European Journal of Plant Pathology **100**: 395–412, 1994.

Subjective components of mildew assessment on spring barley

A. C. NEWTON[1] and C. A. HACKETT[2]
[1] *Mycology & Bacteriology Department, Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, UK;* [2] *Scottish Agricultural Statistics Service, Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, UK*

Accepted 23 August 1994

Key words: mildew, *Erysiphe graminis* f.sp. *hordei*, barley, assessment, keys, DISTRAIN, field trials, linear regression analysis

Abstract. The severity of mildew on barley is usually assessed visually and this leads to variation between different scorers. Field assessments by four assessors were analysed to determine the nature and degree of subjective discrepancies between assessors. Two inexperienced assessors failed to detect a major effect of nitrogen due to differences in the interpretation of a scoring system. A computer-based training programme was evaluated for standardising assessments, and was found to improve assessors' accuracy. Linear regression analysis was used here to resolve the error variance into components representing the accuracy and precision of the assessors. Plots of the cumulative differences between the estimate of disease severity by each assessor and the best estimate were used to display how the discrepancies varied with the level of disease. Some modifications to the barley field scoring system are suggested to improve comparability between assessors.

Introduction

Visual assessments are the most common means of quantifying mildew infection on cereals. However, such methods are highly subjective and depend upon the skill of the assessor. In practical terms we are often required to train new assessors at frequent intervals so comparisons from year to year must take into account this factor in addition to all other variables. There are two approaches to alleviating this problem: either move to more objective assessment techniques, or adapt the training methods and scoring system to minimise variability in standards.

Objective assessment methods can be divided into two classes: non-destructive and destructive. Among the former is image analysis, which can be used effectively on detached leaf material and correlates well with highly objective infection frequency data [Newton, 1989]. Other methods are chlorophyll fluorescence analysis and imaging of the stress induced by disease [Daley et al., 1989], and spectral analysis from various imaging techniques [Nilsson, 1991; Nutter, 1989; Nutter et al., 1990], but such methods are indirect and tend to measure the stress or other secondary effects induced by pathogens. Thus they must be validated by other

measures of pathogen development. They are also relatively expensive high technology methods. Destructive sampling methods include assessing the biomass of the fungus using chromatographic techniques such as sterol quantification [Newton, 1989, 1990] or enzyme-linked immunosorbent assay (ELISA) [Newton and Thomas, 1993; Newton and Reglinski, 1993].

These methods all have several disadvantages that may offset the advantage of objectivity. They can be time consuming and often expensive, especially sterol assays using gas chromatography. They require accurate sampling from each plot, which is time consuming and may involve another subjective element in the choice of the sample of plants. Furthermore, variations in the physiological state of the plants may affect the sensitivity of the assays. However, these methods have the advantage, besides objectivity, that components of fungal growth may be detected which are not visible to the naked eye, such as more branched hyphae giving denser colonies. Such denser colonies may be more damaging to their host by increasing assimilate drain, although they may be more compact and therefore reduce green leaf area less for a given amount of pathogen biomass.

There appears to be no adequate alternative to visual assessments for the majority of the direct assessments of disease levels in the field. The method is rapid, inexpensive and the experiments may be scored repeatedly to follow the progress of disease. Nutter et al. [1991] define accuracy as the closeness of a disease assessment to the true value, and precision as the consistency of repeated assessments of the same units. The consistency of repeated assessments by the same individual is referred to as the intra-rater repeatability and the consistency among several assessors is referred to as the inter-rater reliability. Visual assessments need to be based upon a scoring system that is accurate, precise and reliable for different assessors.

Visual scoring systems are of three general types. There are completely quantitative scales, where the assessor records the disease severity as a percentage, often assisted by using standard diagrams with graded amounts of disease [James, 1971]. An alternative is an ordinal scale, where the disease severity is recorded as a series of levels according to classes in a descriptive key. Thirdly, there are Horsfall-Barratt scales [Horsfall and Barratt, 1945; Lindow, 1983] where there are twelve classes corresponding to ranges of disease severity, with more levels near 0% and 100% and fewer near 50% to compensate for the high variances often found at this level (Horsfall and Barratt, 1945]. Hau et al. [1989] found that inexperienced scorers are less accurate using a completely quantitative scale than using disease classes, but suggested that experienced scorers might be able to use a quantitative scale. O'Brien and van Bruggen [1992] compared two ordinal scales with a Horsfall-Barratt scale for assessing corky root of lettuce and found that no single scale was the best in all situations.

A commonly used scale for assessing mildew is the British Association of Plant Breeders (BAPB) 1–9 rating system. This has descriptions of the

extent of the disease for each class, and also the percentage severity corresponding to the middle of each class. Like the Horsfall-Barratt scale, there are more levels close to 0% but the classes become steadily wider with increasing disease. In this paper variation between four assessors, one experienced and three inexperienced, is investigated in a field experiment using the BAPB rating system.

Various computer-aided training programmes have been devised to overcome differences between assessors. Nutter [1993] found that a one hour training session improves assessments of disease severity. One such programme is DISTRAIN [Tomerlin and Howell, 1988], where the user estimates the severity of disease as the percentage area of lesions on a leaf drawn on a computer monitor, and then receives feedback of the actual severity. This programme was used to assess variation in estimating disease severity among twelve assessors, and to see whether their scoring ability improved after a period of training.

Materials and methods

Field experiment

The design of the field experiment was a split-split plot comparing presence or absence of fungicide disease control (main plot treatment), two levels of nitrogen (sub-plot treatment) and 20 spring barley monocultures or mixtures (sub-sub-plot treatment). No mildew was observed on the plots receiving the fungicide treatment and these plots are not considered in the following analysis. The two nitrogen levels N1 and N2 were 25 and 100 kgN/ha respectively (P and K were 12.5 Kg/ha for N1 and 50 Kg/ha for N2). Five commercial cultivars, Doublet, Tweed, Natasha, Triumph and Camargue, five partial resistance breeding lines, 7163/68/5, 7204/31r/9, 7526/7m/2, 9319/38r/7, 9855/59/2, and some three-component mixtures of each set were grown. The plots were 1.9 × 1.22 m (excluding gaps). The experiment was carried out at SCRI in 1993. One of the early season mildew assessments was carried out by four assessors at the same time. At this point the plants were growing actively and were not suffering from any mineral deficiencies or other symptoms not due to mildew. Only one assessor (B) was experienced; the other three (A, C and D) had never scored disease in field trials. The only training they were given was in comprehension of the BAPB key (Table 1) and a few field plot examples.

The choice of a scale for the analysis of such data has been discussed in general by Agresti [1990] and in the case of plant disease by Forbes and Jeger [1987] and O'Brien and van Bruggen [1992]. The BAPB 1–9 scale is ordinal, but corresponds to percentages of a continuous variable, disease severity. O'Brien and van Bruggen [1992] analyse untransformed ordinal scores and percentages while Forbes and Jeger [1987] found that a logit

Table 1. The BAPB scoring system for mildew

Score	Infection	Description
1	0%	No infection observed
2	0.1%	3 colonies per tiller
3	1%	5 colonies per leaf
4	5%	Lower leaves appear 1/4 infected
5	10%	Lower leaves appear 1/2 infected
6	25%	Leaves appear 1/2 infected 1/2 green
7	50%	Leaves appear more infected than green
8	75%	Very little green leaf tissue left
9	100%	Leaves dead – no green leaf left

The scoring system was extended to include intermediate scores of 1.5, 2.5, 3.5, 4.5, 5.5, 6.5, 7.5, and 8.5 which equate to 0.05%, 0.4%, 3%, 7%, 17%, 37%, 62% and 87%.

transformation of disease percentages gave a closer fit to a normal distribution. Agresti [1990] suggested that a range of scoring systems for the ordinal categories should be examined to see whether the conclusions of an analysis were affected. The data from the field experiment was analysed using four scales, untransformed 1–9 scores, percentages, logit-transformed percentages and angular-transformed percentages. Analysis of variance was used to examine the data for differences between nitrogen levels, cultivars and assessors, using GENSTAT 5 Release 2.2. The residuals were compared to a normal distribution using a probability plot correlation test [Filliben, 1975].

The mean of the four assessments on each plot was calculated and the plots of the experiment were sorted into order of increasing disease on this basis. The cumulative differences between the experienced assessor and the three inexperienced assessors were calculated and plotted against the mean of the four assessments (Fig. 1). If the differences between the inexperienced and experienced assessors are random and the inexperienced assessor is equally likely to under- or overestimate, the cumulative differences will be close to zero. However, if the inexperienced assessor's estimates are consistently higher/lower, the cumulative differences will increase/decrease steadily. Similar graphs of cumulative differences, known as cusum charts, are used regularly in industrial quality control to detect small deviations from a reference point as quickly as possible [Wetherill, 1977].

Computer experiment

The DISTRAIN programme was used to draw leaves with lesions of powdery mildew on a computer monitor. The programme has options for a high, medium, low or random disease severity range. A low range was

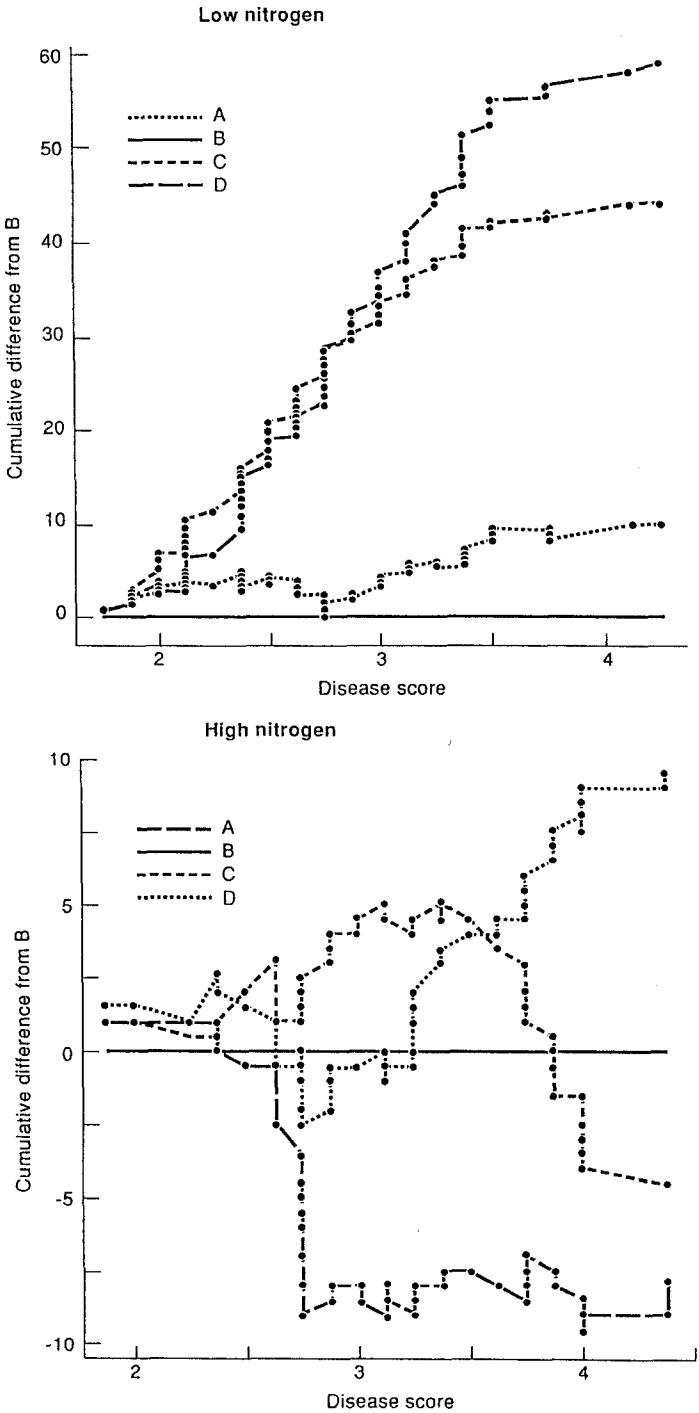

Fig. 1. Cumulative difference from experienced assessor B plotted against mean disease score for the plots receiving low (top) and high (bottom) nitrogen.

chosen to correspond as closely as possible to observations in the field experiment. DISTRAIN provides feedback on the true disease severity of each leaf after this has been assessed but initially this feedback was suppressed. Twelve assessors, with various degrees of experience in scoring barley mildew and other diseases, were each asked to estimate the disease severity (to the nearest integer) of a sample of 50 random leaves. This gave an initial estimate of the assessor's ability to judge disease severity. Each assessor then scored a second sample of 50 leaves, receiving feedback after each assessment. The samples of leaves were different for each assessor. Assessors B, C and D were the same people as in the field experiment. For each assessor, plots of the cumulative difference from the true disease percentage were used to assess improvements in the accuracy of scoring.

Linear regression was also used to evaluate the scores. Nutter et al. [1993] regressed original assessments on repeated assessments to examine intra-rater repeatability, and regressed assessments of one rater on those of another to examine inter-rater reliability. However, this required repeated assessments of the same samples by different assessors and at different times. Here the estimate was regressed on the true severity for each assessor. Hau et al. [1989] discuss this approach and stress that it is important to test the slope and the intercept simultaneously to see whether they are different from the ideal values of 1 and 0. Amanat [1977] derives a permissible range of fitted lines based on an acceptable range of estimates from 25–35% for a true value of 30%, and 3.75–7% for a true value of 5%. For the disease levels found in this computer experiment, we will regard 7.5–12.5% as permissible for a true disease level of 10% and 0.5–1.5% as permissible for a true level of 1%. This gives an upper limiting line with intercept 0.28 and slope 1.2, and a lower limiting line with intercept –0.28 and slope 0.78.

A line with slope 1 and intercept 0 was also fitted for each assessor, and the residual sum of squares was resolved into two components, pure error and lack of fit [Draper and Smith, 1981]. The pure error term measures the variation in the estimates corresponding to each true disease severity e.g. five successive leaves with a true disease severity of 4% may have been estimated as 3%, 4%, 4%, 5%, 7% and their contribution to the pure error term is the sum of squares about their mean of 4.6%. This component measures the precision of the data. The lack of fit component measures the discrepancy between the mean of the estimates (4.6% in the above example) and the fitted value which, in this model, is the true disease severity (4%). This component measures the accuracy of the data. Thus linear regression will detect any improvements in accuracy and precision for each assessor after the period of training.

Results

Field experiment

Analysis of variance of the 1–9 scores, percentages, logit-transformed percentages and angular-transformed percentages gave very similar results. An examination of the residuals indicated that for this experiment the 1–9 scores and the angular-transformed percentages satisfied the requirement of normality while the percentages and logit-transformed percentages did not. The residuals from the analysis of the 1–9 scores had the highest correlation with the normal order statistics and the results from this analysis are given here.

There were significant main effects of barley genotype and nitrogen level ($p < 0.001$) and a possible interaction between these ($p = 0.04$). There were also significant differences between assessors ($p = 0.006$) and significant interactions between assessor and nitrogen level ($p = 0.014$) and assessor and barley genotype ($p = 0.001$). Table 2 shows the mean disease score for each genotype and assessor, indicating those scores which are

Table 2. The mean mildew score for each genotype and assessor, arranged in increasing order of the score of the experienced assessor, B. The SED is 0.26 for differences between assessors and 0.25 for differences within assessors. **underlined bold type** indicates a mean score that is significantly different from that of assessor B

	Assessor			
Genotype	A	B	C	D
Doublet (Do)	2.0	1.6	**2.5**	2.1
Do/Tw/Tr (mix)	2.3	2.1	**2.7**	2.6
7204/31r/9 (72)	2.0	2.2	2.6	2.2
Triumph (tr)	2.1	2.3	**2.9**	2.8
Camargue (Ca)	2.3	2.3	**3.1**	**3.3**
Tw/Tr/Ca (mix)	2.3	2.3	2.5	2.5
Do/Tw/Na (mix)	2.2	2.4	2.9	2.8
Do/Tr/Ca (mix)	2.1	2.5	2.7	2.6
Do/Tw/Ca (mix)	2.3	2.6	2.9	2.9
Natasha (Na)	3.0	2.7	3.1	**3.3**
72/75/93 (mix)	2.8	2.7	2.9	**3.6**
9319/38r/7 (93)	2.6	2.8	3.3	3.3
Tweed (Tw)	2.8	2.8	3.2	**3.6**
7526/7m/2 (75)	3.3	3.1	3.6	**4.4**
71/93/98 (mix)	3.5	3.2	3.6	**3.9**
75/93/98 (mix)	3.3	3.2	3.3	**3.8**
72/93/98 (mix)	3.2	3.4	3.3	3.8
7163/68/5 (71)	3.4	3.5	3.7	**4.3**
71/75/98 (mix)	3.7	3.7	3.4	**4.4**
9855/59/2 (98)	3.9	3.8	3.4	4.3

significantly different from the experienced assessor, B, on whose scores the table is ordered. At low disease levels assessor C tended to record significantly higher scores than B: at higher disease levels D recorded higher scores than B. Table 3 shows the mean disease level for each combination of assessor and nitrogen level. There were no significant differences between B and A. At the low level of nitrogen assessors C and D recorded significantly higher levels of disease than assessors A and B. At the high level of nitrogen there were no significant differences between assessors. A and B found a large difference between the two levels of nitrogen which C and D failed to detect. Data from other experiments [Newton and Thomas, unpublished data] indicate that the nitrogen effect detected by A and B was genuine.

In Table 3 some additional summary statistics are presented which illustrate differences between the assessors for each nitrogen treatment. At the high level of nitrogen, C differed from A, B and D in the lower standard deviation, the lower variance ratio for differences between cultivars and the lower correlation with B. These indicate that C was using a restricted range of the scale. D, who had a slightly higher mean than A, B and C, had a similar standard deviation and variance ratio to A and B and was highly correlated with B. This indicates that D was consistently higher than B across the range of the scale.

At the low level of nitrogen the differences between the assessors were more marked. A had the highest correlation with B and their scores were within 0.5 for 48 out of the 60 plots. D had a lower correlation with B, and a higher mean, standard deviation and variance ratio. C had the lowest correlation with B, a higher mean, and a slightly lower standard deviation and variance ratio.

Figure 1 summarises the differences between the assessors for the low and high levels of nitrogen. For the low nitrogen the lines for assessors A, C and D were all above the zero line, indicating that they tended to have

Table 3. Differences between assessors A, B, C and D in the field experiment. The low and high levels of nitrogen are considered separately. The mean disease score, the standard deviation, the variance ratio for differences between cultivars, the correlation with experienced assessor B and the percentage of scores within 0.5 of B's score are shown. The SED for comparing the means is 0.16

	Low nitrogen Assessor				High nitrogen Assessor			
	A	B	C	D	A	B	C	D
Mean	2.4	2.3	3.0	3.2	3.1	3.2	3.2	3.4
Standard deviation	0.70	0.68	0.61	0.90	0.80	0.68	0.47	0.79
Variance ratio	4.24	6.35	3.49	8.90	8.49	8.42	1.84	8.50
Correlation with B	0.64	–	0.43	0.58	0.71	–	0.61	0.72
% within 0.5 of B	80	–	53	33	82	–	87	87

higher scores than B. For A the difference was slight for disease levels less than 3 but above this A tended to score more highly than B. For D the cumulative difference increased steadily, indicating that estimates were consistently high throughout the range. For C the cumulative difference increased steadily up to a score of about 3.5 but did not increase beyond this i.e. C was similar to B for scores above 3.5. For the higher nitrogen the cumulative differences were generally lower. The line for A was initially close to that for B. The lines then diverged, due to a few plots with disease levels between 2.5 and 3 which A scored lower than B. Above this the lines are roughly parallel, indicating similar assessments. D had similar scores to B for disease levels below 3 but above this the cumulative differences increased steadily, indicating consistently higher scoring. The cumulative differences for C increased over the range below 3 and then decreased steadily, indicating that C's estimates were higher than B's over the range below 3 and lower over the upper range.

Computer experiment

There was considerable variation between the abilities of the twelve assessors to determine disease percentages correctly, both without and with feedback on the true disease severity. Figures 2a–2l summarise the results for each assessor. The estimate, and the cumulative difference between the estimate and the true score are plotted for each assessor, with and without feedback, against the true disease percentage. In the absence of feedback the cumulative difference plots increased steadily for four assessors, B, D, I and M, indicating that these people recorded disease levels consistently higher than the true levels (e.g. a true value of 8% might be scored as 15%, 20%, 30%, 35% respectively). Five assessors, C, F, J, K and L, recorded values consistently lower than the true levels. The cumulative difference plot for H was close to zero except for the highest points, indicating that H was accurate in the lower part of the range but variable at the top of the tested range. E and G generally agreed well with the true results but tended to underestimate disease at the bottom of the range. When feedback was available every assessor improved. The learning effect was investigated by plotting the differences from the true value in order of appearance: these plots showed that the differences tended to be highest for the first five scores and hence these scores are excluded from Figure 2 and from the regression analysis. The estimated and true scores were much closer and the maximum cumulative differences decreased, although some still showed systematic trends, especially D.

Table 4 gives the results of the linear regression analysis and also describes the experience of each assessor in scoring cereal mildew, other crops/diseases and in using the DISTRAIN programme. The slopes and intercepts were tested for significant differences from 1 and 0, using

Student's t-test. However, for this investigation such tests are unsatisfactory because the outcome depends on the residual variation about the fitted line for each assessor. This leads to the conclusion that for assessor F an intercept of −1.1 is significantly different from 0, while for assessor D, with more residual variation, an intercept of −1.8 is not significantly different from 0. A more satisfactory approach is to use the upper and lower bounds determined by the method of Amanat [1977], which are chosen rather subjectively, but are the same for all assessors. Without feedback, only two assessors had fitted lines in the acceptable range; this increased to five with feedback. The percentage variation accounted for by the ideal line, y = x, increased for all assessors except E, the most experienced assessor. Most of the reduction in error variation was attributed to the lack of fit of the ideal line i.e. feedback improved the accuracy of the assessors. Only four assessors (C, D, G and K) showed much decrease in the pure error variation i.e. improvement in precision. In the absence of feedback, the experienced assessors tended to be more accurate than the inexperienced, but the difference was reduced by the period of training.

Discussion

These results demonstrate that there was considerable variation between the visual assessments of mildew by different people. The variation in the field experiment is not due simply to one assessor scoring consistently at a higher level than another, but the significant interaction in the analysis of variance shows that the cultivar and the nitrogen treatment affect the assessment. The high nitrogen treatment is expected to have more disease than the low nitrogen treatment, probably manifested as larger mildew colonies rather than a higher number of mildew colonies. The discrepancies between the assessors were much more noticeable at the lower level of nitrogen, and hence mildew, where assessors C and D overestimate the disease considerably. This would agree with Sherwood et al. [1983] who reported that large numbers of small lesions are overestimated compared to a smaller number of larger lesions occupying the same area, and that overestimation is inversely proportional to log(true disease area). From personal observation of the scorers, C and D took much longer to score the trial than A and B. It is also likely that C and D were more precise in their interpretation of the scoring key, which is highly specific at its lower end but more subjective from 4 (5%) upwards. This suggests that highly specific guidelines for scoring may exclude important subjective factors. Scoring guidelines for keys should have a carefully judged subjective element.

A fundamental problem with all assessments is whether the assessor is measuring the pathogen or the disease [Hau et al., 1989]. In the work reported here only early infection is observed and no damage other than

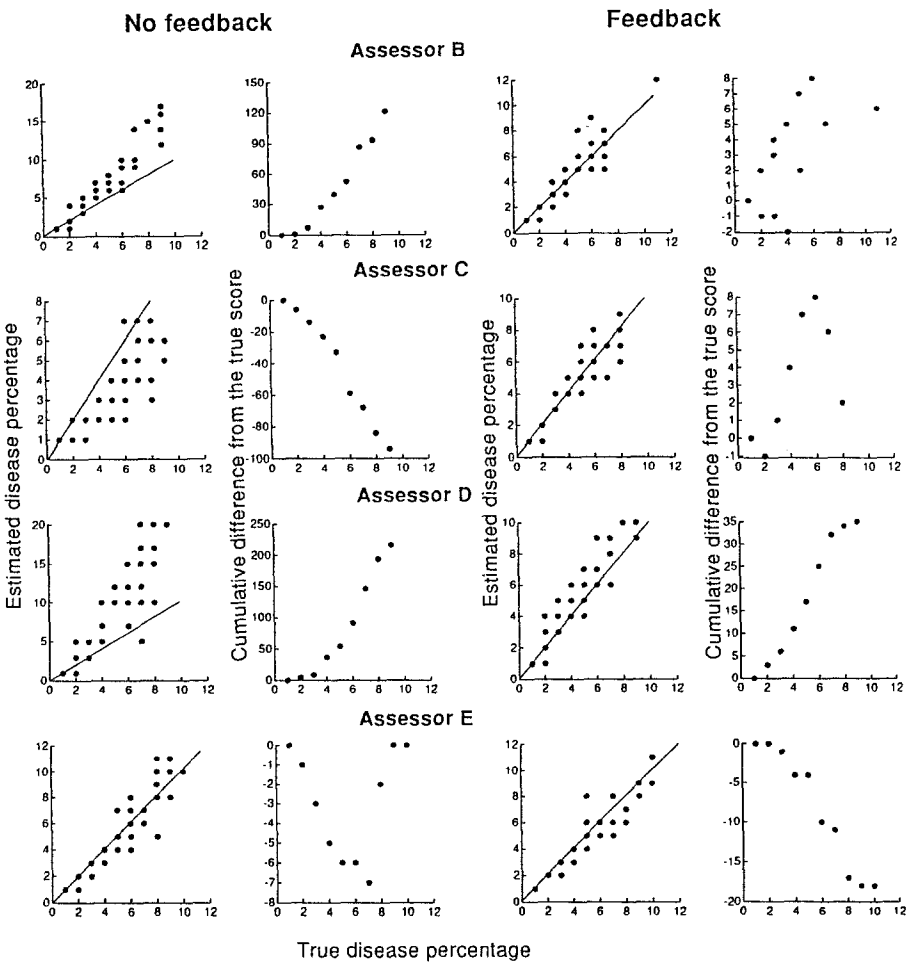

Fig. 2. Plots of estimated disease percentage and cumulative difference from the true disease percentage plotted against the true disease percentage, without and with feedback, for assessors B–M.

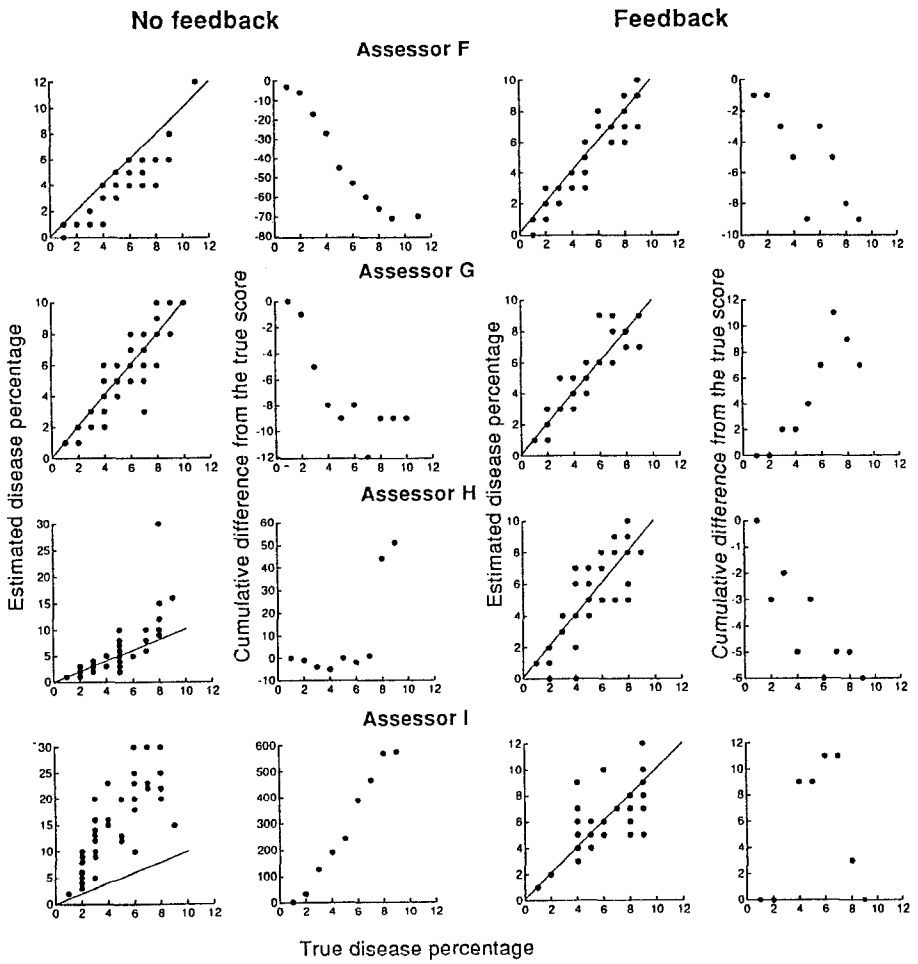

Fig. 2. Continued.

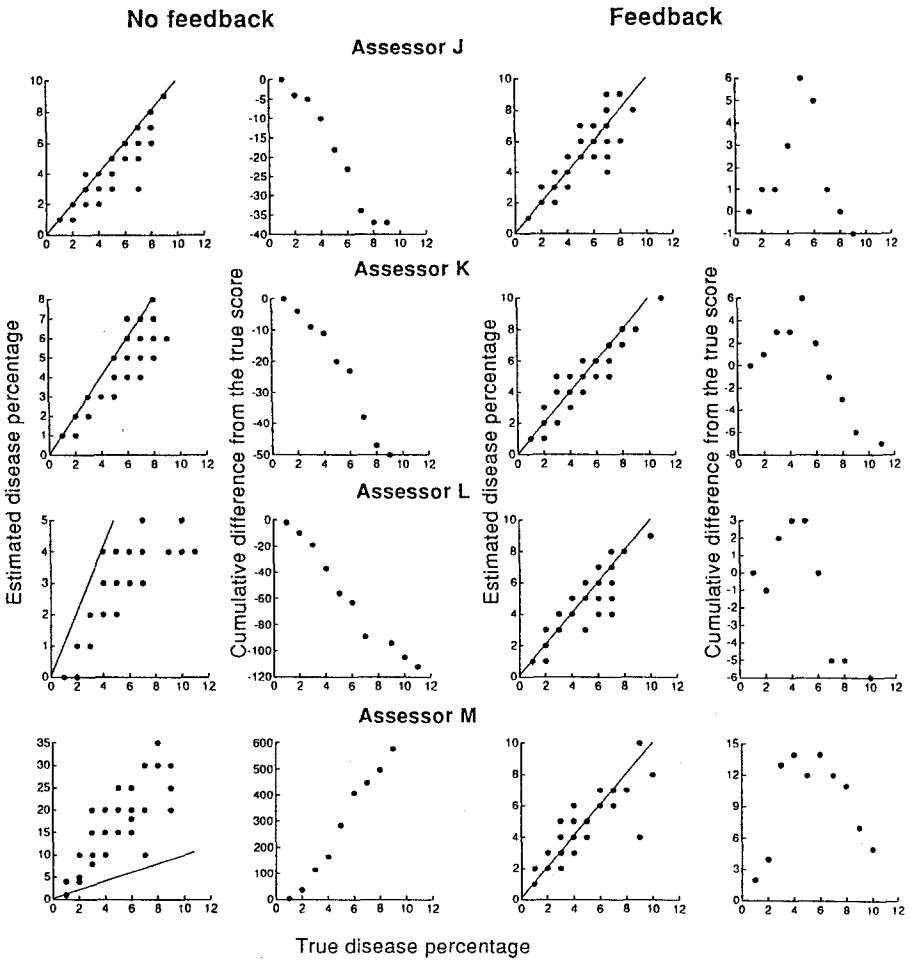

Fig. 2. Continued.

Table 4. The experience of each of the twelve assessors in the computer experiment and linear regression analyses of their performances. Experience with mildew (M), other crops/diseases (O) and DISTRAIN (D) is rated as none (–), some (+) or considerable (++). a is the intercept of the regression line, b is the slope, R^2 is the % variation accounted for by the linear regression y = a + bx, R_0^2 is the % variation accounted for by the line y = x, % P is the % variation due to pure error and % L is the % variation due to lack of fit. *, **, *** indicate an intercept significantly different from 0 or a slope significantly different from 1 with P < 0.05, 0.01 and 0.001 respectively. † indicates that the variation about the line y = x exceeds the variation about the overall mean. **Underlined bold** type indicates fitted lines in the permissible range –0.28 < a < 0.28, 0.78 < b < 1.22

	Experience			No feedback						Feedback					
	M	O	D	a	b	R^2	R_0^2	% P	%L	a	b	R^2	R_0^2	% P	%L
B	++	+	+	–1.5**	1.8***	88	34	11	55	**–0.2**	**1.1**	82	81	14	5
C	+	–	–	0.1	0.7***	65	†	31	†	0.5	0.9	85	83	13	4
D	+	–	–	–1.8	2.2***	77	†	21	†	0.3	1.1	88	78	11	11
E	++	+	+	–0.6	1.1	86	85	13	2	**0.0**	**0.9**	85	81	13	6
F	–	++	–	–1.1***	0.9	85	47	9	44	**–0.2**	**1.0**	87	86	9	5
G	–	–	++	–0.6	1.1	80	79	19	2	**0.2**	**1.0**	86	85	11	4
H	–	–	–	–2.7*	1.8***	57	42	30	28	**–0.2**	**1.0**	72	72	26	2
I	–	+	–	1.9	3.1***	65	†	24	†	0.8	0.8	65	63	30	7
J	+	++	–	**–0.04**	**0.8***	83	68	14	18	0.6	0.9	74	73	24	3
K	–	++	–	**–0.2**	**0.8***	82	56	15	29	0.6	0.9**	89	86	9	5
L	–	++	–	0.1	0.5***	73	†	16	†	0.4	0.9*	81	79	17	4
M	–	–	–	2.5	2.9***	68	†	22	†	1.3***	0.7***	70	59	25	16

observable mildew was present. However, the question becomes particularly important in later plant development stages and where disease is severe, and a measure of disease may correlate with yield loss better than the extent of observable pathogen development. However, the effects of the pathogen may not all be observable as disease symptoms and thus in some cases the pathogen may be more accurately related to yield loss.

There was also considerable variation between disease assessments on the computer-generated leaves. Only three scorers were involved in both the field and computer experiments so conclusions about consistency between field and computer are tentative. However, we note that D consistently over-scored in both experiments. C consistently under-scored on the computer experiment and also in the more quantitative range of the field experiment (5% disease and above). B, the experienced scorer of the field trial, also overestimated the disease level on the computer screen but to a lesser extent than D. If B also overestimated disease in the field trial this would affect the conclusions, but in the absence of an objective field assessment his scores are the best available. Among the assessors in the computer experiment, the most accurate tended to have considerable experience of field disease scoring or previous experience with the DISTRAIN programme. The least accurate had very little experience of

either. All except the two most accurate assessors showed an improvement in their accuracy with feedback on their performance. Four assessors also improved their precision. We intend to investigate whether this improvement is apparent in future field experiments. Parker and Royle [1993] investigated the effectiveness of DISTRAIN briefly but concluded that it did not successfully train two novice observers to provide accurate disease assessments. However, they still propose that this program could be used regularly, or prior to estimates in the field.

Methods of displaying and understanding variation between assessors have generally arisen from sensory experiments, where a panel of assessors evaluates different products. In such cases the ranks of the different products are of interest. In disease scoring, however, we need an accurate measure of the disease severity. Naes and Solheim [1991] discuss simple statistical methods for examining variation and illustrate how different scorers may use different parts of the scale. This occurred in our field data, for example in the plots with high nitrogen where D had a higher mean but similar variance to A and B, while C had a similar mean but lower variance. Hirst and Naes (1993) have used plots of cumulative ranks to identify those parts of the scale where assessors agree/disagree. We adapted this approach to give our plots of cumulative differences, which illustrate clearly the scoring pattern of each assessor.

In the computer experiment some assessors over-scored consistently, while some under-scored. Some authors report the tendency to over-score as more common, except at high disease levels [Forbes and Jeger, 1987; Hock et al., 1992]. Daamen [1986] demonstrated that low disease intensities were particularly difficult to estimate in mildew of wheat, having much higher variance than higher intensities, indicating that particular attention needs to be paid to this part of the scale. Sherwood et al. [1983] and Beresford and Royle [1991] report considerable overestimates of low disease levels using a direct percentage scale. Berger [1980] suggests that direct percentage or proportion assessments should be made rather than using 'arbitrary rating scales' in the interests of accuracy. However, most other evidence points to no increase in accuracy in the use of such direct assessments, indeed the bias from novices may be greater. It was to the goal of enabling novice assessors to avoid high bias that we particularly wish to pay attention and Hau et al. [1989] indicate that inexperienced assessors perform better, making more accurate assessments with class based systems. Stimulus-response curves show that assessments are different using category methods rather than continuous scales, but that neither approach was more accurate [Hebert, 1982]. Therefore, as all scales have disadvantages then the descriptive guidance associated with good scales should be addressed to help overcome the bias. While such bias can be scaled out in analysis, this could also be corrected for when converting from descriptive scale categories to percentage equivalents which could be more accurate if the descriptions were tailored to particular diseases.

From these findings several proposals can be made. First, the scoring system descriptions in the lower part of the key should be modified to direct the scorer to assess percentage infection on the lower leaves (Table 5). This gives a subjective element and eliminates the change in concept between each score category in the previous descriptions. We did not record disease levels beyond 10% so our scoring system remains unaltered above score 5. The new system is compatible with the percentage scale used by DISTRAIN. Our data suggests that DISTRAIN can be used to train and help standardise assessors before field scoring. While there is no substitute for scoring experience in the field, this may reduce the need for duplicating scoring to check the progress of inexperienced assessors and free resources for other, more objective scoring methods such as ELISA. It may be feasible to calibrate the visual assessments by measuring a small sample of plants using ELISA. As scorers A and B were both much faster and at least as accurate scoring the field trial, resources are likely to be better used to score a trial rapidly, but frequently. These results disagree with those of Parker and Royle [1993], who found rapid disease assessments on wheat to be the least precise. Further work is needed to investigate whether experience with the DISTRAIN program, or conventional field training by a colleague, would enable an inexperienced scorer to increase their speed without sacrificing accuracy. Berger [1980] reported that 'some people are reasonably good assessors, others are good but consistently estimate either high or low, and the remainder simply cannot make a satisfactory assessment'. However, our data indicate that training can improve all assessors' accuracy.

Our modified scale retains the original uneven increments between levels of the original descriptions, preferring to change the wording of the descriptions to guide assessors towards a more quantitative evaluation of the plants which in turn will lead to a closer approximation to the class

Table 5. SCRI revised scoring system for mildew

Score	Infection	Description
1	0%	No infection observed
2	**0.2%**	**1% infection on lower leaves**
3	1%	**5% infection on lower leaves**
4	5%	**25% infection on lower leaves**
5	10%	**50% infection on lower leaves**
6	25%	Leaves appear 1/2 infected 1/2 green
7	50%	Leaves appear more infected than green
8	75%	Very little green leaf tissue left
9	100%	Leaves dead – no green leaf left

Intermediate scores of 1.5, 2.5, 3.5, 4.5, 5.5, 6.5, 7.5, and 8.5 equate to **0.1%, 0.6%**, 3%, **7.5%, 17.5%, 37.5%, 62.5%** and **87.5%**.
Changes from Table 1 are in **bold.**

means. Our modified scale (Table 5) also changes descriptive concept in the middle and maintains large categories up to 100%. This needs to be fully investigated with appropriate data in this range. A major factor which our data showed is the effect of nitrogen difference on scores and its interaction with the scorer. This highlights the subjective nature of scales and the need for tailoring class descriptions away from the subjective features which bias the assessment towards more quantitative characteristics. We have addressed this in the changes to our scale.

Acknowledgements

We acknowledge the four field assessors, Damian Cox, David Guy, Sheena Main, and Donald Robertson. We also thank the computer test 'guinea pigs', Tom Connolly, Jim Duncan, George Goleniewski, Julian Harrison, Philip Killington, Bob Lowe and Jess Searle as well as the authors! These are not in order (A–M) to avoid identification with individual data sets! We also thank the Crop Genetic Department and Estate staff who helped sow the trial. This work was supported by funds from the Scottish Office Agriculture and Fisheries Department.

References

Agresti A (1990) Categorical Data Analysis. Wiley, New York

Amanat P (1977) Modellversuche zur Ermittlung individueller und objektabhängiger Schätzfehler bei Pflanzenkrankheiten. Diss. Universität Giessen

Beresford RM and Royle DJ (1991) The assessment of infectious disease for brown rust (*Puccinia hordei*) of barley. Plant Pathology 40: 374–381

Berger RD (1980) Measuring disease intensity. pp 28–31. In: Crop Loss Assessment. Proc EC Stakman Commem Symp Misc Pub 7, Agric Exp Stn, University of Minnesota, St Paul

Daamen RA (1986) Measures of disease intensity in powdery mildew (*Erysiphe graminis*) of winter wheat. 1. Errors in estimating pustule number. Netherlands Journal of Plant Pathology 92: 197–206

Daley PF, Raschke K, Ball JT and Berry JA (1989) Topography of photosynthetic activity of leaves obtained from video images of chlorophyll fluorescence. Plant Physiology 90: 1233–1238

Draper NR and Smith H (1981) Applied Regression Analysis. Second edition. Wiley, New York

Filliben JJ (1975) The probability plot correlation coefficient test for normality. Technometrics 17: 111

Forbes GA and Jeger MJ (1987) Factors affecting the estimation of disease intensity in simulated plant structures. Zeitschrift für Pflanzenkrankheiten und Pflanzenschutz 94: 113–120

Hau B, Kranz J and König R (1989) Fehler beim Schätzen von Befallsstärken bei Pflanzenkrankheiten. Zeitschrift für Pflanzenkrankheiten und Pflanzenschutz 96: 649–674

Hebert TT (1982) The rationale for the Horsfall-Barratt plant disease assessment scale. Phytopathology 72: 1269

Hirst D and Naes T (1993) A graphical technique for assessing differences among a set of rankings. Journal of Chemometrics (in press)

Hock J, Kranz J and Renfro BL (1992) Tests of standard diagrams for field use in assessing the tarspot disease complex of maize (*Zea mays*). Tropical pest management 38: 314–318

Horsfall JG and Barratt RW (1945) An improved grading system for measuring plant diseases (Abstr.) Phytopathology 35: 655

James WC (1971) An illustrated series of assessment keys for plant diseases, their preparation and usage. Canadian Plant Disease Survey 51: 39–65

Lindow SE (1983) Estimating disease severity of single plants. Phytopathology 73: 1576–1581

Naes T and Solheim R (1991) Detection and interpretation of variation within and between assessors in sensory profiling. Journal of Sensory Studies 6: 159–177

Newton AC (1989) Measuring the sterol content of barley leaves infected with powdery mildew as a means of assessing partial resistance to *Erysiphe graminis* f.sp. *hordei*. Plant Pathology 38: 534–540

Newton AC (1990) Detection of components of partial resistance to mildew (*Erysiphe graminis* f.sp. *hordei*) incorporated into advanced breeding lines of barley using measurement of fungal cell wall sterol. Plant Pathology 39: 598–602

Newton AC and Reglinski T (1993) An enzyme-linked immunosorbent assay for quantifying mildew biomass. Journal of Plant Diseases and Protection 100: 176–179

Newton AC and Thomas WTB (1993) Evaluation of sources of partial resistance to mildew in barley using enzyme-linked immunosorbent assay and other assessment methods. Euphytica 66: 27–34

Nilsson HE (1991) Hand-held radiometry and IR-thermography of plant diseases in field plot experiments. International Journal of Remote Sensing 12: 545–557

Nutter FW (1989) Detection and measurement of plant disease gradients in peanut with a multispectral radiometer. Phytopathology 79: 958–963

Nutter FW (1993) Improving the accuracy and precision of disease assessments: Selection of methods and use of computer-aided training programmes. Sixth International Congress of Plant Pathology, Montreal, Abstract S9.3, 11

Nutter FW, Gleason ML, Jenco JH and Christians NC (1993) Assessing the accuracy, intra-rater repeatability and inter-rater reliability of disease assessment systems. Phytopathology 83: 806–812

Nutter FW, Littrell RH and Brenneman TB (1990) Utilization of a multispectral radiometer to evaluate fungicide efficacy to control late leaf spot in peanut. Phytopathology 80: 102–108

Nutter FW, Teng PS and Shokes FM (1991) Disease assessment terms and concepts. Plant Disease 75: 1187–1188

O'Brien RD and van Bruggen AHC (1992) Accuracy, precision and correlation to yield loss of disease severity scales for corky root of lettuce. Phytopathology 82: 91–96

Parker SR and Royle DJ (1993) Sampling and monitoring disease in winter wheat. Home-Grown Cereals Authority Project Report No. 71

Sherwood RT, Berg CC, Hoover MR and Zeiders KE (1983) Illusions in visual assessment of Stagnospora leaf spot of orchardgrass. Phytopathology 73: 173–177

Tomerlin JR and Howell TA (1988) DISTRAIN: A computer program for training people to estimate disease severity on cereal leaves. Plant Disease 72: 455–459

Wetherill GB (1977) Sampling Inspection and Quality Control. Second edition. Chapman and Hall, London