J. Physiol. (1969), **202**, *pp.* 403–419
With 1 *text-figure*
Printed in Great Britain

THE VARIATIONS IN VISUAL THRESHOLD MEASUREMENT

By P. E. HALLETT
From the Department of Physiology, University of Alberta, Edmonton, Alberta, Canada

(*Received* 27 *November* 1968)

SUMMARY

1. This paper presents a statistical analysis of the thresholds of test flashes viewed against zero, steady or transient backgrounds by peripheral vision. The method of threshold measurement is that of the Medical Research Council Report (Pirenne, Marriott & O'Doherty, 1957; Hartline & McDonald, 1943). The data are from the previous paper (Hallett, 1969*b*) and consist of k days' samples of n threshold measurements on an intensity scale of interval 0·087 log. The collection of a sample required 5–6 min and the samples were spread over a 3-month period.

2. The analysis suggests that the nature of biological variations is that the 'instantaneous' frequency-of-seeing curve can occupy a variety of positions, or levels, scattered about its typical position on the log. energy axis. Change of the position of the curve for a given threshold task is most obvious when one compares threshold measurements obtained on different days, but this is not true day-to-day variation; the same sorts of change occur on the same day if the viewing conditions (independent variables) are changed and are perhaps due to shifts in the observer's signal/noise criterion K.

3. Two important consequences of the analysis are (i) the errors of visual threshold functions are worse than one method of calculating error suggests and (ii) it is possible to eliminate part of the biological variations from a particular sort of frequency-of-seeing curve and thus obtain a better estimate of the instantaneous curve which is the physiological function of prime interest.

4. Some possible causes of the biological variations are considered. The design of experiments is discussed. The method of the M.R.C. Report is to be recommended since it can be applied without prior assumptions about the value of the mean threshold or the nature of the errors.

INTRODUCTION

In 1957 the Medical Research Council published a report by Pirenne, *et al.* on 'Individual differences in night vision efficiency'. This report

contains appendices on experimental techniques, one of which, by Hartline & McDonald, deals with 'The frequency of seeing at low illuminations'* and is annotated by Pirenne & Marriott. For brevity the work of the several authors is referred to below as the M.R.C. Report. The report deals very adequately with the definition and measurement of the mean log threshold and correctly demonstrates the calculation and magnitude of the within-sample estimate of the standard error of the mean, which is apparently the same as the between-sample estimate of the S.E. of mean for repeated samples of a *single* threshold task at a single sitting. The purpose of the present paper is mainly to supplement the Report by deriving estimates of the 'day-to-day' variations in the sample mean of a particular type of threshold task and in the 'task-to-task' variation which occurs in a single sitting as the stimulus parameters are changed in order that the experimenter can construct a function relating mean log threshold to the independent variable. This latter variation, if large enough, leads to a poor-looking experimental curve with considerable point-to-point scatter, but the nature and magnitude of this scatter have not been established.

METHODS

The physiological methods were fully described in Hallett (1969*a*). The rod-isolating technique of Aguilar & Stiles (1954) was used, i.e. a 635 nm background entering the centre of the dilated pupil of the left eye and a 530 nm test flash entering at the nasal edge. The test and background were centred 18 degrees nasally to the fixation point. The test subtended 12 min of arc and was usually 1·5 msec duration. The background was 18 degrees subtense and was absent, steady or flashed for either 1·5 or 500 msec.

An experimental error which might be the source of 'day-to-day' variation in threshold was the variable or occasional vignetting of the test focus by the nasal edge of the iris. This was excluded as follows. Each day the alignment of the head in the apparatus was checked by telescope and by direct observation to ensure that the test focus (a 3×1 mm vertical rectangle) was 2·5 mm nasal to the centre of the day light pupil of the left eye. The iris was then paralysed by cyclopentolate hydrochloride when it shrank to an iris of 1 mm or less surrounding a pupil of about 9 mm diameter. When the observer fixates, the test beam should not be visible on the nasal edge of the iris remnant. If the pupil is 9 mm diameter, the centre of rotation of the eyeball is 9·9 mm behind the plane of the iris, and if the test appears dim when about half of the test focus is occluded by the nasal iris then construction shows that the observer should be able to rotate his eye 10 degrees temporally before he notices dimming of the test. On the other hand scarcely any temporal rotation is possible if the test focus happens to be close to the nasal iris remnant. This check was performed each day during the course of an experiment. The telescope procedures were sometimes repeated at the end of the day but there was never any reason to assume that vignetting occurred.

Photometer readings for the apparatus beams were taken twice daily, and for the standard lamp once every other day.

Sampling procedures

In a single *trial* the dark-adapted observer fixated the small red fixation point and triggered the apparatus when he was ready. The observer reported whether or not (**+** or **0**) the

* Also presented as a report to the United States Committee on Aviation Medicine (1943).

test was visible against the prevailing pattern of background lighting and his response was recorded. On the basis of several trials at several intensities the mean log threshold, $\bar{x}_i$, could be estimated as described below. The background was then changed, a new threshold, $\bar{x}_{i+1}$, established and so on, until the whole range of interest of the independent variable was covered. The identical set of independent variable values was studied at two further days' sittings and about one half of all experiments were run with the independent variable increasing throughout the experiment, and the other half with it decreasing, in order to reduce the possibility of within-experiment trend. It is important that the observer detect the flash on the basis of its light energy alone and not from other clues. The flash intensities were therefore randomly ordered and a proportion were, in fact, zero intensity in 0·13 of trials. Repetitions of an experiment were genuinely independent: earlier results were not at hand, the experimenters were usually rotated. Even if the same experimenter did repeat an experiment it was scarcely possible that he remembered the many different apparatus settings.

The experimenter discovered the rough threshold position by presenting the test flash over a range of 3 or so log units of intensity. He then employed a sufficient number of flash intensities, usually 10, which increase in steps of $\Delta(\log I) = 0{\cdot}087$, so that the range of uncertain seeing was convincingly covered. For each trial an intensity was randomly selected without replacement so that after a dozen or so trials, including one or two blanks, the observer's responses when ordered by increasing flash intensity were a *series* of the form (say): blanks (**0, 0**), flashes (**0000+/0+++++**). The vertical line in this example was placed so as to leave as many zeros to the right as there were plusses to the left. The intensity corresponding to the line was a *threshold measurement*, x_i, and was an estimate of the mean of the sigmoid frequency-of-seeing curve, f *v* $\log I$, since x_i satisfies

$$\int_{-\infty}^{x_i} f.\mathrm{d}(\log I) = \int_{x_i}^{+\infty} (1-f)\mathrm{d}(\log I). \qquad (1a).$$

This method of calculation and definition was different from that given in the M.R.C Report but the mean was exactly the same in each case since the geometric representation of (1*a*) gives an x_i which also satisfied the geometric representation of

$$x_i = \log I_H - \int_{-\infty}^{\log IH} f.\mathrm{d}(\log I), \qquad (1b)$$

(where I_H is the highest flash intensity used). Equation (1*b*) is effectively the definition of the M.R.C. Report. The value of f which corresponds to the mean is determined by the shape of the f-seeing curve. For the absolute threshold experiments, f is close to 0·50.

For a single set of conditions $n = 5$ series were collected on a single day in 5–6 min and the mean of the five measurements, x_i, determined. About sixteen–twenty-four sets of conditions were investigated in an experiment of 3 or so hours of observation. The observers worked very hard and very fast—about 1 trial every 6 sec—and rested briefly between each set of five series. There was no evidence of fatigue and the quantum efficiency at the absolute threshold was high.

The sampling procedure above, although different in method of calculation, was nearly identical with that of Appendix C of the M.R.C. Report. Its utility was demonstrated by the fact that only about 1 sample in 500 needed to be rejected for any reason and the method was suited to routine work by semi-skilled technicians.

The viewing conditions

The appearance of a peripherally viewed 1·5 msec background flash is very difficult to describe. The background brightens very rapidly and when fully developed is not completely homogeneous; some observers reported a 2 deg pale yellow spot and some ill-defined tracery. As the sensation died away it seemed to persist longest in the central point. If the test flash preceded the background flash slightly it was seen on the brightening phase

of the background and was then regarded by most observers as being very difficult to see. If the test were coincident with a bright background flash it might be seen on the dying phase of the background sensation and might then be confused with the central remnant of the background. All things considered the set of viewing conditions in a typical experiment was very varied and the errors in the various thresholds are therefore of practical interest to the applied scientist.

Definitions

+, 0 the observer's responses, 'seen' or 'not seen'.

$\Delta(\log I)$ the intensity interval characterizing the 10 or so flash intensities used in determining a threshold.

a series the responses to the various test flash intensities sampled without replacement, and arranged by increasing intensity, i.e. a frequency-of-seeing curve based on one presentation at each intensity.

a sample a set of five series or 5 x_i.

x_i the mean of single series defined by equation (1).

i suffix for a particular value of a statistic for a given type of threshold.

n the number of series or x_i used in finding the daily sample mean $\bar{x}_i$.

$\bar{x}_i$ the daily sample mean which estimates that day's mean, μ with S.E. of mean $= \sigma_f n^{-0.5}$.

σ_f the standard error of the mean x_i for stated $\Delta(\log I)$ obtained either from the variance of the x_i or (see M.R.C. Report) from $f(\log I)$ and $\Delta(\log I)$. If a series is based on i intensities in steps of $\Delta(\log I)$, and if it is assumed that the responses at each intensity are independent events of constant probability, then, from equation (1*b*), σ_f corrected for bias is

$$\sigma_f = \left(\sum_i f_i(1-f_i)\right)^{0.5} \Delta(\log I) \left(\frac{n}{n-1}\right)^{0.5} \qquad (2a)$$

σ_f is a 'within sample' estimate of error in x_i.

$f(\log I)$ observed frequency of seeing on the hypothesis that the frequency-of-seeing curve changes only its mean position μ from sample to sample. The f appropriate to any given $\log I$ can then be approximated by displacing the samples of n series to a common mean and accumulating the +. A correction for bias is necessary.

c' the value of the best fitting Poisson parameter to the experimental f-seeing curve.

p theoretical frequency of seeing from equation (3) and for probability generally.

q $= 1-p$.

$\sigma_{\bar{x}_i}$ the 'between sample' estimate, for stated n and $\Delta(\log I)$, of the standard error of the daily sample mean, obtained from the variance of the $\bar{x}_i$.

$h(\mu)$ the probability density that the true mean assumes a particular value, μ, in a given sample.

σ_μ $= (\sigma^2_{\bar{x}_i} - \sigma^2_f n^{-1})^{0.5}$, used as the S.D. of the distribution of μ in calculating $g(\log I)$ from the convolution of $f(\log I)$ and $h(\mu)$.

m the mean of the distribution of μ.

$g(\log I)$ the observed frequency of seeing when the + at any given I are simply accumulated from daily samples *without* correction for the shifts in the position of the frequency-of-seeing curve.

σ_g the S.E. of the mean x_i for stated $\Delta(\log I)$ obtained from the curve $g(\log I)$

$$\sigma_g = \left(\sum_i g_i(1-g_i)\right)^{0.5} \Delta(\log I). \qquad (2b)$$

j suffix denoting the jth value of the independent experimental variable.

k the number of sample means, $\bar{x}_i$, obtained on separate days.

$Q(\chi^2)$ a measure of goodness of fit: probability that the χ^2 criterion for ν degrees of freedom exceeds the observed value. The null hypothesis is accepted if Q is neither too high nor too low.

ν used generally for degrees of freedom.

F a random variable defined by the variance ratio of two samples with ν_1 and ν_2 degrees of freedom.

$\mathscr{E}$ operator for the mean.

Relations

The relations between the various quantities in this paper are not available elsewhere and are of some practical interest. They are easily derived. From (1*b*),

$$x_i = \log I_H + \Delta(\log I)\left(0{\cdot}5 - \sum_i P_i\right) \tag{3}$$

is the mean threshold of a single series (i.e. a frequency-of-seeing curve based on one presentation at each intensity). If the responses at each intensity are independent events of constant probability then from (3)

$$\text{S.E. of } x_i = \Delta(\log I)\left(\sum_i p_i q_i\right)^{0{\cdot}5}$$

$$= \left\{\Delta(\log I)\sum_i p_i q_i \Delta(\log I)\right\}^{0{\cdot}5}. \tag{4}$$

Now for the standard normal distribution

$$\sum_i p_i(1-p_i)\Delta\left(\frac{x}{\sigma}\right)$$

is 0·56419 if the step width is 1 and 0·56418 if the step width is 0·1. Frequency-of-seeing curves are nearly normal sigmoid curves so (4) becomes

$$\text{S.E. of } x_i = \{\Delta(\log I)\,0{\cdot}564\sigma'\}^{0{\cdot}5}, \tag{5a}$$

where σ' is the S.D. appropriate to any near normal curve $p(\log I)$. This result is very useful and it is appropriate to anticipate some of the experimental results given later.
Thus if $p(\log I)$ is a log plot of a Poisson sum of parameter c then

$$\sigma' = \log_{10}\mathrm{e}.(\text{trigamma}\,\{c-1\})^{0{\cdot}5} \approx \log_{10}\mathrm{e}/(c-0{\cdot}5)^{0{\cdot}5}. \tag{5b}$$

It remains to substitute (5*b*) in (5*a*):

$$\sigma_g = \{0{\cdot}087 \times 0{\cdot}564 \times 0{\cdot}434/(5-0{\cdot}5)\}^{0{\cdot}5} = 0{\cdot}100 \quad \text{if} \quad c_g' = 5,$$

and

$$\sigma_f = \{0{\cdot}087 + (\tfrac{5}{4})^{0{\cdot}5} \times 0{\cdot}564 \times 0{\cdot}434/(13-0{\cdot}5)^{0{\cdot}5}\}^{0{\cdot}5} = 0{\cdot}082 \quad \text{if} \quad c_f' = 13,$$

which values are very close to those given in the text.

Suppose that the position μ of the curve $f(\log I)$ slides along the $\log I$ axis with probability density $h(\mu)$ and S.D. σ_μ generating the curve $g(\log I)$ by the convolution of $f(\log I)$ and $h(\mu)$, then σ' is now nearly

$$\sigma' = (\sigma_\mu^2 + \sigma_f^2)^{0{\cdot}5} = \left(\sigma_{\bar{x}_i}^2 + \frac{n-1}{n}\,\sigma_f^2\right)^{0{\cdot}5}$$

and if this is inserted in (5*a*) σ_g is obtained in terms of $\sigma_{\bar{x}_i}$ and σ_f. Using $\sigma_{\bar{x}_i}$ and σ_f from Table 1, σ_g is found by this method to be 0·107 (D.B.), 0·014 (B.S.) and 0·083 (M.G.) in good agreement with the tabulated values.

It is appropriate to mention here that σ_f can be calculated from the range of the 5 x_i in a sample or from the r.m.s. deviation of the x_i, or from the f-seeing curve $f(\log I)$. The first and second methods cannot be expected to agree exactly for sampling reasons. If the responses at each intensity are not independent events of constant probability then the second and third methods may disagree.

RESULTS

Absolute thresholds

Between sample variance. Table 1 presents a statistical analysis of the absolute thresholds of the three observers. The frequency distribution of the sixty-two daily sample means, $\bar{x}_i$, was found to be acceptably normal, $Q(\chi^2) = 0{\cdot}12$. This is confirmed in Table 1 where it can be seen that the standard error of the daily sample mean, $\sigma_{\bar{x}_i}$, is much the same whether it be estimated from the squared deviations of the sample means, $\bar{x}_i$, or,

TABLE 1. Absolute thresholds

Observer...	D.B.	B.S.	M.G.	Pooled data
Mean of sample means m	2·00 log 100 $h\nu$	2·04 log 110 $h\nu$	1·92 log 83 $h\nu$	1·96 log 90 $h\nu$
Number of samples (k)	20	21	21	62
$\sigma_{\bar{x}i}$ ($\nu = k-1$) from r.m.s. of k sample means	0·219 log	0·205 log	0·119 log	0·186 log
$\sigma_{\bar{x}_i}$ from range of k sample means	0·233 log	0·209 log	0·124 log	0·194 log
σ_f ($\nu = 4k$) r.m.s. of measurements about sample mean	0·088 log	0·095 log	0·074 log	0·086 log
σ_f from mean range of k sets of five measurements	0·084 log	0·087 log	0·073 log	0·081 log
σ_f from f-seeing curve f	0·081 log	0·089 log	0·082 log	0·084 log
σ_g from f-seeing curve g	0·108 log	0·105 log	0·085 log	0·100 log
$\mathbf{F}(\nu_1 = k-1, \nu_2 = k(n-1))$	31***	23***	13***	—
$\mathbf{M}/C$, ($\nu = k-1$)	22	32*	19	—

Significance levels: *5–2·5 %, ** 2·5–1 %, *** 1 % or less (Pearson & Hartley, 1962).

assuming a normal distribution of $\bar{x}_i$, from the range of the $\bar{x}_i$. The estimates of $\sigma_{\bar{x}_i}$ differ for the 3 observers; its magnitude is 0·186 log, it being clear that within sample S.E. of the mean is 0·038 log (sample size, n, is 5 and the intensity interval, $\Delta(\log I)$, is 0·087).

Tests show that the present estimates of sample to sample variation are likely larger than those derived by Solandt & Best (1943) or those that can be derived from the data of Hecht, Shlaer & Pirenne (1942). Solandt & Best found that the between sample standard deviation for 6–8 large samples for each of fifty-two observers was 0·06–0·17 log, which would include the present observer M.G. but not D.B. or B.S. In this paper various estimates of variation are compared, using the same observers and a fixed technique so that it does not much matter if, say, the present estimates of $\sigma_{\bar{x}_i}$ are higher than those for the typical average observer.

Within sample variance. The 310 threshold measurements, x_i, were found to be acceptably normally distributed about the sample mean, $Q(\chi^2) = 0{\cdot}69$. Estimates of the within sample standard deviation, σ_f, based on the squared deviations of the x_i, corrected for bias, and on the range of the x_i are shown in Table 1. The agreement confirms the normality of the distributions of the x_i. The magnitude of σ_f is 0·086 log, for a step width $\Delta(\log I) = 0{\cdot}087$. (For other step widths σ_f should be multiplied by the root of the ratio of the new to the old step widths).

σ_f can also be calculated directly from the probabilities of the frequency-of-seeing curve, f (Fig. 1, top), using the formula (2a) which is also given in the M.R.C. Report. The value is 0·084 log.

These values are in fair agreement with a value of 0·091 log which can be calculated from the data of the M.R.C. Report for a 3 deg subtense 0·2 sec duration flash at 9° eccentricity from the fovea.

Analysis of variance. If the daily samples are derived from a single population of normally distributed measurements then $\sigma_{\bar{x}_i}$, should equal $\sigma_f n^{-0.5} = 0{\cdot}038$ log. This is grossly not the case and the values of Fisher's $\mathbf{F} = n\sigma^2_{\bar{x}_i}/\sigma^2_f$, are very large. The absolute threshold can therefore be subject to very definite biological variations in mean. Does the standard error, σ_f, and hence the shape of the f-seeing curve $f(\log I)$, also vary from sample to sample?

Heterogeneity of sample variance. There are reasons for supposing that σ_f is heterogeneous but it is rather difficult to show clearly that this is the case. Three tests have been applied.

(i) The within sample variance and range can only assume discrete values because the measurement scale is discrete. Now Bartlett's $\mathbf{M}/C$ (Pearson & Hartley, 1962) is infinite if any sample variance is zero. The test has been applied by altering one value of variance zero (for D.B.) to the smallest value appropriate to a sample range of $\frac{1}{2}\Delta(\log I)$. It should be noted that the test is sensitive to any non-normality in the distribution of x_i (which has been excluded already). The value of Bartlett's $\mathbf{M}/C$ for observer B.S. (Table 1) is marginal evidence for day to day variation of σ_f.

(ii) If σ_f does not change from day to day it should not be significantly correlated with the daily sample mean, $\bar{x}_i$, which does vary. The product–moment correlation coefficient was significantly different from zero by the usual tests *only* for observer D.B. (+0·5) but the precise significance level is in some doubt as the σ_{f_i} cannot be assumed to be exactly normally distributed.

(iii) The sample means, $\bar{x}_i$, and the corresponding $\sigma^2_{f_i}$ and blanks (seen/given) for each sample, can be ranked by sample mean for each observer. It is then easy to see that the few samples with blanks seen are not those samples for which $\bar{x}_i$ was very high or very low. There is weak evidence that the sample variance $\sigma^2_{f_i}$ is large when the sample mean $\bar{x}_i$ is high or low: the $\sigma^2_{f_i}$ corresponding to the four highest $\bar{x}_i$ of each observer were pooled, and this procedure was repeated for the four lowest $\bar{x}_i$ and the four median $\bar{x}_i$; the variance ratios of the $\sigma^2_{f_i}$ of the extreme sets to the median set were 1·81 ('high to median') and 1·55 ('low to median') which are significantly high at the *ca.* 2% and *ca.* 5% levels respectively.

These tests (especially (iii)) demonstrate that the present samples are fairly 'reliable' in the sense that there is no evidence that performance deteriorates markedly when the mean sample threshold $\bar{x}_i$ is high or low. The quantum efficiencies of the present observers are considered in the next paper but it is important to point out that the lowest sample threshold (for observer D.B.) is about 35 quanta (507 nm) cornea. If this is the case either the fraction of light observed by the rods must sometimes be considerably greater than the limits set by Rushton (1956), which can scarcely be true, or the frequency-of-seeing curve must be shallower than

usual when the threshold is low: a Poisson parameter c_i' of 4 would do and the corresponding $\sigma^2_{f_i}$ would then be *ca.* 1·8 greater than the usual value, which compares well with the variance ratios given above in (iii).

Stability of the apparatus. The standard errors of the daily photometer readings when the photometer is illuminated by the apparatus beam or substandard lamp is 0·02 log. The daily error in shutter duration is probably no larger than $\pm$ 0·04 log. Thus the unexplained component of variance in $\sigma_{\bar{x}_i}$ is at least $(0{\cdot}186^2 - 0{\cdot}086^2/5 - 2 \times 0{\cdot}02^2 - 0{\cdot}04^2)^{0{\cdot}5} = 0{\cdot}173$ log. Clearly sampling errors in the x_i and daily uncertainties in the apparatus do not account for the size of the between sample variations in threshold.

It is concluded that the absolute threshold for healthy, well paid observers can show definite variations in the mean log threshold μ. There is some evidence (from σ_f) that the spread of the frequency-of-seeing curve varies from sample to sample but for most purposes (*v.i.*)it is reasonable to assume that the nature of the biological variations is such that the *f*-seeing curve $f(\log I)$ slides to and fro along the log intensity axis with little change in shape, as is suggested in the M.R.C. Report.

The next two sections demonstrate the construction of *f*-seeing curves and the effects of biological variation in the true mean log threshold μ.

The average short-term frequency-of-seeing curve, $f(\log I)$. It will be recalled that $n = 5$ series are determined each day for the absolute threshold task, i.e. each flash intensity is presented five times. This is obviously far too few to construct a frequency-of-seeing curve but if it is accepted that the frequency-of-seeing curve does not change its shape each day, but only its position, then the curve can be built up from the sixty-two daily samples from the three observers if the samples are each displaced so as to eliminate variation in the daily sample means, $\bar{x}_i$. This procedure was used in the M.R.C. Report and for Fig. 1 (top) but a source of bias has been corrected in the latter case by increasing the horizontal spread of the

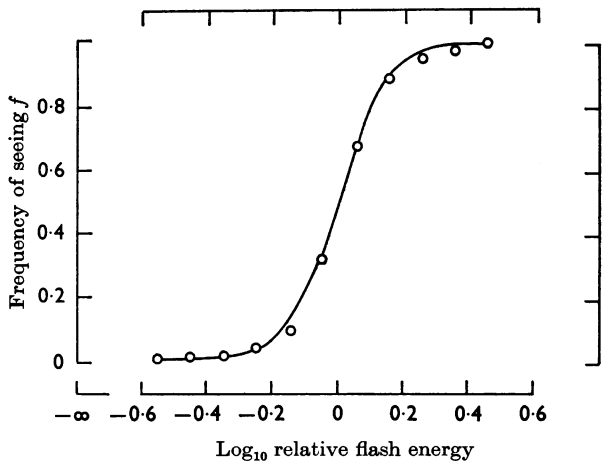
---

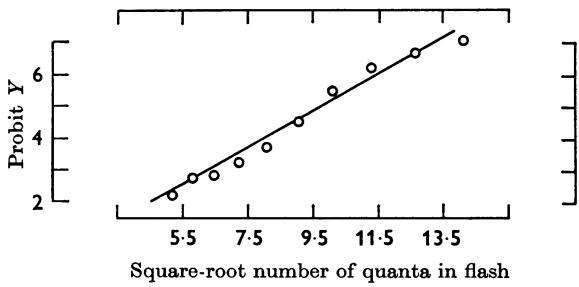
Legend to Fig. 1.

Fig. 1. *Top.* A frequency-of-seeing curve $f(\log I)$ obtained by displacing sixty-two samples of $n = 5$ series to a common mean, accumulating the frequency of seeing f in bins of width 0·087 log and plotting with the abscissa scale expanded by $\{(n/(n-1)\}^{0{\cdot}5}$ to correct for bias. Part of the drift in the mean can be eliminated in this way. The best fitting Poisson sum corresponds to $c' = 13$.

Middle. Barlow's probit transformation of the above curve and the third probit approximation to the regression line. c' is found to be 10·3.

Bottom. A frequency-of-seeing curve $g(\log I)$ which incorporates nearly all sources of biological variation. This is obtained from the same sixty-two samples of five series as the top curve but without any displacement (save that necessary to eliminate observer variation) or correction for bias. The same curve is given by a convolution procedure (see text) or by the Poisson sum for $c' = 5$.





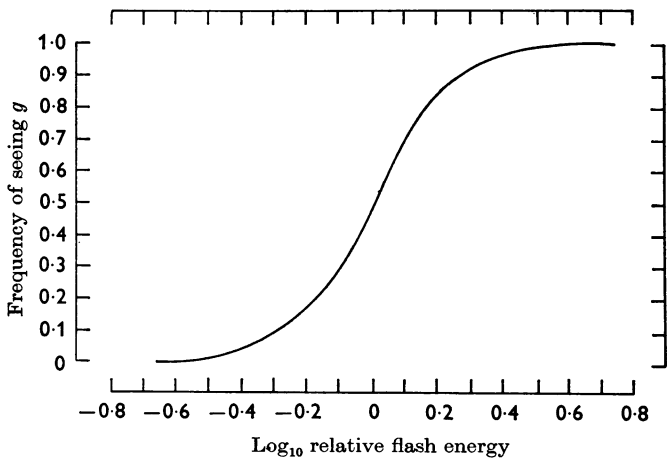


Fig. 1. For legend see opposite page.

points by $\{n/(n-1)\}^{0\cdot5} = 1\cdot12$, otherwise the variance would be underestimated. The curve in Fig. 1 (top) is the Poisson sum,

$$p = \sum_{y=c'}^{\infty} e^{-a} a^y / y!, \tag{6}$$

for $c' = 13$, positioned so that $p = 0\cdot50$ corresponds to log 0 on the abscissa.

The fit over the central range of intensities is acceptable, $Q(\chi^2) = 0\cdot40$, and predicts 161 plusses below log 0 and 143 zeros above. The inequality is trivial, a 0·005 log shift of the curve to lower intensities equalizes the + below and **0** above at the observed value of 152. The fit to the lower tail is not particularly good and would be worse if the tails were correctly estimated. This sort of discrepancy is to be expected for pooled data from an observer population which is heterogeneous for c' or if the observer utters the wrong word (e.g. + instead of **0**) in about 2% of trials. A precise fit to f by p from equation (7) is not to be expected in any case since the contribution of the hypothetical dark light of the eye to $f(\log I)$ has been ignored and the observed parameter c' need not be an integer if either the true c' or the mean threshold varies from trial to trial.

A further indication of the goodness of fit is that σ_f estimated from the observed curve f, using equation (2*a*), is 0·084 log, as mentioned above, and that that from the fitted curve for c' 13 is 0·082 log.

The value $c' = 13$ may be compared with the values of $c' = 5$–7 obtained by Hecht, Shlaer & Pirenne (1942) and of 6–8 by Baumgardt (1960). A difference does exist but is not very large if one considers sampling errors or the errors in the relative intensity axis. A fair estimate of the 0·95 confidence limits for each p_i of an f-seeing curve is $p_i \pm 2(p_i q_i/n)$, where $p_i = 1 - q_i$, assuming that the trials are independent and that the binomial distribution applies. If such limits are attached to a curve with $c = 6$, $n = 50$ (which is a fair approximation to the curves of Hecht *et al.* (1942) then these limits are found to just include the limits appropriate to the curve for $c = 13$, $n = 310$ (which approximates Fig. 1, top), displaced to the same mean. The steepness of frequency-of-seeing curves is considered in more detail in the next paper (Hallett, 1969*d*).

The long-term frequency-of-seeing curve, g. Figure 1 (bottom) shows a shallower frequency-of-seeing curve, $g(\log I)$, which includes the effects of sample to sample variation, and is consequently richer in biological variation than the short-term curve $f(\log I)$. The + at a given intensity for an observer have simply been accumulated *without* displacement of the sixty-two samples to a common sample mean and then the results of the individual observers have been displaced to eliminate the small differences between the observers' overall means. The number of arithmetical operations in Fig. 1 (top and bottom) is rather large and the following check calculation is of interest.

Suppose the f-seeing curve, $f(\log I)$ (Fig. 1, top) slides to and fro horizontally, the mean position μ assuming various values with probability, $h(\mu)$, appropriate to a normal distribution with standard deviation $\sigma_\mu = (\sigma_{x_i}^2 - \sigma_f^2/n)^{0\cdot5} = 0\cdot182$ log and mean $m = 1\cdot96$ log quanta. Then the f-seeing curve $g(\log I)$ can be calculated in a second way as follows. Relabel the $\log I$ axis of the bell-shaped curve $h(\mu)$ as Δ, where $\Delta = \log I - 1\cdot96$, so that $h(\Delta = 0)$ is the peak of the bell-shaped curve. Relabel the $\log I$ axis of the sigmoid curve

$f(\log I)$ as t, where $t = \log I - 1{\cdot}96$, so that $f(t = 0) = 0{\cdot}50$. Then the f-seeing curve, $g(\Delta)$, is given by

$$g(\Delta) = \int_{-\infty}^{+\infty} f(t)h(\Delta - t)\,\mathrm{d}t \qquad (7)$$

Equation (4) was evaluated, using discrete distributions for f and h, with Δ and t in steps of *ca.* $\frac{1}{2}\Delta\log I$. The agreement with the + accumulating procedure was good.

The curve, g, is well fitted by a Poisson sum of parameter, $c' = 5$, displaced so that the $p = 0{\cdot}50$ point corresponds to the mean log threshold $-Q(\chi^2) = 0{\cdot}69$. The observed curve g yields about 262 + below and **0** above the mean. The corresponding values for the fitted curve $c' = 5$ are 286 + below and 283 **0** above. The *over-all* S.E., σ_g, in a mean x_i, from equation (2*b*) above is 0·100 log from either the observed curve g or the fitted curve $c = 5$. Clearly the fit is good.

The effect of sample to sample variation, or drift in the true mean threshold μ, is thus to give a f-seeing curve, $g(\log I)$, of lesser slope than the curve, $f(\log I)$. It should be noted that the curve $g(\log I)$ and the corresponding 'sampling error' σ_g do *not* give a reliable estimate of the accuracy of the over-all mean of k days' samples of size n. The drift in the mean of the short-term f-seeing curve must cause the probability of response at a given flash intensity to vary in the long run so that equation (2*b*) no longer gives the true sampling error. σ_g is useful, however, for identifying the parameter c_g' of the Poisson sum which approximates the curve g.

Other viewing conditions

Considered here are the thresholds for a test presented against a steady background or at various times with respect to the beginnings of 1·5 msec or 500 msec square wave backgrounds (data of Fig. 1, Hallett, 1969*b*). The number of samples of size $n = 5$ for any particular stimulus arrangement is $k = 3$ for each of the three observers, compared with $k = 20$ or 21 in the case of the absolute threshold data. Statistical analysis has been preceded by rather extensive pooling with numerous implicit assumptions. The following discussion has therefore been kept reasonably brief. $\sigma_{\bar{x}_i}$ and σ_f have usually been estimated from mean ranges.

Between sample variance. Variation $\sigma_{\bar{x}_i}$ in the sample means for various viewing conditions is generally of the same magnitude as the pooled absolute threshold estimate of 0·186 (S.E. < 0·012)log.

Now Blakemore & Rushton (1965) and Rushton (1965) have shown that the imaginary light of the eye and real background lights are in some respects additive. A value of the imaginary light for the present three observers is given in Hallett (1969*d*): $-2{\cdot}98\log_{10}$ scotopic trolands. Barlow's (1957) result was equivalent to $-2{\cdot}66\log$ scotopic trolands. These imaginary lights are very weak compared with the real backgrounds used, e.g. -2 to $+2\log$ scotopic trolands. Clearly variations in mean μ for various viewing conditions cannot be due to variation in the imaginary light.

Can sample to sample variation in a single observer be simply described by a filter factor hypothesis (e.g. M.R.C. Report)? Suppose that the observer's characteristics are constant but that there is an attenuating filter in front of his eye which changes from sample to sample. Then the observed between sample variation $\sigma_{\bar{x}_i}$ in the threshold of the observer minus the filter combination will be less for increment thresholds against steady or flashed backgrounds than at the absolute threshold, because in the former case attenuation of the test flash will be partly compensated by attenuation of the background. The observed dependence of the threshold on background is as the *ca.* 0·65 power of the background for the present purpose. If the filter optical density follows a normal distribution with S.D. $= \sigma = 0{\cdot}182$ log, and if the S.E. of the threshold measurement, $\bar{x}_i$, is 0·038 log (within sample estimate), then the S.E., $\sigma_{\bar{x}_i}$ of the daily absolute threshold measurements will be $(0{\cdot}182^2 + 0{\cdot}038^2)^{0{\cdot}5} = 0{\cdot}186$ log and that of the increment thresholds will be about

$$\{(1-0{\cdot}65^2)\ 0{\cdot}182^2 + 0{\cdot}038^2\}^{0{\cdot}5} = 0{\cdot}074 \text{log}.$$

But it has been seen already that the observed standard errors are of the same magnitude. The simple filter factor hypothesis is not substantiated, nor is there much point in proposing as alternative hypotheses that the imaginary filter is present only in the test beam or that it can attenuate the dark light as well; consideration of a well known simple signal/noise hypothesis (Barlow, 1957) shows that the observed biological variations in mean threshold μ could arise in several ways, e.g. from variations in the extent of integration over space or time or from variations in quantum efficiency or criterion, etc. Of these factors variation in the signal/noise criterion would have the most marked effect.

Within-sample variance. Estimates of σ_f from the average within-sample range are always in good agreement with sums of squares estimates and do not differ much from the absolute threshold data already given.

Within-sample trend. There is no evidence of progressive improvement or fatigue. The trend in the average thresholds for the first, third and fifth series in 120 samples from varied thresholds (Hallett, 1969*a*, *b*) was less than the S.E. of the differences in the means (0·012 log).

The nature of day-to-day variations. It has been seen for a variety of conditions that variation in the true sample mean μ is demonstrated by the very real difference between the within-sample estimate of the S.E. of mean, $\sigma_f n^{-0{\cdot}5}$, and the between-sample estimate, $\sigma_{\bar{x}_i}$. The existence of these variations poses very real problems about the accuracy of visual threshold experiments which have not previously received attention.

Do day-to-day variations really exist, such that the thresholds for all viewing conditions on a particular day are high or low as the case may be (hypothesis *A*)? Or, is the variation observed in daily repetitions of a given

viewing condition really another manifestation of variation which also occurs on going from one viewing condition to another on the same day (hypothesis *B*)?

The quality of a single day's experiment according to hypothesis *A* is such that 0·95 of experimental points can be expected to be between $\pm 2\sigma_f n^{-0.5}$ log of a curve, the position of the curve varying from day to day within the limits of, say, $\pm 2(\sigma_{\bar{x}_i}^2 - \sigma_f^2 n^{-1})^{0.5}$ of its mean. Cursory inspection of the results of Hallett (1969*b*) suggests that the quality is worse than this and might well correspond to hypothesis *B*, which is that 0·95 of the experimental points of any day are within $\pm 2\sigma_{\bar{x}_i}$ of a single fixed curve.

A numerical choice between the two hypotheses is most easily made as follows. The experimental design (Hallett, 1969*b*) was that in an experiment twenty-one to twenty-three viewing conditions were repeated on three consecutive working days by each of the three observers. Calling the sample means on days j (= 1, 2, 3), $\bar{x}_{ji}$, the covariance of the differences between the repetitions is defined here as

$$\text{cov. } (\bar{x}_1 - \bar{x}_2, \bar{x}_1 - \bar{x}_3) = \mathscr{E}(\bar{x}_{1i} - \bar{x}_{2i})(\bar{x}_{1i} - \bar{x}_{3i}) - \mathscr{E}(\bar{x}_{1i} - \bar{x}_{2i}).\mathscr{E}(\bar{x}_{1i} - \bar{x}_{3i}),$$

where $\mathscr{E}$ is the expectation taken over the $i = 1, 2, \ldots,$ 21 to 23 various viewing conditions. On hypothesis $\bar{x}_{ji} = a_i + b_j + c_{ji}$, where a_i is the population mean for the ith viewing conditions, b_j is the jth day's deviation from this value and c_{ij} is the deviation (mean 0, variance $\sigma_f^2 n^{-1}$) of the sample mean from b_j. Using the rule for the expectation of the products of independent random variables it follows that cov. $(\bar{x}_1 - \bar{x}_2, \bar{x}_1 - \bar{x}_3)$ is $\sigma_f^2 n^{-1}$. On hypothesis *B*, on the other hand, $\bar{x}_{ji} = a_i + c_{ji}$, where c_{ji} has mean 0 and variance $\sigma_{\bar{x}_i}^2$, and cov. $(\bar{x}_1 - \bar{x}_2, \bar{x}_1 - \bar{x}_3)$ is $\sigma_{\bar{x}_i}^2$.

The data of Hallett (1969*b*) yield twenty estimates of the covariance defined above, ranging from $-0{\cdot}007$ to $+(0{\cdot}309)^2$ log^2 units. The means for the three observers are: D.B., $(0{\cdot}195)^2$, B.S., $(0{\cdot}155)^2$, M.G. $(0{\cdot}196)^2$ log^2 units, i.e. $(0{\cdot}183)^2$ overall. These values are very close to the estimates of $\sigma_{\bar{x}_i}^2$ for the absolute threshold and various viewing conditions given above. and certainly very different from $\sigma_f^2/n = (0{\cdot}038)^2$. Clearly hypothesis *B* is at least true on average: the within sample estimate of error is usually a bad estimate of the accuracy of the shape of experimental functions which is more certainly given by $\sigma_{\bar{x}_i}$, the between sample error.

Problems of this sort are more usually approached by an analysis of variance. A $6 \times 21 \times 3$ analysis of variance table,

(*experiment type*) × (*independent variable*) × (*repetition number*),

was constructed from part of the records, for observer D.B. The residual source of variance was $(0{\cdot}166 \text{ log})^2$ and the between repetition source of variance $(0{\cdot}031 \text{ log})^2$, which is scarcely larger than the likely apparatus variation. This provides further support for hypothesis *B*: true day-to-day variation is small, if indeed it exists at all.

This analysis leads to the important conclusion that the log threshold for a given task can assume a variety of normally distributed levels from a population with variance *ca.* $\sigma_\mu^2 = (\sigma_{\bar{x}_i^2} - \sigma_f^2 n^{-1})$. On a given day the accuracy with which the chosen level is both maintained and measured is reflected in the smallness of σ_f, but if a sample is obtained on another day another level may be chosen. Similarly, on a single day if (say) a high ranking level is chosen for the ith viewing condition the level for the $(i+1)$th condition may be quite different. It seems most plausible that these variations in the chosen level are due to the observer's tendency to change his signal/noise criterion (K in Hallett 1969 *a*, *b*,) whenever the viewing conditions change.

DISCUSSION

The results of this paper are, perhaps, of some interest, since they possibly represent the first serious attempt to pin down the physiological meaning which attaches to the observed variations in mean visual thresholds. It must be admitted the between-sample estimate of variation $\sigma_{\bar{x}_i}$ is likely larger than that reported elsewhere (p. 408), but it should be remembered that some variation between individuals is likely, and that the present data were collected routinely and without any special selection (since it proved impossible to establish any useful criterion for rejection). It is also important to keep in mind the time scale of the measurements. A threshold sample by the present technique takes about 5–6 min but the samples were collected over a 3-month period. Both the absolute threshold measurements and the thresholds for more complicated viewing conditions involving transient and steady backgrounds show that the within-sample estimate of variation $\sigma_f n^{-0.5}$, which represents the average *short-term* variation, is small compared with the between-sample estimate $\sigma_{\bar{x}_i}$, which represents the *long-term* variation. This naturally suggests that the true mean log threshold, μ, is liable to drift, and the analysis of the absolute threshold data shows that the relations between the different types of frequency-of-seeing curve (pp. 412, 413) and various S.E.s of the mean (p. 407) are satisfactorily given on the assumption that the frequency-of-seeing curve does not change shape as it drifts along the $\log I$ axis, although some flattening of the curve must occur, at least when the threshold μ swings low (pp. 409, 410). If the measurements had been restricted to the absolute threshold it would have been natural, but wrong, to call these drifts 'day-to-day variation'. As it is analysis of the other viewing conditions shows that the same sort of drifts occur on a single day when the viewing conditions change and give rise to the scatter of the points which constitute the experimental function relating log threshold to the independent

variable. True day-to-day variation can scarcely be said to exist, since it is no larger than the likely daily variations in the energy output of the apparatus (p. 415). The physiological processes which give rise to the drift are uncertain, but these cannot be described in terms of variation of the hypothetical dark light of the eye (p. 413), or in terms of a simple filter factor hypothesis (p. 414). As will be shown in the next paper, the drifts lower the observed 'over-all' quantum efficiency at the absolute threshold from about 0·1, which represents Rushton's (1956) limit, to about 0·04.

Maximum work load. Visual experiments are tedious and an experiment should be planned so that sufficiently accurate results can be obtained in the shortest time. In this respect it is important that an observer's work load be the highest that he can tolerate but it is not clear what this limit is. Most of my observers have worked for the minimum period of 3 months and although various efforts and inducements have raised the number of flashes in an ordinary experiment of 2–3 hr observation from 500, regarded as large by Pirenne & Marriott (1959), to 1400, this has been partly offset by a reduced number of experiments per week per observer. In retrospect for about a dozen observers the number of flashes per observer per week is 2500 ± 500. Performance does not appear to suffer in the more heavily loaded experiments: a test showing the absence of within sample trend has been given; the quantum efficiency at the absolute threshold is high even though half of the samples were obtained at the beginning of an experiment and half at the end; the scatter of the experimental points seems no worse than usual.

Barlow's 2-point probit method. Barlow (1962) has proposed a method of considerable interest for which $\Delta \log I$ is so large that only two test flash intensities are used, chosen to correspond to expected frequencies of seeing of *ca.* 0·95 and 0·05 respectively.

Barlow's method is primarily intended for the rapid determination of quantum efficiency which may be a more fundamental quantity than the mean threshold, although there can be no doubting the practical usefulness of the latter. The advantages of the method are that it can be easily implemented with fully automatic apparatus and that the accuracy is apparently slightly higher than that of the M.R.C. Report for the same number of flashes. It is, however, implicit in Barlow's method that one has reasonable *a priori* knowledge of the mean threshold of the day (if this is not known experimental time will be lost in roughly assessing it) and that the frequency-of-seeing curve is nearly approximated by a Poisson sum. This cannot be confidently assumed for any novel experimental condition or, indeed, for any new observer, and is most likely to be incorrect when the frequency of seeing is in the extreme regions of 0·05 or 0·95. The probit transformation for the present data gives rise to slightly lower values for c' than do other methods (Fig. 1, middle; Table 1, Hallett, 1969*c*) but practically speaking the difference is small.

The estimation of errors. The within sample estimate of error, σ_f, has been derived above both by accumulating the f-seeing curve, $f(\log I)$, and from the squared deviations of the measurements, x_i, from their sample mean. These methods are thorough but laborious and in many cases the use of the range of the x_i is to be recommended. Pearson & Hartley (1962) show that the range of measurements in small samples from a normal continuous population is a very efficient unbiased estimator of the population standard deviation. The standard error in an estimate based on the average range of several samples is also considered by these authors. In this paper a number of examples have been given to illustrate the practical usefulness of the average range of the x_i in samples of size, $n = 5$, from a discrete measurement scale. The S.E. of the sample means, $\sigma_{\bar{x}_i}$, may also be estimated from the range of the $\bar{x}_i$. The various algebraic relations between frequency-of-seeing curves and standard errors which have been given above may also prove of practical use.

General conclusion. The method of the M.R.C. Report is to be recommended for any new investigation since it can be applied with very few *a priori* assumptions about the value of the mean threshold or the error in that mean. The errors in the threshold are apparently such that there is no special advantage in expending many flashes in a comparison of a few different threshold tasks on a single day if the viewing conditions do not change frequently. The major source of error appears to arise from changing the viewing conditions and may represent a change in the observer's criterion. It is therefore better to change the viewing conditions frequently, e.g. by comparing a large number of threshold tasks on the same day, and then to achieve the desired level of accuracy by repeating the whole experiment on several different days and averaging the repetitions. The level of accuracy attained in any experiment may be rapidly estimated in the ways described above and the values checked if necessary against the data of the M.R.C. Report (appendix A, Table 1) or the present paper. Once the S.E. of mean for a given observer falls below 0·10 log a decision must be made as to whether further accuracy is desirable, in which case Barlow's (1962) probit method may be valuable, or whether it is better to repeat the experiment on other observers, in view of the fact that individual variations in threshold are possibly of this magnitude. Unfortunately, despite considerable research, considerable uncertainty attaches to the published values of variations between individuals since these have not been corrected for variation in the sample means of a *given* individual, as Pirenne (1956) has pointed out.

The work reported in this paper was supported by the Medical Research Council of Canada and the Defence Research Board of Canada, grants MRC MA 1981 and DRB 9310 122. I am indebted to comments from F. H. C. Marriott.

REFERENCES

AGUILAR, M. & STILES, W. S. (1954). Saturation of the rod mechanism at high levels of stimulation. *Optica acta* **1**, 59–65.

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 for determining the over-all quantum efficiency of visual discriminations. *J. Physiol.* **160**, 155–168.

BAUMGARDT, E. (1960). Mesure pyrométrique du seuil visuel absolu. *Optica acta* **7**, 305–316.

BLAKEMORE, C. B. & RUSHTON, W. A. H. (1965). Dark adaptation and increment threshold in a rod monochromat. *J. Physiol.* **181**, 612–628.

HALLETT, P. E. (1969*a*). Rod increment thresholds on steady and flashed backgrounds. *J. Physiol.* **202**, 355–377.

HALLETT, P. E. (1969*b*). Impulse functions for human rod vision. *J. Physiol.* **202**, 379–402.

HALLETT, P. E. (1969*d*). Quantum efficiency and false positive rate. *J. Physiol.* **202**, 421–436.

HECHT, S., SHLAER, S. & PIRENNE, M. H. (1942). Energy, quanta and vision. *J. gen. Physiol.* **25**, 819–840.

PEARSON, E. S. & HARTLEY, H. O. (1962). *Biometrika Tables for Statisticians*, vol. 1. Cambridge: Cambridge University Press.

PIRENNE, M. H. (1956). Physiological mechanisms of vision and the quantum nature of light. *Biol. Rev.* **31**, 194–241.

PIRENNE, M. H., MARRIOTT, F. H. C. & O'DOHERTY, E. F. (1957). Individual differences in night-vision efficiency. *Med. Res. Coun. Spec. Rep. Ser.* no. 294. With a section on 'The frequency of seeing at low illumination' by HARTLINE, H. K. & MCDONALD, P. R.

PIRENNE, M. H. & MARRIOTT, F. H. C. (1959). The quantum theory of light and the psychophysiology of vision. *Psychology: A Study of a Science*, ed. KOCH, S., vol. 1, pp. 288–361. New York: McGraw-Hill.

RUSHTON, W. A. H. (1956) The rhodopsin density in the human rods. *J. Physiol.* **134**, 30–46.

RUSHTON, W. A. H. (1965). Visual adaptation. *Proc. R. Soc.* B **162**, 20–46.

SOLANDT, D. Y. & BEST, C. H. (1943). Night vision. *Can. med. Ass. J.* **49**, 17–21.