ACCURACY OF EYE OBSERVATIONS IN PASTURE TRIALS

By MISS J. G. MILLER, Biometrician, Department of Agriculture, Wellington.

The value of different fertilisers in improving grasslands in New Zealand has been established in no small way by the number of observational topdressing trials carried out in many districts by the Department of Agriculture, virtually since the beginning of this century. Very few of these trials have had production measurements taken and judgment has been passed on the basis of observations and the Department has obviously put great faith in these trials. At the inaugural meeting of this Association A. H. Cockayne said: "The next stage in our deliberations, I think, should be to come to some agreement as to the place that selected or prepared material for observational measurements alone should occupy. This style of work when performed by the Department is generally viewed as quite unsound by the outside worker, but when he does it himself he is not at all sure whether it should be termed so or not, and might even be tempted to call it research."

In the intervening 20 years no doubt many officers of the Department have satisfied themselves that the approach was sound, but I would like to place before, you today a few figures which will help to confound or justify the critics.

Pasture observation trials have been standardised and a set system of marking responses has been adopted. This is a scale of 0 to 5 where 0 is no difference from control, 1 a slight, barely visible response, and 5 an excellent. response-the best that would be expected on that soil type under the given climatic conditions. 2, 3, and 4 are given for fair, good, and very good responses respectively, half marks being allowed. The use of such an apparently vague scale can be justified only if there is agreement among users on the marks which should be given to responses

in particular instances and also consistency from one to another. I shall now quote three cases where this agreement was tested.

In the first trial on a sloping face at Pukerua Bay, 4 observers independently marked 22 plots; a full discussion ensued and an "agreed. mark" was reached for each plot. Among the 22 plots the agreed responses ranged from ½ to 3 ; that is, doubtful to good. For any given plot the points given by the four observers ranged. over 1 point on the average, though on one particular plot there was a range of 2 points, that is, one observer marked the plot as ½ which another marked as 29. When the marks of the individuals were compared with the agreed mark for each of the plots the mean of the deviations was 0.5, that for the separate observers being 0.4, 0.4, 0.5 and 0.6. This mean deviation figure does not take account of the sign of the errors, that is, an error of $+\frac{1}{2}$ and one of $-\frac{1}{2}$ do not cancel out but add. up to a total error of 1 point on the 2 plots, that is, an average error of ½ point per plot.

A second triai, this time of 20 plots, was marked by the same group of observers plus 3 more on the same system, that is independent marking, then a discussion and the setting up of an agreed set of marks. This trial was observed under good conditions and there were responses of all sizes from ½ to 4½. In this trial the average range of marks for any one plot was 1.2, though there was one plot where the marks ranged from 2 to 43. For this plot the agreed value was 4 and the one observer who marked it down to 2 had been consistently down in all his marks, being on the average 0.7 below the agreed values. This highlights one of the important problems. With a certain amount of experience observers will be able to put plots in the same relative order of merit, but opposing temperaments such as the too enthusiastic and the over-cautious may be reflected in a different scale of marks for the same series of plots. However, even including this observer the average deviation from. the agreed values was 0.4 points per reading, the figures for the separate observers being: experienced 0.2, 0.2, 0.2, 0.3, 0.4 ; relatively inexperienced 0.6 ; over-cautious 0.7. If ½ point is added to each reading for this over-cautious observer, his average deviation becomes 0.2, so that

his consistency is equal to that of the most experienced of the group.

A third trial on a larger scale was conducted at Orton near Timaru. This took place during a course in experimental work and 24 plots were marked by 23 observers. Before the group came other plots were marked by three very experienced officers and their figures were taken as the standard values for purposes of comparison. The observers varied greatly in the amount of experience they had, and in analysing the results I omitted the marks given by two Instructors who had been in the Department only a few weeks and whose previous experience in marking plots was nil. The marks of these two observers indicated that the ability to mark plots definitely an acquired skill. In this trial the group was asked first to pick the best plot. All agreed on which this was, although the mark assigned to it varied from 2 to $4\frac{1}{2}$. The "correct" figure was $3\frac{1}{2}$. The average deviations for each of the observers are shown in table 3. Clearly 0.5 can be taken as the overall average figure. In that table two of the observers shown with a mean deviation of 0.5 seemed to be consistently above the standard, including the best plot marks. If $\frac{1}{2}$ point is subtracted from each mark in their series, their mean deviation becomes 0.1 and 0.2 respectively. Similarly the observer with mean deviation of 1.2 was consistently below the standard, his best plot being marked $2\frac{1}{2}$. If then 1 is added to each of his marks, his mean deviation becomes 0.6.

These three trials indicate that where no bias of marking exists, random errors, for partly experienced observers are of the order of 0.5 to 0.7 and for more experienced observers about 0.3 to 0.4 or less on the average'. However, it has also been shown. that occasionally experienced observers can be consistent in their estimates, but can be systematically above or below the average of their colleagues by as much as one point either way. If this is overcome by periodical co-operative markings, it appears that, pointings on the 0 to 5 scale of pasture responses in small plots can be relied on to an accuracy of $\frac{1}{2}$ point or less, provided the observers have had some experience.

A more difficult problem arises in marking different fields, not necessarily adjacent, all on the one scale. It is necessary to do this in connection with soil testing when it is desired to establish correlations between soil analysis levels and vigour of pasture. A scheme

has recently been devised in this connection and an attempt is being made to mark all pastures from which samples are taken on a 0 to 20 scale for fertility index. This is to cover all types from depleted, country with no clovers and only annual grasses at 0 to high producing ryegrass-white clover pastures of about one cow to the acre at 20. In the training courses to launch this scheme some interesting figures on deviations have been obtained from marks given by instructors who were using this system for the first time, but who were reasonably experienced in observing pastures. In this case there was no correct figure, but table 4 shows the standard deviations of instructors' marks for each of a number of fields in two districts. These vary from 1.2 to 2.2, indicating that on this scale the fertility of a field (with the conventional indications) can be judged with an average error of less than two points from what could be considered its "real" value. It was observed without detailed records being kept, that the marks of the senior men and those conducting the course were always within a range of 3 points. I consider that observations on relative fertility of different pastures will be worthwhile if they can be made with the: accuracy here recorded or better, as is likely to be the case when the instructors become more experienced. Where the scale used is 0 to 5 the error is about ½ point and where the scale is expanded by a factor of 4 to become 0 to 20 the error is expanded by almost the same factor to become about 2 points.

A different problem but closely allied with the others is to estimate the ratio of the various components of the sward: grasses, clovers, weeds and bare ground..

Here the further problem arises that it is not sufficient to have all observers able to give the same estimate consistently for the same pasture. An absolute value of the ratio does exist for the pasture and we are required to get an estimate as close to this as possible with no consistent bias up or down. As the B/A/W ratio, as it is termed, should refer to the ground cover provided by the species, an objective estimate can be obtained by the point analysis technique. This method, as you know, uses a frame with a number of needles in it, and a record is made of the species hit when each needle is lowered to the ground. With an, adequate number of points an accurate estimate can be obtained of the relative ground cover provided by each species. To test simply the accuracy of observa-

tion, quite apart from sampling variation, a small area 6ft x 3ft of closely grazed herbage was chosen and 19 observers asked to make an estimate of the B/A/W/BG ratio. There could be doubt that all observers were able to see the same area, but it was their ability to translate their view into figures that was tested. Then a point analysis was made of the same area. Table 5 shows how closely the point analysis figures agree with the mean estimates of all observers, indicating that in these conditions of a somewhat open, hard-grazed sward, observers were at least aiming to measure the same feature. The variation among observers for each constituent is gauged by the standard deviations which are shown in table 6. The absolute errors are larger for the more dominant constituents, but proportionately the errors are larger for the rarer elements, the clover and weeds. This is reasonable, as all eye observations are made only to the nearest 5 per cent.

As the errors for each observer for the four constituents are naturally not independent, a total error figure was calculated for each observer. This was the 'sum of the absolute deviations of his marks from the point analysis *figures* for each constituent. The distribution of these scores is approximately normal, on a casual inspection, with a mean value of 24, which is 6.0 per item. The scores of the three most experienced observers were 16, 16, and 18 respectively and all scores are shown in table 7. It is likely that the lower scores were obtained somewhat by chance by people who were somewhat near the ability of these three. Thus, an average error of about 4 'to 5 per cent. per constituent of an estimate would appear to be the level of accuracy that can be achieved with practice. However, it can be seen that some errors were very high, and some observers were out by as much as 20 on one estimate. It can be assumed that these were inexperienced observers, but this emphasises that B/A/W ratios can be relied on only if the observers are experienced.

To summarise, the position on the validity of eye observations seems to lie between the two extreme opinions of their being the complete answer or of their being not worth a "tin of fish." In the hands of reasonably experienced observers pointings on a 0 to 5 scale can be expected to have an average error of $\frac{1}{2}$ point; on a 0 to 20 scale readings would have an error of somewhat less than 2 points, whether plots. in a field or separate fields are observed.

TABLE 1

Average Deviations of Observers from Agreed Values
Pukerua Bay

Observer	Points
1	0.4
2	0.4
3	0.5
4	0.6
Overall average	0.5

TABLE 2

Average Deviations of Observers from Agreed Values
Silverstream

Observer	Points	
2	0.2	
3	0.7	consistently low
4	0.2	
5	0.6	relatively inexperienced
6	0.4	
7	0.3	

TABLE 3

Spread of Average Deviations from Standard Marks at Orton

Deviations	No. of Observers	No. of Observers after Correction for Bias
0.1	—	1
0.2	1	2
0.3	2	2
0.4	1	1
0.5	10	8
0.6	2	3
0.7	1	1
0.9	1	1
1.2	1	—

TABLE 4

Standard Deviations of Estimates of 12 Observers of Fertility Index for Separate Fields

District	Field	Standard Deviation
Blenheim	1	1.2
	2	2.2
	3	1.4
	4	1.8
	5	1.9
Christchurch	1	1.8
	2	2.0
	3	1.3
	4	1.6
	5	2.1
	6	1.2
	7	1.8

TABLE 5
Estimates of B/A/W/BG ratio (Timaru)

Factor	Point Analysis Figures	Mean of estimate& by all observers
Grass	34	32
Clover	9	7
Weeds	5	8
Bare Ground	52	53
	100	100

TABLE 6
Standard Deviation of Estimates for each Factor

Grass	9.2
Clover	4.5
Weeds	3.5
Bare Ground	11.6

TABLE 7
Distribution of Scores for "Total Error" in B/A/W Estimates

Scores	Number of Observers
5-9	1
10-14	1
15-19	5*
20-24	5
25-29	2
30-34	1
35-39	2
40—44	1

*Includes the 3 most experienced.

DISCUSSION

Q. At Lincoln College we carried out trials concerned solely with the growth-of clover under different fertility treatments. The clover was cut and the weights of yields were recorded. These yields and eye estimations agreed reasonably well though the latter tended to be higher.

A. If "scores" are estimates of production responses it is fair to compare them with production measurements; but to say that measurements although in line with scores were always higher when the scores were arbitrary units simply means that a different scale was uaed for the scores, and a change of scale would bring them into agreement with the measurements.

Hamblyn: From field experience it has been found that eye estimation of yield is liable to be very inaccurate,.

Q. What. is meant bv the concept "thrift"?

A. (Sma'llfield) : In considering- pasture measurement by observation it is necessary to consider what kinds of responses the observer is trying to record. In the fertility index being used in connection with the soil testing service we are aiming to secure some measure. of the value of the pasture for the production of, say. butterfat. This depends on two factors, The first is the composition of the sward which at one end of the scale may be dominated by danthonia or tussock and at the other end by ryegrass and white clover. The second factor is the manner in which these plants are growing. For example a stunted, nitrogen starved pasture would be marked low for thrift and a vigorously growing pasture would be marked high.