Reliability and Accuracy of Visual Estimation of Phomopsis Leaf Blight of Strawberry

M. Nita, M. A. Ellis, and L. V. Madden

Department of Plant Pathology, The Ohio State University, Wooster 44691.
Accepted for publication 3 March 2003.

ABSTRACT

Nita, M., Ellis, M. A., and Madden, L. V. 2003. Reliability and accuracy of visual estimation of Phomopsis leaf blight of strawberry. *Phytopathology* 93:995-1005.

Six different individuals (raters) assessed the severity of Phomopsis leaf blight on strawberry leaflets in five experimental repetitions over 2 years by making a direct visual estimation of the percentage of diseased area of each leaflet or by using the Horsfall-Barratt (H-B) disease scale. Intra-rater and inter-rater reliability and accuracy were determined, and then the relationship between visually estimated severity values and actual severity values was evaluated. Agreement in estimated disease severity values between assessment times by the same raters (i.e., intra-rater reliability), and agreement in disease severity values among raters at a single assessment time (i.e., inter-rater reliability), were both high,

with most correlation coefficients being greater than 0.85. The intra-class correlation for overall agreement among raters ranged from 0.80 to 0.96 for the five repetitions. Based on the concordance coefficient calculated for each rater in each repetition, agreement between estimated and actual severity (i.e., accuracy) was somewhat lower than reliability. The relationship between estimated and actual severity was linear, and there was a slight trend to overestimate disease severity. The H-B scale was not more reliable or accurate than direct estimation of severity, and the linear relationship between estimated and actual severity did not support the principles underlying the H-B scale. Both size of leaflets and number of lesions per leaflet slightly affected the error in estimate of disease severity.

Additional keywords: models, *Phomopsis obscurans*, reproducibility.

Phomopsis leaf blight of strawberry (*Fragaria xananassa* Duchesne) is a disease caused by the fungus *Phomopsis obscurans* (Ellis & Everh.) (10,45). Foliar lesions are irregular, often circular, with purplish halos (10,31). Typically, there are three zones in a lesion: a dark-brown center, tan-to-light-brown surrounding rings, and a reddish or purplish outer zone. Older lesions occurring along veins typically become "V" shaped, with the widest part of the V toward the leaf margin. Although the economic importance of the foliar phase of the disease has not been quantified, it is presumed that infection reduces yield, weakens plants, and affects winter survival (8,22). The fungus also causes fruit rot (9,19). Fruit rot is relatively common in southern states such as Florida, where yield losses can be economically significant (25,30), but it is not common in northern states.

There have been extensive investigations of the spatial component of Phomopsis leaf blight epidemics in commercial fields in Ohio (47,48). In general, the disease is aggregated in small patches, and the degree of aggregation is directly related to mean disease incidence. All of these studies were based on disease incidence. Although severity usually is thought to be a better predictor of crop loss (5,6), the disadvantage of severity as the measure of disease intensity is that there can be considerable measurement error with visual assessments (13,15–17,37,38).

There are many published methods for assessing disease severity (7,16–18,20,21,36), and several references discuss the principles involved in developing and evaluating an assessment scheme (13,15,29,35,38,41,44). A very common approach for visual assessment is use of the Horsfall-Barratt (H-B) assessment scale (5,13,14). With the H-B scale, a rater assigns a sampling unit (e.g., leaf, stem, and root) into 1 of 12 unequal classes based on

the perceived percentage of plant area affected by the disease. The H-B scale is based on two principals: (i) there is a logarithmic relationships between intensity of stimuli (actual diseased area) and sensation (estimated diseased area) (Weber-Fechner law), and (ii) the human eye perceives diseased area below 50% and healthy area above 50% (16,17,37,38). Various authors have questioned the validity of these principals (13,38,40); however, use of the H-B scale remains routine. Alternative assessment methods include direct estimation of the percentage of plant area affected, or assigning sampling units into equal-interval disease classes, with or without the aid of assessment keys or diagrams (13,21,36,40).

Ideally, a disease assessment scheme should be both reliable and accurate (5,38). Reliability can be defined as "the extent to which the same measurements of individuals obtained under different conditions yield similar results" (11). Two kinds of reliability can be considered in the context of disease assessment, intra-rater and inter-rater reliability. Intra-rater reliability (or intra-rater repeatability) (38) is the agreement in visual estimates of diseased area between two different disease assessments, whereas inter-rater reliability (or reproducibility or inter-assessor variability) (13) is the agreement in visual estimates of diseased area between raters. A definition of accuracy is "...how close the value of the statistic is to a supported 'true value'" (23). For disease assessment, accuracy would be the closeness of the visual estimates to the actual severity (39). Although disease severity data are relatively easy to obtain by visual assessment schemes, there have been relatively few studies in which reliability and accuracy have been quantified (13,14,38,40–42). Determination of reliability requires assessment of the same sampling units (e.g., leaves) by different raters, or with different techniques (13,37,38,41). Determination of accuracy, however, is more difficult because true values are difficult or impossible to obtain. However, various authors have shown that one can obtain estimates of severity that can be considered to be close to the actual values, at least relative to visual assessments (13,38,41).

Corresponding author: L. V. Madden; E-mail address: madden.1@osu.edu

For studies on the epidemiology and management of *Phomopsis* leaf blight, it is often desirable to assess severity in addition to incidence of disease, especially in situations when a high percentage of leaves (or leaflets) are infected, but severity varies greatly

among leaves (or leaflets). The objectives of this study were to (i) quantify reliability and accuracy of visual disease assessments for *Phomopsis* leaf blight, (ii) compare direct visual assessment with use of the H-B scale in terms of estimation of percentage leaflet area covered by lesions, and (iii) determine the functional relationship between estimated and actual severity to ascertain if there was evidence supporting either the use of the Weber-Fechner law or the principals of the H-B assessment scheme for this disease.

MATERIALS AND METHODS

Assessment of disease severity. Twenty-five strawberry leaflets were collected from a commercial field for each repetition of the experiment. Leaflets were collected in late June, when a wide range of severities are observed in the field. Leaflets came from different plants and were selected so that a wide range of severity values would be obtained. Selection of leaflets was done by an individual not rating the leaflets for severity. Except for one repetition, each individual leaflet was placed on a sheet of U.S. letter-size white paper and positioned on a lab bench. Six different raters visually assessed the disease severity of each leaflet by making a direct estimation of the percentage of diseased area. For the H-B scale, raters recorded the class number corresponding to the estimated severity percentage, where 0 = 0, 1 = 0+ to 3, 2 = 3+ to 6, 3 = 6+ to 12, 4 = 12+ to 25, 5 = 25+ to 50, 6 = 50+ to 75, 7 = 75+ to 88, 8 = 88+ to 94, 9 = 94+ to 97, 10 = 97+ to 100+, and 11 = 100% (5,17). The midpoint severity value was assigned for each class number for later analyses (5).

There were five experimental repetitions with a different set of 25 leaflets each. The first three were conducted in 1999 and the last two in 2000. In repetitions one, two, and four, raters assessed leaflets twice in the following manner: all 25 leaflets were assessed once, data sheets were collected, and then the same 25 leaflets were assessed again in a different order by the same raters. Raters did not have a time constraint in performing the assessments. Most took less than a total of 10 min for the 25 leaflets. For the fifth repetition, diseased leaflets still attached to plants were assessed in a commercial field. The experimental design for each disease assessment was a random-effects factorial with two factors, rater (six different individuals [levels], or $k = 6$) and leaflet (25 levels, or $n = 25$). Three of the raters (D, E, and F) had substantial experience with disease severity assessments, and three had very limited experience (A, B, and C).

Determination of actual disease severity. After all leaflets were assessed in each repetition, a photograph of each leaflet was taken. From the photograph, the whole leaflet was cut out and weighed. Then, diseased area was cut out from the photograph, and the remaining healthy area was weighed. From the weight difference between a whole leaflet and healthy area, percentage of diseased area was determined. This percentage was considered to be the actual severity (X), and was treated as the “gold standard” (26,28), for comparisons. Two pictures of the same leaflet were cut by different individuals, and average value was used for later analysis.

Fig. 1. Hypothetical examples of relationships between actual (true) and measured values. Accuracy is composed of both precision and bias. Precision is quantified by the correlation coefficient, r (reflecting the scatter around the best-fitting line). Bias is determined by the bias correction factor of Lin (27), C_b , composed of location (height), shift (u), and scale (slope) shift (v). Bias reflects how far the best fitting line is from the concordance line (slope of 1, intercept of 0).

TABLE 1. General statistics, based on mean disease severity assessments across six raters on each leaflet, of direct visual estimation (Dir) and Horsfall-Barratt-based estimation (H-B) of *Phomopsis* leaf blight of strawberry^a

Statistics	Experimental repetition									
	I		II		III		IV		V	
	Dir	H-B	Dir	H-B	Dir	H-B	Dir	H-B	Dir	H-B
Mean	17.3	17.3	10.7	11.5	15.3	17.2	20.6	20.9	22.1	22.6
Median	12.0	12.2	12.7	13.8	12.7	16.2	10.3	12.2	13.5	15.3
Minimum	0.0	0.0	0.03	0.05	0.08	0.25	0.3	0.3	0.08	0.25
Maximum	66.5	68.8	22.7	24.8	34.2	34.3	99.5	99.3	80.0	78.3

^a There were 25 observations (leaflets) per experimental repetition.

For each leaflet, approximate leaf size (length, width, and area in square centimeters), lesion number, lesion type, and relative lesion size were determined for later analyses. For length and width, an average value of measurement by two persons on each leaflet was recorded. Lesion type was classified as “typical” if a majority of lesions had tan centers surrounded by a purple halo or were large and V shaped, and as “atypical” if a majority of lesions were small purple spots or of other appearance. Also, relative lesion size was classified as, small (<3 mm), medium (3 to 10 mm), or large (>10 mm), based on the majority of the lesions on a leaflet.

Data analysis. Reliability. Intra-rater reliability was determined with correlation coefficients, as done in some previous studies (36, 42). A separate correlation coefficient between estimated severity at two assessment times was calculated for each rater in repetitions one, two, and four. Inter-rater reliability was determined in two ways. First, correlation coefficients were calculated between disease assessments done by all pairs of raters. Second, the intra-cluster (or intra-class) correlation (ρ) for all raters combined was calculated (43) based on variance estimates from a two-way random effects analysis of variance (ANOVA) model. Data were transformed using “the angular transformation” or

Fig. 2. Estimated and actual severity values for Phomopsis leaf blight of strawberry for experiment repetitions I and IV. Data are sorted by actual severity values, which was not the order used during actual experiments. Each point represents the estimated severity of a leaflet, and results for all six raters are included. In severity estimations based on the Horsfall-Barratt (H-B) scale, points fell on horizontal lines because there were only 12 possible severity levels (midpoint of each class).

TABLE 2. Intra-rater reliability, determined using the correlation coefficient (r) for each of six raters, for both direct visual estimation of disease severity of Phomopsis leaf blight of strawberry (Direct) and estimation based on the Horsfall-Barratt scale (H-B), for each of three different experimental repetitions^a

Rater	Experimental repetition					
	I		II		IV	
	Direct ^b	H-B ^c	Direct	H-B	Direct	H-B
A	0.92	0.86	0.88	0.90	0.99	0.96
B	0.97	0.91	0.96	0.90	0.99	0.96
C	0.96	0.94	0.77	0.61	0.99	0.96
D	1.0	1.0	0.97	0.90	1.0	1.0
E	0.98	0.85	0.92	0.96	0.99	0.99
F	0.97	0.98	0.98	0.93	0.99	0.94
Mean	0.967	0.925	0.916	0.865	0.991	0.969

^a Raters assessed the same leaflet set twice; they assessed all leaflets once, turned in their notes, and rated these again in a different order. Raters had unlimited time to perform assessments. Raters A to C were inexperienced in assessing severity, and raters D to F had considerable experience.

^b Rater estimated disease to the closest integer percentage.

^c Rater assigned disease leaflets to one of the 12 severity classes using the H-B scale; mid-points were used in analysis.

$Z = \arcsin(\sqrt{\text{estimated}\%/100})$ prior to ANOVA. The ANOVA model is written as $Z_{ij} = \mu + R_i + L_j + \varepsilon_{ij}$, in which Z_{ij} is the transformed severity for the i th rater ($i = 1, \dots, 6$) and j th leaflet ($j = 1, \dots, 25$), μ is the overall mean, R_i is the effect of the i th rater, L_j is the effect of the j th leaflet, and ε_{ij} is the error.

Variance components were calculated using the expected mean squares (43). Then, ρ was determined as $\rho = \sigma^2_{\text{leaflet}} / (\sigma^2_{\text{leaflet}} + \sigma^2_{\text{rater}} + \sigma^2_{\text{error}})$, in which $\sigma^2_{\text{leaflet}}$, σ^2_{rater} , and σ^2_{error} are the variances for leaflet, rater, and error, respectively. For each repetition of the experi-

ment, a separate reliability analysis was done for both the direct estimation of disease severity and the H-B-based estimation.

Accuracy. Accuracy, that is, the agreement in visual estimate of diseased area and the (assumed) actual diseased area ("gold standard") (39), was determined with Lin's concordance correlation coefficient (ρ_c) (27). A concordance correlation coefficient measures the variation of data from a so-called concordance line, which is a 1:1 line with a slope of 1 and an intercept of 0. A hypothetical example is given in Figure 1, which shows measured values versus actual values. The solid line is for a perfect relation (slope = 1, intercept = 0) and the broken line is the best fitting line. The scatter of points around the best fitting line, quantified by the correlation coefficient (r), is a measure of precision (or variation) of the measured variable, and the closeness of the best fitting line to the concordance line is a measure of bias (C_b). The concordance approach is preferred over regression analysis for determining accuracy of a measurement method because the statistical tests in regression analysis for bias of observed values are excessively influenced by observed precision (27,43). In this study, estimated severity values are the measured values, and the actual severity values are the true values (Fig. 1).

Based on r and C_b , ρ_c is defined as $\rho_c = r \times C_b$, in which

$$C_b = \frac{2}{(v+1/v+u^2)}$$

and

$$v = \sigma_1 / \sigma_2 \text{ and } u = (\mu_1 - \mu_2) / \sqrt{\sigma_1 \times \sigma_2}$$

The terms μ_1 and μ_2 are the means for the measurement values and true values, respectively, and σ_1 and σ_2 are their standard deviations. It should be noted that ρ_c is a type of general intra-class correlation (26) and is analogous to Cohen's weighted kappa index for agreement between ordinal data (43). Although there are some other approaches for measuring accuracy, the concordance method has become popular in several disciplines (26). The component term v measures the scale difference between measured and actual values, which essentially is defined by the difference in slope of the two lines. Equal slopes would give a v of 1. The term u is a reflection of a location shift relative to the scale, primarily reflecting height differences in the lines. Equal heights would give a u of 0. A perfectly accurate measurement system is when points are on the concordance line (i.e., $r = 1$, $C_b = 1$ [$v = 1$, $u = 0$], and thus, $\rho_c = 1$). Through this formulation, one can see that increasing variation around the line ($r < 1$) or increasing bias ($C_b < 1$)—in terms of either scale (v) or location (u)—produces a reduction in accuracy. The analysis was done for each rater of each repetition for both the direct estimation and H-B-based estimation.

Functional relationship between estimated and actual disease severity. The functional relationship between estimated and actual disease severity was evaluated using linear and nonlinear regression analysis. This analysis was based on the mean estimate of severity across the six raters for each leaflet (Y) in relation to the actual severity (X). By using the means, the accuracy or inaccuracy of individual raters was minimized in determining relationships. Data were pooled from the five experiment repetitions, giving 125 observations in the analysis. Two types of nonlinear models were considered. The first model is of the form

$$Y = \alpha + \beta X^\gamma \quad (1)$$

in which α , β , and γ are parameters. When $\gamma = 1$, equation 1 reduces to a linear model. As γ goes to 0 in the limit, X^γ reduces to $\log(X)$. This model is consistent with the principles of the Weber-Fechner law (34,46). A functional relationship in the H-B scale is more complicated because the curvature depends on the value of X (13,16,17). A model that may be appropriate for the H-B scale is of the form

$$Y = (\alpha + \beta X^\gamma) \quad \text{if } X \leq 50 \\ Y = 100 - \alpha - \beta(100 - X)^\gamma \quad \text{if } X > 50 \quad (2)$$

Fig. 3. Estimated disease severity of *Phomopsis* leaf blight of strawberry between the first and second disease assessments by each of the six raters (labeled A to F) in experiment repetition I. Plots on the left are based on direct estimation of severity and plots on the right are based on the Horsfall-Barratt (H-B) scale. Numbers in plots are correlation coefficients that show intra-rater reliability.

Depending on the values of α and β , there can be a discontinuity with equation 2, where the first and second expression give different Y s at $X = 50\%$. The discontinuity is avoided by placing a constraint on equation 2 so that both expressions give identical predictions at $X = 50\%$. This is done by replacing β with $\beta = [(50 - \alpha)/50]$. Models were fitted to the severity data, estimated directly or with the H-B scale, using PROC NLIN of SAS. The DUD (or multivariate secant) algorithm was used for parameter estimation. Depending on the estimated parameters, simplified versions of equations 1 and 2, such as a linear model, were fitted to the data. This analysis was extended to determine if errors in estimation ($Y-X$) were related to other measured variables, such as leaflet size, lesion numbers per leaflet, or lesion size.

RESULTS

Over the five experimental repetitions, a wide range of estimated and actual disease severity values were found (Table 1; Fig. 2). At least one of the leaflets was always symptomless, and maximum estimated severity, based on the mean across the six raters for each leaflet (Y), ranged from 22.7 to 99.5% over the five repetitions.

Generally, the severity estimates for the individual raters were scattered about the assumed actual severity values (X), with a slight trend of overestimation. Individual observations for all raters in experimental repetitions I and IV are shown in Figure 2. Within a range of similar actual severity values (e.g., 10 to 25%), variation of the estimates among the six raters were fairly similar for different leaflets. Occasionally, there was a leaflet (leaflet

number 20 in experimental repetition I) in which estimates were far from the actual. Because individual estimates based on the H-B scale can only correspond to 12 possible values (the midpoint of the class intervals), plots of ordered severity estimates showed a "step-like" pattern consisting of a series of horizontal lines.

Reliability. Intra-rater reliability was generally high for all experimental repetitions (Table 2). Results for repetition I are shown in Figure 3. Overall, correlation coefficients were high, with 89% of r values greater than 0.90 for the direct estimation, and 81% greater than 0.90 for the H-B scale. Except for a few cases, r values for the direct estimation of severity values were numerically higher than r values for estimations using the H-B scale. Mean r values per experimental repetition were from 0.02 to 0.06 higher for direct estimation relative to H-B-based estimation.

Inter-rater reliability was generally high based on correlations between pairs of raters (Fig. 4; Table 3). Values of r for direct estimation were greater than 0.8. The relationships were linear in most cases (Fig. 4), and there was less variability, and higher mean r values, with direct estimation than with H-B-based estimation (Table 3). Mean r per experimental repetition for the direct estimation was from 0.03 to 0.09 higher compared with H-B-based estimation.

The random effects ANOVA indicated significant effects ($P < 0.05$) of rater and leaflet on transformed severity for each experimental repetition, based on both methods of assessing disease. The intra-class correlation, ρ , from each experimental repetition was generally above 0.80 (Table 3). Overall, mean ρ of the five repetitions for direct estimation was 0.90 and for the estimation based on the H-B scale was 0.84. In each experimental repetition,

Fig. 4. Matrix plot of estimated disease severity of *Phomopsis* leaf blight of strawberry between pairs of raters (labeled A to F) for the first experiment repetition (I). For instance, the upper left frame shows the agreement in estimated severity between raters A and B. Lower 'triangle' of graphs is based on direct estimation, and upper 'triangle' of graphs is based on use of the Horsfall-Barratt scale. Numbers in each graph are correlation coefficients of estimated disease severity, which indicate inter-rater reliability.

estimated ρ for direct estimation was higher than that for the H-B-based estimation.

Intra- and inter-rater reliability generally was higher for the experienced raters than the inexperienced ones, as expected. However, the differences in correlations were often small (Table 2; Fig. 4).

Accuracy. Accuracy of disease assessment was moderately high, but not as high as reliability (Table 4). In many cases, the concordance correlation coefficient, ρ_c , was higher for direct estimation than for H-B-based estimation. However, the differences often were slight and, in a few cases, ρ_c was higher for H-B-based estimation than for direct estimation (e.g., rater A in repetition I). Mean ρ_c across the repetitions was 0.86 for direct estimation and 0.82 for estimation based on the H-B scale.

Accuracy results for experimental repetition I are shown in Figures 5 and 6. For individual raters, there generally was a linear relationship between estimated and actual severity. These figures demonstrate how accuracy is composed of both precision and bias. Rater A had identical precision for direct and H-B-based estimation ($r = 0.94$); however, there was greater bias with direct estimation ($C_b = 0.88$) than with H-B-based estimation ($C_b = 0.93$), resulting in higher accuracy with the H-B scale than with the direct linear scale for this rater in this experiment ($\rho_c = 0.87$ versus $\rho_c = 0.83$). In the case of rater E, the higher accuracy with direct estimation ($\rho_c = 0.93$ versus 0.87) was due to greater precision ($r = 0.95$ versus 0.89), because bias was virtually non-existent ($C_b \approx 1$). On the other hand, the higher ρ_c with direct estimation by rater D (0.93 versus 0.86) was due to both higher precision and lower bias.

Measured bias could be due to height ($u > 0$) differences between the concordance and best fitting lines (e.g., rater B in

Fig. 5), or differences in slope ($v > 1$) (e.g., rater F in Fig. 6). Based on means of statistics across the six raters for the components of accuracy (r , C_b [u , v]), both imprecision and bias contributed fairly equally to overall accuracy (Table 5). In general, there was a slight overestimation of severity based on mean u values being greater than 0 in most repetitions. There also was a slightly greater slope for the estimated severity than the concordance line, as evidenced by the mean v values greater than 1.

Functional relationship. There was little evidence of curvature in the relationships between mean estimated severity across the six raters (Y) and actual severity (X) (Fig. 7A). This was expected based on the results for individual raters (Figs. 5 and 6; Table 4). Fitting equation 1 to the data based on direct estimation resulted in the following parameter estimates: $\hat{\alpha} = 1.83$, $\hat{\beta} = 1.19$, and $\hat{\gamma} = 0.96$ (Fig. 7A). The R^2 was 0.94. The exponent (0.96) was not significantly different from 1, indicating a linear relationship. Fitting equation 2 to the same data resulted in the parameter estimates: $\hat{\alpha} = 0.12$, $\hat{\beta} = 2.40$, and $\hat{\gamma} = 0.75$ (Fig. 7B). The R^2 equaled 0.95. The exponent (0.75) was significantly different from 1. With the estimated β in the model, there was discontinuity at $X = 50\%$ (Fig. 7B) (as discussed previously). Fitting the model with the constraint on β resulted in the following parameter estimates: $\hat{\alpha} = 1.38$, $\hat{\beta} = [(50 - \hat{\alpha})/50^\gamma] = 1.44$, and $\hat{\gamma} = 0.90$. Although this estimate of γ was significantly less than 1 ($P \leq 0.05$), the predicted values were very similar to the predictions from a linear model (equation 3) over most of range of X . The R^2 equaled 0.94. For both equation 1 and 2, estimated γ was far from 0, which is the expected value for a logarithmic relationship or combined logarithmic relationship. Thus, there was little evidence of the Weber-Fechner or H-B type of stimulus-response relationship.

Based on these results, a linear model was used to describe the data. The resulting prediction equation was

$$\hat{Y} = 2.40 + 1.02X \quad (3)$$

which had an R^2 of 0.94, and is shown in Figure 7 (broken line). The coefficient for X was not significantly different from 1 ($P > 0.05$), but the intercept was greater than 0. This prediction equation was parallel to the concordance line, $Y = X$.

Using the linear model as a foundation, a stepwise regression analysis was conducted to determine if leaflet size (S , square centimeters), number of lesions (LN), typical lesion size (LS), and lesion type (LT) affected the estimate of disease severity. LS and LT were class variables (indicator or factor variables, not continuous variables). The best fitting model was

$$\hat{Y} = -11.0 + 1.03X + 3.33 \log(LN) + 6.56 \log(S) \quad (4)$$

All estimated parameters were significant, and R^2 was 0.96. The negative intercept (-11) has no physical interpretation in this case because $\log(S)$ can never be 0. LS and LT did not significantly

TABLE 3. Inter-rater reliability for both direct visual estimation of Phomopsis leaf blight disease severity of strawberry (Direct) and estimation based on the Horsfall-Barratt (H-B) scale, for each of five experimental repetitions, using two types of analyses^a

Experiment	Mean pair-wise correlation (r) ^b		Intra-class correlation (ρ) ^c	
	Direct	H-B	Direct	H-B
I	0.93	0.88	0.93	0.90
II	0.81	0.72	0.85	0.76
III	0.89	0.81	0.80	0.72
IV	0.97	0.94	0.96	0.94
V	0.95	0.91	0.94	0.89

^a Six different raters assessed disease severity for each of 25 strawberry leaflets in each experimental repetition using Direct and H-B assessment methods. Agreement in assessments between or among raters (i.e., inter-rater reliability) was quantified with correlations and the intra-class correlation.

^b Correlation coefficients (r) were calculated for all 15 pairs of six raters, and the mean r was determined.

^c Intra-class correlation (ρ) was calculated from the estimated variance components of random-effects analysis of variance.

TABLE 4. Coefficient of concordance (ρ_c), a measure of assessment accuracy, for estimates of Phomopsis leaf blight disease severity by six raters in each of five experimental repetitions (of 25 leaflets each), corresponding to direct estimation of disease (Direct) and to estimation based on the Horsfall-Barratt scale (H-B)^a

Rater ^b	Experiment repetition									
	I		II		III		IV		V	
	Direct	H-B	Direct	H-B	Direct	H-B	Direct	H-B	Direct	H-B
A	0.83	0.87	0.51	0.43	0.69	0.60	0.97	0.95	0.91	0.90
B	0.88	0.85	0.80	0.76	0.79	0.70	0.97	0.97	0.88	0.87
C	0.82	0.81	0.79	0.59	0.78	0.63	0.99	0.95	0.96	0.96
D	0.94	0.86	0.81	0.72	0.77	0.66	0.99	0.96	0.95	0.94
E	0.93	0.87	0.81	0.88	0.76	0.68	0.99	0.97	0.96	0.93
F	0.89	0.92	0.80	0.81	0.85	0.81	0.95	0.95	0.97	0.95

^a Accuracy: $\rho_c = r \times C_b$ in which r is the correlation coefficient between estimates and actual disease ("gold standard"), which is a measure of precision, and C_b is the bias estimation factor, reflecting differences in scale and location (27).

^b Raters A to C were inexperienced in assessing severity and raters D to F had considerable experience.

affect severity estimates. Using this equation, the estimation error ($\hat{Y} - X$) was determined by subtracting X from both sides of equation 4, giving $\hat{Y} - X = -11.0 + 0.03X + 3.33\log(LN) + 6.56\log(S)$. The coefficient for X (0.03) was almost zero, indicating that magnitude of the disease estimation error was not affected by actual severity.

Because of the positive coefficients for $\log(LN)$ and $\log(S)$, increase in lesion number and leaflet size resulted in an increase in the error. However, because the increase in R^2 from equation 3 to equation 4 was only ≈ 0.02 , the effect of these variables, although significant, was only slight.

Fig. 5. Estimated disease severity of Phomopsis leaf blight of strawberry, based on direct visual estimation, versus actual severity for the six raters in the first experiment repetition. The solid line is the concordance line, representing perfect agreement between actual and estimated severity. The broken line is the best-fitting line. Accuracy is determined with Lin's (27) concordance correlation coefficient (ρ_c), calculated as the product of the correlation coefficient (r) and the bias correction factor (C_b). C_b is a function of location shift (u) and scale shift (v), indicating changes in line height and slope, respectively.

DISCUSSION

A useful system for the assessment of disease should be accurate and reliable (5,38). Such a system is necessary to quantify disease progress in time and space, predict yield loss, and establish and evaluate disease management practices (5,32). In order to improve accuracy and reliability of disease assessment, various

disease scales and diagrams, such as the H-B scale (16,17) and disease diagrams (14,20,21), have been developed.

In our study, correlation coefficients (r) were used to determine intra-rater reliability and inter-rater reliability for pairs of raters, and the intra-class correlation (ρ) was used for inter-rater reliability among raters. Overall results showed high reliability based on all statistics for each repetition of the experiment. Correlations

Fig. 6. Estimated disease severity of Phomopsis leaf blight of strawberry, based on the Horsfall-Barratt scale, versus actual severity for the six raters in the first experiment repetition. The solid line is the concordance line, representing perfect agreement between actual and estimated severity. The broken line is the best-fitting line. Accuracy is determined with Lin's (27) concordance correlation coefficient (ρ_c), calculated as the product of the correlation coefficient (r) and the bias correction factor (C_b). C_b is a function of location shift (u) and scale shift (v), indicating changes in line height and slope, respectively.

varied among the experimental repetitions (Tables 2 and 3), partly because of the range of severity data within the different repetitions (28). Generally, correlations increase with the increasing range of the data being analyzed. For instance, repetition IV had the greatest range of severity data (Table 1) and the highest correlations. Therefore, direct quantitative comparisons of results should only be made within repetitions and not between them. Others also have found relatively high reliability for visual assessments of severity (14,38,40,41), although quantitative comparisons among studies cannot always be made because of different statistical methods used. Published correlation coefficients, or square-roots of coefficients of determination (14,37,38,40–42), for intra-rater and inter-rater reliability are similar to the ones we have calculated for direct estimation of Phomopsis leaf blight.

Accuracy of assessed plant disease severity is always difficult to determine because of uncertainty of “true” or actual severity. In general, the definition of true diseased area is not always obvious (2). Phomopsis leaf blight has relatively distinct lesions with high contrast between symptom and symptomless areas; therefore, we used a photograph-and-weight method to determine actual diseased area (the “gold standard” for comparisons). However, other diseases, such as powdery mildew on strawberry (31), have less clear margins. Also, as Bastiaans pointed out in his study on leaf photosynthesis, leaves infected by *Pyricularia oryzae*, the causal agent of rice blast, have lower photosynthesis in some of the apparently disease-free areas as well as the areas with obvious symptoms (1). For our study, and in most studies on disease assessment (38,40,41), we took the definition of disease severity as the area of plant tissue visually diseased.

To estimate actual disease area, we followed the practice of Sherwood et al. (41), which is similar to the approach taken by Nutter and Schultz (38), because, at the time of this study, we did not have access to digital imaging facilities. We had an oppor-

tunity to measure several diseased leaflets later with the imaging system WinRHIZO (Régent Instruments Inc., Québec, Canada), and results were very close to those we obtained from weighing of photographs (M. Nita, *unpublished data*). There are many advantages of digital imaging, including rapid processing, efficient storing and managing of the data, and ability to discriminate fine differences between diseased and disease-free area (3,29). However, visual assessment will remain a very common method of measuring disease because of the relative ease of the method and low cost of collecting data (2,5,14). Moreover, digital imaging still requires a decision-making process that relies on human judgment, such as the selection of colors for diseased tissues (2).

Several researchers (13,14,37,40) measured the accuracy of visual disease assessments by using regression analyses. With this approach, the best fitting line is compared with a line with slope of 1 and intercept of 0 (concordance line). Typically, the estimated slope is not significantly different from 1, whereas the intercept is larger than 0, indicating overestimation of disease severity (on average) in the published studies. In our study, we followed the advice of Shoukri and Pause (43) for reproducibility studies and used the concordance coefficient analysis of Lin (27) for determining accuracy. This method is one of several useful approaches that overcome the problems of regression analysis and some other statistical procedures for assessing agreement between a collection of ratings and a “gold standard” (26,43). Through this concordance analysis, one can partition the overall accuracy into the components of bias (C_b) and precision (r), and remove the effect of precision on the bias determination. Although the accuracy of visual estimates of Phomopsis leaf blight disease severity was not quite as high as the reliability, the concordance analysis indicated that the visual estimates of severity by both experienced and inexperienced raters were fairly close to the assumed actual severity. In agreement with other studies, we also found a slight trend of

TABLE 5. Mean concordance coefficient (ρ_c), correlation coefficient (r), bias coefficient (C_b), and components of C_b (v = scale shift and u = location shift) for direct (Dir) estimation of Phomopsis leaf blight of strawberry or estimation based on the Horsfall-Barratt (H-B) scale^a

Estimation factors	Experimental repetition									
	I		II		III		IV		V	
	Dir	H-B	Dir	H-B	Dir	H-B	Dir	H-B	Dir	H-B
ρ_c	0.88	0.86	0.75	0.70	0.77	0.68	0.98	0.96	0.94	0.92
r	0.93	0.91	0.81	0.79	0.88	0.83	0.98	0.97	0.96	0.95
C_b	0.95	0.95	0.92	0.87	0.88	0.81	1.00	0.99	0.98	0.98
v	1.20	1.24	1.16	1.35	0.93	1.08	1.02	1.02	1.12	1.08
u	0.22	0.22	0.32	0.39	0.31	0.44	0.05	0.06	0.13	0.16

^a Six raters visually estimated disease severity for 25 leaflets in each experimental repetition using Dir and H-B assessment methods. Results shown are means of the statistics across six raters.

Fig. 7. Relationship between estimated disease severity of Phomopsis leaf blight of strawberry (Y) and actual severity (X) for data pooled from five experimental repetitions. Points represent means across the six raters for each strawberry leaflet. **A**, Solid line is a fit of equation 1 and **B**, solid line is a fit of equation 2. Broken line in both **A** and **B** is the fit of a linear model (equation 3).

overestimation of severity, based on u values greater than 0. In some cases, v values were greater than 1, indicating a bias that changed with actual severity. However, the location shift (difference between u and 0) generally was greater than the scale shift (difference between v and 1) (Table 5).

There was little, if any, evidence favoring the H-B scale for visually assessing the percentage of leaflet area covered by Phomopsis lesions. Reliability generally was less for the H-B scale than for direct estimation, and accuracy was similar for the two approaches (Tables 2 to 4). The lower intra-rater reliability for the H-B scale was indicative of a problem with this assessment approach. In order to achieve an r of 1.0, a rater simply would have to assign each leaflet to the same severity class for each assessment bout. Because r was less than 1, it was clear that some raters could not repeatedly classify the same leaflets in terms of severity. Presumably, the source of reduced intra- and inter-rater reliability (relative to direct assessment) involved perceived severity near the class cut-off points. For example, with a perceived severity of $\approx 50\%$, an individual rater might assign the leaflet into the 25 to 50% class the first time, but, at the second time of rating, the rater might assign the same leaflet to the 50 to 75% class. Or two different raters might assign this leaflet into two different classes at a given assessment time. Because the midpoints of these classes are far apart (5), there will be large differences in estimated severity values.

The foundation of the H-B scale is the Weber law, which states that perception of stimuli is relative rather than absolute (4), and the Fechner law, which states that sensation is a logarithmic function of the intensity of stimulus (4,34,46). Authors in psychology and psychophysics often have combined these into the so-called Weber-Fechner law (24,46), and it was the belief of Horsfall and Barratt that this law can be applied to the visual assessment of disease severity (17). If this law alone was appropriate, equation 1 (with $\gamma < 1$) would describe the relationships between estimated and actual severity, and a curve similar to the solid line in Figure 7A would be found. Alternatively, the error in estimation would increase with increasing severity. However, several authors have strongly argued that the Weber-Fechner law is not relevant or appropriate for various stimulus-response relationships, such as tactile light and color (34). Even Fechner agreed that the law did not hold for high and low extreme ends of intensity of stimulus (34), and he also suggested that other models of quantifying sensation, such as a power law (i.e., a non-small γ), should be considered (33).

Another principle for the H-B scale is the belief that there are two relevant severity ranges, $<50\%$, where the rater focuses on the

diseased area, and $>50\%$, where the rater focuses on the disease-free area (16,17,38,41). This results in a model such as equation 2 for estimated severity. Alternatively, error in severity estimation would be highest at 50% severity (16,17,38,41). In agreement with several other studies in plant pathology (13,37,41), we found little evidence for the logarithmic (Fig. 7A) or a combination of two logarithmic curves (Fig. 7B) that would be expected for the Weber-Fechner law or H-B scale. In fact, there was strong evidence of a nearly straight-line relationship between estimated and actual severity, based on data from individual raters (Figs. 5 and 6) or means across six raters (Fig. 7A and B). Furthermore, estimation error was not strongly associated with any severity value. Even in tests with computer-generated disease diagrams, there was no evidence for logarithmic or combined logarithmic relationships (38). The largest errors in previous studies generally corresponded to mean severity less than 50% (12,37,40). Moreover, in our study, number of lesions affected the magnitude of the error ($Y-X$) more than the actual severity did (equation 5), although the error was not strongly affected by any variable. Sherwood et al. (41) also showed that lesion number (i.e., number of spots of *Stagonospora* leaf spots of orchard grass) affected the severity estimates.

Because of the observed linear relationship, we tested a new nearly linear scale for Phomopsis leaf blight (M. Nita, unpublished data). The scale was in equal 5% intervals, except for one case: 0, 1, 5, 10, 15, ..., 95, and 100. Data obtained from the direct estimation of disease severity were placed in the appropriate class, and the two types of reliability and accuracy were calculated (Table 6; compare with Tables 3 and 4). In general, results for the 5% increment scale were closer to direct estimation results than to the H-B-based results. In other words, using a 5% scale was more reliable and accurate than using the H-B scale. Thus, we recommend using an equal-interval scale for estimating disease severity over the H-B scale for those who prefer to base their severity estimation on a disease scale. Although there is no real advantage of a scale in terms of reliability and accuracy, it is conceivable that use of a scale could reduce the time to assess disease on each leaflet compared with direct visual estimation of severity (e.g., to the nearest percentage). This was not evaluated in this study. Standard area diagrams (14,21) prepared for equal intervals of severity also would be helpful for visual estimation of disease severity.

ACKNOWLEDGMENTS

This research was supported by state and federal funds appropriated to the Ohio Agricultural Research and Development Center (OARDC), The Ohio State University. We thank L. L. Wilson for valuable technical assistance.

LITERATURE CITED

1. Bastiaans, L. 1991. Ratio between visual and visual lesion size as a measure to describe reduction in leaf photosynthesis of rice due to leaf blast. *Phytopathology* 81:611-615.
2. Beresford, R. M., and Royle, D. J. 1991. The assessment of infectious disease for brown rust (*Puccinia hordei*) of barley. *Plant Pathol.* 40:374-381.
3. Blazquez, C. H., and Edwards, G. J. 1984. Image analysis of tomato leaves with late blight. *J. Imaging Technol.* 11:109-112.
4. Bringmann, W. G., Luck, H. E., Miller, R., and Early, C. E. (eds.) 1997. *A Pictorial History of Psychology*. Quintessence Publishing Co., Inc., Chicago.
5. Campbell, L. C., and Madden, L. V. 1990. *Introduction to Plant Disease Epidemiology*. John Wiley & Sons, New York.
6. Cooke, B. M. 1998. Disease assessment and yield loss. Pages 42-68 in: *The Epidemiology of Plant Diseases*. D. G. Jones, ed. Kluwer Publisher, Dordrecht, the Netherlands.
7. Dixon, G. R., and Doodson, J. K. 1971. Assessment keys for some diseases of vegetable, fodder, and forage crops. *J. Natl. Inst. Agric. Bot.* 23:299-307.

TABLE 6. Results of intra-rater reliability, inter-rater reliability, and accuracy of (nearly) 5% increment severity scale for estimation of Phomopsis leaf blight of strawberry^a

Repetition	Reliability ^b			Accuracy ^c
	Intra-rater (r)	Inter-rater (r)	Inter-rater (ρ)	
I	0.96	0.92	0.92	0.88
II	0.91	0.82	0.83	0.75
III	... ^d	0.87	0.79	0.77
IV	0.99	0.90	0.96	0.98
V	... ^d	0.94	0.92	0.94

^a Severity values from direct estimation of Phomopsis leaf blight were converted into the 5% increment scale, then calculation of three types of reliability and accuracy was made.

^b Intra-rater reliability determined with a mean correlation coefficient for the repeated assessments of the leaflets by the same raters (mean statistics across the six raters). Inter-rater reliability was determined with both a mean correlation (of the 15 pairs of ratings for the six raters) and the intra-class correlation (ρ) from analysis of variance.

^c Accuracy determined with the concordance coefficient of Lin (27). Mean statistics across the six raters are shown.

^d Leaflets only assessed once, so intra-rater reliability could not be determined.

8. Ellis, M. A. 1998. Strawberry leaf spot diseases. Ohio State Univ. Ext. Fact Sheet HYG-3015-95. Ohio State University, Columbus.
9. Ellis, M. A., Nita, M., and Madden, L. V. 2000. First report of *Phomopsis* fruit rot of strawberry in Ohio. *Plant Dis.* 84:199.
10. Eshenaur, B. C., and Milholland, R. D. 1989. Factors influencing the growth of *Phomopsis obscurans* and disease development on strawberry leaf and runner tissue. *Plant Dis.* 73:814-819.
11. Everitt, B. S. 1999. *Dictionary of Statistics*. Cambridge University Press, Cambridge.
12. Forbes, G. A., and Jeger, M. J. 1987. Factors affecting the estimation of disease intensity in simulated plant structures. *Z. Pflanzenkrankh. Pflanzenschutz* 94:113-120.
13. Forbes, G. A., and Korva, J. T. 1994. The effect of using a Horsfall-Barratt scale on precision and accuracy of visual estimation of potato leaf blight severity in the field. *Plant Pathol.* 43:675-682.
14. Godoy, C. V., Carneiro, S. M. T. P. G., Imauti, M. T., Pria, M. D., Amorim, L., Berger, R. D., and Filho, A. B. 1997. Diagrammatic scales for bean diseases: Development and validation. *Z. Pflanzenkrankh. Pflanzenschutz* 104:336-345.
15. Hebert, T. T. 1982. The rationale for the Horsfall-Barratt plant disease assessment scale. (Abstr.) *Phytopathology* 72:1269.
16. Horsfall, J. G., and Barratt, R. W. 1945. An improved grading system for measuring plant disease. (Abstr.) *Phytopathology* 35:655.
17. Horsfall, J. G., and Cowling, E. B. 1978. Pathometry: The measurement of plant disease. Pages 119-136 in: *Plant Disease: An Advanced Treatise*, Vol. 2. How Disease Develops in Populations. Academic Press, New York.
18. Horsfall, J. G., and Heuberger, J. W. 1942. Measuring magnitude of a defoliation disease of tomatoes. *Phytopathology* 32:226-232.
19. Howard, C. M., and Albrechts, E. E. 1973. A strawberry fruit rot caused by *Dendrophoma obscurans*. *Phytopathology* 63:419-421.
20. James, W. C. 1974. Assessment of plant diseases and losses. *Annu. Rev. Phytopathol.* 12:27-48.
21. James, W. C. 1971. An illustrated series of assessment keys for plant diseases: Their preparation and usage. *Can. Plant Dis. Surv.* 51:39-65.
22. Kerkhoff, K. L., Williams, J. M., and Barden, J. A. 1988. Effects of defoliation on growth and yield of 'Redchief' strawberries. *Adv. Strawberry Prod.* 7:26-28.
23. Kotz, S., and Johnson, N. L. (eds.) 1982. *Encyclopedia of Statistical Sciences*. John Wiley & Sons, New York.
24. Krasil'nikov, N. N. 1993. Mathematical model of neural adaptation of human visual system to image luminance. *Sens. Syst.* 7:100-109.
25. Legard, D. E., Whidden, A. J., and Chandler, C. K. 1997. Incidence and occurrence of strawberry diseases in Florida from 1991-1996. *Adv. Strawberry Res.* 16:35-47.
26. Lin, L., Hedayat, A. S., Sinha, B., and Yang, M. 2002. Statistical methods in assessing agreement: Models, issues, and tools. *J. Am. Stat. Assoc.* 97:257-270.
27. Lin, L. I.-K. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 45:255-268.
28. Lin, L. I.-K., and Chinchilli, V. 1997. Rejoinder to the letter to the editor from Atkinson and Nevill. *Biometrics* 53:777-778.
29. Lindow, S. E., and Webb, R. R. 1983. Quantification of foliar plant disease symptoms by microcomputer-digitized video image analysis. *Phytopathology* 73:520-524.
30. Maas, J. L. 1985. New symptoms of strawberry leaf blight disease. *Adv. Strawberry Prod.* 4:34-35.
31. Maas, J. L., ed. 1988. *Compendium of Strawberry Diseases*. 2nd ed. The American Phytopathological Society, St. Paul, MN.
32. Madden, L. V., and Nutter, F. W., Jr. 1995. Modeling crop losses at the field scale. *Can. J. Plant Pathol.* 17:126-137.
33. Murray, D. J. 1990. Fechner's later psychophysics. *Can. Psychol.* 31:54-60.
34. Murray, D. J., and Ross, H. E. 1988. E. H. Weber and Fechner's psychophysics. Pages 79-86 in: *G. T. Fechner and Psychology*. Passavia Universitätsverlagund-Druck GmbH, Passau, Germany.
35. Newton, A. C., and Hackett, C. A. 1994. Subjective components of mildew assessment on spring barley. *Eur. J. Plant Pathol.* 100:395-412.
36. Nutter, F. W., Jr. 2002. Disease assessment. Pages 312-323 in: *Encyclopedia of Plant Pathology*. O. C. Maloy and T. D. Murray, eds. John Wiley & Sons, New York.
37. Nutter, F. W., Jr., Gleason, M. L., Jenco, J. H., and Christians, N. C. 1993. Assessing the accuracy, intra-rater repeatability, and inter-rater reliability of disease assessment systems. *Phytopathology* 83:806-812.
38. Nutter, F. W., Jr., and Schultz, P. M. 1995. Improving the accuracy and perception of disease assessments: Selection of methods and use of computer-aided training programs. *Can. J. Plant Pathol.* 17:174-184.
39. Nutter, F. W., Jr., Teng, P. S., and Shokes, F. M. 1991. Disease assessment terms and concepts. *Plant Dis.* 75:1187-1188.
40. Parker, S. R., Shaw, M. W., and Royle, D. J. 1995. The reliability of visual estimates of disease severity on cereal leaves. *Plant Pathol.* 44:856-864.
41. Sherwood, R. T., Berg, C. C., Hoover, M. R., and Zeiders, K. E. 1983. Illusions in visual assessment of *Stagonospora* leaf spot of orchard grass. *Phytopathology* 73:173-177.
42. Shokes, F. M., Berger, R. D., Smith, D. H., and Rasp, J. M. 1987. Reliability of disease assessment procedures—A case-study with late leafspot of peanut. *Oleagineux* 42:245-251.
43. Shoukri, M. M., and Pause, C. A. 1999. *Statistical Methods for Health Science*. 2nd ed. CRC Press, Boca Raton, FL.
44. Slopek, S. W. 1989. An improved method of estimating percent leaf area diseased using a 1 to 5 disease assessment scale. *Can. J. Plant Pathol.* 11:381-387.
45. Sutton, B. C., and Waterson, J. M. 1970. *Phomopsis obscurans* in: *C.M.I. Description of Pathogenic Fungi and Bacteria No. 227*. Commonwealth Mycological Institute, Kew, Surrey, England.
46. Titchener, E. B. 1928. *A Text-Book of Psychology*. The Macmillan Company, New York.
47. Turechek, W. W., Ellis, M. A., and Madden, L. V. 2001. Sequential sampling for incidence of *Phomopsis* leaf blight of strawberry. *Phytopathology* 91:336-347.
48. Turechek, W. W., and Madden, L. V. 1999. Spatial pattern analysis of strawberry leaf blight in perennial production systems. *Phytopathology* 89:421-433.