



# The use of an implicit standard for measuring discrimination thresholds

M.J. Morgan<sup>a,b,\*</sup>, S.N.J. Watamaniuk<sup>c</sup>, S.P. McKee<sup>d</sup>

<sup>a</sup> *Institute of Ophthalmology, University College London, 11-43 Bath Street, London EC1 9EV, UK*

<sup>b</sup> *Department of Anatomy, University College London, London, UK*

<sup>c</sup> *Department of Psychology, Wright State University, Dayton, OH, USA*

<sup>d</sup> *Smith-Kettlewell Eye Research Institute, San Francisco, CA, USA*

Received 9 July 1998; received in revised form 27 January 2000

---

## Abstract

We measured thresholds for comparing the separation between lines, using either the method of constant stimuli (MCS) or the method of single stimuli (MSS). In the MCS an explicit standard is presented on each trial, whereas in the MSS the standard is the mean of the set. The thresholds for the MSS procedure were nearly identical to those with the MCS procedure, whether or not feedback was used. A statistical model is presented showing how the threshold error estimated by MSS varies according to the number of past stimuli used by the observer to calculate the mean of the set. If the model is an accurate representation of human processing, our observers were averaging over the last 10–20 trials to estimate the implicit standard. Our results show that the explicit standard in the MCS procedure is generally superfluous. Provided that the test range is small, and that the observer is given some practice trials, thresholds measured with MSS procedure are just as precise as those measured with the traditional MCS procedure. © 2000 Elsevier Science Ltd. All rights reserved.

*Keywords:* Psychophysical procedures; Psychometrics; Discrimination thresholds

---

## 1. Introduction

In the traditional MCS, an observer is shown the standard or reference stimulus in the first of two intervals, and a test stimulus, chosen from a range of values centred on the standard, in the second interval. The observer is asked to judge whether the test is above (more intense) or below (less intense) the standard stimulus. This judgement can be made in one of two ways. The observer can take the difference between the two intervals to assess which stimulus is larger. Alternatively, the observer can ignore the first interval and simply judge whether the test stimulus is above or below some criterion value as in the ‘Yes–No’ task of signal detection theory. An ideal observer will always choose the second strategy because, assuming that the two intervals are independent the variance of the difference judgement is double the variance of the ‘Yes–No’ task (Sorkin, 1964; Vogels & Orban, 1986a).

Although the MCS resembles a 2AFC task, there is an important distinction between the two procedures: a given test stimulus always appears in the second interval in the MCS procedure, but it appears at random in either interval in the true 2AFC task. According to signal detection theory, this difference in stimulus arrangements leads to different predictions about sensitivity. Let us assume that the standard stimulus and a given test stimulus generate two noisy internal distributions, and that the observer makes a judgement by taking the difference between the first and second interval. In the 2AFC task, the means of the *difference* distributions, i.e. (standard – test) or (test – standard), that determine the observer’s decision, are separated by double the distance between the noisy internal distributions. So, although the variance of the difference distributions is also double in the 2AFC task, the observer’s sensitivity in  $d'$  units is actually increased by  $\sqrt{2}$ , when compared to his performance in a Yes–No task (Green & Swets, 1966; Viemeister, 1970; Jesteadt & Sims, 1975; Macmillan & Creelman, 1991). In the MCS arrange-

---

\* Corresponding author. Fax: +44-20-76086846.

E-mail address: m.j.morgan@ucl.ac.uk (M.J. Morgan).

ment, the separation between means of the difference distributions is equal to the separation between the internal distributions for the standard and test stimuli, so taking a difference degrades sensitivity by a  $\sqrt{2}$  (see Appendix A for details). Thus, the ideal strategy for an MCS task is to make all judgements on the basis of the information in the second interval alone. Of course, the ideal observer has perfect knowledge of the underlying internal distributions and of the optimal criterion, and thus, has no need to use the standard stimulus as a reference. What about the real observer? Which strategy do human observers choose in the MCS?

It has been known since the 1920's that, in some tasks, real observers can perform as well without a reference as with one (Wever & Zener, 1928; Fernberger, 1931; Pfaffman, 1935; Woodworth, 1938; Woodworth & Schlosberg, 1954; Viemeister, 1970; Westheimer & McKee, 1977). If, on each trial, an observer is shown only the test stimulus and is asked to judge whether it is above or below the mean of the test range, the precision of his or her judgements, i.e. threshold, is nearly indistinguishable from that obtained with the standard MCS. We shall call the reference-free procedure by its traditional name, 'the

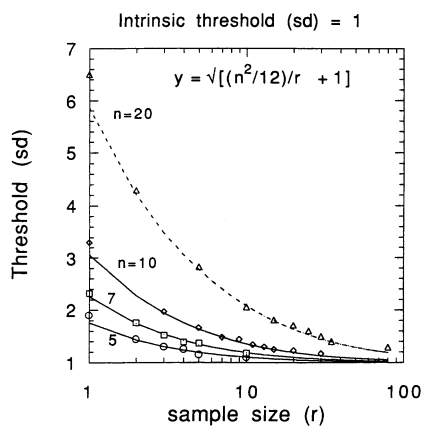


Fig. 1. A model of the behaviour of an observer classifying stimuli in the MSS. The observer encodes the magnitude of the stimulus on the current trial with an error equivalent to unit variance, and compares the encoded magnitude with the remembered mean of the last  $r$  stimuli. The stimulus on each trial is drawn randomly with equal probability from a set of  $n$  equally spaced stimuli. The observer decides whether the stimulus on the current trial is greater or smaller than the mean of the last  $r$  stimuli. The observer's threshold (vertical axis) is calculated as the standard deviation of the psychometric function relating the stimulus magnitude to the probability of the observer's classifying it as greater than the mean. The symbols were obtained by a Monte-Carlo simulation obeying the rules just described; the solid curves were obtained from the formula given in the figure. The formula is based on the principle of the additivity of variances under convolution. Thus the observer's variance is the sum of the encoding variance (unity) and the square of the standard error of the mean. The observer's threshold is the square root of the observer's variance (since threshold is defined as the standard deviation). The different curves refer to different values of  $n$ , the number of stimuli in the set.

method of single stimuli' to distinguish it from the MCS.

To make a judgement in the MSS, observers must use some representation or memory of the stimuli they have been shown before. It was recognised in Fechner's time that sequential effects could alter perceived stimulus magnitude. In his famous adaptation level theory, Helson (1947, 1948) proposed that the perceived mean of a set of stimuli depended on the weighted geometric mean of all stimulus presentations up to and including the current trial. Many subsequent studies have shown that asymmetrical stimulus sequences can introduce a bias (constant error) in observers' judgements, although the exact value of the measured shifts are not in perfect agreement with Helson's theory (Parducci, 1956, 1964; Masin, 1987). While these studies revealed how the human observer created a reference system in the absence of a fixed standard, they usually did not examine sequential effects on *sensitivity* (threshold), which is represented by the slope, rather than by a shift, of the psychometric function. Thresholds; are measurements of variability (random error or noise) in the observers' sensory responses and decision-making processes. It seems reasonable to expect that sequential perturbations in the 'implicit standard' would introduce additional noise in the judgements, resulting in higher thresholds for the MSS procedure. Why doesn't the explicit reference in the MCS procedure improve the real observer's performance by stabilising the reference?

Treisman and colleagues (Treisman & Faulkner, 1984; Treisman & Williams, 1984; Lages & Treisman, 1998) have examined the general problem of criterion-setting in all types of psychophysical tasks that require observers to categorise stimuli. In particular, they explored how criterion noise affects threshold sensitivity in the 'Yes-No' paradigm. Here we examine how an observer can use a criterion-setting approach in MSS to determine the 'implicit standard', and what the predicted effect of this approach is on the observer's threshold.

Suppose that the observer bases his or her judgement on a comparison of each stimulus with the mean of the last  $r$  stimulus presentations. The stimuli are randomly sampled from a set of  $p$  exemplars of each of  $n$  stimuli drawn from a uniform distribution  $\{1, 2, \dots, n\}$  without replacement. The functions in Fig. 1 show the predicted thresholds (the standard deviation of the overall error function) as  $r$  is varied, for different values of  $n$ , the number of stimuli in the test set. We assume that there are two sources of error contributing to the observer's judgement. The first of these is the discrepancy between the observer's estimate of the mean, and the true mean of the stimulus set. The second is the observer's error in encoding the current stimulus value, which we assume to be Gaussian with unit variance. The error in the estimate of the mean is given by the standard error of

the mean based on  $r$  observations. Note that the standard deviation of a uniform distribution  $\{1, 2, \dots, n\}$  is given by  $n/\sqrt{12}$ ,<sup>1</sup> and thus the standard error based on  $r$  observations will be  $n/\sqrt{(12*r)}$ . The error in stimulus encoding is, by definition, unity. The theoretical curves are based on the assumption of the additivity of variances under convolution, so that the variance of the overall error function ( $y$ ) is given by:

$$\text{Var}(y) = (n^2/12r) + 1 \quad (1)$$

and the predicted threshold is the square-root of this variance. To clarify:  $r$  is the number of observations (previous trials) in memory, and  $n$  is the number of stimulus values in the test set. The data points in Fig. 1 were obtained by Monte Carlo simulation as a check on the reasoning.

The simulation of each point tested a series of trials on each of which: (a) a stimulus was chosen randomly without replacement from a discrete, uniform distribution  $\{1, 2, \dots, r\}$ , (b) Gaussian noise with unit variance was added to the stimulus value, (c) the mean of the last  $r$  stimulus presentations was calculated and (d) a response value of 0 or 1 was selected depending on whether the noisy stimulus value was greater or less than the mean. At the end of the formulated list, a psychometric function was calculated relating the probability of a '1' response to the actual stimulus value, and the standard deviation of the function was derived from probit analysis (Finney, 1971).

In the simulations, a decision had to be taken about the first  $r$  trials of each run, in which the simulated observer had insufficient information about the mean of the stimulus values. If the response were based on the stimuli seen so far, there would be wild errors on the first one or two trials, which would result in meaningless error functions. In actual applications of MSS, this problem is usually taken care of by giving the observer practice trials to familiarise them with the set of test stimuli before data are recorded. We simulated the situation by setting the internal mean to the actual mean of the stimulus set at the start of each run, and then changing it appropriately as each new stimulus arrived.

The simulations show that performance rapidly approaches an asymptotic value (the observer's threshold) as the number of remembered stimuli increases; note that this assumed threshold is set equal to the internal noise arising from both sensory and decision sources of variance. With a test set of seven stimulus values and a memory of the last ten stimuli, performance would

probably be experimentally indistinguishable from a threshold estimated by other procedures involving hundreds of trials. Indeed, this threshold estimate would be elevated by only 50% with as little as three remembered stimuli. The experiments described below explore how much memory observers actually use to estimate the implicit standard.

## 2. Experimental results

### 2.1. Comparing explicit and implicit standards

For these experiments, we first confirmed the earlier studies showing that the MSS produces the same estimate of threshold as the traditional MCS. We used the separation discrimination task often used in hyperacuity studies (e.g. Westheimer & McKee, 1977). In this task, the observer judges the distance separating a pair of lines. Separation thresholds were measured in the same observers under the same conditions with the two psychophysical procedures.

### 2.2. Methods

In the MCS procedure, the standard separation, equal to the mean to the test set, was always presented in the first interval. The test separation, presented in the second interval, was randomly chosen from a set of five separations centred on the mean (two smaller, two larger and one equal to the standard). In the nomenclature of our simulations,  $n = 5$ . Observers judged whether the test was smaller or larger than the standard (mean). The time between intervals was 500 ms. In the MSS procedure, the test alone was presented in a single interval; it was randomly chosen from the same set of five separations used for the MCS judgements. Observers were instructed to judge whether the test was larger or smaller than the implicit mean of the set. We used this instruction to make the two procedures comparable, since then observers would be judging the test stimuli against a representation of the mean in both tasks. For this experiment, the standard separation was 25 arcmin and the incremental step size was 0.87 arcmin (3.5% steps). A practice set, consisting of 150 MCS trials, was run before the experimental data were taken, but no practice trials were used in any subsequent experimental measurement. Each of the subsequent experimental blocks also consisted of 150 trials. Auditory feedback (a short beep in response to errors) was provided for the MCS condition and for one MSS condition. A second set of MSS thresholds were measured without feedback.

The target consisted of a pair of bright horizontal lines, 14 arcmin long, presented against a dark background, in the centre of a 1332A Hewlett-Packard

<sup>1</sup> Strictly speaking, the formula is not accurate for the discrete case with small  $n$ . But the error is less than 3% for the smallest value of  $n$  we used ( $n = 5$ ) and decreased to a 0.2% for  $n = 20$ . The small error for  $n = 5$  accounts for discrepancy between the theoretical and the Monte-Carlo curves in Fig. 1.

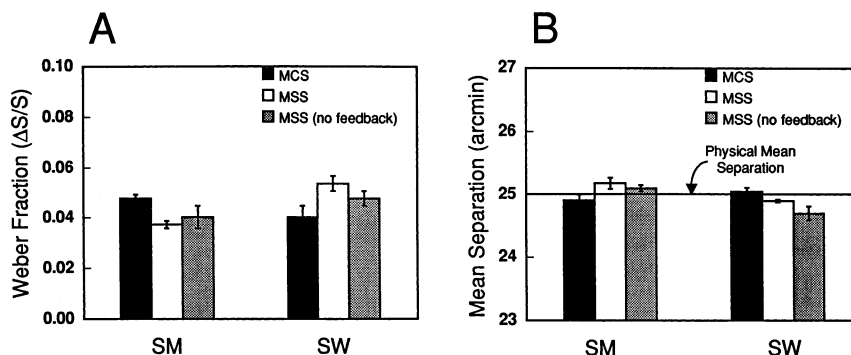


Fig. 2. (A) Weber fractions for separation discrimination measured with the MCS and the MSS in two observers. Standard is 25 arcmin. Data based on the average of three 150 trial sessions. Audible error feedback provided for MCS condition (black bars) and one for MSS condition (white bars). In another MSS condition (grey bars), no feedback was given. (B) PMFs (stimuli corresponding to the 50th percentile on the psychometric function) for the same three conditions. Horizontal line shows the true physical distance separating the target lines, i.e. the standard.

monitor that was equipped with P4 phosphor. Target duration was 1 s. A fixation pattern, formed of four small right-angle brackets, outlining a square 90 arcmin on a side, was presented for 800 ms preceding the target. The brackets disappeared at target onset. Target luminance was 6 cd/m<sup>2</sup>. Observers viewed the screen binocularly at a distance of 1.5 m, and wore appropriate corrective lenses for that viewing distance. The two observers had participated in a wide variety of psychophysical experiments employing both MCS and MSS procedures. However, observer SW had never been an observer in a separation discrimination task prior to this study.

We plotted the percentage of trials on which the observer judged the test 'larger', as a function of the physical separation between the target lines. A cumulative normal function was fitted to the data by probit analysis. For all conditions, 'threshold' was defined as the incremental change in separation that produced a percentage change in the response equal to the standard deviation of the cumulative normal function (a change from 50 to 84%). The values shown in Fig. 3 are based on the average of three experimental blocks of 150 trials each; standard errors were estimated from the measured variance among blocks, i.e. the experimental error.

### 3. Results

The thresholds for the MSS procedure were nearly identical to the thresholds measured with the standard MCS procedure (Fig. 3A). Moreover, the precision of MSS judgements was equally good with and without feedback. Feedback may be useful for the first few trials (Tabachnick & Parducci, 1973), but its beneficial effects were not apparent in these long runs. Vogels and Orban (1986b) also found that feedback did not affect thresholds in the MSS procedure. The 50th percentile

of the psychometric functions corresponds to the 'perceived median frequency' (PMF) for traditional MCS judgements, or to a bias in the implicit standard for MSS judgements. Fig. 3B shows that the stimulus value corresponding to the 50th percentile was close to the true physical standard for all three conditions. The absence of bias in the MSS procedure shows that observers can use instructions, feedback, and their tendency to use the two response categories equally, to track the true mean when presented with a balanced stimulus distribution centred on the true mean. In short, estimates of human performance measured by the MSS procedure can be both accurate (small bias) and precise (low thresholds) (see Fig. 2).

According to our simulations, the two procedures will produce the same thresholds only if the observer averages (or remembers) a substantial number of trials in the MSS procedure. If the simulations are an accurate representation of human processing, our observers are averaging over the last 10–20 trials to estimate the implicit standard.

#### 3.1. The development of the implicit standard

If observers are using such a large number of trials to make their estimate, the effect on thresholds of the development of the implicit standard should be evident during the early part of an experimental run. In their analysis of the MSS procedure, Vogels and Orban (1986b) found that the observer's threshold had reached asymptote by the end of ten trials. As they used a set size of five stimulus values ( $n=5$ ), their results are consistent with the simulations shown in Fig. 1 for a small set size. However, our simulations also indicate that thresholds should be elevated during the first 10–20 trials, relative to their asymptotic values, when there is a large number of stimulus values in the test set, i.e. when  $n=20$ . To study this development, we ran 20 experimental sessions of 50 trials each and binned the

data by sequential blocks of ten trials each. We then summed the separate bins over sessions and estimated the threshold for the first ten trials, the second ten trials and so forth. Separation discrimination was measured with the MSS procedure for test sets containing five stimulus values and 20 stimulus values ( $n = 5$  or 20, respectively). No feedback was given.

To prevent transfer from one 50 trial session to the next, we changed the mean separation every session. In the conditions used for our experiments, increment thresholds are known to be proportional to mean separation for distances ranging from 10 to 200 arcmin: an example of Weber's Law (Burbeck, 1987; McKee, Welch, Taylor & Bowne, 1990).

Therefore, the binned data from different mean separations could be summed together after the incremental steps in the test set were converted into percentages. The Weber fractions for separations  $> 10$  arcmin are between 3 and 5%, so we used a step size of approximately 3% of all separations, and chose the mean value for each 50-trial session at random from a range spanning 10–40 arcmin.

We divided the threshold values for each ten-trial bin by the MCS threshold shown above in Fig. 3A. We assumed that the MCS threshold represented an asymptotic estimate of the internal noise. As we were trying to determine the amount of additional noise contributed by the observer's estimate of the implicit standard during the early trials, it seemed reasonable to represent this additional error as a ratio of the MSS threshold to the asymptotic MCS threshold. These threshold ratios are plotted in Fig. 3 as a function of the average number of trials seen before the judgement was made. For example, on average, the observer has seen 5 stimulus presentations before making a judgement during the first ten-trial block. The threshold for the test set composed of five stimulus values has nearly reached asymptote after five trials, confirming the results of Vogels and Orban (1986b). Our results show that threshold based on 20 stimulus values requires many more trials to reach a stable value, demonstrating that the larger test set increases the additional threshold noise associated with uncertainty about the reference.

The continuous curves shown in Fig. 3 are taken from the simulations in Fig. 1. We also calculated best-fitting curves to the data,<sup>2</sup> using the same formula, but allowing the observer to use a proportion  $p$  of the previous trials. The model had just one free parameter, the proportion  $p$ . Results (dashed curves) show very little difference between the real and theoretical observers for the case of  $n = 20$  (set size of 20). The best-fitting values of  $p$  for SM and SW were 0.84 and 0.8, respec-

tively, indicating that they used 80% of the available information. The data for  $n = 5$  are too flat and noisy to permit definite conclusions. Those for SM do not exclude the possibility that she used all the available trials. The single discrepant point for SW meant that the best fit used only a proportion 0.2 of the available trials.

### 3.2. Multiple implicit standards

One common method for controlling observer variables such as fatigue, practice, attention or adaptation state is to intersperse conditions during a test run, presenting different conditions on different trials at random. The MSS procedure can be used for interspersed conditions, provided that the observer is given unambiguous information about which implicit standard to apply to a given test trial. Using the MSS procedure, Morgan (1992) asked observers to make precise separation judgements for four different, interspersed mean separations. The implicit standard for a given trial was indicated by the orientation of the target in one experiment, and by its position in another; the stimulus values for the four sets, corresponding to the four different implicit means, were drawn from overlapping ranges. The resulting thresholds were not significantly different from those obtained when the threshold for each separation was measured in an isolated block. Apparently, observers can maintain independent estimates of several implicit standards simultaneously. There is a limit to this capability. Morgan found that increasing the number of interspersed implicit standards to eight produced some loss in the precision of the separation judgements, when compared to the blocked measurements.

The results shown in Fig. 3, in agreement with our simulations, imply that observers maximise the precision of their judgements by taking a running average of most or all of the presented stimuli. However, Morgan's results indicate that the observer can *select* the stimuli that are relevant to a particular judgement, and ignore the others. Parducci (1956) studied the effect of shifted ranges on constant error; his results also suggest that observers decide whether the current range of stimuli should be combined with the previously tested range or not. How well can trained observers select among similar targets when asked to generate multiple implicit standards? To study this question, we used the separation task described above (see Section 2) combined with Morgan's approach using interspersed standards. However, unlike the original Morgan study, all the targets had the same orientation, and were presented in the same position. Instead, a symbolic cue was presented before each trial to designate the appropriate implicit standard to be used for that trial. Note that we are primarily interested in the effect of these

<sup>2</sup> An two-factor ANOVA was run on the data pooled across subjects; it showed that the effects of both the number of trial blocks ( $r$ ) and the set size ( $n$ ) were significant.

interspersed standards on the *precision* of thresholds, not on the constant error.

We estimated thresholds for nine different mean separations, ranging from 21 to 39 arcmin, from interspersed trials in a single experimental run of 1000 trials.

Test sets of seven separations ( $n = 7$ ; step size =  $\sim 4\%$ ), centred on each of the nine implicit means, were drawn from overlapping ranges. Immediately before each trial, a number was briefly flashed indicating which of the nine implicit standards should be used for judging the

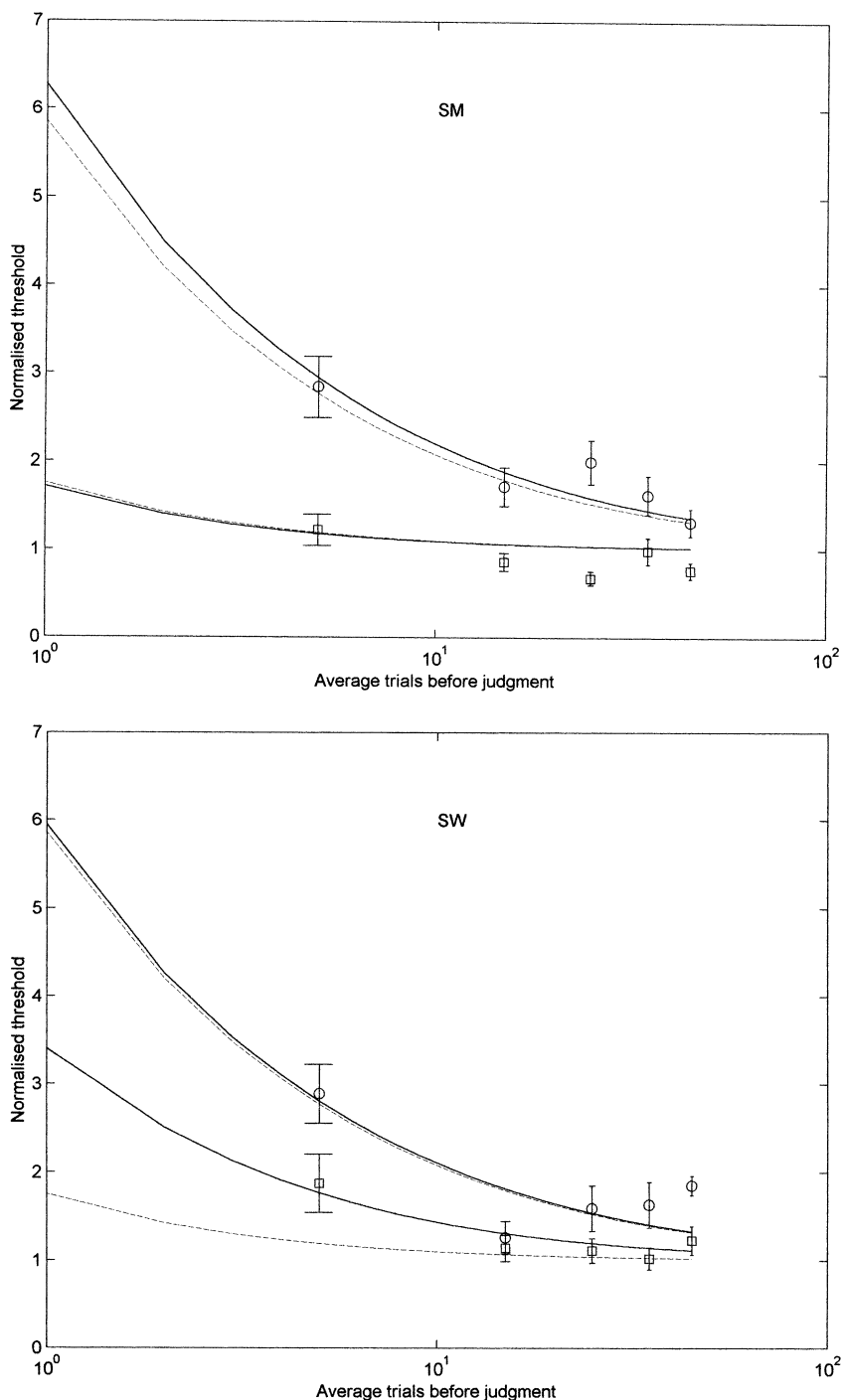


Fig. 3. Thresholds for sequential ten-trial session using MSS procedure. Plotted thresholds were divided by MCS Weber fractions for each observer. Open circles: 20 stimuli ('steps') in test set; filled circles: five stimuli in test set. The development of the implicit standard follows theoretical curves taken from Fig. 1, indicating that the observer averages or remembers a substantial number of prior trials. Two curves are shown for each set of data. The continuous curve assumes that the observer used all of the preceding trials to calculate the internal standard. The dashed curve is the one-parameter best fit to the real observer data, assuming that a proportion  $p$  of the previous trials are used. For both the real and ideal fit we use Formula 1 in the test. Panel A shows the results for observer SM and Panel B for SW.

# MULTIPLE IMPLICIT STANDARDS

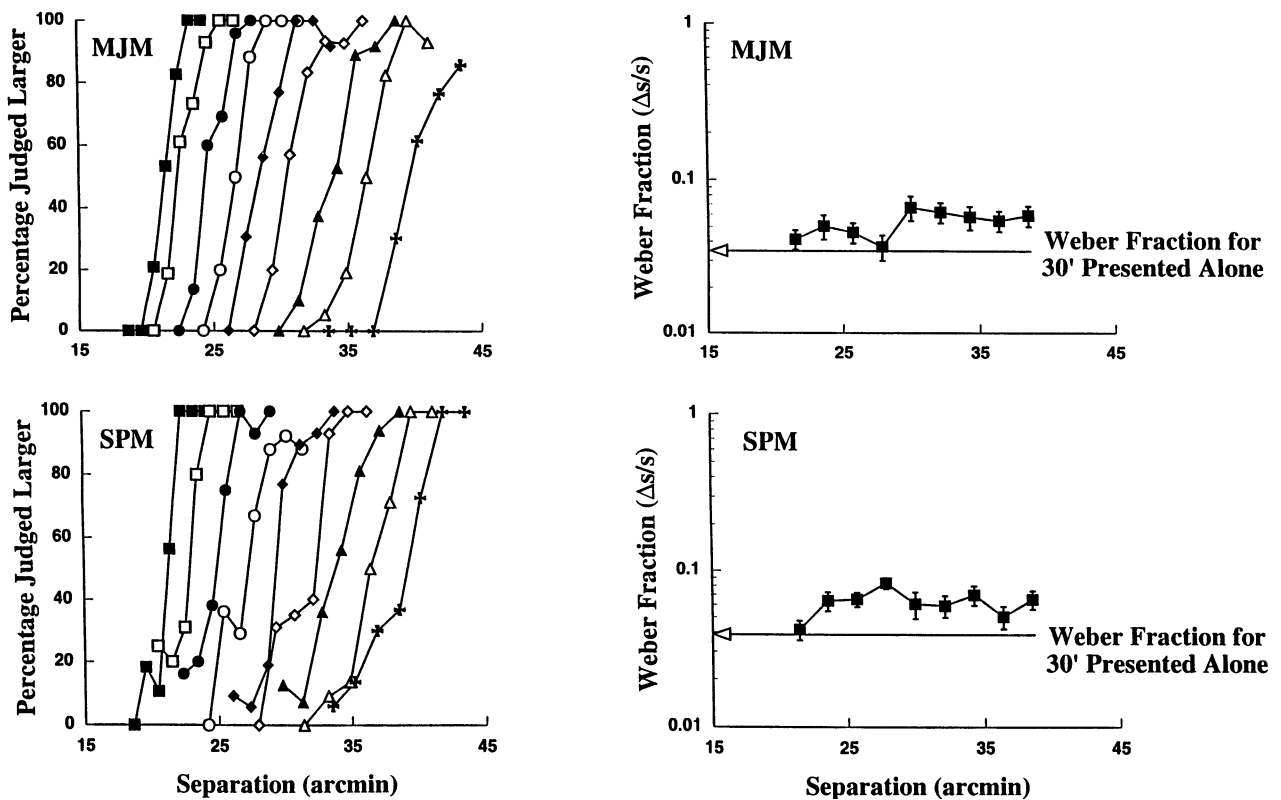


Fig. 4. Left side: Psychometric functions for nine interspersed implicit standards. Observers were shown a symbolic cue (a number between 1 and 9) which designated the implicit mean separation to be used in judging the test separation presented in the succeeding trial. Right side: Weber fractions estimated from the psychometric function shown on the left. The horizontal arrow shows the Weber fraction for a single separation tested in a block. The thresholds from the interspersed test sets are about a factor of two higher than the blocked condition.

test separation presented in the next trial. These symbolic cues were arranged in order, so that 1 corresponded to the smallest mean separation, 2 corresponded to next smallest mean separation, and so forth, up to 9, which corresponded to the largest mean separation. Each observer was given 200 practice trials with these nine interspersed standards, followed by the main experiment of 1000 trials. As in the experiments above, audible feedback was provided.

The left side of Fig. 4 shows psychometric functions for the nine interspersed separations. The functions are fairly discrete, indicating that observers can use the symbolic cues (the flashed numbers) effectively to judge each test trial against a different implicit mean. The Weber fractions on the right side of the Fig. 4 show some loss in precision. Thresholds are elevated by, at most, a factor of two when compared to the Weber fraction for single separation presented alone (block testing). Apparently, there is no obligatory combination of similar stimuli. The observer can decide on the basis of a symbolic cue which implicit reference is the correct one with only a moderate loss of precision, probably due to some confusion between adjacent standards.

## 4. Discussion

Thresholds measured with the MSS are as precise as thresholds measured with the MCS, and the PMFs are equally unbiased. As our simulations indicate, the estimation of the implicit standard during the first few trials in the MSS procedure contributes little additional noise to thresholds, provided that the range of test stimuli is relatively narrow and the total number of trials is large. If the observer can maintain a steady criterion based on a running average of the previous trials, the 'standard' stimulus commonly used in the MCS procedure is superfluous. Our results show that well-trained observers can estimate and store an accurate representation of stimulus parameters by sampling stimulus information over as many as 20 trials. Given that there is no cost in precision, the MSS procedure may be preferable to the MCS procedure. For one thing, the MSS procedure saves time over the two temporal intervals commonly used in MCS procedures, or, alternatively, it simplifies display arrangements over the simultaneous presentation of standard and test.

Theoretical considerations aside, it may still seem somewhat surprising that human observers use an internal criterion with the same precision as an external reference. Indeed, there are conflicting reports on the effect of reference stimuli in the extensive literature on this subject. Viemeister (1970) compared intensity discrimination measured with a 2AFC procedure to discrimination measured with a single-interval (Yes–No) procedure, with or without a reference stimulus preceding the single interval presentation. The reference stimulus had no effect on Yes–No performance. Viemeister also found that the  $d'$  for the Yes–No procedure was a factor of two worse than the uncorrected  $d'$  for 2AFC, i.e. a  $\sqrt{2}$  worse than predicted by signal detection theory. He argued that observers were taking a difference, either between the test and reference stimuli, or between the test and an internal criterion. As described in the introduction, this differencing strategy increases the variance by a factor of two, producing an extra  $\sqrt{2}$  reduction in Yes–No performance. Since performance was the same with or without the reference stimulus, Viemeister concluded that the noise or variance of the internal criterion was equal to the sensory noise associated with measuring the external reference.

Viemeister's conclusion may explain the equivalence of the MCS and MSS procedures. However, there have been many studies over the last three decades that show that an explicit reference actually improves performance (Jersteadt & Bilger, 1974; Jersteadt & Sims, 1975; Creelman & Macmillan, 1979). Far from degrading performance relative to 2AFC, the use of a reference brings Yes–No performance closer to the signal detection theory prediction. The reference stimulus appears to reduce internal criterion noise, serving as a 'reminder' (Macmillan and Creelman, 1991). So, why does the reference stimulus not improve performance in the MCS procedure?

We speculate that observers use an implicit standard for both MCS and MSS procedures. The only difference between a standard Yes–No procedure, in which the observer chooses between two possible stimuli, and the MCS — MSS procedures is that the latter employ multiple stimulus levels. With multiple test stimuli, the test stimulus from the preceding trial is often more discriminable from the current test stimulus than is the MCS reference. It might seem likely that in both MSS and MCS procedures, observers simply take the difference between the current test stimulus and the preceding one, but theoretically, this is a poor strategy. Vogels and Orban (1986b) simulated this strategy for the MSS procedure, and found that the predicted thresholds were generally *higher* than the observed thresholds. They concluded that the observer is judging each test stimulus against an internal criterion.

But given Viemeister's conclusion, why is this a better strategy? There is compelling evidence that the

internal criterion for a real observer is not a fixed value with zero variance (Treisman & Faulkner, 1984; Treisman & Williams, 1984; Lages & Treisman, 1998), but the internal criterion may usually have a smaller variance than the sensory response to an external reference stimulus. Over the last half-century, numerous studies have shown that the observer uses information from many trials to estimate the implicit standard. Our own study, showing that threshold precision improves over the first 15–20 trials, provides additional evidence of this sequential strategy. Thus, the range of stimuli used in these two venerable procedures may stabilise the internal criterion (implicit standard) and reduce its variance so that an explicit reference contributes no additional benefit in contrast to its stabilising effect in the standard Yes–No task.

### Acknowledgements

The research was supported by NEI grant RO1-EY06644, and by a grant from the Biological and Biotechnology Science Research Council (BBSRC) of Great Britain. We thank Roger Watt and Preeti Verghese for valuable discussion.

### Appendix A

To show why the MCS and 2AFC procedures lead to different predicted sensitivities according to signal detection theory, we will consider only two hypothetical test stimuli in the MCS test set. Typically, a MCS test set consists of five to seven test values that symmetrically straddle the value corresponding to the standard stimulus, and includes a test stimulus which is equal to the standard. At the end of a MCS test run, a psychometric function is fitted to the proportion of trials labelled 'greater' for each test value. The threshold is estimated as the difference between an estimated stimulus value corresponding to the median ( $P = 0.5$ ) and a value corresponding to some criterion proportion, such as  $P = 0.84$  (equivalent to  $d'$  of 1). Suppose fortuitously that these two proportions (0.5 and 0.84) correspond to two members of the test set, namely the test stimulus (S1) which is equal to the standard and so has a sensitivity in  $d'$  units = 0, and another higher test stimulus (S2) which has a corresponding sensitivity in  $d'$  units = 1.

In signal detection theory, each stimulus generates an internal probability distribution with a mean and variance. The means for these two stimuli are equal to their two  $d'$  values (0, 1) and their variances are assumed to be equal. Now consider how these two stimuli are tested in the two psychophysical procedures. In the MCS procedure, the 'standard' is always presented in



the first interval and the test stimulus in the second. Thus, the observer is presented with S1, S1 on some trials, and with S1, S2 on other trials. If we assume that the observer makes his or her judgement by taking the difference between the two intervals, then the decision to choose 'first' or 'second' is based on a difference distribution. For these two stimuli, the mean of the difference distribution in  $d'$  units is 0 for S1, S1, and 1 for S1, S2. However, the variance of each difference distribution is double the variance of the internal probability distribution associated with the individual stimuli, assuming that the two test intervals are independent. Thus, the standard deviation of the difference distributions is  $\sqrt{2}$  larger than the standard deviation associated with the ideal performance in a Yes–No task, meaning that the effective  $d'$  for the difference distributions in MCS is  $[1 - 0]/\sqrt{2} = d'/\sqrt{2}$ , relative to the ideal Yes–No performance.

In the 2AFC procedure, the observer is presented with S1, S2 on some trials, and S2, S1 on other trials. Again, if we assume that the observer is taking the difference between the two intervals, the mean of the difference distribution is  $-1$  for S1, S2, and  $+1$  for S2, S1, so the separation between the two different distributions is *double* the separation between the internal probability distributions for the two stimuli. So, although the variance for the 2AFC task is also double, the  $d'$  for the difference distributions in 2AFC is  $[1 - (-1)]/\sqrt{2} = 2/\sqrt{2} = \sqrt{2}d'$ , meaning that the effective  $d'$  in a 2MC task is  $\sqrt{2}$  better than the ideal performance in a Yes–No task.

## References

- Burbeck, C. (1987). Position and spatial frequency in large-scale localization judgements. *Vision Research*, 27, 417–427.
- Creelman, C. D., & Macmillan, N. A. (1979). Auditory phase and frequency discrimination: a comparison of nine procedures. *Journal of Experimental Psychology: Human Perception and Performance*, 5, 146–156.
- Fernberger, S. W. (1931). On absolute and relative judgements in lifted weight experiments. *American Journal of Psychology*, 43, 560–578.
- Finney, D. J. (1971). *Probit analysis*. Cambridge: Cambridge University Press.
- Green, D. M., & Swets, J. A. (1966) *Signal detection theory and psychophysics*. New York: John Wiley & Son
- Helson, H. (1947). Adaptation-level as frame of reference for prediction of psychophysical data. *American Journal of Psychology*, 60, 1–29.
- Helson, H. (1948). Adaptation-level as a basis for quantitative theory of frames of reference. *Psychology Review*, 55, 297–313.
- Jersteadt, W., & Bilger, R. C. (1974). Intensity and frequency discrimination in one- and two-interval paradigms. *Journal of the Acoustical Society of America*, 55, 1266–1276.
- Jersteadt, W., & Sims, S. L. (1975). Decision processes in frequency discrimination. *Journal of the Acoustic Society of America*, 57, 1161–1168.
- Lages, M., & Treisman, M. (1998). Spatial frequency discrimination: visual long-term memory or criterion setting? *Vision Research*, 38, 557–572.
- Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. Cambridge: Cambridge University Press.
- Masin, S. C. (1987). Different biases in the methods of constant and single stimuli. *Bulletin of the Psychonomic Society*, 25, 379–382.
- McKee, S. P., Welch, L., Taylor, D. G., & Bowne, S. F. (1990). Finding the common bond: stereoacuity and other hyperacuties. *Vision Research*, 30, 879–891.
- Morgan, M. J. (1992). On the scaling of size judgements by orientational cues. *Vision Research*, 32, 1433–1445.
- Parducci, A. (1956). Direction of shift in the judgement of single stimuli. *Journal of Experimental Psychology*, 51, 169–178.
- Parducci, A. (1964). Sequential effects in judgement. *Psychology Bulletin*, 51, 169–178.
- Pfaffman, C. (1935). An experimental comparison of the method of single stimuli in gestation. *American Journal of Psychology*, 47, 470–476
- Sorkin, R. D. (1964). Extension of the theory of signal detectability to matching procedures in psychoacoustics. In J. A. Swets, *Signal detection and recognition by human observers*. New York: John Wiley & Sons Inc.
- Tabachnick, B., & Parducci, A. (1973). Improved recognition with feedback: discriminability and range-frequency effects. *Bulletin of Psychonomics Society*, 1, 56–58.
- Treisman, M., & Faulkner, A. (1984). The setting and maintenance of criteria representing levels of confidence. *Journal of Experimental Psychology: Human Perception and Performance*, 10, 119–139.
- Treisman, M., & Williams, T. C. (1984). A theory of criterion setting with an application to sequential dependencies. *Psychology Review*, 91, 68–111.
- Viemeister, N. F. (1970). Intensity discrimination: performance in three paradigms. *Perception and Psychophysics*, 8, 417–419.
- Vogels, R., & Orban, G. A. (1986a). Decision processes in visual discrimination of line orientation. *Journal of Experimental Psychology: Human Perception and Performance*, 12, 115–132.
- Vogels, R., & Orban, G. A. (1986b). Decision factors affecting line orientation judgements in the method of single stimuli. *Perception & Psychophysics*, 40, 74–84.
- Westheimer, G., & McKee, S. P. (1977). Spatial configurations for hyperacuity. *Vision Research*, 17, 941–947.
- Wever, E. G., & Zener, K. E. (1928). The method of absolute judgement in psychophysics. *Psychology Review*, 35, 560–578.
- Woodworth, R. S. (1938). *Experimental psychology*. New York: Henry Holt & Co.
- Woodworth, R. S., & Schlosberg, H. (1954). *Experimental psychology*. New York: Henry Holt & Co.