Vision Res. Vol. 23, No. 8, pp. 811-820, 1983
Printed in Great Britain.

0042-6989/83 $3.00 + 0.00

THE PRECISION OF NUMEROSITY DISCRIMINATION IN ARRAYS OF RANDOM DOTS

A. BURGESS* and H. B. BARLOW
Physiological Laboratory, Cambridge CB2 3EG, England

(*Received* 9 *August* 1982)

Abstract—The precision of making discriminations between the numbers of dots in a pair of irregular arrays was measured. The results fit the assumption that the observer adds intrinsic variance to whatever variance is present in the numbers displayed, the errors depending upon the sum of the two. We found no evidence for incomplete use of the sample of information presented, other than this observer variance. Its value increases as about the 0.75 power of the mean number of dots in a display, except for numbers up to about 20 where it changes much more rapidly. Decreased irregularity in the arrangement of the dots decreases observer variance, but it is little affected by large variations in average density of dots per unit area, and is also little changed by making the dots vary irregularly in brightness.

INTRODUCTION

Experimental psychologists have been interested in the ability to judge numbers of objects in a field of view, first because this number was thought to give a measure of the mind's "span of apprehension", and second because the experiments were particularly amenable to quantitative treatment. Jevons (1871) was one of the first to attack the problem, and he showed that errors began to be made for 5 objects, and were over 50% for 10 objects. Taves (1941) and Kaufman *et al.* (1949) drew a sharp distinction between what the latter called "subitizing", when the number was small, and the mechanism for judging larger numbers. This distinction was based on the absence of errors, greater subjective confidence, and shorter reaction times for numbers below about 7. Hunter and Sigler (1940) also showed that there was a reciprocal relation between the intensity of illumination and the duration of exposure required for correctly judging the number when this was below 8, whereas for larger numbers the reciprocity broke down and more time was required to judge the number correctly, no matter how high the intensity: the subjects presumably required several fixation positions and probably had to count in order to obtain the correct answer. These studies are well reviewed by Woodworth and Schlosberg (1954).

Our own interest in the problem also arose from the ease of quantification together with the fact that judging numbers must involve processes more complex, and presumably accomplished at higher levels in the brain, than those involved in psychophysical tests of the eye's elementary visual functions. We take a physicist's viewpoint, regarding the human observer as an instrument that forms some measure of dot number, and the first question we naturally ask concerns the precision of these measures. We concentrate on random errors and play little attention to the consistent errors for much the same reason that a designer of clocks is more interested in the reproducibility of the daily error than in its actual value: random errors tell one much about the quality of the mechanism's design, whereas consistent errors merely betray faulty calibration. This is also the reason we use comparative judgements in our experiments, asking the subject which of two fields has more dots rather than what the actual number is, for in this experimental design consistent errors tend to cancel out.

In a task such as judging the number of dots scattered over a large field it is difficult to understand how the brain gathers together the specific items of information required to allow reliable judgements to be made, while avoiding disturbance from irrelevant features. The central problem is that of achieving high signal/noise ratios at an enormous range of perceptual tasks, as argued elsewhere (Barlow, 1980). Because of this difficulty it is appropriate to express performance in such a way that one can see at once how well the task of collecting together the relevant information has been done, and the measure of statistical efficiency proposed by Fisher (1925) does this. He suggested that degradation of performance, from whatever cause, be expressed *as if* it resulted from using a smaller sample of information than was actually available. Since the coefficients of variation (standard deviation/mean) of estimates are usually inversely proportional to the square root of the sample size on which they are based, accuracy will vary as the square root of efficiency; we have followed this assumption.

Notice that efficiency is an overall measure of per-

*Present address: Department of Diagnostic Radiology, Faculty of Medicine, University of British Columbia, 10th Avenue and Heather Street, Vancouver, BC, Canada V52 1M8.

formance and does not depend upon any specific model: although expressed *as if* the available sample had been incompletely utilised, many other factors would cause a loss of efficiency, as is also the case when efficiencies are used in mechanics and thermodynamics. Fisher was concerned mainly with degradation of performance resulting from inappropriate calculation, but the measure can be used whenever a statistical estimate or decision is involved, and in previous studies (Barlow, 1978; Barlow and Reeves, 1979; van Meeteren and Barlow, 1981) it has been shown that human subjects make perceptual decisions with efficiencies up to 50%. The principle purpose of that work was to extend the use of the absolute measure of statistical efficiency from cases where quantum fluctuations are involved (Rose, 1948, 1973; Barlow, 1977) to a situation involving higher levels of perception and other sources of noise. In those studies no attempt was made to investigate the cause of lost efficiency. Does it actually result from failure to use part of the sample of information available, or is the pattern of imperfect performance better explained in some other way?

In considering why a human subject may make unnecessary errors in judging the number of dots two possibilities naturally come to mind. As suggested above, the subject may really fail to take in all of the sample of information available to him; this type of inefficiency would be analogous to a failure to absorb photons in a light-meter, or to throwing away a proportion of the measurements when deriving a mean, and the efficiency measure expresses the results as if this was the only cause. But efficiency is a model-free overall measure of performance, and just as noise in a photo-cell limits its performance at low intensities, so noise in the human "dot-number estimator" may reduce its efficiency. This is a second possible cause of efficiency loss to consider. In this paper we try to separate the two cases by measuring performance as a function of the amount of added noise; that is, the standard deviation of the numbers of dots presented in the target stimulus was varied. The main result is to show that a source of intrinsic random error in dot number estimation may account for all the loss of efficiency.

Some results are also given on subsidiary problems, such as a comparison of the two-alternative forced-choice psychophysical procedure with the Yes–No or two-point method that has been mainly used in previous studies on this problem from this laboratory.

Definitions

N = The average number of dots in a field (one side of the display)

ΔN = the number of extra dots added to a field. In the forced-choice procedure the subject must decide which field has ΔN; in the two-point method the subject must decide whether the left field has N or $N + \Delta N$

ΔN_T = the value of ΔN_T for which a subject achieves d'_E of 1

$\Delta N_T(0)$ = the value of ΔN_T when added variance, σ_N^2, is zero

σ_N^2 = the variance of the number of dots

σ_O^2 = observer variance, which is assumed to be added to the measure of dot number upon which the subject's judgements are made. When this is taken into account the subject will perform as if the variance of dot number is $\sigma_N^2 + \sigma_O^2$, instead of σ_N^2

d'_E = experimental value of an index of discriminability between two populations. It is defined as the difference between the means of two populations divided by the standard deviation of one of them, usually the smaller

d'_I = highest value of d' that could theoretically be obtained in the situation under consideration

F = Fisher's measure of absolute statistical efficiency. If the full sample available is reduced by the factor F, then d'_I will be reduced to a value equal to that actually obtained. Hence $F = (d'_E/d'_I)^2$, where d'_I is here the value appropriate to the full sample. It is assumed that the coefficient of variation (standard deviation/mean) of estimates varies inversely with the square root of the sample used

F_N = the value of statistical efficiency that is obtained when allowance is made for observer variance; a possible cause would be that the subject only makes use of an average fraction F_N of the sample available.

THEORY

Suppose a subject is presented with samples from two populations placed sidy-by-side as in Fig. 1A. Both parent populations have gaussian probability distributions of the number of dots, with standard deviation σ_N, but one population has mean N and the other $N + \Delta N$. The population with the excess, ΔN, is placed on either side with equal prior probability and the subject's task is to identify which side. The subject maximizes his probability of being correct by selecting the side that appears to have the greater number of dots, and care is taken to avoid other cues to help him. Since the mean difference between the half-fields is ΔN and the variance of the difference is $2\sigma_N^2$, the highest possible percentage correct that could be obtained would correspond to a normal equivalent deviate, or Z score, of $\Delta N/(2\sigma_N^2)^{\frac{1}{2}}$, which corresponds to d'_E of $\Delta N/\sigma_N$ in the two-alternative force-choice test. It has been shown that, in a very similar task, subjects do not reach this ideal performance, but achieve efficiencies of about 50% under quite a wide

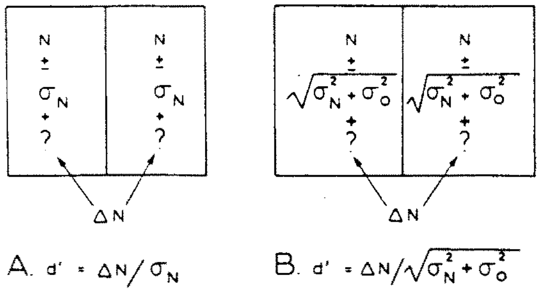

Fig. 1. In the two-alternative forced-choice method the subject has to decide whether an excess, ΔN, of dots has been placed on the left half-field or the right. In (A) the background number, N, is subject to independent random variation of $\pm\ \sigma_N$ on both sides. The maximum attainable value of d', calculated from the percent correct, would be $\Delta N/\sigma_N$. We postulate that the observer adds his own intrinsic variance, σ_O^2, to the experimentally controlled variance, σ_N^2, so that the situation is as shown in (B); the maximum attainable d' becomes $\Delta N/\sqrt{\sigma_N^2 + \sigma_O^2}$.

range of conditions (Barlow, 1978). That is

$$d'_E = F^{\frac{1}{2}}d'_I = F^{\frac{1}{2}}\Delta N/\sigma_N \quad \text{and} \quad F = 0.5 \qquad (1)$$

where d'_E is the experimental discriminability achieved, d'_I is the ideal performance described above, and F is the statistical efficiency as defined by Fisher (1925). Tanner and Birdsall (1958) use the square of the ratio of input signal/noise ratios for ideal and real observers giving the same output performance; in some situations, but not all, this is identical with our definition of F.

We want to explore the possibility that the loss of performance is due to a source of noise intrinsic to the observer, rather than to his failure to make use of the full sample of information available to him. Since we are interested in the effects of noise on the estimate of dot number, this step is exactly analogous to expressing the dark current of a photocell in terms of the illuminance that would cause it, or the intrinsic noise of the retina in terms of "dark light" (Barlow, 1956; 1957). We thus postulate that this intrinsic observer noise adds variance σ_O^2 to the estimate of the number of dots seen. The new situation is shown in Fig. 1B. The best value of d'_E the subject can now achieve is $\Delta N/(\sigma_N^2 + \sigma_O^2)^{\frac{1}{2}}$, and if the intrinsic observer variance was the only limit to his performance he would achieve this. However, for generality we want to preserve the possibility that performance is degraded by other factors as well as by adding noise, and we therefore define F_N; this quantity is analogous to the overall efficiency F, but calculated after taking account of the observer variance. We thus predict that the experimentally determined d'_E will be

$$d'_E = F_N^{\frac{1}{2}}\Delta N/(\sigma_N^2 + \sigma_O^2)^{\frac{1}{2}} \qquad (2)$$

If the only other factor causing degradation was incomplete use of the sample of information available, then F_N will be the average fraction utilized. d'_E can be measured experimentally, ΔN and σ_N are controlled by the experimenter, but there are still two unknowns to determine from the data, namely F_N and σ_O. Our procedure for separating these factors is to vary σ_N, which was always given a value of $\sqrt{N}$ in the previous study (Barlow, 1978). Rewriting equation (2) as

$$(\Delta N/d'_E)^2 = (\sigma_N^2 + \sigma_O^2)/F_N \qquad (3)$$

one sees an expected linear relation with σ_N^2. The slope of this relation is $1/F_N$, and the negative intercept with the horizontal axis depends on σ_O^2, as illustrated in Fig. 2.

The above formulation can be readily adapted for other psychophysical methods. If a threshold is measured this will be the value of increment ΔN_T that gives a fixed value of d'_E, which we have chosen to be 1. Equation (3) then becomes

$$\Delta N_T^2 = \sigma_N^2/F_N + \sigma_O^2/F_N. \qquad (4)$$

We have often used the two-point method in which the subject has to decide from which of two populations of stimuli already known to him the members of a series of unknown stimuli were drawn. For this situation equations (3) and (4) are as above. We label the value of ΔN_T when the added variance is zero $\Delta N_T(0)$. Thus

$$\Delta N_T(0) = \sigma_O^2/F_N.$$

METHODS

The patterns were generated and displayed by a PDP 11/40 computer with GT40 display. The subject sat 1.4 m from the screen, at which distance the full display subtends 2.2° and the smallest step of the 10 bit D/A converters causes a spatial displacement of

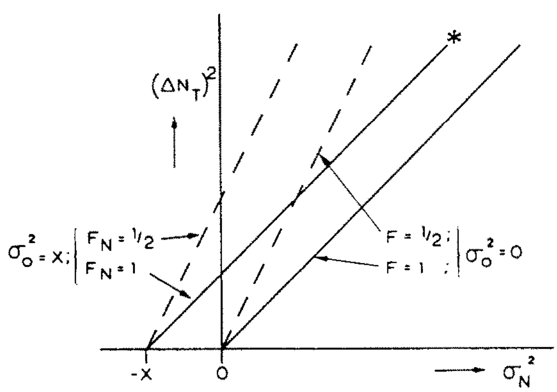

Fig. 2. The effect of observer variance, σ_O^2, can be differentiated from incomplete utilization of the sample by measuring threshold, ΔN_T, with different values of experimental variance, σ_N^2, of the dots in each hemifield. The lines converging at $\sigma_N^2 = -X$ show the behaviour predicted with observer variance of X, and full utilization of the sample (solid line, $F_N = 1$) or incomplete utilization (dashed line, $F_N = \frac{1}{2}$). With zero observer variance these lines are shifted to the right and converge on the origin. Note also that $(\Delta N_T)^2$ at $\sigma_N^2 = 0$ is a good estimate of σ_O^2 when $F_N = 1$. Results of Fig. 4 fit best the line marked at the top with an asterisk.

just under 8 sec of visual angle. The program for generating the patterns and analysing the responses were written in BASIC and used a pseudo-random number generator for positioning the dots, as described previously (van Meeteren and Barlow, 1980). The two authors were the main subjects; both are experienced as psychophysical subjects and wore their normal correcting glasses. The viewing time was usually self-paced and was about 1 sec. This procedure was adopted in order to reduce observer fatigue. Some results using 200 msec exposures are shown in Fig. 9, and where the conditions are otherwise similar the self-paced results are only marginally more precise than those with 200 msec.

Two psychophysical methods were used, and one of the objects of the experiments was to compare them. In the two-alternative forced-choice method two patterns were generated from populations with gaussian probability distributions of equal standard deviation, σ_N, and means N and $N + \Delta N$. The subject's task was to say which side contained the sample from the population with the higher mean. The gaussian distributions were truncated at $\pm 3\sigma$, which necessitated a small correction to the calculations. Correct responses were signalled by extinguishing the display immediately, whereas for errors this was delayed about $\frac{1}{2}$ sec; note that the populations overlapped, so the side with more dots was not always the "correct" one to select. The other method used a similar display, but the pattern on the right side was always the same in any given run of 100 unknown presentations. The pattern on the left either had the excess, ΔN, or did not have it, and the subject's task was to decide from which population it came. Before a run he gained experience of the two populations by seeing examples from each on demand, and during the run correct and incorrect answers were signalled as before. We call this the two-point method (Barlow, 1962); it is similar to the classical Yes–No method, but the criterion used for distinguishing the two populations is not necessarily the same as a threshold. At first it might appear that the fixed comparison field on the right makes this into a version of the two-alternative forced-choice task, with the subject simply deciding which field has the greater number. However, the subject pays much more attention to the varying left field, and the unimportance of the right field is reinforced by the fact that it can actually be omitted without affecting the results to any great extent.

With either two-alternative forced-choice or two-point methods a "threshold" ΔN_T was arbitrarily defined as the value of increment that can be discriminated with a $d'_E = 1$; many determinations were made with zero added noise, and this is $\Delta N_T(0)$.

RESULTS

Comparison of two-alternative forced-choice and two-point methods

Table 1 shows the results for both subjects of a comparison of the two-alternative forced-choice method with the two-point method, based on 8 or 10 runs of 100 trials each. The agreement is obviously very satisfactory, and the result also agrees with previous results on the same test (Barlow, 1978). For comparing variability, the results of both subjects were taken together since the means agree and the total number of trials was only 18. The empirical scatter of the results is rather less than that calculated from the expected binomial variability of the experimentally determined probabilities, but this divergence is not unexpected because of the crudity of the approximations involved in calculating sampling variance.

One subject (H.B.) did a series with the two-point method in which the fixed comparison field was omitted together and he had to maintain his criterion for distinguishing the two populations without it. d'_E was 1.30 ± 0.11 (8), which is barely lower than the result with the reference field present.

Both subjects found that the two-alternative forced-choice task required fewer preliminary learning trials than the two-point method, and involved a judgement that seemed marginally easier. The predicted sampling errors are also slightly less, because it requires inspection of two fields containing in all double the

Table 1. Comparison of results using two methods on two subjects

		A.B. Mean ± SE	(N)	H.B. Mean ± SE	(N)	SD's Empirical	SD's Calculated
Two-alternative forced-choice	d'_E	1.46 ± 0.04	(10)	1.50 ± 0.09	(8)	0.19	$\sqrt{2/nw} = 0.22$
	F	0.54 ± 0.03	(10)	0.55 ± 0.07	(8)	0.14	$\sqrt{4d'_E/nw(d'_I)^2} = 0.19$
Two-point	d'_E	1.43 ± 0.09	(10)	1.49 ± 0.04	(8)	0.22	$\sqrt{4/nw} = 0.28$
	F	0.53 ± 0.06	(10)	0.56 ± 0.03	(8)	0.15	$\sqrt{8d'_E/nw(d'_I)^2} = 0.24$

The two-alternative forced-choice method had a mean of 100 dots ± 10 (SD) in one hemifield and 120 ± 10 the other. The two-point method had either 100 ± 10 or 120 ± 10 in the left hemifield and 110 in the right; the arrangement of the dots in the right hemifield was unchanged throughout the 100 trials of each run, but varied from run to run. The eight or ten runs for each condition were done within a few days of each other. In comparing SE's, note that A.B. was more accustomed to the two-alternative method, H.B. to the two-point. Expected standard deviations were calculated from the approximate expressions in the final column; n is the total number of observations, w is the weighting coefficient (Finney, 1947) and d'_I, d'_E are the ideal and experimental discriminabilities.

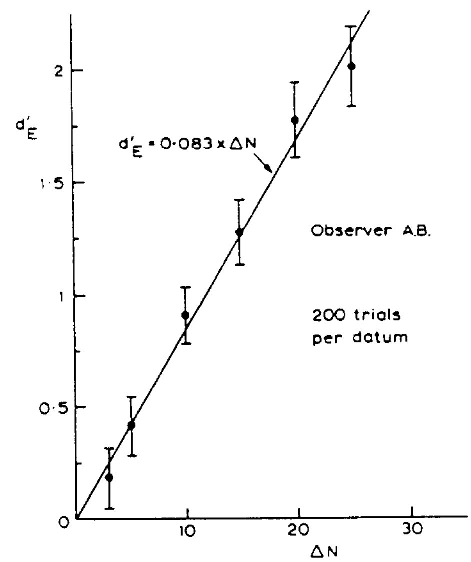

Fig. 3. Experimental discriminability, d'_E, plotted as a function of the added dots, ΔN, in a two alternative forced choice task with 100 dots in each hemifield and zero added variance, σ_N^2. Bars indicate ±1 SE, based on 200 trials per point. The hemifields were 2.2° × 1.1°. Solid line is the least squares fit constrained to pass through the origin, and is clearly adequate.

number of dots. However, it takes longer to produce the patterns and to look at them, because each hemifield has to be separately inspected, and for some tests the fact that two uncontrolled fixation positions are required would be disadvantageous.

Changing the variance of the number of dots

As described under theory and illustrated in Fig. 2, changing the variance of the number of dots enables one to distinguish between two causes of efficiency loss. If ΔN_T^2 is plotted against added variance, failure to use the full sample would increase the slope of the line, whereas a constant added variance would shift the line upwards by causing an equal elevation at all imposed variances.

We define threshold, ΔN_T, as the number of dots that must be added to a background containing N dots in order to achieve an experimental d'_E of 1. This corresponds to 76% correct in the forced choice method, or 69% correct Yes's and No's in the two-point method. Apart from the difficulty of adjusting ΔN to obtain these rates, it is also more accurate (Barlow, 1962) and subjectively more satisfactory to use higher values of ΔN where the error rates are lower, but in order to interpolate from the measured d'_E values to obtain ΔN_T it is then necessary to know if the relation between ΔN and d'_E is linear. Figure 3 shows that this is so for the forced-choice method under our conditions, and Fig. 3 of Barlow (1978) shows it also holds for the two-point method under similar conditions. ΔN_T can thus be estimated by measuring d'_E in the convenient range where it has a value 1.5 to 2.5.

ΔN_T was determined by the above method for different values of added variance, and Fig. 4 shows these results. Subject A.B. used the forced-choice

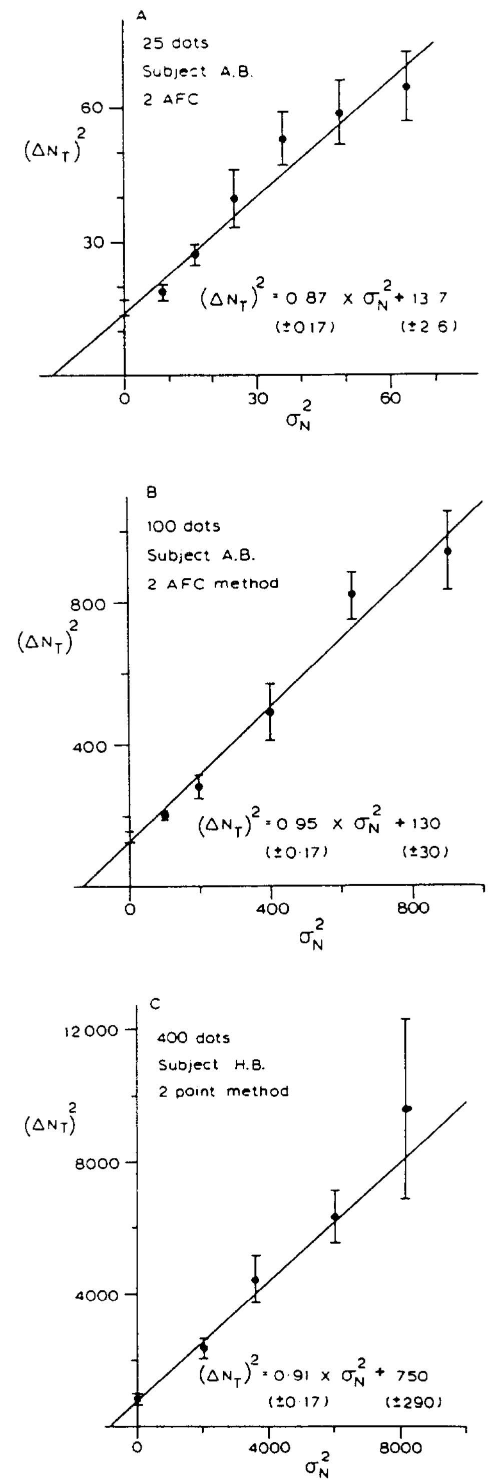

Fig. 4. The effect of added variance on the threshold defined as the number of added dots, ΔN_T, required to achieve a $d'_E = 1$. The results fit best the assumption that avoidable errors result from observer variance rather than incomplete use of the sample of information provided by each pattern (see theory and Fig. 2).

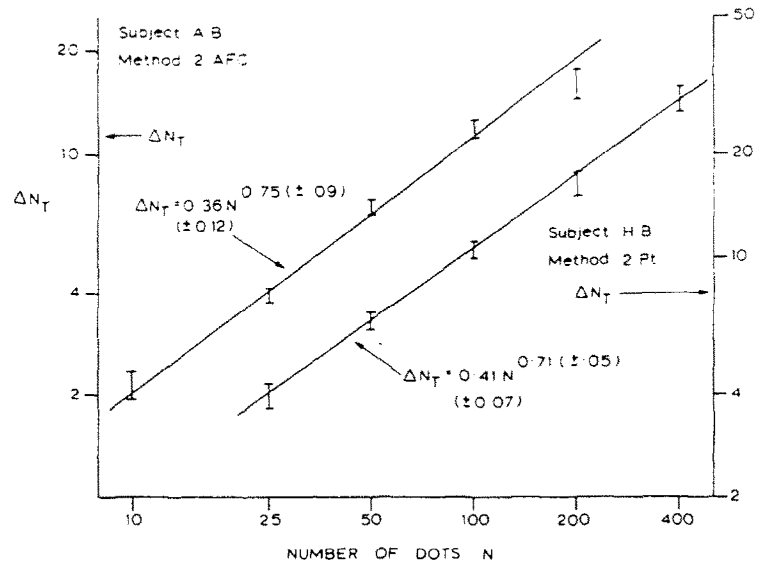

Fig. 5. The relation between the number, N, of dots in each hemifield of the display and the threshold number of extra dots ΔN_T. ΔN_T varies approximately as the 0.75 power of N (but see also Fig. 9). Because the slopes in Fig. 4 are close to unity, ΔN_T is a good estimate of σ_0, the intrinsic observer noise. The numbers in parentheses indicate the 90% confidence limits of the parameters they are adjacent to. Note that H.B.'s results are shifted down by a factor of 2 to prevent confusion.

method with background numbers of dots of 25 and 100 per hemifield. H.B. used the two-point method and 400 dots per hemifield. The slopes of the best fitting straight lines through the data do not differ significantly from unity. Thus the result favours the hypothesis that the observers' performance is degraded by the addition of variance, and that this variance does not change when the experimentally imposed variance is changed. On the alternative hypothesis that a constant proportion of the sample is discarded or ineffectively utilized the lines would have slopes greater than unity (see Fig. 2): all the slopes are actually less than unity, though not significantly.

Effect of number of dots on observer variance

Since a constant additive observer variance seemed to account for the whole of the departure from ideal performance in this experiment we decided to find how it varied with the number of dots in the background field. The results are shown in Fig. 5. The experiments were done with no added variance, and ΔN_T was estimated in the usual way. Although it is σ_0 that interests us, we have preferred to plot the experimentally determined values; consideration of expressions (3) and (4) shows that $\sigma_0 \simeq \Delta N_T(0)$ when $F_V \simeq 1$, which was shown to be nearly true in Fig. 4. The two subjects used different methods and somewhat different ranges of dot numbers, but the relations found are very similar.

We cannot find any easy explanation for the exponent of about 0.7 in this relationship. Since the variance of the number of dots in each small area is Poisson distributed and therefore proportional to dot density, it would be relatively easy to explain an exponent of 0.5; indeed such a relationship could be explained in a number of ways. There are also some grounds for expecting an exponent of unity, since this is the prediction of Weber's Law. All that can be said is that the empirical value lies between these "explainable" values.

Effect of varying dot density

The variation of observer variance with dot number might be related to the associated change in the average separation of dots, or number of dots per unit area. The experiment of Fig. 6 shows the effect of varying the overall size of the display on ΔN_T with zero added variance. When each hemifield subtended about 1/4 × 1/8 deg $\Delta N_T(0)$ was slightly elevated, but above this it was unaffected by a 16 fold change of linear dimensions. Dot density, therefore, does not seem to be an important variable.

Effect of making distribution more regular

Subjectively, the difficulty in estimating which of two hemifields has more dots appears to be related to the unevenness of the random distributions. An example in which the distribution is, by chance, rather even, seems easier to judge than one in which there are accidental clusters. A program was therefore written in which the display area was divided into a regular array of nearly square cells equal in number to the number of dots to be displayed. Each dot was then placed at a random position within each cell. The

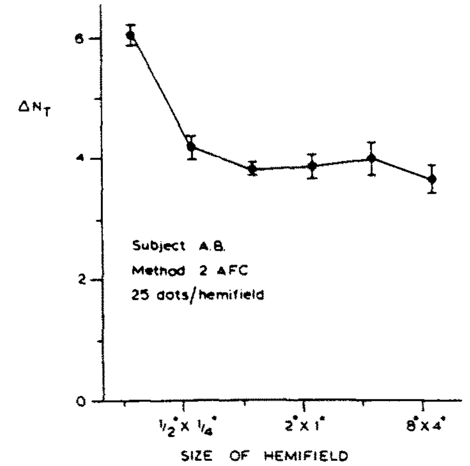

Fig. 6. The effects of reducing the size of the display and thereby increasing the average density of dots. There were 25 dots per hemifield. Clearly size and density have little effect until the hemifields are reduced to about $1/2° \times 1/4°$.

patterns appeared less irregular than for the non-cellular patterns, but it was not obvious how they had been made more regular. Figure 7 shows the effect on $\Delta N_T(0)$ for various dot numbers. As expected, $\Delta N_T(0)$ is reduced compared with the totally random displays. The change in exponent (slope on this log/log plot) may not be significant.

Effect of varying luminous intensity as well as number of dots

It is possible that there are cells in the visual system that respond to the total luminous flux emanating from the dots in one hemifield, and that the estimation of dot number is based on the responses of such cells. If this was the case the accuracy of estimation would be seriously perturbed by brightening randomly selected dots in each hemifield. We have not done exhaustive trials of this type, but the following example suggests that it is possible to disregard variations of luminance in a display of dots and judge their number independent of these variations.

The background number of dots was 25, and this was incremented by 3 to 7 dots on one side or the other in order to estimate $\Delta N_T(0)$ in the usual way. Of the dots in each hemifield 12 were enhanced by over-writing them 4 times. The value of $\Delta N_T(0)$ was 3.89 ± 0.26, which should be compared with a value of 3.83 ± 0.27 when all the dots were at a single intensity level. Apparently the number of dots was judged equally well when some were dim and some 5 times more intense as when they were all at the same luminance. The number of enhanced dots was then made variable, a mean of 12 ± 4(SD) in each hemifield being enhanced, the number being independently selected for each hemifield. The value of $\Delta N_T(0)$ obtained was 4.20 ± 0.18, which is not a significant difference. If the judgement had been based on luminance, the independent enhancement of 12 ± 4 dots in each field would have introduced an easily detectable elevation of $\Delta N_T(0)$.

Detection of incremental dots in an undelimited area

In the experiments described so far the dots have been displayed in two rectangles separated by a small space and a dividing line. The subject's task of deciding whether a particular dot should be included in the left side or the right side was thus made very easy. The question arises whether a subject can arbitrarily delimit an area within a display and decide if there is an excess of dots within that area, or whether he is unable to do so and will thus suffer a loss of performance caused by an error in matching the counting area to the target area. Such a loss of performance would cause a change of slope when ΔN_T^2 is plotted against σ_N^2 and would thus change F_N, rather than observer variance, σ_0^2. It was previously suggested that such matching errors lay behind the figure of approximately 50% efficiency that was found for a variety of tasks (Barlow, 1978).

This turns out to be a difficult hypothesis to test, and since our results are inconclusive they will be described only briefly. The target area can form only a small fraction of the total area, and therefore can contain only a small fraction of the total number of dots. Since we cannot handle conveniently more than about 200–400 dots the average number in the target area was limited, and in the two tests described below it was 25 and 6.25. In order to determine the effect of changing the experimental variance one must be able to change it by a large factor, but one cannot impose a greater standard deviation on the population than $N/3$, using our normal distribution truncated at ± 3 standard deviations, for even this will occasionally cause displays of zero dots. Since the normal standard

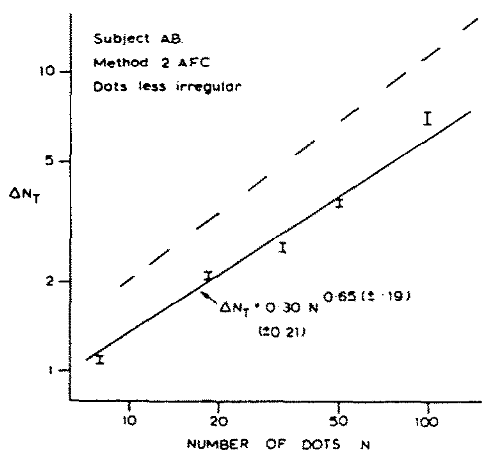

Fig. 7. The dots were arranged more regularly by dividing each hemifield into a grid with the required number of rectangles, and putting each dot at a random position in its rectangle. The dependence of ΔN_T on N is moved downwards compared with the previous results with completely random positioning; the dotted line is the fit to A.B.'s results in Fig. 5.

deviation of the 25 dots in the target area is ±5, this can only be raised to ±8, which is an insufficient range to test the hypothesis.

In the first attempt to overcome this difficulty we divided the display area into 12 cells of equal area, each containing an average of 25 dots. The standard deviation of the number within each cell could be arranged to vary from 0 to ±8. Then an increment was added to the central square of either the left or the right hemifield and the subject had to decide which. The results for the two subjects were concordant, and the best fitting line through the results had a slope of 0.92 ± 0.15 (90% confidence limits). There was thus no evidence at all for any matching error (see Figs 2 and 4). The intercept at zero added variance gave a value of $\Delta N_T(0) = 4.7$, which is only slightly higher than the expected value of 4.0 with 25 dots.

One difficulty with this experiment was that the cellular design became apparent when the imposed variance was high, for neighbouring cells had widely different dot numbers and their borders therefore became apparent; this in turn might be used as a cue to the target area, thus endangering the experiment.

Our second attempt to increase the range of standard deviations in the target area used a technique for positioning the background dots in clusters, rather than each dot being positioned independently. This only achieved slightly more than a twofold increase in standard deviation, and again there was no evidence of departure from a slope of unity, but $\Delta N_T(0)$ was slightly greater than expected from the average number of dots in the target area.

Our conclusion from these experiments is that more powerful techniques are required to test whether a loss of performance results from mismatch between target area and counting area. Since observer variance by itself accounts for loss efficiency under the present conditions, the reasons for postulating a matching error under similar conditions (Barlow, 1978) have vanished. However one obviously cannot conclude that matching errors never occur under any conditions.

Other factors influencing observer variance

Performance at detecting symmetry in random dot arrays can be very substantially improved by interposing a diffusing screen between the pattern and the eye (Barlow, 1979). The blurred pattern lacks high spatial frequencies, and for reasons that are unknown this improves performance, as it does in other cases (Harmon and Julesz, 1973). We therefore tried the effect of blurring on the discrimination of dot number, with the results shown in Fig. 8. Under these conditions blurring has no beneficial effect, but is deleterious when carried to extremes.

The measurement of $\Delta N_T(0)$ for $N = 25$ was repeated without a diffusing screen, with an increased separation of the hemifields, and without the usual dividing line. To our surprise there was a substantial

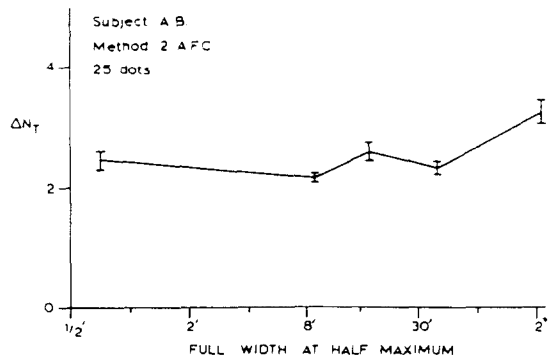

Fig. 8. The effect of blurring dot patterns. The subject viewed the dots through a diffusing screen placed at various distances from the oscilloscope screen. The amount of blurring is expressed as the full-width at half-height of a line on the oscilloscope face seen by the subject through the screen. No significant improvement results, but even very severe blurring does little to impair performance.

decrease of $\Delta N_T(0)$, from 3.8 to 2.5. This suggests that the dividing line and close proximity of the two hemifields were in some way interfering with the numerosity discriminations.

Subitizing and observer variance

Figure 5 shows that observer variance decreases as the number of dots being estimated decreases. Clearly, as the error becomes small compared with unity it should only rarely lead to an incorrect judgement; is this what is happening when the number lies in the range 1–7 and is frequently judged correctly, without error, in a single glance?

Calculations based on the best fitting regression lines of Fig. 5 do not support this notion. If $\sigma'_O = \frac{1}{2}$ one might expect the estimated number to be closer to the real number than any others on 68.3% of occasions, and extrapolating the relation of Fig. 5 suggests that this should not happen until the number lies between 1 and 2; Jevon's data shows that one gets 70% correct even with 6 or 7.

The results of Fig. 9 help to resolve this problem. The cross-hatched rectangles are the result of a two-point experiment in which the subject had to judge from which of two populations (with $\sigma_N = 0$) a sample pattern presented for 200 msec had been drawn. The edges of each rectangle represent the lower and higher numbers, and top and bottom represent the mean value of log $\Delta N_T(0)$ obtained, ± 1 SE The stippled rectangles represent the previous measurements at higher dot numbers on this subject (see Fig. 5) using a self-paced viewing time. Duration of viewing appears to make little difference, but it is clear that the relation between log $\Delta N_T(0)$ and log N becomes steeper when N drops below about 20. The absence of errors, greater subjective confidence, and shorter reaction times led Kaufman *et al.* (1949) to suggest that a different mechanism called "subitizing" is involved with numbers below 7, and the steeper slope may reinforce this. However the interpretation of this result is made difficult by the fact that we do not have any theoretical explanation of the dependence of $\Delta N_T(0)$ on N.

DISCUSSION

Considering the obvious complexity of the perceptual mechanisms called into action when inspecting a field of random dots, there are two surprising aspects of our results. The first is that a single postulated quantity, observer variance, accounts for the errors made, and the second is that, when this is taken into account, the subjects appear to make full and efficient use of all the information available in the display. The validity and generality of these conclusions need to be checked by experiments of the type reported by van Meeteren and Barlow (1980) in which the same set of patterns is presented several times and the lack of consistency of the responses examined: at this point the two methods of estimating observer noise seem to be in approximate agreement.

It will also come as a surprise to some that the two-point method, with its requirement for holding a steady criterion, gives as high performance as the two-alternative forced-choice method, where the simultaneous comparison of two fields is intended to make the task easier. In the two-point method good performance is even maintained when there is no comparison field at all and a steady criterion has to be maintained for the duration of the run.

It does not seem profitable to speculate about the details of the mechanism for judging dot number, but it is safe to say that these judgements must be based on the number of impulses fired by a neurone or groups of neurones, or upon some other measure of the pattern of firing rates among a group of neurones. Observer variance then corresponds to the variability of this measure, converted back into dot-number according to the rules that relate the measure to the number of dots in the field. In terms of this sketchy model our results tell us, first that this variability is the main factor limiting performance, and they also show how variability increases with the number of dots (Figs 5 and 9), how it is decreased by arranging them more regularly (Fig. 7), and how it is little affected by changing the scale (Fig. 6) or by blurring (Fig 8).

We cannot explain why observer variance changes with dot number. The steepening of this relationship below 20 dots could be taken as evidence for a different mechanism ("subitizing") coming into operation and decreasing observer variance in this range, and this does receive support from the different subjective experience: one says to oneself "there were 4, or 6, or 9 dots in that field", as opposed to "more, or less, than usual". But since our model does not explain the effect of dot number on observer variance it may be preferable simply to accept the empirical relationship shown in Fig. 9. Subitizing, on this view, would be the operation of the dot-number estimating mechanism in the range where it gives an unambiguous answer. Other characteristics of subitizing, such as greater subjective certainty and shorter latency, might be secondary results of this absence of ambiguity.

In previous papers (Barlow, 1978, 1980; Van Meeteren and Barlow, 1980) performance at detecting patterns in random dots has been expressed in terms of efficiency, that is the proportion of the statistical sample of information that the subject appears to make use of. In none of those cases were efficiencies above 50% found, and it seemed that there might be some natural limit to the capability of the human system at that figure. In all those cases the arrangements of the dots followed Poisson statistics, so that the variance of the number of dots in any region of the target was equal to the mean number. In the present series this has not been the case and the effects of changing the variance to figures above or below the mean have been explored. The results suggest that efficiency loss is due to intrinsic noise rather than failure to use the full sample of information; thus the apparent limit at 50% efficiency is readily broken if the variance is increased above the mean.

Another way of expressing these results would be to say that the intrinsic variance of the observer is about

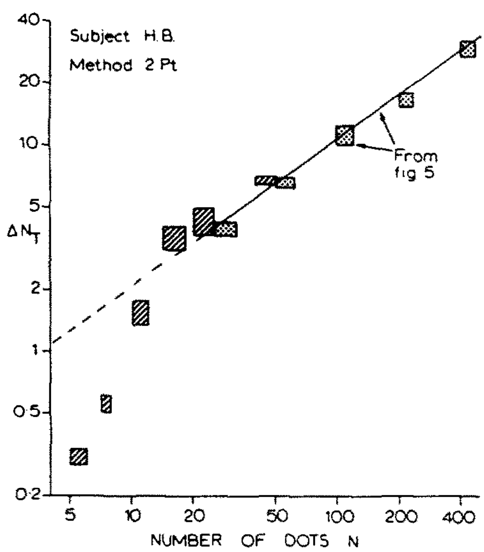

Fig. 9. Threshold increment with small numbers of dots. The cross-hatched rectangles show ΔN_T calculated from the discriminability of displays containing numbers of dots given by the two edges: top and bottom of the rectangles show ± 1 SE. The stippled rectangles are a replot of the results with higher dot-numbers in Fig. 5. The comparison field containing an unvarying number of dots (see Table 1) was omitted for the new results with small numbers, and an exposure duration of 200 msec was used instead of an unlimited viewing time that was usually about 1 sec. The dashed extrapolation of the line fitted to the stippled rectangles lies well above the new points for numbers below 10, possibly because a new mechanism ("subitizing") is used in this range. Note that viewing time makes little difference to performance with 20 to 50 dots.

equal to the extrinsic variance of the dot patterns when these are constructed according to Poisson statistics; when the extrinsic variance is artificially raised above the mean number of dots, it becomes the dominant factor limiting performance, and when the patterns are made more regular, then intrinsic variance becomes dominant. Perhaps the pressure of natural selection has lowered intrinsic observer variance to the point where the mechanism deals adequately with the estimation of the number of truly randomly disposed objects, but has not lowered it beyond this point. If reducing intrinsic variance is costly in terms of added nerve cells or increasing their size or separation, then reducing it to the point where it is about equal to N would be an economic strategy in a world dominated by Poisson statistics.

Our experiments are consistent with the view that the observer can correctly match his inspection area with the signal area even though it is not delimited, but as explained already this should be regarded as a technical failure, not as proof that there is no matching error. There must be conditions under which the mechanism loses efficiency because of such a mismatch, and subsequent experiments have confirmed this (Burgess *et al.*, 1981; Watson *et al.*, 1983).

Acknowledgements—We are indebted to the Nuffield Foundation for the travel grant that made Dr Burgess's visit possible.

REFERENCES

Barlow H. B. (1956) Retinal noise and absolute threshold. *J. opt. Soc. Am.* **46**, 634–639.

Barlow H. B. (1957) Increment thresholds at low intensities considered as signal noise discriminations. *J. Physiol.* **136**, 469–488.

Barlow H. B. (1962) A method of determining the overall quantum efficiency of visual discriminations. *J. Physiol.* **160**, 155–168.

Barlow H. B. (1977) Retinal and central factors in human vision limited by noise. In *Vertebrate Photoreception* (Edited by Barlow H. B. and Fatt P.), Chap. 19 pp. 337–358. Academic Press, New York.

Barlow H. B. (1978) The efficiency of detecting changes of density in random dot patterns. *Vision Res.* **18**, 637–650.

Barlow H. B. (1980) The absolute efficiency of perceptual decisions. *Phil. Trans. R. Soc. Lond. B* **290**, 71–82.

Barlow H. B. and Reeves B. C. (1979) The versatility and absolute efficiency of detecting mirror symmetry in random dot displays. *Vision Res.* **19**, 783–793.

Burgess A. E., Wagner R. F., Jennings R. J. and Barlow H. B. (1981) Efficiency of human visual signal discrimination. *Science* **214**, 93–94.

Finney D. J. (1947) *Probit Analysis*. Cambridge Univ. Press, Cambridge.

Fisher R. A. (1925) *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh.

Harmon L. D. and Julesz B. (1973) Masking in visual recognition: effects of two-dimensional filtered noise. *Science* **180**, 1194–1197.

Hunter W. S. and Sigler M. (1940) The span of visual discrimination as a function of time and intensity of stimulation. *J. exp. Psychol.* **26**, 160–179.

Jevons W. S. (1871) The power of numerical discrimination. *Nature* **3**, 363–372.

Kaufman E. L., Lord M. W., Reese T. W. and Volkmann J. (1949) The discrimination of visual number. *Am. J. Physiol.* **62**, 498–525.

Rose A. (1948) The sensitivity performance of the human eye on an absolute scale. *J. opt. Soc. Am.* **38**, 196–208.

Rose A. (1973) *Vision Human and Electronic*. Plenum Press, New York.

Tanner W. P. Jr. and Birdsall T. G. (1958) Definitions of d' and η as psychophysical measures. *J. accoust. Soc. Am.* **30**, 922–928.

Taves E. H. (1941) Two mechanisms for the detection of visual numerousness. *Arch. Psychol.* **37**, 1–47.

Van Meeteren A. and Barlow H. B. (1980) The efficiency of detecting sinusoidal modulations of dot density in random figures. *Vision Research* **21**, 765–777.

Watson A. B., Barlow H. B. and Robson J. G. (1983) What does the eye see best? *Nature* **302**, 419–422.

Woodworth R. S. and Schlosberg H. (1954) *Experimental Psychology* 3rd edn. Methuen, London.