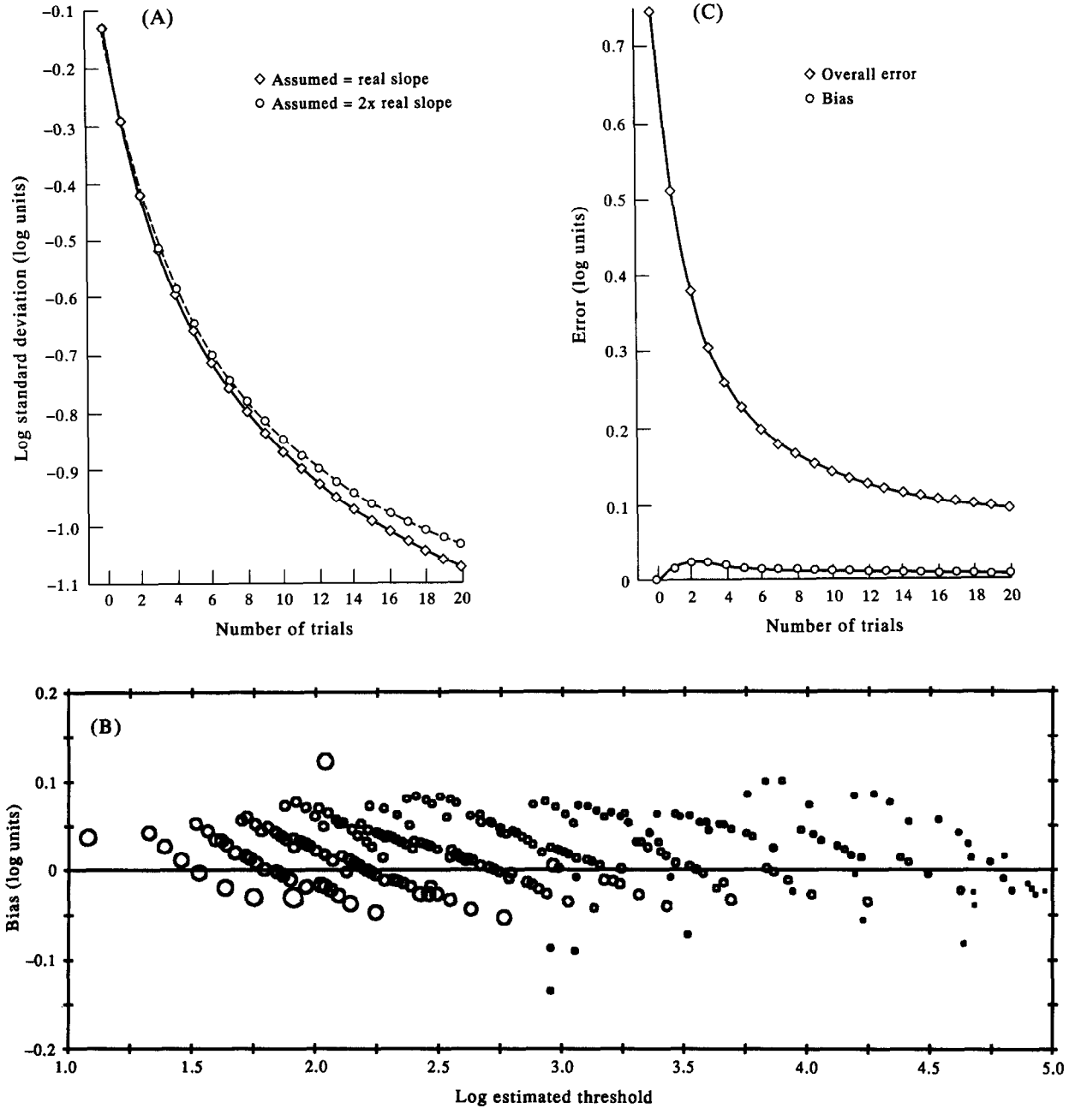

FIGURE 15. Effects of an error in the slope of the psychometric function. (A) Overall r.m.s. error plotted as a function of number of trials for a yes–no ZEST method ($\epsilon = 0$). For both diamonds and circles, the real slope of the psychometric function was $\beta = 1.75$—i.e. half that of previous simulations. For the diamonds, the assumed slope was equal to the real slope, whereas for the circles, the assumed slope was twice the real slope (and so equal to the values in previous simulations, $\beta = 3.5$). In the latter case, the r.m.s. error is determined from the mean-square error using the *real* final p.d.f. and the value of threshold calculated from the *assumed* slope (see text). Other details are in caption to Fig. 3. (B) A scatter plot of interpretation bias as a function of log estimated threshold for all 256 possible sequences of responses using eight trials. The relative probability of each sequence is given by the area of the corresponding circle (as in Fig. 10). (C) Diamonds are the overall r.m.s. error replotted from the circles in (A) on a linear (rather than log) scale. Circles give the average interpretation bias [cf. (B)].

the "ideal" performance when the assumed slope is equal to the real slope, 1.75, and the overall weighted standard deviation is calculated by the standard technique described in the Methods section. Circles show the performance when the assumed slope is the standard value, $\beta = 3.5$, and so is twice the real slope; in these simulations, separate p.d.fs for "real" and "assumed" slopes were calculated (see previous section). The error in assumed slope causes a modest reduction in performance; after 20 trials with the wrong slope, the overall weighted error is about equal to that after 17 trials with the correct slope.

Figure 15(B) gives a scatter plot of interpretation bias (due to this error in assumed slope) as a function of log estimated threshold [as in Fig. 14(B)]. Each point represents one of the 256 possible sequences in an experiment with eight trials; as in Fig. 14(B), the probability of any sequence is represented by the area of the

corresponding circle. For most sequences, the magnitude of the bias is less than 0.1 log unit. The average bias is 0.0146 log units; this is relatively small compared to the overall r.m.s. error of 0.1642 log units. In Fig. 15(C), overall error and average bias are plotted as a function of the number of trials; average bias is always relatively small compared to the overall error of the threshold estimate.

It should be noted that this error in the assumed slope of the psychometric function causes a considerable error in the calculated standard deviation of the log threshold estimate; for example, the calculated overall standard deviation after eight trials was 0.116 log units, whereas the r.m.s. deviation of real thresholds about the estimated threshold values was 0.164 log units [Fig. 15(A) circles].

Precision and efficiency

The efficiency of a threshold measurement may be determined by comparing its variance to a prediction based on the ideal sweat factor [equation (10)]. More specifically, if $K_{\min}$ is the ideal sweat factor [minimum value of $K(X)$ in equation (10)], then efficiency is given by

$$\eta = (K_{\min}/N)(1/V_N - 1/V_0) \quad (27)$$

where V_N is the overall weighted variance at the end of N trials [equation (23)] and V_0 is the variance of the initial p.d.f. (Taylor, 1971). [It should be noted that "efficiency" has a different meaning in automated perimetry (e.g. Johnson & Shapiro, 1989). Also, one authority, (Pelli, personal communication) proposes that, for the brief threshold runs considered here, the above definition of efficiency should be replaced by one based on comparison with the Ideal Psychometric Procedure; in our opinion, both Taylor's and Pelli's definitions of efficiency are of value, and a new name, e.g. "relative efficiency", should be used for Pelli's measure.] Using equations (9) and (10), the values of $K_{\min}$ for yes–no and 2AFC experiments are 0.0259 and 0.0596 respectively; thus, for our assumptions and for ideal conditions (all stimulus intensities very close to the intensity which gives the ideal sweat factor) a yes–no experiment would need only 0.0259/0.0596 = 43.5% of the number of trials needed by a 2AFC experiment to achieve a given accuracy.

It is convenient to define "precision", R, as the reciprocal of variance, (Taylor, 1971), so that equation (27) may be rewritten

$$\eta = K_{\min}(R_N - R_0)/N \quad (28)$$

where R_N and R_0 are the final and initial precisions; thus, efficiency is proportional to the precision added to the threshold estimate by the threshold method, divided by the number of trials. Equation (28) may be rearranged to yield

$$R_N = R_0 + \eta N/K_{\min}. \quad (29)$$

Thus, for an efficiency of 100%, a plot of R_N vs N would be a straight line of slope $1/K_{\min}$. In typical threshold measurements, the efficiency is considerably less than 100% and the slope of such a plot will be reduced in proportion to the efficiency.

Figure 16(A) is a plot of precision [reciprocal of overall weighted variance, $1/V_N$, from equation (23)] as a function of number of trials for yes–no experiments using the Minimum Variance Method (one-trial look-ahead). Line AB corresponds to 100% efficiency [cf.

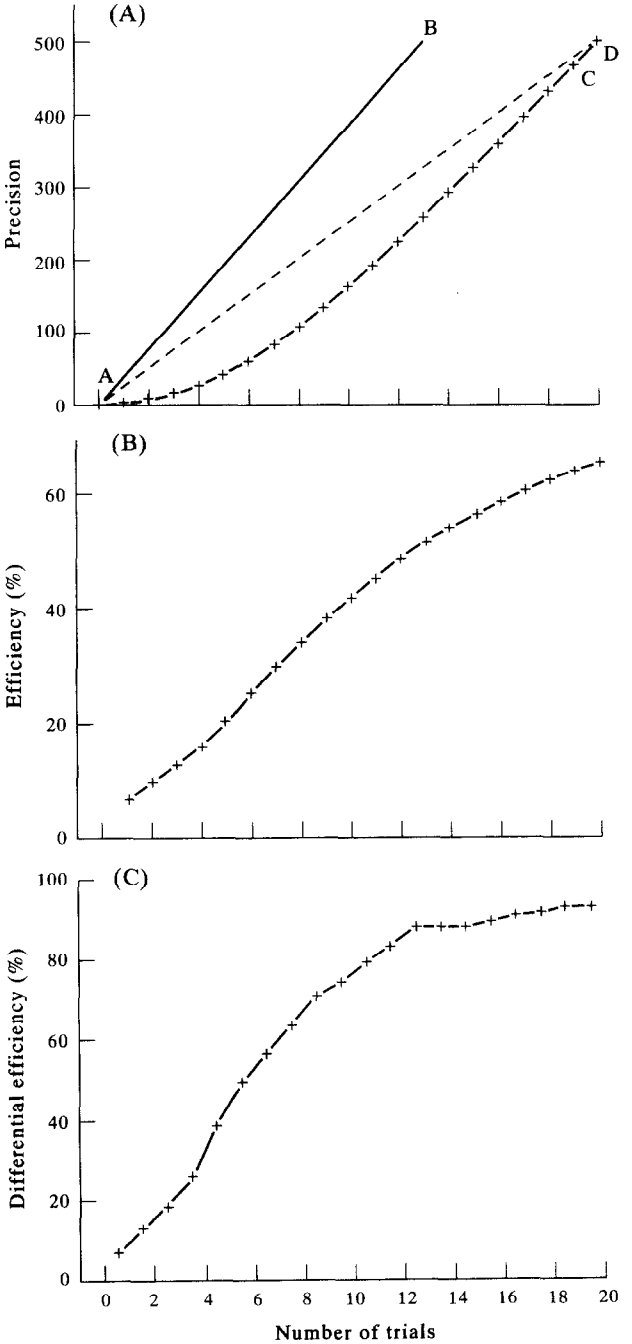

FIGURE 16. (A) Precision (reciprocal of variance) plotted as a function of number of trials for yes–no simulations using the Minimum Variance Method (one-trial look-ahead, +). The slope of line AB corresponds to 100% efficiency (see text); thus the ratio of the slope of AD to that of AB gives the efficiency using 20 trials and the ratio of the slope of CD to that of AB gives the differential efficiency (see text). See caption of Fig. 3 for details. (B) Efficiency as a function of number of trials. (C) Differential efficiency as a function of number of trials.

equation (29)]. Thus, it can be seen from the slope of the simulated data, that efficiency is relatively low for the first few trials, but improves with increasing number of trials; for example, after 20 trials, efficiency is the ratio of the slope of the dashed line, AD, to the "ideal" slope of AB. Efficiency may be derived in this way, or by using equation (28), and it is plotted as a function of number of trials in Fig. 16(B); it reaches about 34% after eight trials and 65% after 20.

"Differential efficiency", η', is a measure of efficiency for any given trial in an experiment and is given by an equation similar to equation (28), i.e.

$$\eta' = K_{\min}(R_{N+1} - R_N). \tag{30}$$

It may be shown that this corresponds to the slope of the line segment connecting two data points in Fig. 16(A) (e.g. CD), divided by the slope for 100% efficiency (AB). Differential efficiency is plotted as a function of number of trials in Fig. 16(C). Differential efficiency reaches over 90% after 20 trials, corresponding to the fact that in Fig. 16(A), the slope of CD is nearly as great as the ideal slope of AB.

Corresponding plots for 2AFC simulations are shown in Fig. 17 (note the differences in vertical scales from the corresponding yes–no simulations—Fig. 16). In Fig. 17(A), it is seen that the precision obtained from 2AFC simulations is much smaller than that from yes–no simulations [Fig. 16(A)]; the added precision, $(R_N - R_0)$, from 20 trials using 2AFC is 47.2 which is only 9.4% of the added precision (503) from 20 trials using yes–no [Fig. 16(A)]. In Fig. 17(A), the slope of the plot of precision vs number of trials is always much less than the slope for 100% efficiency (AB); for example, the slope AD is much less than that of AB [cf. Fig. 16(A)], implying a relatively low efficiency. Figure 17(B) gives a plot of efficiency as a function of number of trials; efficiency rises to only about 14% after 20 trials compared to about 65% for yes–no simulations [Fig. 16(B)]. Differential efficiency is also relatively small; thus, in Fig. 17(A), the slope of CD is considerably less than that of the ideal, AB [cf. Fig. 16(A)]. A plot of differential efficiency vs number of trials is given in Fig. 17(C); it reaches only about 30% after 20 trials, which is again much less than the corresponding value of over 90% for yes–no simulations [Fig. 16(C)].

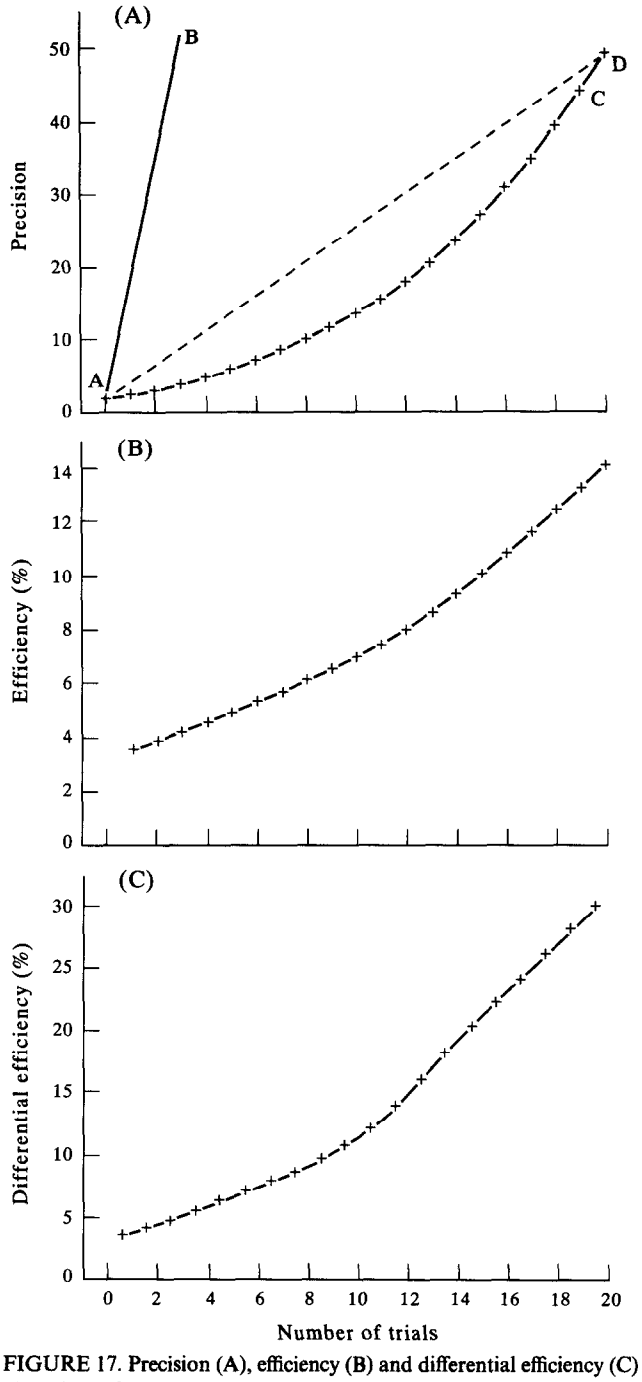

FIGURE 17. Precision (A), efficiency (B) and differential efficiency (C) plotted as a function of number of trials for a 2AFC method using the Minimum Variance Method (one trial look-ahead). See captions of Figs 3 and 16 for details.

An evaluation of the accuracy of the ZEST method

In many of our color-mixture threshold measurements (cf. Grigsby *et al.*, 1991), two independent threshold determinations were made for a certain color-mixture. The "observed" standard deviation of these measurements can be estimated from the formula

$$\sigma = \sqrt{[\Sigma(E_1 - E_2)^2/2n]} \tag{31}$$

where E_1 and E_2 are such a corresponding pair of threshold estimates and the summation is over n pairs of measurements. This observed standard deviation can be compared with the corresponding "calculated" standard deviations, s_1 and s_2 from equation (22); an overall estimate of calculated standard deviation was derived from

$$S = \sqrt{[\Sigma(s_1^2 + s_2^2)/2n]} \tag{32}$$

where, again, the summation is over all n pairs of threshold measurements. Methods were those described in the Appendix with eight trials per threshold measurement. We excluded from equations (31) and (32) any pair of thresholds in which the subject never responded during one or both measurements.

For 967 pairs of threshold measurements, the observed standard deviation [equation (31)] was 0.112 log units, whereas the calculated standard deviation

[equation (32)] was 0.098 log units; thus the observed standard deviation was 14% higher than the calculated value. A more detailed comparison is given in Fig. 18(A), where the data have been divided into different ranges of mean threshold, $(E_1 + E_2)/2$. Ranges were usually 0.1 log units except for very high and low thresholds where ranges were enlarged so as to include at least 10 pairs; the number of data pairs in each range is plotted against the mean threshold for that range in Fig. 18(B).

It is seen from Fig. 18(A) that there is reasonable agreement between observed and calculated standard deviations for intermediate intensities (log threshold from 0.5 to 1.75); random variation in the relative values of these observed and calculated standard deviations may be largely due to the small samples used to derive the observed standard errors [Fig. 18(B)]. However, the observed standard deviation is higher than the calculated value for high (and perhaps for low) thresholds.

The accuracy obtained with the ZEST method depends on the slope of the psychometric function and on the assumption that the threshold does not vary from trial to trial; if the real slope of the psychometric function is less than the assumed value and/or if the threshold is time-varying (which would have a similar effect to reducing the slope—Leek, Hanna & Marshall, 1991), the observed standard deviation of the measurements will be greater than the calculated value. The reasonable agreement between observed and calculated standard deviations at intermediate threshold values indicates that the assumed slope, $\beta = 3.5$, is about correct for these measurements. However, the fact that observed standard deviations are higher than the calculated values at high thresholds, indicates either that the real slope of the psychometric function tends to be less than the assumed slope (Weber & Rau, 1992) and/or that the threshold varies from trial to trial in these cases (Leek *et al.*, 1991). A possible explanation for the observed discrepancy is that these high thresholds correspond to subjects with severely impaired vision, whose thresholds may vary from trial to trial due to temporal fluctuations of visual sensitivity and/or spatial variations of sensitivity (coupled with unsteady fixation). If optimum conditions are required for the study of subjects with impaired vision, it would probably be advantageous to use a lower value of slope than our value of $\beta = 3.5$.

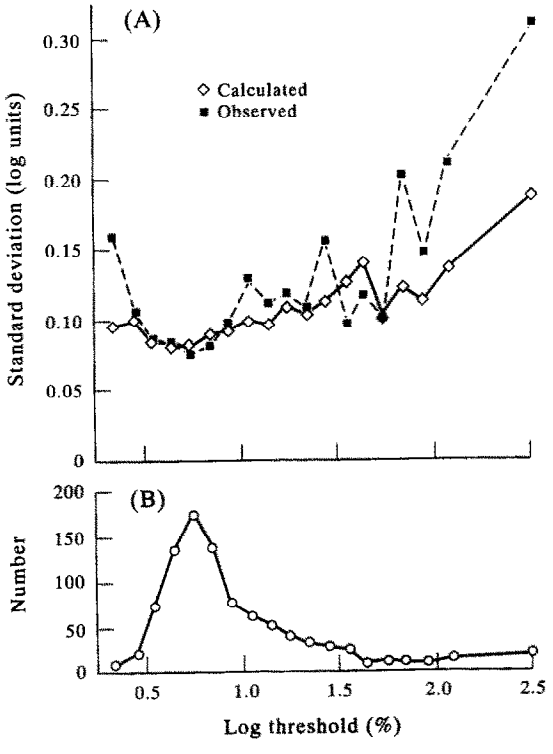

FIGURE 18. A comparison between the standard deviation of the threshold estimate calculated from the ZEST method (diamonds) and the observed standard deviation of measurements determined from duplicate threshold measurements (squares); data were averaged over ranges of estimated threshold of 0.1 or more log units and are plotted as a function of log threshold. The number of pairs of threshold measurements for each threshold range is given in the lower panel. A color television display system was used, whose initial p.d.f. is given in Fig. 2(B). Thresholds were determined in eight trials using a yes–no method with $\epsilon = 0$. See text and caption to Fig. 3 for other details.

DISCUSSION

Comparison of QUEST variations

Two general recommendations can be derived from these simulations for optimizing the precision of the QUEST method:

1. Using the ZEST modification—i.e. setting the stimulus intensity to the *mean* of the current p.d.f. provides greater precision than using the *mode* or *median*; this is true for both yes–no [Figs 4(A), 5, 6(A) and 7(A)] and 2AFC [Figs 4(B), 6(B) and 7(B)] simulations. This is in agreement with previous analyses by King-Smith (1984) and Emerson (1986).

2. For the relatively short experiments which are simulated here (up to 20 trials), optimum performance is not given by the "ideal sweat factor" but by a rather different, optimized, threshold criterion (Figs 5, 6 and 7). Harvey (1986) has come to a similar conclusion.

Thus, relatively good performance can be obtained from the QUEST method by setting the next intensity to the *mean* of the current p.d.f. and using the *optimum threshold criterion*. A slightly better precision can generally be obtained by the Minimum Variance Method (King-Smith, 1984; Pelli, 1987) for both yes–no and 2AFC simulations (Fig. 7). In deciding whether to use the Minimum Variance Method, the extra programming complexity and the additional computational time should be kept in mind. For these reasons, we have not, as yet, implemented the Minimum Variance Method experimentally.

Pelli (1987) describes an "Ideal Psychometric Procedure" which provides optimum performance under the assumptions made by the QUEST method. It is essentially the same as the Minimum Variance Method discussed here, but, instead of looking only one or two trials ahead, the computation is performed for many trials ahead—until the end of the experiment; in fact, the optimum strategy (e.g. stimulus intensity to use at any stage of the experiment, after any sequence of "yes" and

"no" responses) could be determined before the first trial. Although theoretically ideal, the method is particularly complex and time-consuming; each additional stage of "look-ahead" would increase the computation time by at least a factor of three. For this reason, look-ahead was limited to one or two trials in our simulations. It may be noted that two-trial look-ahead provided negligible advantage over one-trial look-ahead for yes–no simulations; it seems probable that extending the look-ahead to many trials would provide little further advantage. However, for 2AFC simulations, there was a rather greater advantage of using two-trial look-ahead relative to one-trial, so it is possible that multiple-trial look-ahead and the Ideal Psychometric Procedure would provide a worthwhile improvement in performance.

Comparison with other threshold methods

Simulations by Watson and Pelli (1983) indicated that, using a forced choice method, QUEST is more efficient than both the original version of PEST (Taylor & Creelman, 1967) and Hall's (1981) modified version of PEST; however, Watson and Pelli note that the conditions of the simulations were not identical. In a more direct comparison using both yes–no and 2AFC threshold methods by Madigan and Williams (1987), QUEST was again found to be more efficient than the standard version of PEST. Computer simulations by Pentland (1980) showed that, for yes–no experiments, a maximum-likelihood method was more efficient than a standard staircase method and two versions of PEST—the original version (Taylor & Creelman, 1967) and Findlay's (1978) modification; similarly, Lieberman and Pentland (1982) showed that the same maximum likelihood technique is more efficient than a modified staircase method (Wetherill & Levitt, 1965; Corwin, Kintz & Beaty, 1979). This maximum-likelihood method differed from QUEST in that it did not take advantage of information about the initial p.d.f. of thresholds; thus one would expect QUEST to be even more efficient than the maximum-likelihood method and so it would also be more efficient than both variations of PEST and both staircases. As shown in Figs 4 and 6, ZEST should, in turn, be more efficient than QUEST.

Shelton *et al.* (1982) used experimental determinations of thresholds to compare a standard staircase, PEST and a maximum-likelihood method; they found little difference in efficiency between the three threshold methods. The difference between this conclusion based on experimental data and the above conclusions based on simulations may be due to the more extensive data available by simulation; for example, it is relatively easy to simulate 500 or more experimental runs using a Monte-Carlo method (Findlay, 1978; Green, 1993; McKee *et al.*, 1985; Pentland, 1980) whereas, in their experimental study, Shelton *et al.* (1982) used, say, 48 runs of 80 trials for each threshold method. It may be noted again that QUEST, which takes advantage of knowledge of the initial p.d.f. should be more efficient than the maximum-likelihood method used by Shelton *et al.* (1982), and ZEST should be more efficient than QUEST.

In earlier yes–no simulations, King-Smith (1984) found similar results to those reported here, namely that ZEST yielded a higher efficiency than the QUEST method. Emerson (1986) simulated a method which is similar to the ZEST method (i.e. used the mean of the p.d.f. rather than the mode), but did not take into account an initial p.d.f.; for 2AFC, he found this technique to be more efficient than a maximum likelihood method (i.e. using the mode), as well as having less measurement bias.

In conclusion, the available evidence suggests that, given the assumptions made by the QUEST (and ZEST) methods (see Introduction), ZEST is more efficient than QUEST which is more efficient than PEST and standard staircases. We do not know of any studies which would rate the efficiency of either APE or MOBS relative to this sequence; both methods (APE—Watt & Andrews, 1981; MOBS—Tyrrell & Owens, 1988) appear to be superior to the method of constant stimuli, but the latter, being a non-adaptive method, is probably less efficient than most adaptive methods (Watson & Fitzhugh, 1990). The current simulations indicate that the Minimum Variance Method (King-Smith, 1984; Pelli, 1987) is typically slightly more efficient than ZEST and one would expect that the Ideal Psychometric Procedure (i.e. the Minimum Variance Method with look-ahead to the end of the experiment—Pelli, 1987) would have the highest efficiency of all. Taking account of the complexity of the calculations involved, this last procedure might be called BEST (Behemothic Estimation by Sequential Testing).

As a final point, it should be noted that [apart from the experimental studies of Shelton *et al.* (1982) which were rather indecisive] these conclusions have been derived from simulations on the assumptions that individual trials are statistically independent and that threshold does not vary from trial to trial. Taylor, Forbes and Creelman (1983) show that the assumption of statistical independence is not strictly true for real subjects (see Appendix; Leek *et al.*, 1991). They emphasize that, for certain types of measurement (e.g. determining the best possible detection performance of a subject), one technique, such as PEST, may have advantages compared to the others.

Measurement and interpretation bias

In this simulations, the final estimate of threshold is taken to be the *mean* of the final p.d.f. As argued previously, if the assumptions made are correct, this estimate of threshold should be free from *interpretation* bias, i.e. the log threshold estimate should equal the mean of the possible values of log real threshold which could give rise to the observed sequence of yes and no responses. This lack of interpretation bias in the threshold estimate is a major advantage of the techniques which are simulated in this paper. It may be noted that using the *mode* of the final p.d.f., as in the standard version of the QUEST method (Watson & Pelli, 1983) would cause interpretation bias if the final p.d.f. is skewed; likewise, the Probit method for estimating thresholds from a measured psychometric function, is

based on a maximum likelihood calculation (i.e. the mode of a p.d.f. of threshold) and may also be subject to a similar bias (Finney, 1971; McKee *et al.*, 1985; O'Regan & Humbert, 1989).

While the current techniques are free from interpretation bias, they do suffer from *measurement* bias (see Introduction and Fig. 11). We know of no way of eliminating measurement bias, in techniques like QUEST and ZEST which are based on a finite number of binary responses; however, this bias is typically small (Fig. 11) and becomes increasingly unimportant with increasing number of trials per threshold.

Measurement and interpretation bias are important in different situations. When a *single* threshold measurement is made, *interpretation* bias indicates whether the estimated value deviates from the mean of the possible log real thresholds. When an average of *many* threshold estimates is taken, *measurement* bias indicates how this average differs from the observer's real threshold. In this situation, the current methods could suffer from a serious measurement bias (one which is not small compared to the random measurement error). Rather than simply averaging the log threshold estimates, a better method might be to record the intensity and response for all trials in the series of runs; a threshold estimate could then be made by performing the calculations [e.g. equation (7)] as if all these trials had been collected in a single run (in this way, M runs each of N trials are treated like one run with MN trials, which would have no interpretation bias and relatively little measurement bias). However, it should be noted (Pelli, personal communication) that thresholds often change significantly between runs; therefore the above analysis, while theoretically correct, may be based on an incorrect assumption (see Introduction) that the threshold does not vary from trial to trial.

Precision of yes–no and forced choice simulations

Our simulations are consistent with previous simulations (McKee *et al.*, 1985; Madigan & Williams, 1987) and with experimental data (Gerr & Letz, 1988; Hesse, 1986; Pierce & King-Smith, 1992) which show that, for a given number of trials, yes–no experiments provide a considerably higher precision than 2AFC. For example, Pierce and King-Smith (1992) found that measurement error for yes–no experiments, derived from simultaneous, independent visual threshold estimates (ZEST method, 30 trials/run) was only 37% of that for 2AFC. This is reasonably consistent with the current simulations (e.g. Fig. 7) given that these are for a smaller number of trials. Our simulations illustrate two reasons for the higher precision of yes–no measurements:

1. The ideal sweat factor for our yes–no simulations is less than that for 2AFC by a factor of about 2.3.
2. The efficiency of yes–no experiments is higher than for 2AFC [compare Fig. 16(B) and 17(B)]. For example, after 20 trials using the Minimum Variance Method and one-trial look-ahead, the efficiency of yes–no simulations is about 4.8 times higher than for 2AFC.

The added precision, $R_N - R_0$, of yes–no experiments is greater than for 2AFC due to both of the above factors [cf. equation (28)]; for the example given above, the added precision of yes–no simulations is 2.3×4.8 i.e. about 11 times greater—this can be confirmed by comparing Fig. 16(A) and 17(A). With increasing number of trials, the efficiency of 2AFC experiments would improve relatively more than that for yes–no experiments; for very large numbers of trials, where both efficiencies should approach 100%, the precision of yes–no experiments would still be greater than that of 2AFC by a factor of 2.3, due to the difference in ideal sweat factors.

The precision of forced-choice experiments may be improved (for a given number of trials) by increasing the number of alternatives (Green *et al.*, 1989; Kollmeier *et al.*, 1988; Pelli, Robson & Wilkins, 1988; Schlauch & Rose, 1990; Shelton & Scarrow, 1984). However, increasing the number of alternative intervals in a forced-choice experiment will also increase the duration of the experiment; consequently, Schlauch and Rose (1990) estimated that, for their conditions and for a given experimental duration, the highest precision for forced choice experiments would be given by three alternative intervals.

The effects of criterion in yes–no experiments

A disadvantage of yes–no experiments is that threshold may be affected by the subject's criterion (i.e. his willingness to respond "yes" when he is doubtful about the occurrence of a stimulus). The importance of the subject's criterion was emphasized by Higgins, Jaffe, Coletta, Caruso and de Monasterio (1984) who measured contrast sensitivity functions for 20 normals on two occasions, using both 2AFC and subject-setting; the latter thresholds depend on the subject's criterion, as do results for yes–no experiments. For subject-setting, there were changes in average threshold, on retest, of up to 0.4 log units, while the shape of the contrast sensitivity function was relatively unchanged; for 2AFC experiments, similar large changes in average threshold did not occur.

However, results for subject-setting are not necessarily applicable to yes–no measurements. Pierce and King-Smith (1992) measured thresholds for test spots in 19 normals in three sessions at weekly intervals; in each session, eight threshold runs of 30 trials were performed for both yes–no and 2AFC methods. Test–retest reliability (derived from session means) for yes–no measurements was not significantly worse than for 2AFC [standard deviations between sessions, within subjects: yes–no, 0.0506 log units: 2AFC, 0.0496 log units: $F(38,38) = 1.04$, $P > 0.1$]. The apparent discrepancy with the results of Higgins *et al.* (1984), may indicate that subject-setting has poor test–retest reliability for naive subjects, while yes–no measurements may have better reliability which is comparable to 2AFC. However, additional analysis of the Pierce and King-Smith (1992) data has demonstrated that the ratio of yes–no to 2AFC thresholds is negatively correlated with false positive rate (Spearman $R = -0.557$, $v = 17$,

$P < 0.02$) indicating that the subject's criterion does have a significant effect on yes–no thresholds.

In concluding the comparison of yes–no and 2AFC methods the following points can be made:

1. If the major information required is the *relative* threshold for different stimuli (e.g. high vs low spatial frequency, chromatic vs achromatic thresholds), the higher precision of yes–no measurements makes it the method of choice.
2. When an approximate absolute (rather than relative) value of threshold is required (e.g. standard error of 0.1 log units or more), the greater speed of the yes–no method makes it preferable to 2AFC.
3. When a more accurate absolute value of threshold is required (e.g. for comparing thresholds on different occasions or from different populations), 2AFC thresholds are preferable because of their freedom from the subject's criterion. However, it should be noted that two-interval forced choice judgments can be difficult for some subjects (even when they see the stimulus clearly, they may have difficulty remembering the interval). Also 2AFC measurements are typically quite time-consuming. Higgins *et al.* (1984) used 50 trials for each spatial frequency and the total time taken for nine spatial frequencies ranged from 18 to 40 min; for comparison, for yes–no measurements using eight trials at each of 11 spatial frequencies, we measure a contrast sensitivity function in under 5 min (including 16 blank trials).

Conclusions

1. Stimulus intensity selection. Setting the next intensity equal to the *mean* of the current p.d.f. (ZEST) gives higher efficiency than the standard QUEST method of using the *mode* (Watson & Pelli, 1983). Choice of the optimal threshold criterion can also improve efficiency relative to the criterion which gives the "ideal sweat factor". The Minimum Variance Method can provide even higher efficiency than ZEST.

2. Initial p.d.f. A modified hyperbolic secant can be used to fit a histogram of thresholds; using this initial p.d.f. increases efficiency and reduces bias.

3. Final threshold estimate. This is derived from the mean of the final p.d.f. (rather than the mode of a likelihood function used in the standard QUEST procedure). Again, this increases efficiency and reduces bias.

4. Bias. A distinction is made between *measurement* bias, derived from repeated measurements (or simulations) on one subject, and *interpretation* bias, derived from all possible thresholds which could give rise to one threshold estimate. Our final threshold estimates are free of interpretation bias (but not measurement bias).

5. Yes–no vs 2AFC. The considerably higher precision of yes–no measurements, compared to 2AFC, is due to both a lower "ideal sweat factor" and a higher efficiency.

6. Practical implementation. ZEST is a flexible procedure in that it can be readily modified to circumvent certain experimental problems; it can also be enhanced to estimate such factors as the slope of the psychometric function. This flexibility is discussed in the Appendix.

REFERENCES

Blackwell, H. R. (1963). Neural theories of simple visual discriminations. *Journal of the Optical Society of America, 53*, 129–160.

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

Corwin, T. R., Kintz, R. T. & Beaty, W. J. (1979). Computer-aided estimation of psychophysical thresholds by Wetherill tracking. *Behavioral Research Methods and Instrumentation, 11*, 526–528.

Crozier, W. J. (1936). On the sensory discrimination of intensities. *Proceedings of the National Academy of Sciences, 22*, 412–416.

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

Findlay, J. M. (1978). Estimates on probability functions: A more virulent PEST. *Perception & Psychophysics, 23*, 181–185.

Finney, D. J. (1971). *Probit analysis* (pp. 50–51). Cambridge: Cambridge University Press.

Gerr, F. E. & Letz, R. (1988). Reliability of a widely used test of peripheral cutaneous vibration sensitivity and a comparison of two testing protocols. *British Journal of Industrial Medicine, 45*, 635–639.

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

Green, D. M. (1993). A maximum-likelihood method for estimating thresholds in a yes–no task. *Journal of the Acoustical Society of America, 93*, 2096–2105.

Green, D. M., Richards, V. M. & Forrest, T. G. (1989). Stimulus step size and heterogeneous stimulus conditions in adaptive psychophysics. *Journal of the Acoustical Society of America, 86*, 629–636.

Grigsby, S. S., Vingrys, A. J., Benes, S. C. & King-Smith, P. E. (1991). Correlation of chromatic, spatial and temporal sensitivity in optic nerve disease. *Investigative Ophthalmology and Visual Science, 32*, 3252–3262.

Hall, J. L. (1968). Maximum-likelihood sequential procedure for estimation of psychometric functions. *Journal of the Acoustical Society of America, 44*, 370.

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

Harvey, L. O. (1986). Efficient estimation of sensory thresholds. *Behavioral Research Methods, Instruments and Computers, 18*, 623–632.

Hays, W. L. (1988). *Statistics*. Fort Worth, Tex.: Holt, Rinehart & Winston.

Hesse, A. (1986). Comparison of several psychophysical procedures with respect to threshold estimates, reproducibility and efficiency. *Acustica, 59*, 263–273.

Higgins, K. E., Jaffe, M. J., Coletta, N. J., Caruso, R. F. & de Monasterio, F. M. (1984). Spatial contrast sensitivity: Importance of controlling the patient's visibility criterion. *Archives of Ophthalmology, 102*, 1035–1041.

Johnson, C. A. & Shapiro, L. R. (1989). A comparison of MOBS (Modified Binary Search) and standard staircase test procedures in automated perimetry. In *Noninvasive assessment of the visual system* (1989 Technical Digest Series, Vol. 7, pp. 84–87). Washington, D.C.: Optical Society of America.

Johnson, C. A. & Shapiro, L. R. (1990). RIOTS (Real-time Interactive Optimized Test Sequence): A heuristic software test strategy for automated perimetry. *Investigative Ophthalmology and Visual Science (Suppl.), 31*, 191.

King-Smith, P. E. (1984). Efficient threshold estimates from yes–no procedures using few (about 10) trials. *American Journal of Optometry and Physiological Optics, 61*, 119P.

King-Smith, P. E., Grigsby, S. S., Vingrys, A. J., Benes, S. C. & Supowit, A. J. (1991). Evaluation of four different variations of the QUEST procedure for measuring thresholds. *Investigative Ophthalmology and Visual Science (Suppl.), 32*, 1267.

Klein, S. A. (1981). Rapid determination of psychometric functions. *American Journal of Optometry and Physiological Optics, 58*, 1038.

Kollmeier, B., Gilkey, R. H. & Sieben, U. K. (1988). Adaptive staircase techniques in psychoacoustics: A comparison of human data and a

mathematical model. *Journal of the Acoustical Society of America, 83*, 1852–1862.

Laming, D. & Marsh, D. (1988). Some performance tests of QUEST on measurements of vibrotactile thresholds. *Perception & Psychophysics, 44*, 99–107.

le Grand, Y. (1968). *Light, colour and vision* (p. 481). London: Chapman & Hall.

Leek, M. R., Hanna, T. E. & Marshall, L. (1991). An interleaved tracking procedure to monitor unstable psychometric functions. *Journal of the Acoustical Society of America, 90*, 1385–1397.

Leek, M. R., Hanna, T. E. & Marshall, L. (1992). Estimation of psychometric functions from adaptive tracking procedures. *Perception & Psychophysics, 51*, 247–256.

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. *Behavioral Research Methods and Instrumentation, 14*, 21–25.

Madigan, R. & Williams D. (1987). Maximum-likelihood procedures in two-alternative forced-choice: Evaluation and recommendations. *Perception & Psychophysics, 42*, 240–249.

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

Nachmias, J. (1981). On the psychometric function for contrast detection. *Vision Research, 21*, 215–223.

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.

Pelli, D. G. (1987). The ideal psychometric procedure. *Investigative Ophthalmology and Visual Science (Suppl.), 28*, 366.

Pelli, D. G. & Farell, B. (1994). A user's guide to psychophysical methods. In *Handbook of optics* (2nd ed.). New York: McGraw-Hill. In press.

Pelli, D. G., Robson, J. G. & Wilkins, A. J. (1988). The design of a new letter chart for measuring contrast sensitivity. *Clinical Vision Sciences, 2*, 187–199.

Pentland, A. (1980). Maximum likelihood estimation: The best PEST. *Perception & Psychophysics, 28*, 377–379.

Pierce, G. E. & King-Smith, P. E. (1992). Yes–no or two alternative forced choice. Which is the better clinical threshold technique? *Investigative Ophthalmology and Visual Science (Suppl.), 33*, 964.

Robbins, H. & Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics, 22*, 400–407.

Schlauch, R. & 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. & Scarrow, I. (1984). Two-alternative versus three-alternative procedures for threshold estimation. *Perception & Psychophysics, 35*, 385–392.

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.

Swanson, W. H. & Birch, E. E. (1992). Extracting thresholds from noisy psychophysical data. *Perception & Psychophysics, 51*, 409–422.

Taylor, M. M. (1971). On the efficiency of psychophysical measurement. *Journal of the Acoustical Society of America, 49*, 505–508.

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

Taylor, M. M., Forbes, S. M. & Creelman, C. D. (1983). PEST reduces bias in forced choice psychophysics. *Journal of the Acoustical Society of America, 74*, 1367–1374.

Tyrrell, R. A. & Owens, D. A. (1988). A rapid technique to assess the resting states of the eyes and other threshold phenomena: The Modified Binary Search (MOBS). *Behavioral Research Methods, Instruments and Computers, 20*, 137–141.

Watson, A. B. & Fitzhugh, A. (1990). The method of constant stimuli is inefficient. *Perception & Psychophysics, 47*, 87–91.

Watson, A. B. & Pelli, D. G. (1983). QUEST: A Bayesian adaptive psychophysical method. *Perception & Psychophysics, 33*, 113–120.

Watt, R. J. & Andrews, D. P. (1981). APE: Adaptive probit estimation of psychometric functions. *Current Psychological Reviews, 1*, 205–214.

Weber, J. & Rau, S. (1992). The properties of perimetric thresholds in normal and glaucomatous eyes. *German Journal of Ophthalmology, 1*, 79–85.

Weibull, W. A. (1951). A statistical distribution function of wide applicability. *Journal of Applied Mechanics, 18*, 292–297.

Wetherill, G. B. & Levitt, H. (1965). Sequential estimation of points on a psychometric function. *The British Journal of Mathematical and Statistical Psychology, 18*, 1–10.

Acknowledgements—This work was supported by NIH grant EY-04948, the Ohio Lions Eye Research Foundation and the Ohio Supercomputer Center. We thank Drs A. M. Brown, P. L. Emerson, K. E. Higgins, C. A. Johnson, S. A. Klein, D. T. Lindsey, S. P. McKee, D. G. Pelli, A. P. Pentland, G. E. Pierce and D. Y. Teller for their comments on an earlier version of the manuscript and Drs M. W. Browne, K. Kornacker, L. Krueger and M. L. Moeschberger for statistical and theoretical advice.

APPENDIX

Practical Implementation of the ZEST Method

For several years, we have used a variation of the ZEST method to measure normal and clinical visual thresholds. A yes–no method is used, and (as implied by the definition of the ZEST method), the next intensity is set to the *mean* (rounded to the nearest 0.05 log units) of the current p.d.f. P.d.fs. are calculated directly (i.e. linear probability rather than the logarithm of probability used by Watson & Pelli, 1983); this makes it easier to calculate the mean and standard deviation of any p.d.f. The final estimate of log threshold is taken to be the (unrounded) mean of the final p.d.f. [equation (15)] and the standard error in log threshold is estimated to be the standard deviation of this p.d.f. [derived from the square root of variance, given by equation (22)]. By integrating the final p.d.f., it should be possible to derive a 95% confidence range for the final threshold estimate (we have not yet implemented this idea).

P.d.fs. are calculated over a 5 log unit range of threshold with a step size of 0.05 log units. An 8 bit (North Star Horizon) microcomputer provides adequate speed when used in conjunction with a floating point processor and machine code subroutines for calculating a p.d.f. [equation (7)] and its mean; with a more modern microcomputer, it should be possible to achieve adequate speed using a high level language such as BASIC, C or Fortran.

The threshold criterion factor, ϵ in equation (9), is set to 0; this is close to the value yielding optimum performance for thresholds based on 8–16 trials [Fig. 6(A)]. The slope of the psychometric function, β in equation (9), has been assumed to be 3.5 (as in Watson & Pelli, 1983) and the reasonableness of this assumption is considered in the discussion of Fig. 18. As in the current simulations, false positive and false negative rates are assumed to be 0.03 and 0.01 respectively.

Choice of the initial p.d.f. is, in the first instance, a matter of experience and intuition. Watson and Pelli used a Gaussian function of log threshold, but a more satisfactory alternative may be a modified hyperbolic secant [equation (8)—see also, Harvey, 1986]; we have found that a hyperbolic secant provides better performance than a Gaussian when a threshold is relatively far from the peak of the initial p.d.f. (e.g. for a subject with very elevated thresholds). It may be noted that an error in the initial p.d.f. has a relatively small effect on estimated thresholds after eight or more trials of a yes–no method (Fig. 14). After some representative threshold data have been obtained, it is possible to derive a better initial p.d.f. by fitting a curve [e.g. equation (8)] to a histogram of thresholds, as in Fig. 2. It should be possible to increase the precision of the ZEST method by using a different initial p.d.f. for each stimulus (e.g. for each spatial frequency of a contrast sensitivity function) but we have not tried this idea; it should be particularly advantageous for experiments where the thresholds for

different stimuli vary over many log units (e.g. for spectral sensitivity measurements over a wide range of wavelengths).

Our experiments are terminated after a fixed number of trials for each threshold measurement. We have found 8 trials per threshold measurement to be satisfactory for many studies, yielding a standard deviation of about 0.1 log units [Figs 6(A) and 18]; if higher precision is required, the number of trials can be increased. Different stimuli are interleaved; for example, for a contrast sensitivity function using 11 spatial frequencies, all 11 frequencies, plus two blank trials, are first presented in random order and then this cycle of 13 trials is repeated a further seven times in different random orders (thus making a total of eight trials for each frequency plus 16 blank trials). The subject is asked to repeat the experiment if the false positive rate is over 15% (i.e. three or more out of 16). Alternative termination rules would be to end each threshold measurement when the estimated standard deviation of threshold falls below a certain value or when the 95% confidence range (calculated directly by integrating the final p.d.f.) falls below some criterion; these are roughly equivalent to Watson and Pelli's (1983) proposal based on a chi-square calculation, and make no assumptions about the form of the final p.d.f.

Circumventing problems

A major advantage of QUEST and its variations (e.g. ZEST), is that the choice of stimulus intensity can readily be altered under special circumstances; the final p.d.f. of log threshold provides an unbiased estimate of log threshold and its standard deviation even if the intensities used are not all derived in the standard way (e.g. as the mean of the current p.d.f.). For example, suppose that the threshold is close to or above the maximum intensity which can be generated by the equipment; the mean of the current p.d.f. can then exceed this maximum intensity. In this circumstance, the log intensity [x_i in equation (7)] can be set to maximum (rounded down to the nearest 0.05 log unit step); after the subject responds, a new p.d.f. can be calculated using this value of x_i in equation (7). In this way, thresholds can be estimated even when they are somewhat above the maximum intensity available from the apparatus; we consider any threshold estimate to be "valid" if the subject responded at least once, but the threshold estimate will, of course, be relatively inaccurate if the subject responded only once, and this will be reflected in the estimated standard error.

A second example of this flexibility of ZEST occurs during the first "cycle" of an experiment—e.g. during the first presentations of all 11 frequencies for the contrast sensitivity measurements discussed above. In the standard ZEST method, all these stimuli would be presented at a fixed contrast—the mean of the initial p.d.f.; a normal subject may see all or most of these initial presentations, whereas a subject with high thresholds may see none or few. This is both inefficient and disturbing to the subject who may begin to wonder if the apparatus is working correctly. A more satisfactory alternative is to vary the contrast during this first cycle, so that, for example, contrast is increased if the subject responds infrequently. This can be done by running an independent "pseudo" ZEST method where all the trials in the first cycle are treated as if they were the *same* stimulus. A corresponding "pseudo" p.d.f., $q'_K(T)$, is calculated from an equation analogous to equation (7) (after k trials in this first cycle); the next intensity is set to the mean of this p.d.f. At the same time, p.d.fs. for each frequency are calculated in the standard way [equation (2)] and these p.d.fs. are used for the second and subsequent cycles.

Taylor, Forbes and Creelman (1983) have criticized the QUEST method because it assumes that the subject's threshold is stationary (i.e. it does not vary during the experiment) whereas they show that this is typically not strictly true (cf. Leek *et al.*, 1991). Variation in threshold during the experiment would effectively reduce the slope of the psychometric function. A more serious problem may be that a lapse in the subject's performance could produce a relatively long sequence of unseen stimuli, particularly in the later stages of a long run, when the stimulus intensity is close to threshold and it increases in relatively small steps after a "no" or incorrect response; such a sequence of unseen stimuli would be exacerbated by the lack of a clearly detected stimulus, to remind the subject of the characteristics which are to be detected. Again, the flexibility of the QUEST and related methods can be used to counteract this problem. For example, Watson and Pelli (1983) suggest adding a random "jitter" to the stimulus intensities of up to plus or minus 0.1 log unit; an alternative method might be that, after a sequence of responses indicating that the subject is not responding above chance level, one rather stronger stimulus could be used to remind the subject of the stimulus characteristics; in either case, the response to these "nonstandard" intensities can be incorporated into the p.d.fs. in the normal way, using equation (7).

Enhancements to ZEST and the minimum variance method

A limitation of the current methods is that assumptions must be made about the psychometric function—e.g. in most of our simulations and in our experiments, we assume that the slope, $\beta = 3.5$ [equation (9)]. However, if the variation of β in the population is known (or a reasonable guess can be made), the current methods can take this into account by working with p.d.fs. which are a function of both log threshold, T, and slope β, i.e. by replacing $q(T)$ by $q(\beta, T)$. For example, after trial i, the Bayesian multiplication of equation (7) becomes

$$q_i(\beta, T) = p(r_i, x_i, \beta, T) q_{i-1}(\beta, T). \tag{33}$$

Integration over both slope and log threshold is required for determining such things as the next intensity, the final threshold estimate and its variance; for example, the final threshold estimate of equation (15) becomes

$$E_j = \left[\iint T q_{Nj}(\beta, T)\, d\beta\, dT\right] \Big/ \left[\iint q_{Nj}(\beta, T)\, d\beta\, dT\right]. \tag{34}$$

An estimate of slope, β'_j, can be derived from a similar equation:

$$\beta'_j = \left[\iint \beta q_{Nj}(\beta, T)\, d\beta\, dT\right] \Big/ \left[\iint q_{Nj}(\beta, T)\, d\beta\, dT\right]. \tag{35}$$

If the assumed form of the initial p.d.f., $q_0(\beta, T)$, is correct, this gives an estimate of slope which is free of interpretation bias, for the same reasons that equations (15) and (34) give unbiased estimates of log threshold (see discussion of interpretation bias in the Introduction). It may be noted that the initial p.d.f., $q_0(\beta, T)$, can take into account the possibility that the slope of the psychometric function might be less when the threshold is high (see discussion of Fig. 18; Weber & Rau, 1992). In theory, false positive and false negative rates, γ and δ in equation (9), can also be incorporated as extra dimensions in the p.d.fs. in the way discussed above for the slope, β; however, the time taken for calculations using such three or four dimensional p.d.fs. may be prohibitive, even with current microprocessors.

Similar considerations may be applied to corrections for drift in thresholds during an experimental run (cf. Leek *et al.*, 1991). For example, suppose that it is suspected that the subject's log threshold, T, drifts in a linear way during a run so that at trial i

$$T = T_0 + \tau i \tag{36}$$

where T_0 is the log threshold at the start of the run and τ gives the rate of drift. If the population variation of τ can be estimated (or guessed) then it can be incorporated in an initial p.d.f. $q_0(\tau, T_0)$ and the Bayesian multiplication of equation (7) becomes

$$q_i(\tau, T_0) = p(r_i, x_i, T_0 + \tau i) q_{i-1}(\tau, T_0). \tag{37}$$

The final estimate of T_0 is given by

$$E_j = \left[\iint T_0 q_{Nj}(\tau, T_0)\, d\tau\, dT_0\right] \Big/ \left[\iint q_{Nj}(\tau, T_0)\, d\tau\, dT_0\right]. \tag{38}$$

An estimate of drift is given by

$$\tau'_j = \left[\iint \tau q_{Nj}(\tau, T_0)\, d\tau\, dT_0\right] \Big/ \left[\iint q_{Nj}(\tau, T_0)\, d\tau\, dT_0\right]. \tag{39}$$

The significance of this drift can be estimated by generating a (one-dimensional) final p.d.f. of drift from

$$P(\tau) = \int q_{Nj}(\tau, T_0)\, dT_0. \tag{40}$$

By integrating $P(\tau)$, one may check whether zero drift, $\tau = 0$, lies within, say, the 95% confidence range of τ.

The current methods may also be modified to take advantage of correlations between the thresholds of two or more stimuli which are being measured simultaneously (e.g. at two locations, A and B). Let T_A and T_B be log thresholds for two such stimuli, and let $q_0(T_A, T_B)$ be their joint initial p.d.f. (which takes account of any correlation between these thresholds in the population). Suppose that r_1 is the response to the first trial of log intensity x_1, which is, say, for stimulus A. Then the joint p.d.f. after this trial becomes

$$q_1(T_A, T_B) = p(r_1, x_1, T_A) q_0(T_A, T_B). \tag{41}$$

The advantage of this strategy is that, if there is correlation between T_A and T_B, this multiplication shifts the center of gravity of the new p.d.f., q_1, along not only the T_A axis but also the T_B axis, so that the first intensity chosen for stimulus B takes advantage of information from the response to stimulus A; similar considerations apply at any later stage of the experiment.

In principle, this technique can be extended to measuring many thresholds simultaneously (as in automated perimetry), but evidently, with more than two or three stimuli, the calculations become very time-consuming, even for modern microprocessors. It remains a challenge for the future to develop more efficient, practical algorithms for measuring many thresholds simultaneously, but the flexibility of the ZEST and Minimum Variance Methods, illustrated in the above examples, indicate that they should form a good basis for further developments in this area.