Euphytica (2007) 155:15–26
DOI 10.1007/s10681-006-9296-z

Are ordinal rating scales better than percent ratings? a statistical and "psychological" view

K. Hartung · H.-P. Piepho

Received: 16 May 2006 / Accepted: 17 October 2006 / Published online: 28 November 2006

Abstract Disease incidence and severity are often assessed by either an ordinal rating scale, e.g., with scores from 1 to 9, or a percentage rating scale. This paper compares three different rating scales regarding accuracy, precision, and time needed for scoring. Pictograms of mildew diseased cereal leaves were generated following a right skewed beta-distribution. Persons with different rating experience were asked to rate the leaves on three different scales: two different percentage scales [1%-steps (P1) and 5%-steps (P5)] and an ordinal 9-point rating scale (R9) where thresholds followed a logarithmic pattern with respect to the underlying percentage scale. A transformed value of the estimated disease severity as well as the transformed time needed to estimate per leaf was documented and evaluated using mixed models. In most cases both percent ratings performed better than the ordinal rating scale. For the time needed per leaf by the untrained group, method R9 was better. With the trained group P5 performed better than both other methods. The raters mostly preferred R9, especially when untrained. Nevertheless, the results suggest that P5 can be recommended in terms of accuracy.

K. Hartung · H.-P. Piepho (✉)
Institute für Pflanzenbau und Grünland (340c), Universität Hohenheim, Fruwirthstr. 23, 70599 Stuttgart, Germany
e-mail: piepho@uni-hohenheim.de

Keywords Accuracy · Disease severity · Mixed model · Percentages · Precision · Visual assessment

Introduction

Plant disease severity is often visually scored using either a percentage scale or an ordinal scale. It is not always obvious which scale is preferable. There are some problems with ordinal rating scales, e.g., the Horsfall–Barratt (H–B) scale (Horsfall and Barratt 1945) or a 1–9 rating scale (Bundessortenamt 2000). Thresholds for these scales are rarely accurately defined but mostly descriptive and may change during time (more or fewer or even different thresholds). Often, there is an underlying percentage scale with clearly defined class thresholds, but the true class means on that underlying scales are usually unknown. If, e.g., on a linear percentage scale the lower and upper thresholds are 10 and 20%, respectively, one might consider the central value of 15 as the class mean, but the real mean of observed values in the class might be 12 or 18. Also the transformation of ordinal ratings back to percentages or absolute values is difficult. The compatibility of two scales is not given, e.g., with two different 1–4 rating scales the same disease value might fall in two different classes. Finally, most statistical methods as used for metric data are not strictly

valid, but several nonparametric methods are available (Shah and Madden 2004).

Percentage ratings have many advantages. Most problems associated with ordinal ratings do not occur, even though one needs to account for heteroscedasticity and nonnormal distribution of the data (Piepho 1999; Shah and Madden 2004). Furthermore one uses a larger number of values with percentages than with ordinal ratings (e.g., 100 vs. 9), which is expected to result in more accurate disease assessment. If 1%-steps are assumed to be the smallest distinguishable unit, estimation error does not lead to wrong classification with 1%-steps and therefore does not bias the class mid-points as could occur when thresholds are used. As James (1974) and Duveiller (1994) mentioned, if scales are based upon percentages, upper and lower thresholds are uniquely defined, scales can be divided, interpolations, and transformations are possible, and scales can be used universally.

In practice there are often immense psychological barriers to overcome if raters are forced to use percentage scales, especially with 1%-steps and smaller. The rater may feel overcharged with this duty and needs to overcome his inhibitions to decide on a definite number, while he may feel more secure when needing to decide only on a range of values as implied by an ordinal rating. As a result, most investigations are done using ordinal ratings with underlying percentage scales, even in phytopathology, where it is generally well known that direct percentages or metrical data should be prefered.

With a percentage scale many raters tend to use values that are multiples of 5 or 10% (Hau and Kranz 1989; Schumacher et al. 1995), which leads to pseudo-classes. The wider the step sizes the more similar the scale becomes to an ordinal rating scale, and therefore it may suffer from the same problems. Another widely held belief is that there is the disadvantage of additional time required to directly estimate percentages.

This paper compares three rating scales regarding their accuracy, precision, and time needed for scoring. The three scales were two different percentage scales (1%-steps (P1) and 5%-steps (P5)) and an ordinal 9-point rating scale (R9). We assessed the relative performance when the methods are employed by persons not used to do ratings versus persons with a certain degree of experience in rating. Methods were compared using mixed model analysis.

Materials and methods

Assessment of disease severity

An MS Access program was developed to collect ratings of simulated mildew infected cereal leaves. Pictorial representations of leaves were generated in a form analogous to those in the program DISTRAIN (Tomerlin and Howell 1988).

The Access program presented the rater—in succession—with 100 cereal leaves of the same shape and size with different disease severities of mildew. With the first ten leaves, the real disease value was shown as a help to ''calibrate'' the rater, so only 90 data points per rater and method were available for analysis. The sequence of leaves shown was identical for each of the three ratings and for every rater. Three different rating methods were compared:

- P1: 1%-steps (0–100%).
- P5: 5%-steps (0, 5, 10,..., 95, 100%).
- R9: rating scale (1–9, defined on a logarithmic percentage scale; see Table 1).

To avoid transformation problems, no leaves were generated that were not diseased at all or completely damaged. Therefore, values of zero or 100 were excluded from the percentage scales and the value of 1 from the 1–9 rating scale. R9

Table 1 Definition of the ordinal rating scale thresholds in percent and its corresponding midpoint

Score	Range (%)	Mid-points (%)[a]
1	0	0
2	> 0–2	1
3	> 2–5	3.2
4	> 5–8	6.2
5	> 8–14	10.6
6	> 14–22	17.5
7	> 22–37	28.5
8	> 37–61	47.5
9	> 61–100	78.1

[a] For scores 3–9 mid-points are the geometric mean of the class thresholds

followed the ''Guidelines for the testing of the value for cultivation and use (VCU) of agricultural crops'' of the Bundessortenamt (BSA: Federal Plant Variety Office, Hannover, Germany; www.bundessortenamt.de) for mildew. The scale involves a logarithmic division of the underlying percentage scale. To obtain leaves for every class, the leaf disease severity was simulated according to a right skewed beta-distribution with parameters $\alpha = 1.5$ and $\beta = 3.9$ for all methods. Additionally, unrecognized by the rater, the time needed to input and verify the rating per leaf was recorded. If the time required was more than 140 s (this happened 14 times), time was treated as ''missing value.'' Values over 140 s were exceedingly large and so it was assumed that the rater was disturbed by external influences, e.g., a telephone call.

To be able to compare the ordinal rating scale with the ratings in percent, the back-transformed logarithmic class midpoints of the ordinal scale were used. The arithmetic mean of class boundaries on the logarithmic percentage scale equals the geometric mean of lower and upper threshold (L and U, respectively) on the untransformed percentage scale, i.e.,

$$\text{class midpoint} = \sqrt{L \times U}$$

where L is lower threshold of class of interest and U is upper threshold of class of interest.

Exceptions to this definition were ratings equal to 1, which were equated to zero percent, and ratings of 2, for which the mid-point was set to 1% (see Table 1).

The raters

Two groups of persons rated the simulated leaves. Group A consisted of students untrained to rate, while group B was a heterogeneous group with different levels of experience in the use of rating scales. In the following, data collected for a combination of method and rater are referred to as data record.

Group A (untrained)

Fifteen students of a fourth semester course in crop protection were introduced to the Access program and then randomly divided into three groups. Every rater of each group was asked to rate the diseased area of all leaves using the assigned method. After a break of 15 min they were asked to rate again using a different method than before. This second method to be used was randomized within the first group. In addition eight students of a sixth semester course in crop protection were introduced to the Access program and also divided into the three groups. They only did one rating. All together 38 combinations of method and rater (=data records) were available from untrained raters (P1: 13 data records, P5: 12 data records, R9: 13 data records).

Group B (trained)

A heterogeneous group of 16 persons, all used to do ratings up to different levels of training, were requested to rate leaves by all three methods. The order of the three methods was randomized per rater. A total of 43 (16 + 13 + 14) data records were available from trained raters. Additionally, from this group the four best raters were analyzed separately.

Every rater, who did more than one rating, was asked to fill in a questionnaire (available from the authors upon request) to obtain subjective information about her or his perception of the rating scales. Issues of interest were: which rating was found to be easiest, whether there were any problems, barriers or inhibitions, and which rating scale they would prefer.

Accuracy and precision

In a phytopathological context, accuracy and precision are attributes of disease assessment, where accuracy describes the closeness of a sample estimate (E) to the true value (T), whereas precision refers to the repeatability (Campbell and Madden 1990). These terms are closely related to variance (precision) and bias (accuracy) in statistics, but they are rarely defined rigorously in statistical terminology when used in publications appearing in plant science journals. Variance and bias determine the mean squared

error (MSE), which in statistics is frequently used to assess the performance of an estimator. The MSE of an estimator E is defined as:

$$\mathrm{MSE}(E) = \mathrm{variance}(E) + [\mathrm{bias}(E)]^2$$

(see Rice 1995, pp. 126).

Descriptive statistics

For visual analysis of the data, histograms of the estimated values for P1 were plotted separately for groups A and B, and the four best raters from group B (Fig. 1a–d). We looked at the scatter plots of the differences of estimated minus true disease severity (E–T) versus the true disease severity (T) for P1 and groups A and B, respectively (Fig. 2a–b) as was also done in Hock et al. (1992) and Forbes and Korva (1994). This difference (E–T) is also known as accuracy (O'Brien and van Bruggen 1992; Forbes and Korva 1994). The standard deviation of accuracy ($S_{(E-T)}$) per rater and method was calculated. For the estimated value (E), the smallest possible standard deviation (SPS) was computed supposing that 1%-steps are the most precise differences that could be determined. Assuming that every leaf is rated optimally (assigned to the class it really belongs) with every method, the standard deviation of the difference of the optimally rated value minus true disease severity (T) leads to the SPS. The value of SPS for P1 had to be 0.00, due to the assumption that 1%-steps are the highest possible precision, for P5 it was 1.45 and for R9 it was 4.80. These values were compared with the lowest observed standard deviation ($S_{(E-T)\mathrm{min}}$) and the

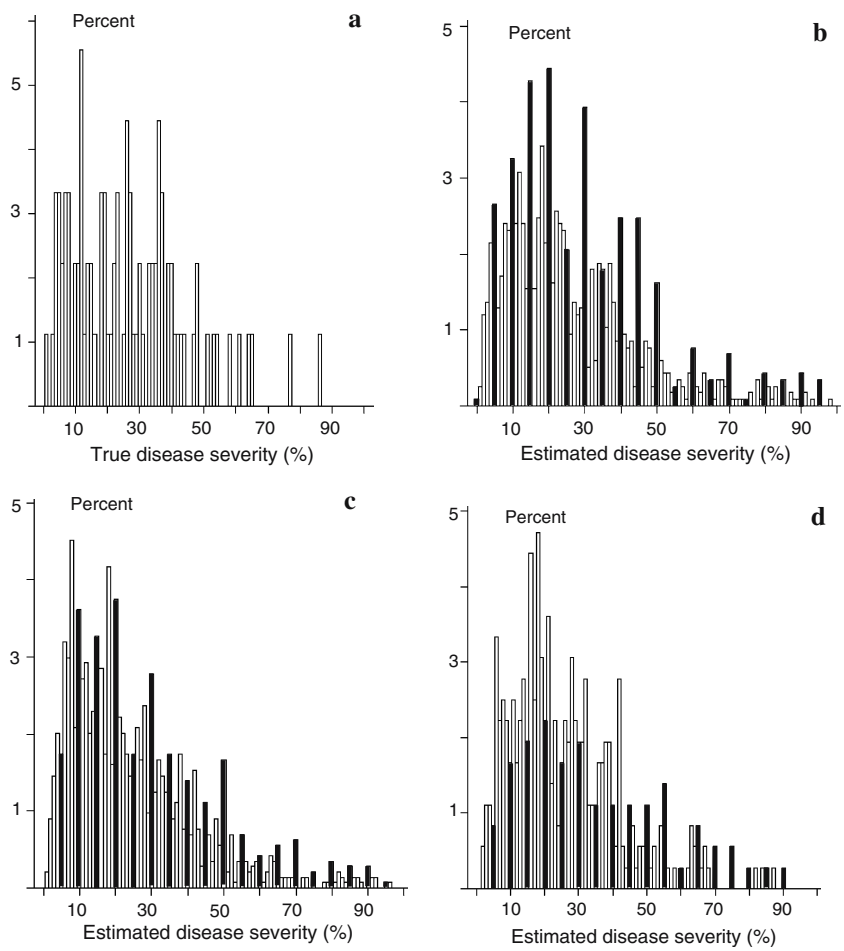

Fig. 1 Histogram in percent of values of (a) right skewed beta-distributed true disease severity used in this investigation, (b) estimated disease severity with Group A (untrained), (c) estimated disease severity with Group B (trained), (d) estimated disease severity with the four best raters. Black bars used in (b) to (d) refer to values that are multiples of five

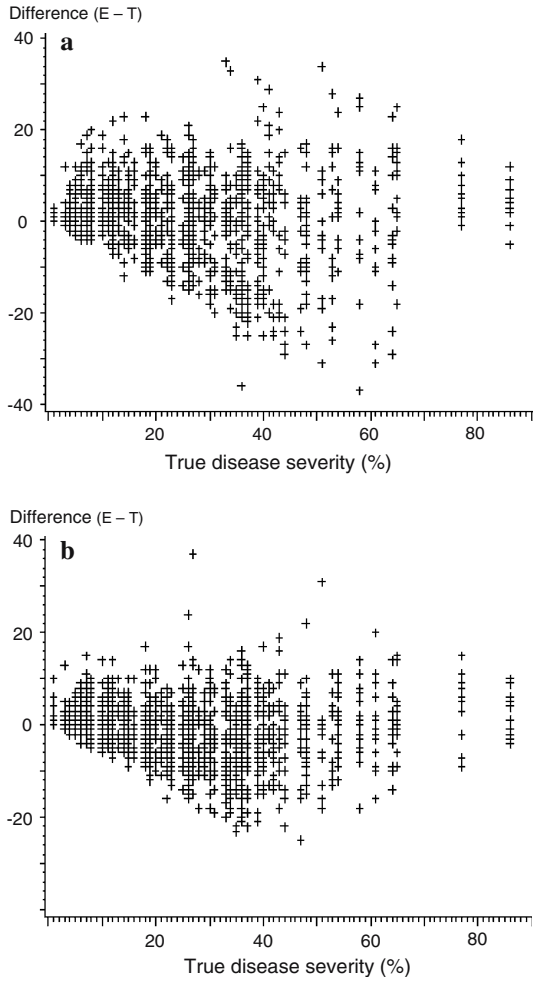

Fig. 2 Scatter plot of difference between true and estimated disease severity (E–T) with method P1 versus true disease severity (a) in Group A (untrained), (b) in Group B (trained). One cross can represent more than one observation

highest observed standard deviation ($S_{(E-T)\max}$) found per group and method (Table 3). Additionally, $S_{(E-T)}$ was also analyzed by simple ANOVA. Box-and-whisker plots were generated using the standard deviation ($S_{(E-T)}$) of the difference of estimated minus true disease severity per group and method. The $S_{(E-T)}$ is equal to the square root of Variance(E). The box length shows the distance between 25 and 75% quartile, the line indicates the median and the cross the represents mean. Values more than 1.5 times the length of the box are interpreted as outliers and marked separately by a square (Fig. 3).

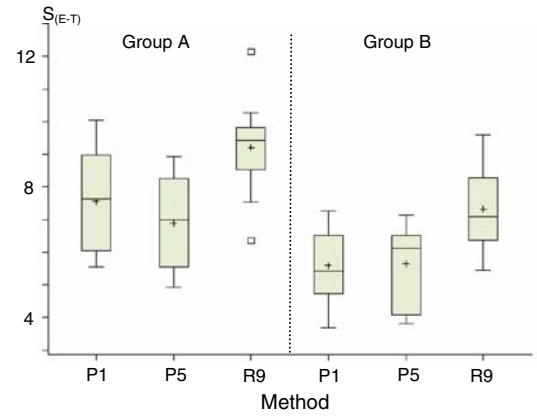

Fig. 3 Box-whisker plots per group and method (of standard deviations ($S_{(E-T)}$) of estimated (E) minus true disease severity (T) per person); Box: 25–5% quartile, Line: median; +: mean, square: values more than 1.5 time the length of the box (interpreted as outliers), whisker 0–5% and 75–100% quartile if all values within 1.5 times the length of the box, else last value within this distance

Error of estimation and time requirement

Error of estimation (D') was defined as

$$D' = \frac{E - T}{\sqrt{T(1-T)}},$$

where E is the estimated disease severity and T is the true disease severity. The difference E–T is here scaled by $\sqrt{T(1-T)}$, assuming that the variance function of estimation errors E–T is proportional to that of a binomial distribution (McCullagh and Nelder 1989, p. 328). This transformation led to residuals with better variance homogeneity. There were no leaves not diseased at all or completely damaged presented to the rater, so there were no transformation problems with percentages of zero and 100. Also, it was not allowed to input zero or 100 into the form. The mean of D' assesses estimation bias of a method, while variance in D' is due to random error. Time needed per rating was logarithmically transformed, which also led to residuals with better normal distribution and variance homogeneity.

Mixed model analyses

With mixed model analysis, a response (also called dependent variable) is modeled by explan-

atory factors plus a residual error that represents all the variability of the response not accounted for by the other terms. A factor can either be fixed or random. A factor is fixed if the levels of the factor were selected with the purpose of comparing the effects of the levels to one another. A factor with random levels represents a single population from which the levels being investigated are a random sample. Interest may lie in the variability within the population from which the sample came (variance component), or perhaps in a prediction of the mean of a particular level. Units that are observed repeatedly through time without new randomization are termed repeated measurements. If, e.g., the effect of the rater on the rating value depends on whether or not the previous leaf was infected more seriously, so that there is a difference in slopes among at least two factors, then there is said to be an interaction between these factors.

The data records were analyzed separately for transformations of error of estimation (D') and time needed per leaf for groups A and B as well as for the four best raters from group B. The four best raters from group B were analyzed separately because they were relatively experienced, thus providing insights as to the potential of training. It is obvious that the full model needs to comprise effects for factors method, rater, leaf, and their interactions. Leaf here was assumed to be fixed as we used the same leaves with each rater and method, which allows to compare the same leaf within the three rating methods. Because one might tend to overestimate leaf disease severity when the previous leaf has small disease severity and to underestimate when the previous leaf is highly diseased, the true value of the leaf previous to the actual estimated leaf (True Previous) was used as a covariate. The interactions of covariate and leaf and method, respectively, were not taken into account. Additionally every rater could have individual abilities related to the three methods and respond differently to the value of the previous leaf, requiring random effects $(\alpha\beta)_{ij}$ and δ_{ij}, respectively. Also, the two random effects $(\alpha\beta)_{ij}$ and δ_{ij}, might be correlated. Different correlation structures were tested, i.e., first-order factor analytic [FA0(1)], and unstructured (UN) (see appendix). Also it is very likely that ratings of different leaves done by the same rater will be serially correlated. Therefore, serial correlation structures were fitted for the residual error, i.e., autoregressive [AR(1)] and autoregressive-moving-average [ARMA(1,1)]. For both [$(\alpha\beta)_{ij}$ and δ_{ij},] and residual error (e_{ijl}) we allowed for heterogeneity among methods.

The full model was:

$$y_{ijl} = \mu + \alpha_i + \beta_j + \eta_l + \varepsilon x_l + (\alpha\beta)_{ij} + (\alpha\eta)_{il} + (\beta\eta)_{jl} + \tau_j x_l + \delta_{ij} x_l + e_{ijl},$$

where y_{ijl} transformed value of error of estimation (D') and disease severity or log-transformed value of time needed per leaf per rating, depending on the analysis,
μ overall mean, fixed,
α_i effect of ith method, fixed,
β_j effect of jth rater, fixed,
η_l effect of lth leaf, fixed,
x_l true disease severity of leaf previous to lth leaf [True Previous],
ε regression coefficient of True Previous, fixed,
$(\alpha\beta)_{ij}$ interaction between method and rater, random,
$(\alpha\eta)_{il}$ interaction between method and leaf, fixed,
$(\beta\eta)_{jl}$ interaction between rater and leaf, fixed,
τ_j regression coefficient of rater on True Previous, fixed,
δ_{ij} regression coefficient of δ_{ij} (rater and method) on True Previous, random $\begin{pmatrix} (\alpha\beta)_{ij} \\ \delta_{ij} \end{pmatrix} \sim N(0, \Sigma)$
e_{ijl} residual error, with $e_{ijl} \sim N(0, \sigma_i^2)$ (the SAS codes are available from the authors).

To determine the best of the various possible correlation models, different models were tested using Akaike's Information Criterion (AIC). The idea behind AIC is to examine the complexity of the model together with goodness of fit to the observed data, and to produce a measure which balances between the two. The model to prefer is the one with the smallest AIC value (Burnham and Anderson 1998; Garrett et al. 2002).

Using a model including the fixed main effects and interactions (see above), the covariance structures for random effects and residual error were tested separately. First, the random effects were tested, assuming the residual error covariance structure to be AR(1). Then the ARMA(1,1) covariance structure for the residual error was tested, with the best model determined for the random effect. After the best model for each covariance structure was selected, all nonsignificant fixed effects tested by type 1 hypotheses were excluded except for main effects that were included in interactions of the random or repeated part of the model (for further information, see, e.g., McCullagh and Nelder 1989; Garett et al. 2002; Schabenberger and Pierce 2002; Piepho et al. 2003).

For error of estimation 2 (D'), the model for group A was

$$y_{ijl} = \mu + \alpha_i + \beta_j + \eta_l + \varepsilon x_l + (\alpha\beta)_{ij} + \tau_j x_l + e_{ijl},$$

while for group B the model was

$$y_{ijl} = \mu + \alpha_i + \beta_j + \eta_l + e_{ijl},$$

with the same covariance structures for residual errors [e_{ijl}] (see Table 2). The data of the four best raters of group B were again analyzed using the same model as for group B.

Analyzing the log-transformed time needed per leaf, the optimal model chosen for group A was

$$y_{ijl} = \mu + \alpha_i + \beta_j + \eta_l + \varepsilon x_l + (\alpha\beta)_{ij} + \delta_{ij} x_l + e_{ijl}.$$

For group B the selected model was

$$y_{ijl} = \mu + \alpha_i + \beta_j + \eta_l + \varepsilon x_l + e_{ijl}.$$

Covariance structures for both groups are shown in Table 2.

Table 2 Selected covariance structures for the transformation of error of estimation (D') and time needed per rating with respect to the effects in the mixed model and group of raters

Response	Effect	Group A	Group B
Estimation error2 (D')	Random effect	–	–
	Repeated error	ARMA(1,1)	ARMA(1,1)
Time per rating	Random effect	FA0	–
	Repeated error	ARMA(1,1)	ARMA(1,1)

Results

Descriptive statistics

Plotting the estimated values for P1 in a histogram (Fig. 1) separately for every group, differences can be seen in the frequency of multiples of five. Among the untrained there is an obvious tendency to prefer values of 5, 10, 15,...,95. A similar but slighter tendency can be found within the trained group and no such tendency with the four best raters.

The typical elliptical form of variance heterogeneity known from percent ratings was found with all ratings, where the greatest differences between estimated and true value of disease level (E–T) were found with intermediate true values (T) (Fig. 2).

To get an impression of the range of differences between estimated and true disease severity (E–T), we generated box-and-whisker plots (Fig. 3) of standard deviation of accuracy ($S_{(E-T)}$). With the same method, the mean is always smaller with group B. With group A and method R9 there are two raters (small squares in Fig. 3) strongly differing from the rest of the group.

The ''SPS'' per method as well as the lowest and highest values for the standard deviation per group and method can be found in Table 3. When comparing the differences of lowest standard deviation ($S_{(E-T)\min}$) to SPS, group B always has smaller differences than group A. The range between highest ($S_{(E-T)\max}$) and lowest standard deviation ($S_{(E-T)\min}$) is always smallest for P5 and with the same method always smaller in group B.

Table 3 Relation of smallest possible standard deviation (*SPS*), lowest, and highest observed standard deviation by group of raters and rating method

Method	SPS	Group A		Group B	
		$S_{(E-T)\min}$	$S_{(E-T)\max}$	$S_{(E-T)\min}$	$S_{(E-T)\max}$
P1	0	5.54	10.06	3.68	7.25
P5	1.45	4.94	8.93	3.81	7.14
R9	4.80	6.36	12.14	5.46	9.59

SPS smallest possible standard deviation, $S_{(E-T)\min}$ lowest observed standard deviation, $S_{(E-T)\max}$ highest observed standard deviation

The smallest standard deviation is found with group B and P1 (Table 3).

Mixed model analyses

Model selection for error of estimation (D') with AIC for group A led to a model without δ_{ij} and for group B to a model without $(\alpha\beta)_{ij}$ and δ_{ij}, both times with ARMA(1,1) as correlations structure for the residual error (AIC values not shown).

With time needed per rating and group A, the covariance structures had the same AIC value, except for UN, where there was no convergence because too many likelihood evaluations were needed, so a factor-analytic model (FA0) was chosen. With group B a model without a random effect worked best. Testing covariance structures for the residual error ARMA(1,1) was best with both cases (AIC values not shown).

Error of of estimation (D')

With D' the residual variance was smallest, i.e., precision was highest, with either P1 or P5 for all three groups and decreased with increasing rating experience (Table 4).

With type 1 testing of the fixed effects of D' there were different results for groups A and B. With D' and group A the effects of Leaf and the interaction of True Previous and Rater were significant. With group B there was no significant effect of True Previous or interaction. With the four best raters of group B the Leaf and the Rater effect were significant (Table 5).

There were significant differences in mean estimation error (D') with group A between P5 and R9. With group B and with the four best raters P1 and P5 were significantly different (Table 6). The mean of D' gives information about the rater's bias for each method.

Table 4 Residual variance parameter estimates for groups A and B and the four best raters of group B

Method	Variance estimate of D'		
	Group A	Group B	
		All	4 best
P1	0.0260	0.0141	0.0081
P5	0.0231	0.0167	0.0072
R9	0.0355	0.0417	0.0242

Table 5 p-Values of type I F-test for fixed effects of error of estimation (D')

Effect	p-value of D'		
	Group A	Group B	4 Best
Leaf	<0.0001	<0.0001	<0.0001
Rater	0.2163	<0.0001	<0.0001
Method	0.0596	0.0028	0.0507
True previous	0.6985	–	–
True previous × rater[a]	<0.0001	–	–

[a]Interaction between True previous and rater; – not used with this mixed model

Time needed per ratings

The residual error variances [in squared log (seconds)] differed between both groups. The variance was 0.3313 for group A and P1, 0.3168 for P5, and 0.2925 for R9. For group B and P1 the variance was 0.3648, for P5 it was 0.3601, and for R9 the variance was 0.3328. There were significant influences of fixed effects in group A for Method and Leaf and in group B for all four single effects and the interaction of True Previous and Method (Table 7).

Group A showed significant mean differences among P1 and R9, R9 being faster. With group B there were significant differences between P5 and

Table 6 Least square (LS) means of method for groups A and B and four best rater of group B

Group	Method	LS means of D'	
		Mean[a]	SE
A	P1	0.0335 ab	0.0426
	P5	0.0951 a	0.0443
	R9	–0.0613 b	0.0407
B	P1	–0.0184 a	0.0041
	P5	0.0124 b	0.0072
	R9	–0.0510 ab	0.0339
4 Best	P1	0.0267 a	0.0055
	P5	0.0510 b	0.0065
	R9	–0.0166 ab	0.0382

[a] Means for a group followed by the same letter are not significantly different according to a t-test

SE standard error

Table 7 *p*-Values of type I *F*-test for fixed effects of time needed per rating

Effect	*p*-value	
	Group A	Group B
Method	0.0546	0.0171
Rater	0.1529	<0.0001
True Previous	0.5750	0.0239
Leaf	<0.0001	<0.0001

Table 8 Least square (*LS*) means of methods for both groups of raters for time needed per rating

Method	LS means	
	Group A[a]	Group B[a]
P1	1.7472 a	2.1413 a
P5	1.5570 ab	1.7672 b
R9	1.4054 b	1.9894 a

[a] Means in a column followed by the same letter are not significantly different according to a *t*-test

both others, P5 being faster than P1 and R9 (Table 8).

Questionnaire

Not every rater who did more than one rating returned the questionnaire, so that 14 were available for group A and 10 for group B. Table 9 shows the preferred method in comparison to the distribution of best and worst standard deviation per rater and method. Raters of group A clearly prefer the rougher of the two methods they tested (P5 or R9 compared to P1 and R9 compared to P5), except for three which prefer P1. This is in contrast to the fact that they did better with the more precise rating method they tested, except one who rated better with P5 ($S_{(E-T)}$ = 6.8) than with P1 ($S_{(E-T)}$ = 10.4). With the trained raters (group B) the preferred rating method was equally distributed among the three methods. With group A the percent ratings were better than the ordinal rating scale: four raters did better with P5 than with P1 and two did worst with P5 but best with P1. It is also interesting to note that one rater using R9 ($S_{(E-T)}$ = 8.3) in practice, was actually best with P5 ($S_{(E-T)}$ = 6.3). The ANOVA results show that there are significant differences only between R9 and both P1 and P5 (each *p*-value < 0.0001). For the difference between P1 and P5 the *p*-value was 0.4456. The estimates for the least square means are given in Table 10.

Psychological barriers and problems as mentioned in the questionnaire were not homogenously judged within one method and between methods. Contrasting assessments were given, as is detailed below. It was stated with respect to P1 and P5 that explicitly assessing values (like 18 or 77) is difficult, particular for values between 20 and 80%. Also, many felt that having all leaves available at once or having more aids (e.g., assessment keys) would have facilitated assigning every leaf the right value, and that shapes and sizes of diseased areas might influence the rated value. With P1 it was mentioned that ''it is guessing'' for most leaves, but that it is accurate if values are under 10 or over 90%. Also, for some raters P1 was easier than R9 because one need not consult the thresholds of classes in a table (Table 1). By some, P5 was experienced as more difficult than R9. It was assumed that P5 is easier to analyze later on because it has fewer classes than P1, and it was mentioned to be convenient and quick. R9 was characterized as imprecise and having a strange classification, which was hard to accustom to. To

Table 9 Preferred (as per questionnaire), best, and worst method measured by variance per group and method given in number of raters

Method	Preferred method as per questionnaire		Best method (smallest residual variance per rater)		Worst method (highest residual variance per rater)	
	Group A	Group B[a]	Group A	Group B	Group A	Group B
P1	3	3	7	6	1	0
P5	6	3	7	4	2	2
R9	5	3	0	0	11	8

[a] One missing value

Table 10 Least square (*LS*) means of method and group for ''standard deviation per rater and methods''

Factor	Level	LS Mean[a]
Method	P1	6.5649 a
	P5	6.2809 a
	R9	8.2656 b
Group	A	7.8936 g
	B	6.1807 h

[a] Means for a factor followed by the same letter are not significantly different according to a *t*-test

avoid errors, the classes needed to be checked frequently, which was time-consuming. Others assumed that one makes less error because classes are wider.

Discussion

This investigation provided an overall comparison between three rating methods for disease severity. Following the statistical results we recommend the more precise percentage scales over the 1–9 rating scale. With the percentage scales, we propose to use the 5%-scale, because the 1%-scale is only little more precise and 5%-scales are more convenient to the rater. A detailed analysis of individual rater performance as done by Nita et al. (2003) or Nutter and Schultz (1995) might be interesting for single individuals, but this was not examined in the present paper.

To give an advice which method should be used, several aspects should be reconsidered. Generally the closer the scale of collected data is to a ratio scale with normal distribution, the more powerful methods are available for analysis. Therefore, percentages perform better than ordinal ratings and more ordered classes are better than fewer. If interpretation is to be possible even after years or over locations, percentages are unique and therefore are more informative especially if statistical analysis is required, while the definition of the ordinal rating scales might change over time and later be forgotten.

The H–B scale, which is based on the Weber–Fechner law, is discussed widely in the literature (Horsfall and Barratt 1945; Jenkins and Wehner 1983; Forbes and Korva 1994; Hau and Kranz 1989; Nita et al. 2003), and it is expected that estimation error is highest at 50% disease severity due to physiological factors. Additionally there is a mathematical reason for the typical elliptical form of the scatter plots of $S_{(E-T)}$ versus true disease severity (Fig. 2a, b). This is the form of variance heterogeneity from percent ratings, which is well known in statistics (McCullagh and Nelder 1989). This supports our use of the binomial variance function $T(1–T)$ in the standardization for D'.

Considering the results shown in Table 3, the differences between SPS and $S_{(E-T)\min}$ of group B are small with R9 compared to the same differences of P1 and P5, so it seems easier to improve the $S_{(E-T)\min}$ of P1 and P5, e.g., with training. Hence, percent ratings should be recommended.

The higher frequency of multiples of five among observed percentage ratings was expected from the literature (Hau and Kranz 1989; Schumacher et al. 1995) and previous experiments by the authors. Knowing the problem of preferring to use multiples of five might reverse the problem, so that multiples of five are underrepresented. This problem should be looked at using an appropriate approach. The level of training reduces this problem as well as that of pseudo-classes, especially with P1. Despite the problem of pseudo-classes, raters will, at least to a certain degree, use the entire scale of P1. Hence, the difference between true and estimated value can get smaller than with less divisions of the scale (P5 and R9).

The variance (Fig. 2) increases toward the true value of 50% or not, depending on whether one looks at absolute or relative differences. In absolute terms a difference of 20 is bigger than a difference of one, but in relative terms a change from 20 to 40 is the same as a change from 1 to 2. Statistically the relative value can be analyzed by standardizing the estimated values as is done in this paper. If a small absolute difference is of interest, training on a percent rating scale is suggested.

Another finding favoring percent ratings is that the smallest residual variance per rater (see Table 9) was never found with R9, but with R9 often the highest residual variances was found. In

contrast to our finings, which favor the more precise percentage scales, the raters often prefer R9, especially when untrained. This preference of R9 seems to be based on the internalized belief that one has to give correct answers (here: values). The deeply felt wish to ''do it right'' shows up in the stated request of raters to compare all leaves before rating. The chances to give the right answer are also felt to increase with decreasing number of possible answers (here: 9 vs. 100). The internalized belief leads to psychological barriers, which lead to preference of R9 and to time differences between rating methods. One decides faster when there are fewer choices due to wider distances between class thresholds. On the other hand, if these thresholds have a sound basis, being derived from a logarithmic scale, but ''not plausible,'' since logarithmic thresholds are not as natural or common to human imagination as are values of 25% (a quarter) or 33% (a third), the time spent per rating increases due to the time needed to look up thresholds in a table. Moreover, raters feel uncomfortable with these uncommon thresholds and tend to linearize scale intervals (Forbes and Korva 1994). From a statistical point of view, the variance of ratings decreases, when using percentages and small step sizes. And it is obvious that training leads to better results (Nutter and Schultz 1995).

In view of the results of our analysis as well as the raters subjective perception of the three ratings, the 5% rating scale seems to be a good compromise. This conclusion is in agreement with Nita et al. (2003). If more precise information for a specific range of values is needed—e.g., 0–10 and 90–100%—1%-steps can be combined with 5%-steps.

For many diseases, there are only standard disease assessment keys available on paper, showing just a few different intensities of disease severity for many plant diseases. So it is difficult to train rating of these diseases. Therefore, it would be helpful to have a program like DISTRAIN, with which one can train rating of different diseases and different plants species and their organs. Specialized programs for peanut (Disease.Pro), alfalfa (Alfalfa.Pro), barley (Barley.Pro), and corn (Corn.Pro) are available (Nutter and Schultz 1995).

The smaller the intervals of the used scale the better the statistical properties of the resulting data. One may have to overcome one's inhibitions to decide on a definite number, but the possible estimation error is usually smaller than with rougher methods, as shown in this paper. So directly rating percentages whenever possible leads to smaller overall estimation errors, and with proper training accuracy and precision can be further improved.

Acknowledgments We are indebted to Mr. Hilbert (Rechenzentrum Universität Hohenheim) for the support in developing the MS Access program. We also pay thanks to all raters for their help and time. The comments from Prof. Dr. Zebitz and Dr. K. Emrich are greatfully acknowledged. This research is supported by the German Research Foundation (DFG), grant number PI 377/5.

Appendix

Covariance structures for $[\ (\alpha\beta)_{ij},\ \ \delta_{ij}]$ were

FA0(1): $\mathrm{var}\begin{pmatrix}(\alpha\beta)_{ij}\\ \delta_{ij}\end{pmatrix} = \begin{bmatrix}\lambda_1^2 & \lambda_1\lambda_2\\ \lambda_1\lambda_2 & \lambda_2^2\end{bmatrix}$,

UN: $\mathrm{var}\begin{pmatrix}(\alpha\beta)_{ij}\\ \delta_{ij}\end{pmatrix} = \begin{bmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{bmatrix}$ (SAS Institute Inc 1999).

References

Bundessortenamt (2000) Richtlinien für die Durchführung von landwirtschaftlichen Wertprüfungen und Sortenversuchen. Landbuch Verlag, Hannover, Deutschland

Burnham KP, Anderson DR (1998) Model selection and inference: a practical information-theoretic approach. Springer-Verlag, New York, USA

Campbell CL, Madden LV (1990) Introduction to plant disease epidemiology. Wiley, New York

Duveiller E (1994) A pictorial series of disease assessment keys for bacterial leaf streak of cereals. Plant Dis 78:137–141

Forbes GA, Jeger MJ (1987) Factors affecting the estimation of disease intensity in simulated plant structures. J Plant Dis Prot 94:113–120

Forbes GA, Korva JT (1994) The effect of using a Horsfall-Barratt scale on precision and accuracy of visual estimation of potato late blight severity in the field. Plant Pathol 43:675–682

Garett KA, Madden LV, Hughes G, Pfender WF (2002) New applications of statistical tools in plant pathology. Phytopathology 94:999–1003

Hau B, Kranz J (1989) Fehler beim Schätzen von Befallss tärken bei Pflanzenkrankheiten. J Plant Dis Prot 96:649–674

Hock J, Kranz J, Renfro BL (1992) Test of standard diagrams for field use in assessing the tarspot disease complex of maize (*Zea mays*). Trop Pest Manage 38:314–318

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

James WC (1974) Assessment of plant diseases and loss. Annu Rev Phytopathol 12:27–48

Jenkins SF Jr, Wehner TC (1983) A system for the measurement of folia diseases of cucumber. Cucurbit Genet Coop Rep 6:10–12

McCullagh P, Nelder J (1989) Generalized linear models, 2nd edn. Chapman and Hall, London

Nita M, Ellis MA, Madden LV (2003) Reliability and accuracy of visual estimation of phomopsis leaf blight of strawberry. Phytopathology 93:995–1005

Nutter FW, Schultz PM (1995) Improving the accuracy and precision of disease assessments: selection of methods and use of computer-aided training programs. Can J Plant Pathol 17:174–184

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

Piepho HP (1999) Analysing disease incidence data from designed experiments by generalized nonlinear mixed-effect models. Plant Pathol 48:668–674

Piepho HP, Büchse A, Emrich K (2003) A hitchhiker's guide to the mixed model analysis of randomized experiments. J Agron Crop Sci 189:310–322

Rice JA (1995) Mathematical statistics and data analysis, 2nd edn. Wadsworth Publishing Co Inc., Duxbury Press, Belmond, CA

SAS Institute Inc (1999) SAS user's guide, version 8. Cary, NC

Schabenberger O, Pierce J (2002) Contemporary statistical models for the plant and soil sciences. CRC Press, New York

Schumacher E, Bleiholder H, Thöni H (1995). Methodische Untersuchungen zur biometrischen Analyse von Boniturwerten aus Freilandversuchen mit Herbiziden. In: Proceedings of the 9th EWRS symposium, vol 1. Budapest, pp 283–290

Shah DA, Madden LV (2004) Nonparametric analysis of ordinal data in designed factorial experiments. Phytopathology 94:33–43

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

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