Brit. J. Pharmacol. (1949), 4, 366.

PERSONAL ERROR IN PENICILLIN ASSAY

BY

PETER A. YOUNG

From the Wellcome Research Laboratories, Beckenham, Kent

(Received October 15, 1949)

In many forms of biological testing the intrinsic errors of the methods used are so great as to render insignificant any manipulative errors on the part of the assayers. Some assays, however, have such small inherent errors as to make the personal factor of considerable importance. This paper will consider the problem as met with in the course of four years' routine penicillin estimations.

METHOD OF ASSAY

The test employed was that described by Pope and Stevens (1945–6). Standard penicillin was dissolved accurately to 5 u./ml. in phosphate buffer *p*H 6.0; 2.0 ml. of this solution were pipetted into 20 ml. of test broth and a range of six volumes delivered from the broth dilution into six tubes, each containing 20 ml. of broth. Three series of six standard tubes were set up for every thirty unknowns.

A typical series of volumes would be:

1.3	1.2	1.1	1.0	0.9	0.8
1.25	1.15	1.05	0.95	0.85	0.75
1.2	1.1	1.0	0.9	0.8	0.7 ml.

The procedure for unknown samples was to dilute each sample in buffer to approximately 5 u./ml., further dilute this 1/11 in broth, and deliver volumes of this dilution into a set of six tubes. The range of volumes employed may of course be varied. In addition to the 10 per cent ranges illustrated above (the percentage indicates the average dilution increment between adjacent tubes) we have commonly used:

20 per cent range	2.0	1.6	1.3	1.1	0.9	0.7 ml.
30 per cent range	2.0	1.5	1.1	0.85	0.65	0.5 ml.

and, less often:

90 per cent range	6.0	2.0	1.4	0.7	0.4	0.2 ml.

After addition of the penicillin to the test broth the latter was inoculated from a suspension of *Staph. aureus* (Oxford strain). Usually all the day's tests were set up with penicillin before beginning inoculation, although this order was reversed for a few months only (see below). Tests were incubated overnight at 37° C. in a controlled temperature room, and were then read visually by moderate indirect illumination. The end-point was taken as the minimum trace of growth detectable by eye, and the volume permitting this amount of growth was estimated to 0.05 ml. and recorded as "end-point volume." With practice it was possible to recognize degrees of growth corresponding to 5, 10, and 15 per cent less penicillin than that producing an end-point. Thus in a range of tubes with a dilution factor of 1.2 (20 per cent range) it was possible to place the end-point to within 5 per cent (approx. 0.05 ml.) should it fall between two adjacent tubes. The computation of results was completely covered by a system of tables which allowed for standard and unknown end-point volumes, and preliminary buffer dilution of the unknown.

RESULTS

The sources of personal error in the technique were in the preparation of dilutions, delivery of volumes, and in the reading of the tests. In skilled hands and with the 10 per cent ranges, the standard deviation of a single test has often been of the order ±3–4 per cent. Such proficiency could only be obtained after much experience in assay, however, and even then the frame of mind of the worker was of great importance if maximum accuracy were to be achieved.

The problem which confronted this laboratory at the time of its inception in November, 1944, was one of training half a dozen assistants in the technique of penicillin assay. Only one had previous laboratory experience (L.B., male); the rest were girls whose ages ranged from 16 to 18, of whom all except one had matriculated and were fresh from school. After only one week of general training, including a few practice assays, it became imperative that some routine tests should be undertaken, although it was realized that these could not be expected to be of great accuracy. The demand for assays at that time was steadily mounting, and in order to make the assayers "accuracy-conscious" some simple form of "scoring" their work was sought. The method adopted was to test each sample submitted for assay four to eight times, the arithmetical mean of the results of these tests being issued as the final answer. Each result was then expressed as a percentage of the mean, and these percentages plotted for the assayer responsible. When the number of tests performed by a worker exceeded forty in one week, only that number taken at random was treated in this way from the week's work. Fig. 1 reproduces this chart over a period of five weeks ; each chart was commenced during the assayer's first week. The improvement in the later weeks is very marked for some of the workers, all of whom entered into the competitive spirit which this comparison so easily fostered. However, it did not seem desirable to continue the charts after the five-week period, as the workers by then had more or less reached their maximum accuracy.

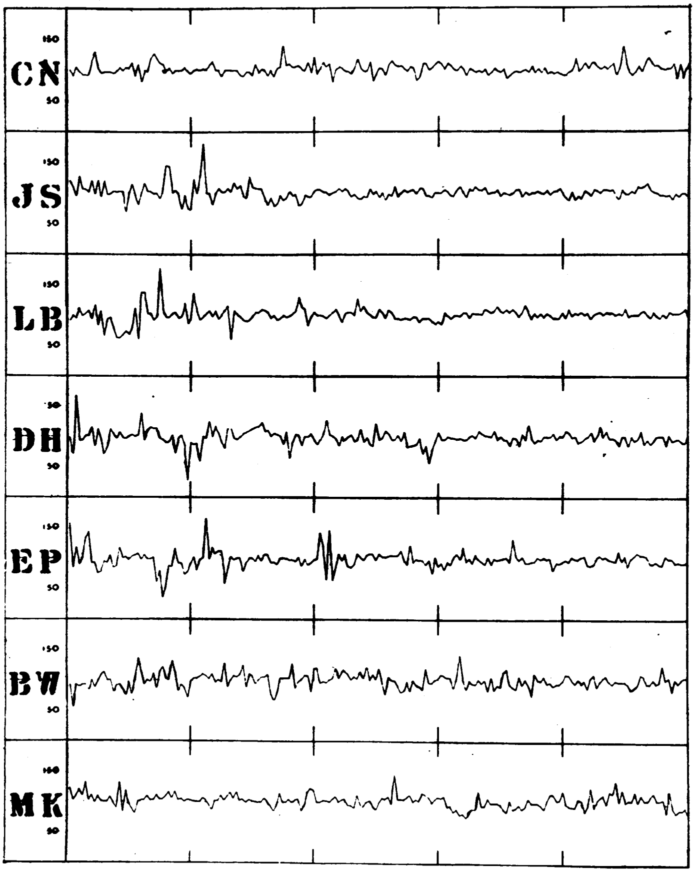

FIG. 1.—Worker deviation chart. Results obtained by each worker plotted as percentages of the means for all workers. Forty results at random taken from each week's tests for each assayer. The horizontal scale represents weeks. Charting commenced for each worker during the first week's work.

The percentages obtained for the preparation of the charts were also used to calculate the variance each week for each worker, and on the basis of the previous week's variances each assayer was assigned a weight in inverse proportion to his or her variance. For simplicity these weights were reduced to simple digits. The computation of the mean for each sample was then made by appropriately weighting each worker's values. This appeared to be particularly valuable when new staff were engaged to perform assays, as only one week's practice testing was required before they could be assigned a weight, and from then on they could enter into the routine assays without fear of their inexperience unduly influencing the results issued from the department. Further experience taught us, however, that week-to-week fluctuations in individual variance were unaccountably high. Consequently it was not valid to assess one week's work in terms of the previous week's variance, and the system of weighting was discontinued.

A study of the mean variance for all workers each week over the period, November, 1944, to March, 1945, revealed some interesting trends (Fig. 2). Initially, the mean variance was very high (S.D. for single tests ±12.5 per cent). On the third week, however, this fell, and by the fourth week had reached quite a satisfactory level (S.D. ±8 per cent). The value remained fairly steady for the next three weeks, but rose again to an alarming extent during the eighth week, only to fall during the succeeding weeks to an even lower figure than previously obtained. The point of interest is that the eighth week was in fact the three working days after Christmas. We appeared, therefore, to be recording the after-effects of festivities which, for five out of the seven staff, fell after only seven weeks' employment. Some of the workers had displayed greater excitement than the others about this time, and these were the ones whose testing demonstrated the greater variation. No such general rise in variance was observed around Christmas, 1945, by

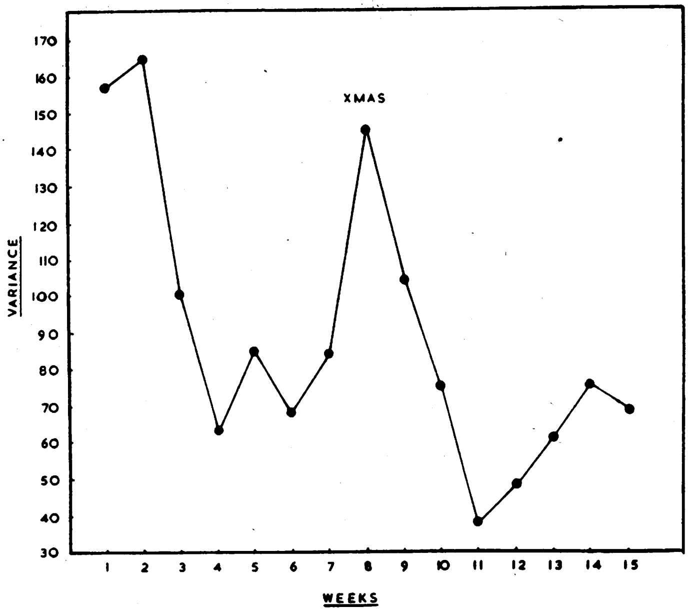

FIG. 2.—Weekly variance; mean values for seven workers. Each estimate plotted is associated with about 200 degrees of freedom.

which time the staff had naturally " settled down " to their employment.

From the eleventh week onwards some deterioration of the accuracy of our tests was noted. This period corresponded with the introduction of samples from the production unit; previously we had been assaying samples derived from small-scale laboratory experiments. The effect of this change was twofold: First, it was necessary to obtain a final answer for the production samples as quickly as possible. Second, whereas the laboratory experiments were conducted with great precision, on accurately assayed material, and the samples submitted for assay could very often have their potency predicted to some extent, the penicillin content of the samples from the large-scale plant was, for some time, quite unpredictable. The effect was to oblige us to test the plant samples on the wider ranges, 20, 30, and even 90 per cent steps being used. The relative accuracies of these ranges will be dealt with later, but it will be clear that the overall effect at the outset was to raise the general variance.

After fourteen weeks' work in the department it became too laborious to calculate individual variances each week. Instead the general variance was obtained by taking a hundred results at random from the week's work, each result being expressed as a percentage of the mean value for the sample from which it was obtained, and the variance of these percentages was calculated. The value so obtained was subject to a bias, particularly when only three or four results were used to obtain the mean value for a sample. Nevertheless, this statistical device seemed suitable as a general guide to the current standard of testing. When new staff came into the laboratory they were allowed to spend their first week in general training with the apparatus used, and thereafter were put on to the testing of samples which had already been assayed by the more competent workers. Their results on these samples were expressed as percentages of the values already assigned and the variance of these percentages determined. Usually after two to three weeks of such " duplicate " testing they were considered reliable enough to enter the routine of the laboratory. The weekly variances for six new arrivals are listed in Table I together with the general mean variance for the established staff over the first five weeks of their work. Fig. 3 shows the mean variance for thirteen assayers for each of their first five weeks of testing. It demonstrates very clearly that accurate testing could not be expected during the first fortnight, after which the variance drops rapidly until, at four to five weeks, the curve markedly flattens out, and, at five weeks, reaches the value of 80 for the variance, which was exactly the average value for the

TABLE I

WEEKLY VARIANCE OF TRAINEES

Each estimate is derived from 40 results drawn from the week's work. The general mean (for the established staff) is the mean of the 5 variances for the respective weeks, each derived from 100 results.

Worker	Week of testing					General mean for weeks 1–5
	1st	2nd	3rd	4th	5th	
RK ...	461	540	327	149	182	103
SG ...	388	396	91	120	—	85
NB ...	228	73	87	58	84	104
KD ...	167	199	111	101	89	111
SR ...	184	179	120	98	48	115
JP ...	110	66	79	117	77	103

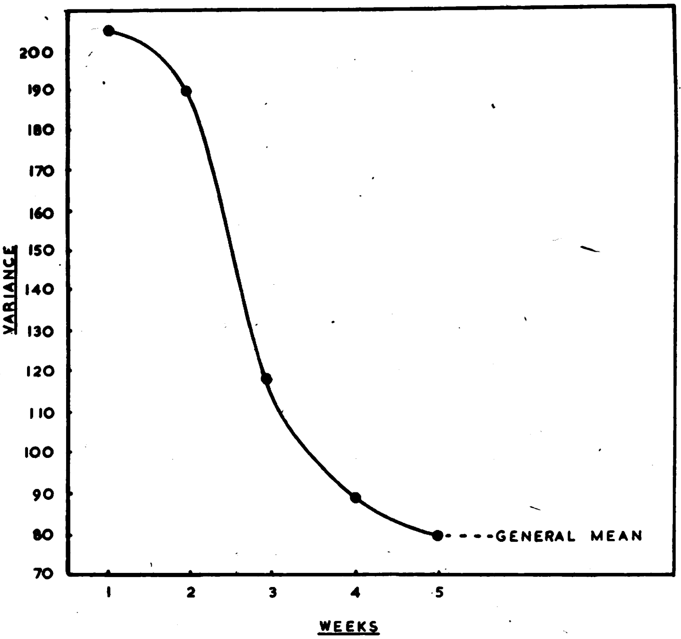

FIG. 3.—Weekly variance during training period; mean values for 13 workers. The general mean is that for the established staff for the 5th week. Each estimate plotted is associated with about 500 degrees of freedom.

general mean variance corresponding to the fifth week.

It should be noted at this stage that the procedure in the laboratory was to record details of samples received in a log book, and to label the samples with a sample number. Test sheets were then prepared showing the sample numbers, the buffer dilution required for each, and the volumes in ml. to be delivered to the test broth. One worker would then prepare the buffer dilutions of the samples and another, the " tester," would be responsible for delivering the test volumes. The latter worker would also be responsible for reading her own tests after overnight incubation. Thus the variance in the results would be shared between the worker preparing the dilutions and the " tester." It has always been felt that the contribution of dilution preparation to assay variance was relatively insignificant, a fact more or less borne out by consideration of the effect of the range used on the variance of results ; this effect is illustrated in Fig. 4, where the variance associated with different levels of titre (u./ml.) is plotted. In general the samples of lower titre were tested on the wider ranges and these tests were inevitably liable to greater variance. The titre range 5–100 u./ml. is given the variance for dried products which were dissolved at concentrations falling within these limits of potency. These solutions were tested on the 20 per cent range. The unexpectedly low figure for the 90 per cent range was brought about by the fact that some of the results taken for its computation were derived from closer range tests, it not being possible from the records available to say exactly which range had been used. The variances of 51 and 85 for the two highest titre groups reflect an increase due to the necessity for serial buffer dilution in the preparation of samples in the higher of the two groups. The increase, however, cannot be directly assessed as the result of two buffer dilutions compared with one in the lower groups because the physical properties of the concentrated penicillin solutions make their measurements far more difficult.

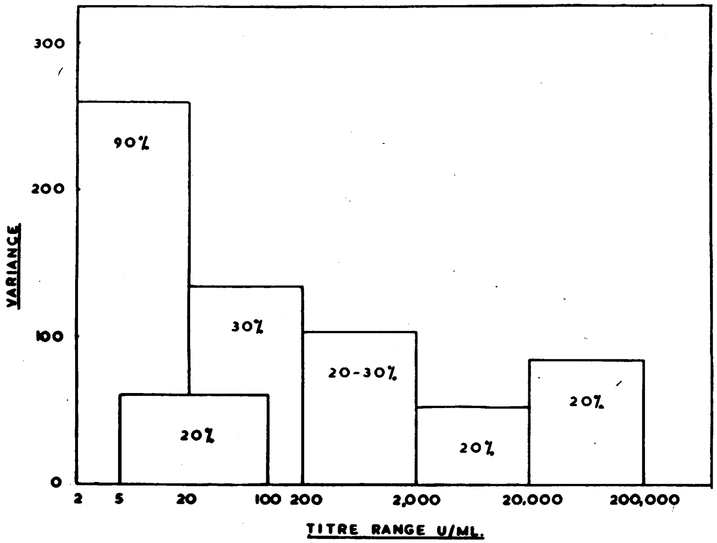

FIG. 4.—Variance related to titre of the sample and range employed for testing. Each variance is derived from 80 observations.

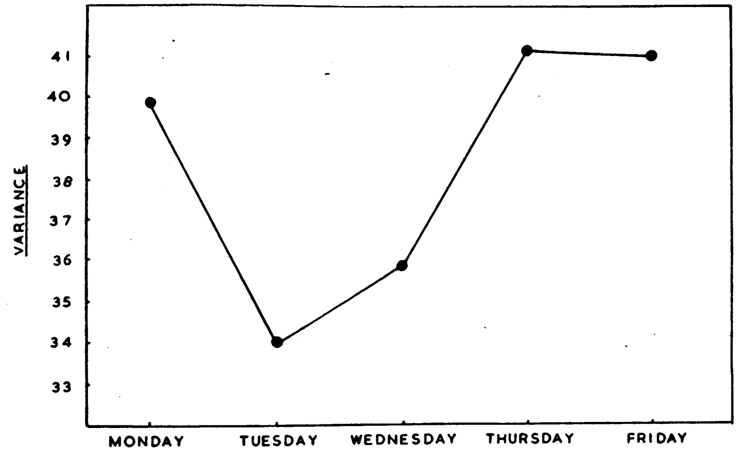

FIG. 5.—Daily variance; mean results for 3 girls; 280 observations per day

The general variance each week was particularly of use when, for any reason, a change in technique became necessary. It was found, for example, in October, 1945, that the variance had risen to an undesirably high level, and the cause of this rise was not known. By an internal reorganization of the department it was arranged that each tester should perform only half the number of tests as hitherto, since it was believed that the prevalent inaccuracy might be due to an excess of work. This arrangement did, in fact, reduce the S.D. for a single test from ±14.2 to ±11.8 per cent, which figure was still considered unsatisfactory. The next step was to introduce delivery pipettes in place of blow-out pipettes for the setting up of tests. The immediate effect of this was to raise the S.D. to ±13.7 per cent, but after a week's use of the new type of pipette the figure fell again to ±11.8 per cent for the second week and ±12.2 per cent for the third week. It appeared that the high variance first obtained with the new pipettes was due to unfamiliarity with their calibration, but that after some practice no advantage was gained by their use. Finally it was decided to discontinue the practice, which we had adopted some months previously, of inoculating the test broth with a suspension of *Staph. aureus* before introduction of the penicillin volumes. We found that by inoculating the broth after penicillin had been added the S.D. immediately dropped to ±9.0 per cent; this was considered satisfactory for the type of assays in hand and no further modifications seemed desirable. It may be interesting to record that when, some four months later, considerable impending reductions in staff were announced, the effect was to raise the S.D. from about ±9.7 to ±15.0 per cent, but that after one week it again fell to below ±10.0 per cent. This again reflected the emotional interference with the accuracy of the work.

In view of the personal factor involved in this form of assay I was interested to find out whether the day of the week on which tests were carried out might influence the accuracy of testing. Variances were, therefore, calculated on results obtained from tests performed on each day of the week. At first a hundred results were examined for each day taken at random over a 12-week period. The variances so obtained did in fact show a tendency to reach a minimum at the middle of the week and to average at more or less the same high level both at the beginning and at the end of the week. Over the period involved there were seven assayers who averaged seventy tests each day. A similar analysis was made on results eighteen months later, when over a period of ten weeks three girls averaged forty-four tests each day. For each day of the week two hundred and eighty observations were made; the variances obtained are plotted in Fig. 5. Again the middle of the week would seem optimal from the point of view of accuracy, although here Tuesday is even better than Wednesday. Since each sample was always assayed on two successive days, the variances attributed to one day would be influenced to some extent by the preceding and following days.

Since the mid-week improvement in accuracy shown above appeared to be real it was natural to inquire whether any similar relationship held between variance and the time at which tests were carried out. It was possible to obtain assessments of the variance figure corresponding to tests set up at varying times under three different sets of conditions. First, during the peak period of penicillin assay, when testing took up three and a half hours in the afternoon, and approximately one and a half hours in the morning devoted to reading the tests. The setting up of tests during this period was at the rate of approximately one test every two minutes. Later, when owing to changed circumstances penicillin testing occupied only one hour or less of the day, it was possible to obtain variances at short time intervals: (*a*) When the total number of tests performed on one day was not greater than thirty-two, and (*b*) when the total number was greater than thirty-two, and averaged about sixty. These tests were all performed at the rate of about one per minute, and were set up on a 10 per cent range, while the first series were set up on a 20 or 30 per cent range. Fig. 6 shows the variance plotted against the time in minutes after commencement of testing. It will be seen that the variance always rose in the course of testing, and fell again towards its completion. This would imply a subconscious timing by the tester, which would give the beginning and end of her work a greater significance than the middle, and would entail greater concentration to the measurements involved. Such an argument would apply equally to the setting up and to the reading of tests. Since occasional check readings have usually agreed closely with the originals, it would appear that the chief errors must enter during the setting up of the tests. It is probable that a form of rhythmical pipetting might commence once the testing had been started ; this would particularly affect the two shorter series of assays since the volume ranges in use at that time were invariably the same for all samples. Since the final increase in accuracy did not appear to be affected by the duration of testing, it would seem that no question of fatigue was involved. Moreover the mean variance for thirty tests (45.4) did not differ significantly from that for sixty tests (43.9). The mean variance for the longest series (71.1) was, without a doubt, higher by reason of the wider ranges used.

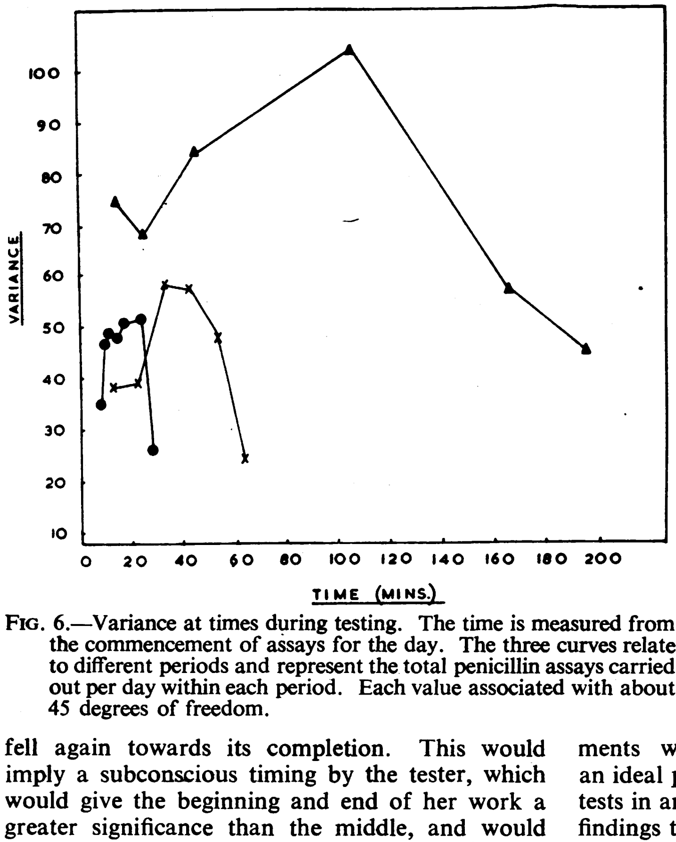

FIG. 6.—Variance at times during testing. The time is measured from the commencement of assays for the day. The three curves relate to different periods and represent the total penicillin assays carried out per day within each period. Each value associated with about 45 degrees of freedom.

DISCUSSION

It must be admitted that many of the effects described above are partially confounded ; for example, the error of one worker's testing may be assessed only by comparison with the results of other testers, and is, therefore, subject to error on that account. Unfortunately, too, the analysis of errors between days, and within days, was only completed after the laboratory had ceased to assay penicillin samples. Consequently it was not possible to impose any design upon the assay procedure which would have yielded results free from confounding. Nevertheless the paucity of literature on the subject of personal error in routine occupations has prompted the author to present his findings.

Much information has been published on tests designed to detect accident-proneness among varied groups of workers. Of this, some experiments carried out at the request of the Flying Personnel Research Committee appear to be most pertinent to the present problem. Outlined by Bartlett (1943) the work is fully reviewed by Davis (1948) in an Air Ministry publication. The purpose of the experiments was to enumerate the deviations from an ideal pattern of behaviour of pilots subjected to tests in an experimental cockpit, and to relate their findings to the subsequent accident history of individual pilots. In the first instance, therefore, it was the variance of behaviour that was beirrg assessed. These authors found that the number of deviations exceeding arbitrary limits within fixed time periods rose from the beginning of the test, reached a maximum at about the middle, and fell off towards the end. If the number of deviations they observed be taken to be a function of the variance, then such observations would parallel the results shown in Fig. 6.

Whether the psychological interpretations of their results put forward by Bartlett and Davis could be applied to the simple routine of penicillin assay will not be discussed here at length. Nevertheless, it seems likely that there is a mental process common to these extremely different types of behaviour, since in each the disorganization of skill appears to be independent of physical fatigue.

It has already been stated that the variance for individual workers may change considerably from week to week and from day to day. The magnitude of this " variance of variance " is probably a better guide to the efficiency of the worker than the mean variance over a long period. The latter may, in fact, fail to reveal any significant differences between workers. Thus it has been observed that a worker with a mean variance lower than that of her colleagues had a variance of variance (variance measured weekly over a twelve-week period) considerably greater. This type of worker has usually been observed to be subject to " moods," or periods of depression, between which their work is excellent, but during which it is most unreliable. For none of the female workers has a monthly cycle been observed in individual variance, however, but sickness, possibly of nervous origin, has been associated with an increase in errors. One girl subject to migrainous headaches was quite incapable of reading her tests during attacks.

Summary

A method is described for the estimation of the variance of results obtained in the course of routine penicillin assays. In particular this is applied to investigate the effect of personal error on the results of the tests. Its use in the control of laboratory technique is also indicated.

I would like to thank Miss Dawn Hayles and Miss Joyce Peacock for their assistance in many of the computations involved in this study, and to acknowledge the encouragement of my colleagues on the staff of the Wellcome Penicillin Unit in whose service the assays were carried out.

References

Bartlett, F. C. (1943). *Proc. Roy. Soc.*, **131 B**, 247.

Davis, D. R. (1948). *Pilot Error*, Air Ministry A. P. 3139A. London: H.M. Stationery Office.

Pope, C. G., and Stevens, M. F. (1945-6). *Bull. Hlth Org. L.O.N.*, **12**, 274.