NZ ISSN 0048-0134

NEW ZEALAND
JOURNAL OF FORESTRY SCIENCE

New Zealand Forest Service,
Forest Research Institute, Rotorua

Acting Joint Editors: P. D. Gadgil, J. M. Harris

VOLUME 9 DECEMBER 1979 NUMBER 3

AN EARLY PROGENY TRIAL IN *PINUS RADIATA*

2. SUBJECTIVE ASSESSMENT OF CROOKEDNESS

M. H. BANNISTER

Forest Research Institute, New Zealand Forest Service, Rotorua

(Received for publication 7 September 1979)

ABSTRACT

Five independent observers assessed more than 1600 stems of **Pinus radiata** D.Don for crookedness on a 0-9 scale. Inconsistencies in the scoring of individual observers resulted in erratic changes in the mean and the variance as the work progressed; this was probably the origin of four distinct interactions, which contributed a small but statistically significant part of the total variance. The error variance of the individual observer was about 0.5 and constituted about 32% of the total variance.

The frequency distributions of the errors generally showed significant departures from the normal, but for no consistent reason. In some there was skewness, in some positive kurtosis, and most showed apparently anomalous frequencies in some classes. As expected, the mean score of five observers per tree was greatly superior in its statistical properties to the single scores, the departure of its frequency distribution from the normal falling well short of the 5% significance level.

The variance of the errors showed significant heterogeneity and were to some extent correlated with the means. Attempts to eliminate these undesirable features by transformations were only partly successful; but, despite rather severe changes brought about by the transformation, the analyses of variance made before and after transformation gave essentially the same results.

The data were analysed as five separate sets (one from each observer) and as a single set combining the five scores for each tree. The results consistently

indicated the presence of a substantial families component in the total variance ($P < 0.001$). The best estimate of heritability was 0.44 (90% confidence limits about 0.28 and 0.87).

In ranking the means of 26 open-pollinated families, the five observers showed good agreement (Kendall and Babington Smith's coefficient of concordance $W = 0.81$; $P < 0.001$). There were also differences in the amount of tree-to-tree variation within families. When the families were ranked according to their variances, the observers were again found to have displayed a highly significant concordance ($W = 0.55$; $P < 0.001$).

It is concluded that a single observer, under conditions like those experienced in this study, could score crookedness accurately enough if the purpose were solely the ranking of group means; but in general the assessment of crookedness should be based on the sum of scores by two or more independent observers per tree. This is particularly true when the data are to be used for statistical work involving measures of dispersal, as in the analysis of variance and covariance.

INTRODUCTION

There is general agreement that straightness in sawlogs and pulp logs is highly desirable. On casual inspection many stems may look straight, and in practice these may be just as valuable as if they were absolutely straight in the geometrical sense. In truth, however, one may doubt whether any stem is perfectly straight and, rather than judging from casual inspection, it is more realistic to regard every stem as crooked, more or less. Moreover, there is often a noticeable variation in the crookedness of trees within stands and of the mean crookedness of different stands. Therefore, if crookedness can be expressed quantitatively its variation should be amenable to statistical analysis, and with a proper experiment it should be possible to gauge the relative importance of the hereditary and environmental components of that variation.

A photogrammetric technique for measuring crookedness has been developed (Shelbourne and Namkoong, 1965), but apparently neither that nor any other objective method has proved practicable for assessing large numbers of trees. Instead, subjective methods are used in many parts of the world. These depend on visual inspection, judgment using an arbitrary scale, and the allocation of a numerical score to each tree.

The assessment of crookedness, as described here, was part of a much larger study of phenotypic and genetic variation in a wind-pollinated progeny trial of *Pinus radiata.* The results of the work as a whole are being reported in a series of papers, the first of which (Bannister, 1969) should suffice as the main introduction to all its successors. The main purpose of this second paper is to examine some of the statistical properties of more than 7500 subjective scores for crookedness, and to consider the effectiveness of the technique. The genetic information that is presented here will be discussed more fully in the later papers in the series.

MATERIALS AND METHODS

The experiment involved wind-pollinated progenies, from 26 female parents, arranged in a randomised-block design with nine replications, 6.5 km from Wakefield in the Nelson district. The trees were planted in 1950 at 1000 stems/ha, spaced at 3-m intervals in single-row plots of 10 trees each, in rows 4 m apart. Heavy mortality in the first year contributed to a natural reduction in stocking to 727 stems/ha by the

fourteenth year; out of a potential 234 plots, 24 were missing. At the time of the assessment the trees had a mean height of 19.5 m.

Before attempting the main assessment, four inexperienced untrained observers were taken into an ordinary stand of *P. radiata,* 19 years old, about 1 km from the progeny trial. They were told to judge each stem as a whole, from a point 2 m above ground to the limit of visibility in the crown. They were to ignore any gradual curvature affecting the whole stem, but any other deviations from a straight line they were to judge as contributions to crookedness. In general, they were to adopt the concept that most of the trees would be on, or close to, the average for the stand, and would be given a score of 4 or 5; that some trees, seemingly more crooked than average, would be given a score of 6 or more; and that others, seemingly less crooked than average, would be given a score of 3 or less. The permitted range of scores comprised the 10 integers from 0 to 9.

Next, the five observers (four novices and their instructor) together examined and discussed a selected sample of stems in an attempt to standardise their scoring. Finally, it was emphasised that each stem to be scored had to be examined from several positions, through an arc subtending an angle of at least 90° at the base of the stem, and from distances ranging from 10 cm or less to 3 m or more.

As a preliminary test of the method they then independently scored a random sample of 112 trees. The five scores for each tree were clearly concordant, and an analysis partitioned the variance as follows:

Observers	0.0426
Trees	0.9342
Residual	0.4984

From this the estimate of phenotypic variance for trees scored by a single observer was 0.9342 + 0.4984, i.e., 1.4326; that for trees scored by five observers, using the mean of five scores for each tree, was 0.9342 + 1/5 (0.4984), i.e., 1.0340.

They then assessed the trees in the progeny trial. Each observer began in a different part of the experiment and worked independently through one randomised block at a time, taking care that the completion of a block coincided with the end of work for the day. On completing the assessment for the whole experiment, each observer went back to his starting point and scored all the trees of one block a second time. In addition, one observer scored another block for a second time, so that for the one block the scoring was repeated by two observers.

The three stems represented in Fig. 1 indicate the range of variation.

ORIGINS OF ERRORS

Analyses of variance in the original scores yielded estimates of error variances of several kinds, depending on their composition. These may be considered under four headings:

(1) *Error variance for two observers who both scored the same sample of trees twice*

One set of data was of this type. It was based on 26 plots of one replication, comprising 190 trees (Table 1). This was the only analysis in which one could test

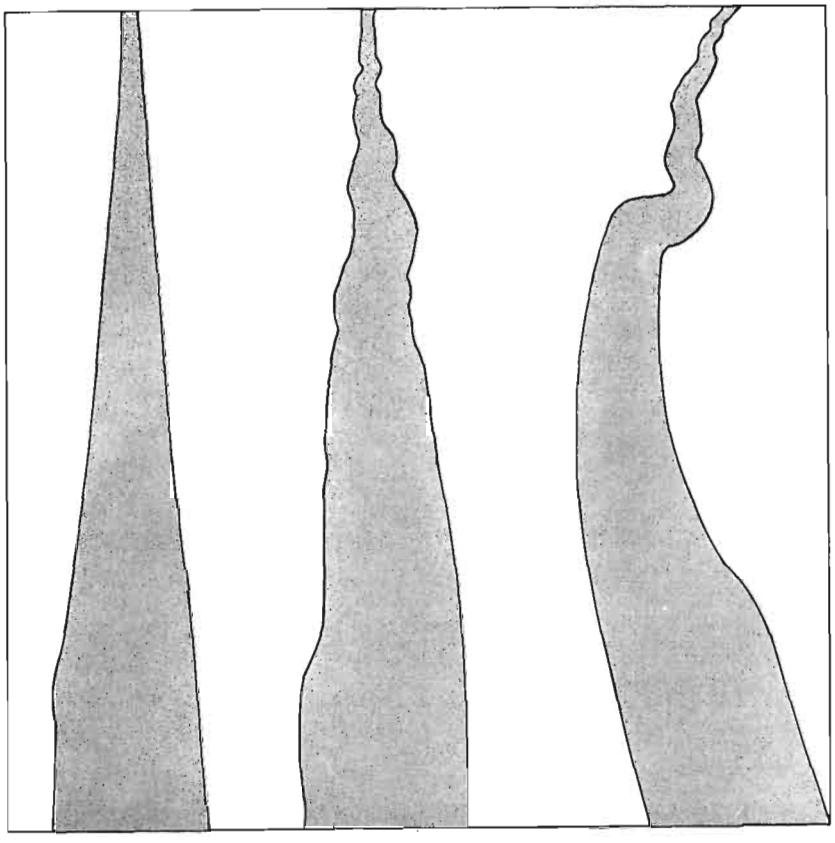

FIG. 1—These outlines show the range of variation in crookedness, and were drawn from photographs taken at eye level with a 35 mm camera. Assessed by five observers, these stems had mean scores as follows: left — 1.2; centre — 4.4; right — 7.2.

the significance of observers × trees-in-plots and estimate the error and observers × trees components separately. Excluding the main effect of observers, this analysis may be summarised as follows:

Variance component	*Estimate*	*Percentage of total variance*
Plots	0.1377	7.7
Observers × plots	0.0804	4.5
Trees in plots	0.8561	48.1
Observers × trees-in-plots	0.1710	9.5
Error	0.5355	30.2
Total	1.7807	100.0

TABLE 1 - Analysis of variance of crookedness scores:
190 trees x 2 observers x 2 assessments

Source of variation	d.f.	Mean square	F ratio	Expectations of mean squares(1,2)
Observers	1	18.64	9.10**	$\sigma^2_w + 2\sigma^2_{st:p} + 2k\sigma^2_{sp} + 380S^2$
Plots	25	7.97	2.01**	$\sigma^2_w + 4\sigma^2_{t:p} + 4k\sigma^2_p$
Observers x plots	25	2.05	2.33***	$\sigma^2_w + 2\sigma^2_{st:p} + 2k\sigma^2_{sp}$
Trees in plots	164	3.96	7.33***	$\sigma^2_w + 4\sigma^2_{t:p}$
Observers x trees-in-plots	164	0.88	1.64***	$\sigma^2_w + 2\sigma^2_{st:p}$
Error	380	0.54		σ^2_w
Total	759			

** $P < 0.01$

*** $P < 0.001$

(1) A mixed model was adopted, with S^2 representing observers as a fixed effect and all other effects random, the components being:

σ^2_p - plots

σ^2_{sp} - observers x plots

$\sigma^2_{t:p}$ - trees in plots

$\sigma^2_{st:p}$ - observers x trees-in-plots

σ^2_w - error

(2) The coefficient k = 7.28 (*see* Snedecor, 1956 p.268) and represents effective number of trees per plot.

(2) *Error variance for any one observer who scored the same sample of trees twice*

Each observer interpreted the variation in a somewhat different way, and each showed some inconsistency from day to day or from one part of the experiment to another. Examples of this are shown in Fig. 2; statistics for duplicate scores by all five observers are given in Table 2.

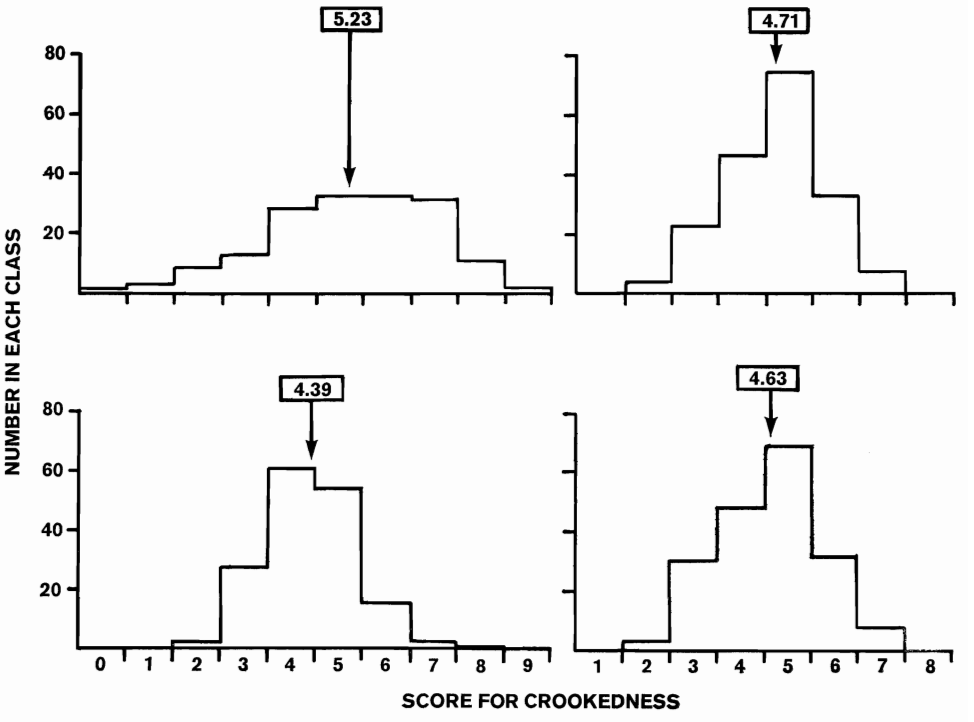

FIG. 2—The two upper histograms show the frequency distributions of the scores allotted by observers A (left) and B (right) when they assessed a replication for the first time. The two lower ones represent a second assessment of the same trees, carried out seven days later by the same observers (A, left: B, right). Arrows indicate means.

TABLE 2—Statistics of duplicate scores by each of five observers

Observer	No. of trees	Mean 1st time	Mean 2nd time	Variance 1st time	Variance 2nd time	Error variance	Consistency*
A	166	5.23	4.39	3.14	1.06	1.24	0.46
B	190	4.71	4.63	1.18	1.24	0.44	0.63
C	193	4.52	4.35	1.89	1.36	0.55	0.66
D	197	3.76	4.05	2.37	0.94	0.61	0.64
E	196	4.15	4.09	2.05	1.30	0.50	0.67

* In this context the consistency of an observer's scoring may be expressed as:
$\sigma^2_b / (\sigma^2_b + \sigma^2_w)$
where σ^2_b and σ^2_w are respectively the variance-component estimates for "Between Trees" and "Error", derived from a one-way analysis of variance. The derivation of Consistency is therefore that of an intra-class correlation (cf. Harris, 1913) and of repeatability with two scores per tree (cf. Falconer, 1960 pp. 142-9). Note that the Error includes any subconscious shift in the observer's standard between the first and the second assessments.

(3) *Error variance for any one observer who scored all the trees once*

Five sets of data came under this heading. Taking each observer separately, the analyses were carried out in two stages: the first was a one-way analysis to divide the sums of squares into two parts — i.e., "between plots" and "within plots" (cf. Kendall, 1948 pp. 178-81; Scheffé, 1959 p. 362; Bannister, 1969). The second stage will be dealt with later; for the moment the only results to note from these analyses are the estimates of the "within-plots" variance. This is a composite term, which may comprise three possible components: trees-in-plots, observer × trees-in-plots, and the true error of the individual observer. In the symbols defined at the foot of Table 1, it may be expressed as:

$$\begin{aligned}\text{Error variance} &= \text{within-plots variance}\\ &= \sigma^2_{w} + \sigma^2_{st:p} + \sigma^2_{t:p}\end{aligned}$$

The five estimates, each with 1308 degrees of freedom, are:

Observer	*Estimate of composite error*
A	1.2938
B	1.2024
C	1.4174
D	1.0773
E	1.2491

(4) *Error variance when the variate is the sum (or mean) of the scores for each tree, each observer having scored each tree once*

To make comparisons easier, let it be assumed that the variate considered here is the mean, rather than the sum. Its error variance, like that discussed under (3) above, is composite also. It contains the same tree-in-plots component, but the composition of the remainder is rather uncertain. In a fully random model it would consist of a fraction of the true error and a similar fraction of the observers × trees-in-plots interaction. Thus, if s denotes the number of observers:

$$\begin{aligned}\text{Error variance} &= \text{within-plots variance}\\ &= \frac{1}{s}(\sigma^2_{w} + \sigma^2_{st:p}) + \sigma^2_{t:p}\end{aligned}$$

On the other hand, if observers were better treated as a fixed effect, we should have:

$$\text{Error of variance} = \frac{1}{s}(\sigma^2_{w}) + \sigma^2_{t:p}$$

Either way, the estimate of the error variance for analyses based on the mean of five scores per tree is 0.8142.

STATISTICAL PROPERTIES OF ERRORS

The errors examined under this heading were derived from the deviations of the scores for individual trees from their plot means. As explained above, these deviations were confounded with the errors of one or more observers.

Frequency Distributions

The statistics for kurtosis, skewness. and goodness of fit to the normal distribution are shown in Table 3. The distributions for the five observers, taken separately, showed distinct significant aberrations from the normal distribution: these included positive and negative skewness and positive kurtosis, but no consistent pattern could be discerned. The only more-or-less general feature seemed to be the presence of a few anomalous frequencies in each distribution. In contrast to this, the mean of five scores per tree showed a frequency distribution that conformed well with that calculated from the normal curve.

TABLE 3—Tests for normality of distributions of crookedness scores: (a) original (b) after transformation(1)

Observer	Departure from normal (χ^2)	Skewness(2)	Kurtosis(2)
A (a)	43.4**	0.1844**	0.2118
(b)	54.2***	0.2089***	—0.0433
B (a)	44.5**	—0.2702***	—0.0326
(b)	42.9**	—0.2500***	—0.0271
C (a)	28.0	0.0748	0.3452**
(b)	51.5***	0.0786	0.3890**
D (a)	108.4***	0.0830	0.3673**
(b)	71.2***	0.0642	0.0547
E (a)	63.9***	0.3280***	0.6349***
(b)	54.9***	0.2982***	0.3882**
Mean of five observers (a)	16.1	0.0511	0.2102
(b)	19.2	0.0493	0.1047

(1) Described later under the heading "Effects of Transformations".
(2) The values for skewness and kurtosis are for the statistics g_1 and g_2, calculated as described by Snedecor (1956).
** $P < 0.01$
*** $P < 0.001$

Heterogeneity of Variances

Taking the scores by separate observers as the variate, within-plot sums of squares and degrees of freedom were pooled to give an error variance for each observer, each family, and each block. Results from Bartlett's test were:

Errors of observers (4 d.f.): $\chi^2 = 26.3$, $P = 0.0001$;
Errors of families (25 d.f.): $\chi^2 = 105.3$, $P = 0.0000$;
Errors of blocks (8 d.f.): $\chi^2 = 96.4$, $P = 0.0000$.

The same test, applied to errors based on the mean of five scores per tree, gave:

Errors of families: $\chi^2 = 35.8$, $P = 0.0745$;
Errors of blocks: $\chi^2 = 31.4$, $P = 0.0002$.

Considering each observer separately, there were clearly significant differences among the variances. Taking the mean of five scores for each tree, the variances for families became practically homogenous and those for blocks, although varying much less than they did for separate observers, still differed significantly.

Relationship Between Variance and Mean

Graphical examination of the 45 observer-block means and their error variances suggested that they might be positively correlated (Fig. 3A), and this was strongly supported by the related data from the means of five scores per tree (Fig. 3B). Contrariwise, no relationship whatsoever could be discerned between the family means and their variances.

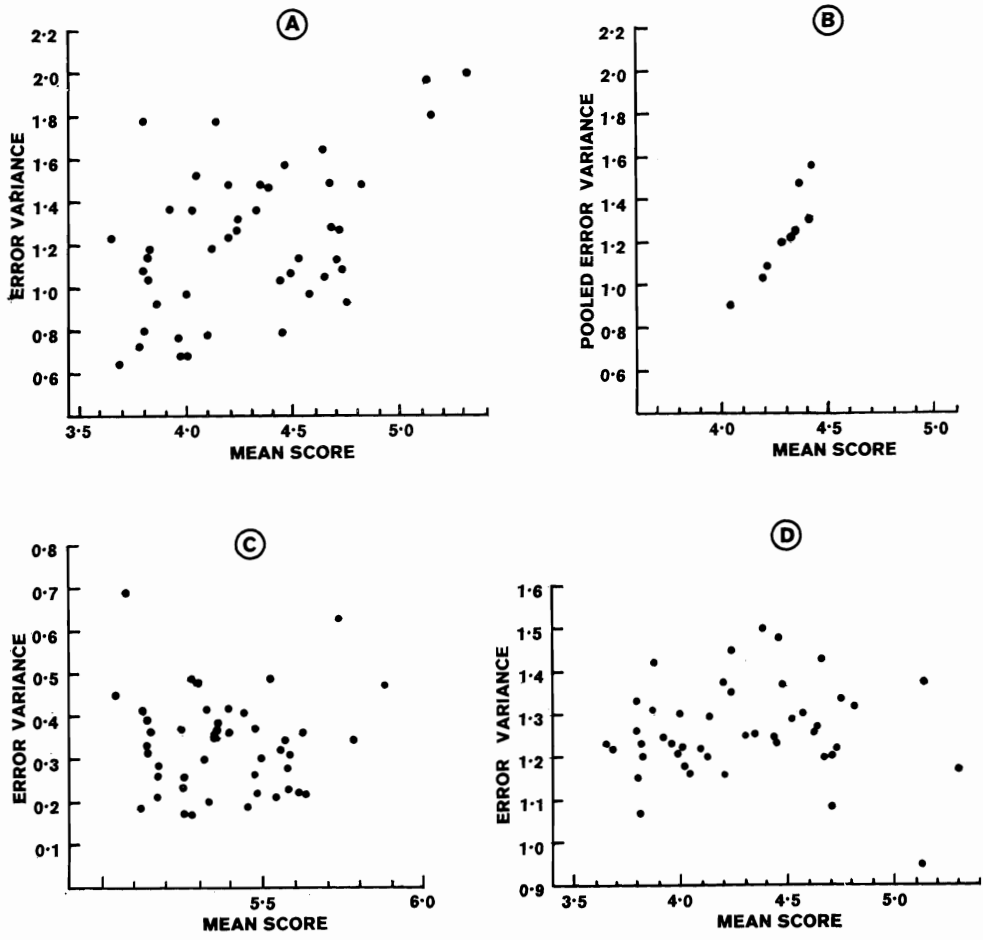

FIG. 3—Relationship between variance and mean for observer-block combinations:

A — Original scores;

B — Original scores pooled over observers within blocks;

C — After transformation;

D — After adjustment by special procedure to stabilise variance.

Effects of Transformations

In view of the apparently linear relationship revealed by Fig. 3 (A, B), the ordinary square-root transformation might have seemed to offer a means of getting the data into a form more satisfactory for analysis. However, the graphs provided no information about the relationship where the mean lay close to the limits 0 and 9, and these regions might well have been critical. All one can say for certain is this: for hypothetical samples with means 0 and 9 the variance would be 0; a straight-line regression, therefore, while it might have described the relationship well enough for part of the range, would have been a very poor fit for the whole of the range. (For a somewhat similar problem, involving subjective scores with finite upper and lower limits, *see* Hopkins, 1950.)

A solution was sought on the lines indicated by Kendall (1948, pp. 205-6), who noted that if the relationship between the variance, v, and the mean, m, of an original variate, x, can be determined as $v = f(m)$,

$$\text{then } g(x) = \int^{x} \frac{1}{(f(m))^{\frac{1}{2}}}\, dm$$

is a transformation that will stabilise the variance and probably normalise the distribution as well. (After performing the integration, each original observation x is substituted for m in the right-hand side of the equation, which then provides g(x), the transformed value.) In the data considered here, the graphical trial of several equations that might have portrayed the relationship of the variance to the mean led to the choice of

$$v = \frac{m^2(9-m)}{13-m}$$

as one that gave the kind of curve requred. Manipulation of this according to Kendall's formula was not entirely satisfactory, because strict application of the results would have led to an original score of 0 being transformed to $\log_e 0$, which is $-\infty$. It was realised later that this difficulty might have been obviated by adding an arbitrary constant such as $\frac{1}{2}$ to each score before transforming. However, graphical examination of various possible relationships between the original and the transformed values suggested that precise mathematics were not required. Instead, it was deemed sufficient to draw a free-hand curve, closely following the one obtained by Kendall's formula, that would stretch the intervals suitably towards the ends of the range and compress them in the neighbourhood of 6 (Fig. 4). It is worth noting, in passing, that the square-root transformation happens to be merely one special case yielded by Kendall's general formula; it can be represented by a curve very like the one in Fig. 3 over the range 0-6, or even 0-7, but no further.

The transformation shown in Fig. 4 was applied to the data. It effectively eliminated the dependence of the variance on the mean (Fig. 3), but at the same time — far from having a stabilising effect — it aggravated the heterogeneity of the variance. The situation seemed to call for a fresh approach.

In the early stages of this work it had been noted that some abrupt changes had occurred, sometimes in the mean and sometimes in the variance, as an observer moved

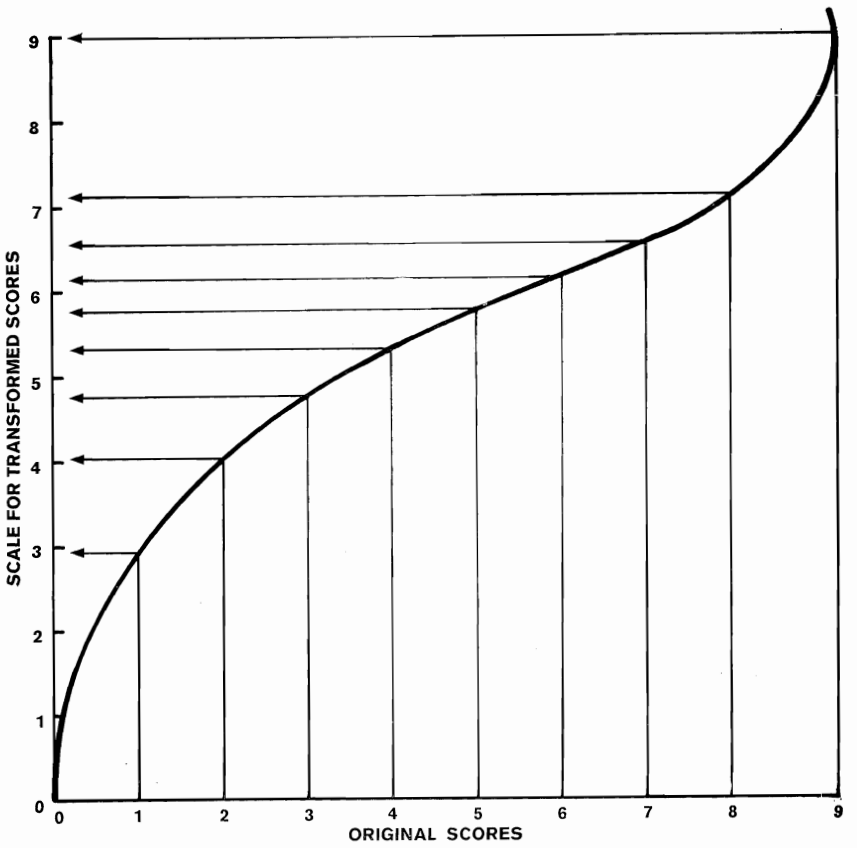

FIG. 4—Replica of the free-hand curve used to transform the original scores. The scores 0 and 9 remained unchanged; the intermediate values were read from the ordinate at the points indicated by the arrows.

from one block to another, whereas within each block each observer seemed to have scored consistently. It was supposed, therefore, that one might consider the scores to be based on a unique scale for each observer-block combination. and that converting them to a single, standard scale would make them more amenable to the analysis of variance. A conversion was carried out by the formula

$$g(x) = B + \frac{D}{S}(x - B)$$

where $g(x)$ = transformed score

x = original score

B = over-all mean of observer-block combination containing x

S = over-all standard deviation of observer-block combination containing x

D = pooled within-groups standard deviation for 45 observer-block combinations.

As with the previous transformation, this method of adjustment eliminated the correlation between the variance and the mean (Fig. 3D). Heterogeneity of variance also was eliminated for observers, blocks, and observer-block combinations; but for families the adjustment aggravated the heterogeneity. For the original scores the ratio of the highest to the lowest variance, from a total of 26 families, was 2.16 : 1; after the adjustment the ratio was 3 : 1. For all the variances obtained after the adjustment, the results from Bartlett's test were as follows (these may be readily compared with those pertaining to the original scores, presented earlier):

Errors of observers (4 d.f.): $\chi^2 = 4.8$, $P = 0.31$;
Errors of families (25 d.f.): $\chi^2 = 268.8$, $P = 0.0000$;
Errors of blocks (8 d.f.): $\chi^2 = 11.3$, $P = 0.19$.

Working with the mean of five adjusted scores per tree, the test gave:

Errors of families: $\chi^2 = 61.1$, $P = 0.0001$;
Errors of blocks: $\chi^2 = 7.1$, $P = 0.53$.

The distributions of the errors for the transformed scores showed generally very little change from those of the original scores (Table 3).

ANALYSES OF VARIANCE: ORIGINAL SCORES

For the analysis and interpretation of the results, it was assumed initially that although the original data did not have the error distributions that are demanded in theory for the analysis of variance to be completely valid, the discrepancies were not great enough to cause serious bias in estimation or gross errors in tests of significance (cf. Cochran, 1947).

Two-way Analyses

These analyses were done by a method that involves two stages. The first stage, as mentioned earlier, yields an unbiased estimate of the "within-plots" or error variance; the second stage uses the unweighted plot means, augmented by estimated values for the missing plots (Yates, 1934; Kendall, 1948 pp. 228-33; Snedecor, 1956 pp. 312-3; Scheffé, 1959 p. 362; Bannister, 1969).

This method was used to analyse the variance of the scores of each observer separately and that of the mean of the scores of five observers per tree. The results for the five observers separately were in some ways unanimous, e.g., they all found a strong families effect (the variance-ratio tests all gave $P < 0.001$), and they all indicated the presence of a small but significant interaction for families × blocks ($F_{176,1308} = 1.25$ to 1.59; $P < 0.05$ or $P < 0.01$). But in other ways they showed striking contrasts — most noticeably in the blocks effect, for which $F_{8,176}$ ranged from 1.4 ($P > 0.05$) to 14.0 ($P < 0.001$).

The analysis based on the mean of five scores per tree is shown in Table 4.

Components of variance were estimated from the equations representing the expected composition and the calculated numerical values of the mean squares (Fig. 5). In this diagram there appear to be great differences between the estimates by the five observers, but apart from those for error — where their significance has already been demonstrated — no exact statement can be made about their possible importance. For blocks there are only 8 degrees of freedom, so that in the absolute sense the estimates of that component shown in Fig. 5 are quite unreliable; even so, they may

TABLE 4—Analysis of variance in crookedness, based on mean of five scores per tree

Source of variation	d.f.	Mean square	F ratio
Families	25	9.7354	8.10***
Blocks	8	2.0020	1.67
Families × blocks	176	1.2019	1.48**
Error	1308	0.8136	
Total	1517		

** $P < 0.01$
*** $P < 0.001$

have some heuristic value. For the other three components, some idea of reliability can be gained by considering the 90% confidence intervals. The limits of these were calculated by methods described by Henderson (1959 pp. 40-1). For families and families × blocks the results are based on the distribution of F, and are approximate; those for error are based on the distribution of χ^2 (Table 5).

TABLE 5—Estimates of components of variance in crookedness, from original scores by five observers, with 90% confidence limits (point estimates in parentheses)

Observer	Families			Families × blocks			Error		
	Lower		Upper	Lower		Upper	Lower		Upper
A	0.09	(0.18)	0.34	0.10	(0.17)	0.25	1.21	(1.29)	1.38
B	0.10	(0.18)	0.34	0.04	(0.09)	0.15	1.12	(1.20)	1.28
C	0.05	(0.11)	0.21	0.02	(0.07)	0.14	1.32	(1.42)	1.51
D	0.08	(0.15)	0.28	0.05	(0.09)	0.15	1.01	(1.08)	1.15
E	0.17	(0.28)	0.51	0.01	(0.05)	0.11	1.17	(1.25)	1.33
Mean of five	0.09	(0.16)	0.29	0.03	(0.06)	0.10	0.76	(0.81)	0.87
Mean of four	0.08	(0.14)	0.26	0.04	(0.07)	0.12	0.81	(0.87)	0.93

Compared with the confidence intervals, the differences between the five observers' estimates may appear unimportant, but it should be remembered that for any one of the variance components the five estimates were *not* derived from five independent samples from the same population. On the contrary, they were derived from the same trees, growing in the same environments, with the same genotypes and the same developmental history, for all five observers. Differences in the estimates, therefore, cannot be ascribed to the random effects that typically accompany the study of small samples. What they do reflect is the individuality of five distinct human beings: each of the observers interpreted the tree-to-tree variation privately, and it is most improbable that in doing so any two of them, let alone all five of them, thought exactly alike.

A specific example of individualistic thinking is provided by observer E's estimates (Fig. 5 and Table 5). His estimate of phenotypic variance conformed well with the estimates of the other four observers, but his estimate of the families component was by far the highest. There is a strong suspicion that this was not a fortuitous result: observers A-D were complete newcomers, and presumably as unbiased as anyone could be; but E had a prior familiarity with the experiment, could recognise some of the

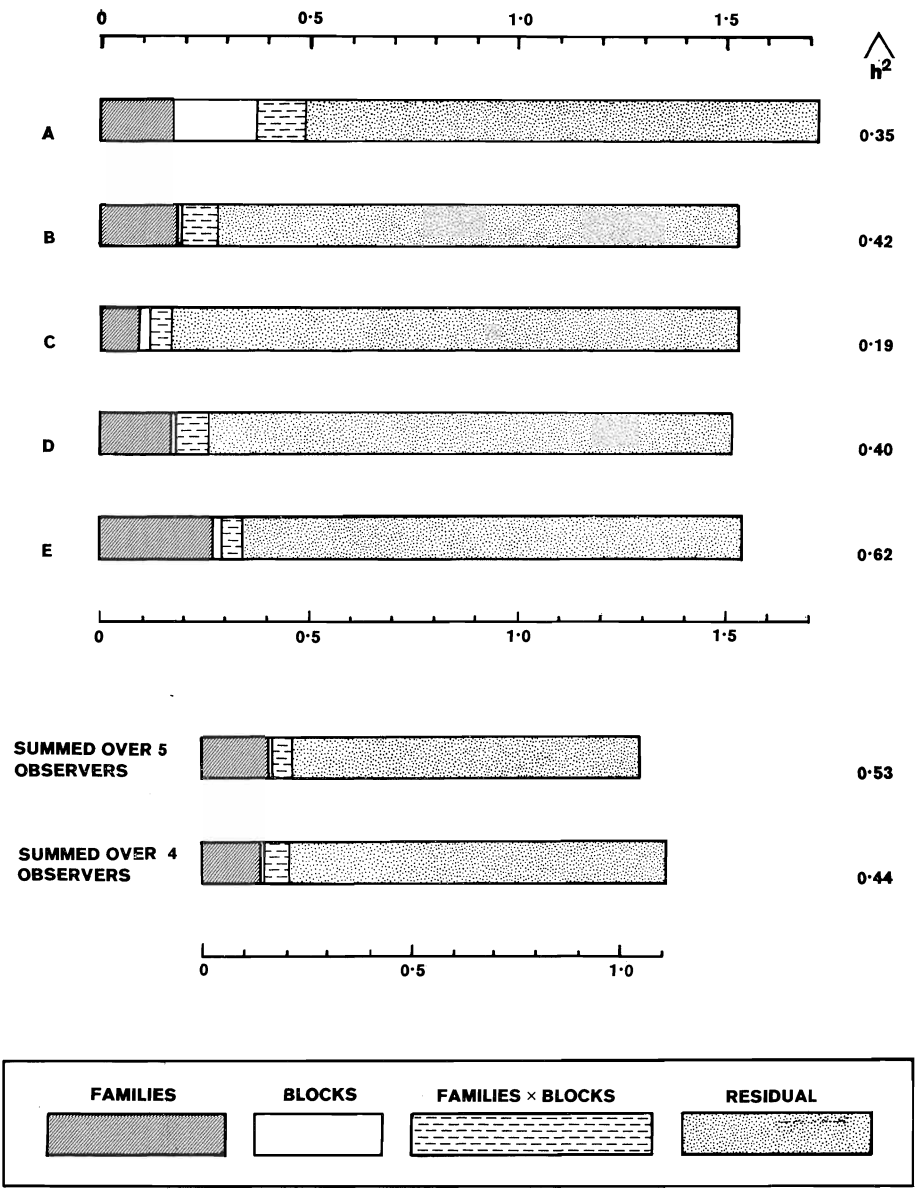

FIG. 5—Diagram based on scores for crookedness, showing the partitioning of phenotypic variance into four components. The symbol h^2 stands for "point estimate of heritability" (see Bannister, 1969). A-E represent analyses of the scores of five observers taken separately. For the scores summed over observers the variances have been divided by 25 and 16 respectively to make all the scales equivalent.

families on sight, and was therefore almost certainly influenced subconsciously so that his scores exaggerated the families effect. For this reason, analyses of variance were carried out on both the mean of five scores per tree (observers A-E) and the mean of

four scores per tree (observer E excluded). Exclusion of observer E made no difference to the essential conclusions, but it did lower the estimate of the families component a little (Table 5).

Three-way Analysis

It would have been instructive to see the results of a single, comprehensive analysis, based on a model with three main effects — i.e., observers, families, and blocks. With missing trees and missing plots, however, the data were far from orthogonal, and that sort of analysis was impracticable. Instead, a search was made for one or more orthogonal three-way models that could be matched in full by real data. Several were found and, although even the best of them used no more than one-third of the information, it still provided useful results (Table 6). Where comparisons are possible, this analysis generally shows good agreement with the one shown in Table 1; a notable feature is the high significance of the interactions involving observers. Three of these interactions (Table 6) together contributed 9.5% of the total variance.

The exclusion of observer E's scores had only slight effects on the outcome — the relative importance of families and of families × blocks was reduced a little, and that of observers × blocks and of observers × families × blocks was increased a little.

ANALYSES OF VARIANCE: TRANSFORMED SCORES

Each of the transformations described earlier was applied, and the adjusted data were subjected to the same two-way and three-way analyses as for the original scores. Variance ratios, heritability estimates, and even probability levels for the significance of minor components, were changed very little by the transformations. *In all three sets of analyses, therefore, the variance was partitioned in virtually the same proportions.*

It was not feasible to test the data thoroughly for non-additivity, but family × block classifications of the plot means were examined by the method of Tukey (1949). The results, with 1 degree of freedom for non-additivity and 175 for the remainder, were:

Original scores (mean of 5): $F = 3.31$, $P > 0.05$;
Original scores (mean of 4): $F = 2.72$, $P > 0.05$;
After second transformation (mean of 4): $F = 0.31$.

Evidently, at least at this level in the analyses, non-additivity was unimportant. At the same time, any non-additivity that might have been present in the original scores of observers A-D was apparently eliminated by the transformation.

RANK CORRELATION

The parametric analyses show that, although the observers were rather inconsistent and erratic in their individual judgments, there was nevertheless a good deal of agreement between them. A statistic that is convenient for expressing the strength of such agreement is W, the coefficient of concordance (Kendall and Babington Smith, 1939). This depends on listing a series of objects according to their ranking by two or more judges working independently. If there is no agreement between the judges, $W = 0$; if there is perfect agreement, $W = 1$.

The family means, calculated from the original scores of each observer separately, were used to list the 26 families according to the rankings that had been conferred on

TABLE 6—Analysis of variance for crookedness scores of 486 trees, assessed by five observers[1,2]

Source of variation	d.f.	Mean square	F ratio	Probability	Estimate of component	Percentage of total variance[1]
Observers[1]	4	74.2366	12.54	0.0000	(0.1419)[1]	—
Families	17	33.7364	6.80	0.0000	0.2132	12.8
Blocks	8	9.2458	1.86	0.0882	0.0159	1.0
Observers × families	68	1.2161	1.70	0.0005	0.0185	1.1
Observers × blocks	32	4.7595	6.63	0.0000	0.0749	4.5
Families × blocks	136	4.9603	1.26	0.0565	0.0682	4.1
Observers × families × blocks	544	0.7175	1.37	0.0004	0.0640	3.9
Trees in (families × blocks)[2]	324	3.9374	7.49	0.0000	0.6824	41.0
Observers × trees in (families × blocks)	1296	0.5254			0.5254	31.6
Error	0					
Total	2429				1.6625	100.0

(1) The model for this analysis treats observers as a fixed effect; the estimate of the observers' main effect is therefore not a variance component, and it has been excluded from the total variance shown in the last column.

(2) The F-ratio test and variance estimate for trees-in-(families × blocks) are valid only if the (observers × trees) variance equals zero.

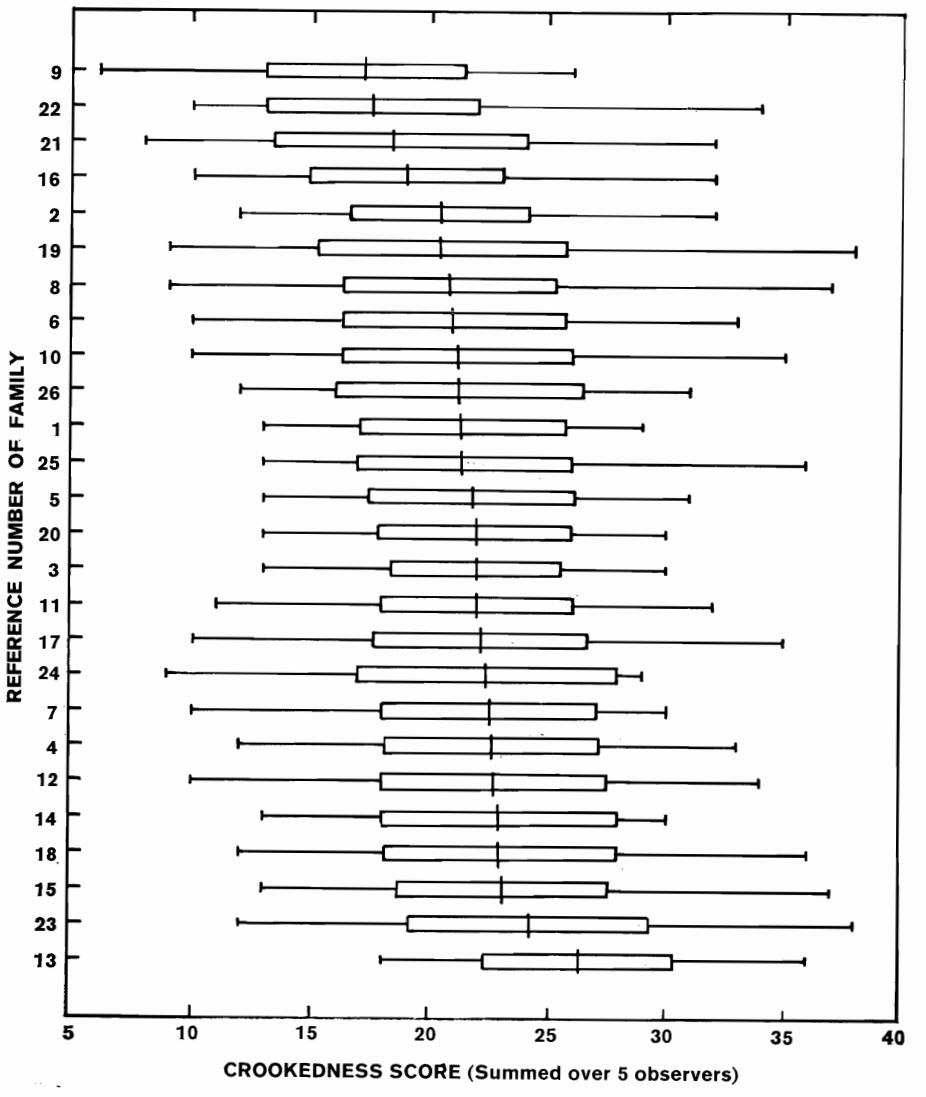

FIG. 6—Variation of the 26 open-pollinated families ranked in order of increasing crookedness. For each family the central vertical shows the mean, the hollow horizontal bar shows one standard deviation above and one below the mean, and the single horizontals show the range (based on original scores).

them quite subconsciously. The coefficient of concordance was 0.81; the test for the significance of this gave $z = 1.42$, which is about three times the value of z for $P = 0.001$. The observers, therefore, showed a strong concordance when their judgments were expressed by the criterion of the family means. The pattern of variation within and between families, based on five scores per tree, is shown in Fig. 6.

The error variance within families, which showed greater heterogeneity after the second transformation than before it, was also used as a criterion to rank the 26 families (i.e., from the one with the highest error variance down to the one with the lowest). For the original scores, the coefficient of concordance was 0.43 ($P < 0.001$); after the transformation it was 0.55 ($P < 0.001$). This suggests that the observers were quite sensitive to deviations from a mean, and that they agreed to a significant extent about the relative sizes of the deviations.

DISCUSSION

Consistent scoring by any one of the individual observers, and a high degree of concordance in their judgments, do not in themselves imply accuracy on the part of any one observer or, for that matter, of all five jointly (cf. Kendall and Babington Smith, 1939). To determine the relationship between the subjective scores and the deviations of each tree-trunk from a straight line, one would require objective physical measurements as well as scores. Unless these can be provided, the subjective method must remain somewhat questionable, whether it is used for estimating genetic parameters or as a basis for selection. Yet, although it is partly qualitative, it seems to depend quite heavily on quantitative estimation. If this is true, one may suppose that the eye can distinguish degrees of crookedness, and can relate them to a continuous scale that represents some more-or-less normally distributed variate, intrinsic to the trees themselves. The results of this study suggest, therefore, that in favourable circumstances the scores of a single observer could suffice to rank a series of genotypes — either indirectly by means of progenies or directly from clonal means — provided that enough individuals were scored in each progeny or clone. In other situations, however, scores by only one observer might suffer from serious inaccuracies. Increasing the number of observers, having each of them score the same trees independently, and using the mean score for each tree as the assessed value would increase the discriminative power of the technique. It would also improve the estimation of phenotypic and genotypic components of variance for individual traits and of covariance between traits — all of which are necessary for the construction of a commonly used selection index. Just how many observers should be used is not easy to determine. Generally the number would have to be a compromise between the desire for statistical reliability and the need to keep down costs; intuitively it is suggested that two or three might be suitable, whereas Hopkins (1950), in his work with flavours, used 30 or more. Some of the effects that one would expect from a variation in the number of observers, with respect to the crookedness data, are shown in Table 7.

The heterogeneity displayed by the error variances of families deserves special attention. Unlike that for blocks, it could not be regarded as dependent on the means and, again unlike that for blocks, its response to the transformation — expressly designed to stabilise the variance — was to emerge stronger than before. These observations, the concordance in the judgments about the variation within families, and the fact that the plots had been properly randomised, seem to rule out some possible explanations for the heterogeneity — e.g., that it was brought about by the environment, that it resulted from peculiar scoring by one or more of the observers, or that it arose quite by chance. The correct explanation may well be that it reflects

TABLE 7—Example of statistical effects produced by varying the number of observers assessing each tree (based on estimates of variance components from this study)

No. of observers	Phenotypic variance	Repeatability(1)	Heritability
1	1.64	0.55	0.29
2	1.27	0.71	0.38
3	1.15	0.79	0.42
4	1.10	0.83	0.44
5	1.05	0.86	0.46
6	1.03	0.88	0.47
7	1.01	0.90	0.48
8	1.00	0.91	0.48
9	0.99	0.92	0.49
10	0.98	0.93	0.49
20	0.94	0.96	0.51
30	0.93	0.97	0.52

(1) Repeatability measures the correlation between repeated measurements of the same individual. It may be defined as the ratio of the between-individual (i.e., trees) component of variance to the total variance (Falconer, 1960 pp. 142-9).

genetic phenomena. As explained previously (Bannister, 1969) the families came from diverse physical backgrounds, which may have brought about wide differences in mating patterns. That was perhaps one contributing factor. Another possibility is that the genotype, insofar as it affects crookedness, varies in such a way that the gametes produced by one tree have a much greater genetic diversity than those from another. If this were true, it would suggest that the inheritance of crookedness depends heavily on a few individually powerful genes, rather than on a nebulous "polygenic" system.

The results of the analyses of variance remain somewhat obscure, because in no case did the data conform with the mathematical model. This is assuredly a very common problem, which has been considered by several people (e.g., Cochran, 1947: Scheffé, 1959); but there appears to be no satisfactory technique, at least for biologists, for determining how much one can trust results that have been obtained by riding rough-shod over specific imperfections of fit. According to Scheffé (1959), abnormal kurtosis can have serious effects on inferences about variances; unequal error variances, combined with imbalance in the sub-classes, may also cause trouble. Since most of the crookedness analyses were based on severely imbalanced data with heterogenous errors, and since the method using the unweighted plot means may lead to biased estimates even when the error variances are homogeneous, the results could be doubly suspect. But suspicion should be allayed by the following:

(a) The results of analyses before and after transformation were practically the same, despite considerable changes in the variance;

(b) The results of an orthogonal analysis, using part of the data, were in good agreement with those based on complete, but imbalanced, sets of data;

(c) To check the adequacy of the approximate method of analysis, two different sets of data from the same experiment were analysed in that way and by the method of fitting constants. Here too the results were in good agreement (e.g., *see* Bannister, 1969).

Apparently, therefore, the combination of unequal numbers and heterogeneity of variance did not seriously disturb or invalidate the analyses.

One other theoretical requirement underlying the analysis of variance is that the experimental errors should be independent. The question of whether or not the data in this study satisfy that requirement has not been examined statistically; but it must be admitted that, since each observer scored all the trees within a plot consecutively, the scores within any observer-plot combination may have had an artificial, positive correlation bestowed on them unwittingly. Such a serial correlation can sometimes have very serious effects on the outcome of an analysis of variance (Scheffé, 1959); but in this study it seems that if it were present it would affect only the observer-related part of the composite error; the trees-in-plots part should be free from correlation, because that arose from genotypes arranged at random within the family. And if that were true, in the analyses based on the mean of five scores per tree only about one-tenth of the error would have been affected. Correlation, therefore, may have led to slight under-estimation of error variance, and consequently to such things as a slight over-emphasis on differences between families, but it would not be a justifiable reason for rejecting the main conclusions. In retrospect, it is realised that in an experiment of this kind scoring should ideally be directed so that each observer encounters a completely random sequence of individual trees, rather than a random sequence of plots with several trees in each. Such a requirement would be easily satisfied by randomised blocks in which the experimental units are single-tree plots.

CONCLUSIONS

The general conclusions from this work are in accord with much that has been learned before, in other fields of research:

(1) Subjective scoring can quickly give results of practical value, at least for some attributes under suitable conditions;

(2) Statistical analyses using the subjective scores of only one observer are potentially misleading (cf. Snedecor, 1946);

(3) If one can make an analysis based on the means of several observers, a failure on the part of individual observations to conform with a recognisable statistical model should cause little concern. The following comments by Finney (1949) are apt:

> ". . . general statistical theory teaches that means of several independent observations, under very wide conditions, have distributions more nearly normal than the distributions of individuals . . . it is the means . . . rather than the individual responses that are important. Again, experience has indicated that modification of the statistical analysis . . . in order to allow for the inconstancy of the variance . . . adds much to the labour but does not make any appreciable difference to the conclusions unless the variation in variance is very great."

For *P. radiata* in particular, the subjective assessment of crookedness is a practicable method, at least in stands where a wide range in potential crookedness has been able to express itself. The technique appears to be very good for discriminating and ranking groups of trees, and rather less satisfactory — although still useful — for studies that depend on the analysis of variance.

ACKNOWLEDGMENTS

I wish to thank Mr I. A. Andrew for testing frequency distributions and for clarifying various points; Mr R. M. H. Scott and Dr R. D. Burdon for helpful discussion; Professor B. I. Hayman for advice; Mr S. Lowe for solving a problem in integration; and Messrs R. Wiblin, D. Moug, R. Newcombe, and E. Kearns for assistance in the field.

REFERENCES

BANNISTER, M. H. 1969: An early progeny trial in **Pinus radiata**. 1. Juvenile characters. **New Zealand Journal of Botany 7:** 316-35.

COCHRAN, W. G. 1947: Some consequences when the assumptions for the analysis of variance are not satisfied. **Biometrics 3:** 22-38.

FALCONER, D. S. 1960: "Introduction to Quantitative Genetics". The Ronald Press Co., New York.

FINNEY, D. J. 1949: The choice of a response metameter in bio-assay. **Biometrics 5:** 261-72.

HARRIS, J. A. 1913: On the calculation of intra-class and inter-class correlations when the number of possible combinations is large. **Biometrika 9:** 446-72.

HENDERSON, C. R. 1959: Design and analysis of animal husbandry experiments. Pp. 1-55 **in** "Techniques and Procedures in Animal Production Research", American Society of Animal Production, Beltsville, Maryland.

HOPKINS, J. W. 1950: A procedure for quantifying subjective appraisals of odour, flavour, and texture of food stuff. **Biometrics 6:** 1-16.

KENDALL, M. G. 1948: "The Advanced Theory of Statistics" Vol. 2, pp. 205-6. Griffin, London.

KENDALL, M. G. and BABINGTON SMITH, B. 1939: The problem of m rankings. **Annals of Mathematical Statistics 10:** 275-86.

SCHEFFÉ, H. 1959: "The Analysis of Variance". John Wiley & Sons, New York.

SNEDECOR, G. W. 1946: **In** Queries and answers. **Biometrics Bulletin 2:** 123-4.

——— 1956: "Statistical Methods" (5th ed.). Iowa State University Press, Ames.

SHELBOURNE, C. J. A. and NAMKOONG, G. 1965: Photogrammetric technique for measuring bole straightness. **Proceedings 8th Southern Conference on Forest Tree Improvement:** 131-6.

TUKEY, J. W. 1949: One degree of freedom for non-additivity. **Biometrics 5:** 232-42.

YATES, F. 1934: The analysis of multiple classifications with unequal numbers in the different classes. **Journal of the American Statistical Association 29:** 51-66.