

The Horsfall-Barratt scale and severity estimates of citrus canker

C. H. Bock · T. R. Gottwald · P. E. Parker ·
A. Z. Cook · F. Ferrandino · S. Parnell ·
F. van den Bosch

Received: 29 August 2008 / Accepted: 26 February 2009 / Published online: 22 March 2009
© KNPV 2009

Abstract Citrus canker assessment data were used to investigate effects of using the Horsfall-Barratt (H-B) scale to estimate disease compared to direct estimation to the nearest percent. Twenty-eight raters assessed each of two-hundred infected leaves (0–38% true diseased area). The data were converted to the H-B scale. Correlation (r) showed that direct estimates had higher inter-rater reliability compared to H-B scaled data ($r=0.75$ and 0.71 for direct estimates and H-B scaled data, respectively). Lin's concordance correlation (LCC, ρ_c) analysis showed individual rater estimates by direct estimation had better agreement

with true values compared to H-B scaled data. The direct estimates were more precise compared to H-B scaled data ($r=0.80$ – 0.95 and 0.61 – 0.90 , respectively), but measures of generalised bias or accuracy (C_b) were similar for both methods (0.38 – 1.00). Cumulative mean disease and cumulative variance of the means were calculated for each rater on a leaf-by-leaf basis. Direct estimates were closer to the true severity 59.5% of the time, and to the cumulative true sample mean 53.7% of the time, and to the cumulative true sample mean variance 63.6% of the time. Estimates of mean severity for each leaf based on estimates by 3,

C. H. Bock (✉)
University of Florida,
2001 S. Rock Rd.,
Ft. Pierce, FL 34945, USA
e-mail: clive.bock@ars.usda.gov

T. R. Gottwald
USDA-ARS-USHRL,
2001 S. Rock Rd.,
Ft. Pierce, FL 34945, USA
e-mail: tim.gottwald@ars.usda.gov

P. E. Parker · A. Z. Cook
USDA-APHIS-PPQ,
Moore Air Base,
Edinburg, TX 78539, USA

P. E. Parker
e-mail: paul.e.parker@aphis.usda.gov

A. Z. Cook
e-mail: amanda.cook@tx.usda.gov

F. Ferrandino
Department of Plant Pathology and Ecology, Connecticut,
Agricultural Experiment Station,
New Haven, CT 06511, USA
e-mail: 365874@po.state.ct.us

S. Parnell · F. van den Bosch
Rothamsted Research,
Harpenden, Herts. AL5 2JQ England, UK

S. Parnell
e-mail: stephen.parnell@bbsrc.ac.uk

F. van den Bosch
e-mail: frank.vandenbosch@bbsrc.ac.uk

5, 10, 20 and 28 raters were compared to true disease severity. LCC showed that rater-means based on more raters had better agreement with true values compared to individual estimates, but H-B scale data were less precise, although with means based on ≥ 10 raters, agreement was the same for both assessment methods. Magnitude and dispersion of the variance of the means based on H-B scaled data was greater than that by direct estimates. H-B scaling did not improve reliability, accuracy or precision of the estimate of citrus canker severity compared to direct visual estimation.

Keywords Epidemiology · Disease scales · Disease incidence · Infection · Crop loss · Disease management

Introduction

Monitoring the status and spread of citrus canker (*Xanthomonas citri* subsp. *citri*) on citrus requires estimation of the incidence and severity of the disease (Gottwald et al. 1993; Gottwald and Timmer 1995). Incidence is a more straightforward measure of disease and relates to the number of units infected, while the severity of disease is a measure of the amount of disease per unit, and has greater subjectivity in estimation (Kranz 1988; Madden et al. 2007; Nita et al. 2003; Bock et al. 2008a, b, 2009). Various studies including epidemiological analyses and the ranking of plant breeding material benefit from accurate and precise assessments of disease severity, and should be based on methods that reflect the true disease (Kranz 1988; Nutter et al. 1993; Nita et al. 2003; Parker et al. 1995; Shaw and Royle 1989; Forbes and Korva 1994; Bock et al. 2008a, b, 2009).

Apart from estimating area diseased to the nearest percent, various scales have been developed to address perceived problems with direct estimation, or to simplify or speed the process of disease assessment (Madden et al. 2007). The most widely used and cited scale is that developed by Horsfall and Barratt (1945). The Horsfall and Barratt (H-B) scale was developed in the early 1940s as the science of plant pathology was becoming more quantitative (Horsfall 1986). The scale has been used widely, and is still used by plant pathologists (Mila et al. 2005; Lang et al. 2007; Stevenson et al. 2007), and has been adopted in other disciplines (Pernezny et al.

2008; Bussotti et al. 2003; Chappelka et al. 2003). The impact of the H-B scale can be further established by it being noted as a “Citation Classic” in *Current Contents* (Horsfall 1986). However, over the last 25 years it has been subject to controversy regarding the rationale for its psychophysical basis, and what it actually achieves in accuracy and precision (Herbert 1982; Forbes and Korva 1994; Nutter and Schultz 1995; Nita et al. 2003; Nutter and Esker 2006; Madden et al. 2007).

The H-B scale is an interval scale—dividing percent severity into categories. Horsfall and Barratt (1945) developed the scale in the presence of tested theory (the Weber law), and untested theory (the Fechner law; Baird and Norma 1978, Birnbaum 1994; Nutter and Esker 2006). Weber’s law states that the physical size of a just noticeable difference is a constant proportion of the value of the standard for a given dimension; Fechner’s law, which assumes Weber’s law, adds that the subjective value is a logarithmic function of the physical value. On this basis Horsfall and Barratt divided the percent area disease from 1–100% into a series of logarithmically increasing, then decreasing categories symmetrical around 50% (Table 1). The symmetry was based on the assumption that the eye reads infected tissue below 50% disease, and healthy tissue above 50% disease (Horsfall and Barratt 1945). Thus the H-B scale was designed to take into account the (unproven) assumption that estimated disease was logarithmically related to true disease (Horsfall and Barratt 1945; Horsfall 1945; Horsfall and Cowling 1978; Herbert 1982; Nutter and Esker 2006).

The way the H-B scale was to be applied is described by Horsfall and Barratt (1945). This was later expanded on (Horsfall 1945; Horsfall and Cowling 1978), and by Redman et al. (1969). First an appropriate category on the scale (Table 1) is allotted to the sample leaf or plant based on the estimated percent area diseased, and in a similar manner several replicate units (“several plants (20 or more)”) are assessed for that plant or plot. For each estimate, the percent range mid-point is taken (Table 1) directly, or using the Elanco conversions (Redman et al. 1969). Finally, the percent mid points for all replicates are summed and averaged (sum of midpoint percents \div number of replicates) to arrive at the estimated mean percent disease severity for that plant, plot or field. The mid-point conversion is necessary prior to taking

Table 1 The Horsfall-Barratt scale showing the categories, the ranges for the categories and the category mid-points

Horsfall-Barratt category	Percent ranges	True range	Midpoint for conversion (ELANCO formula ^a)
0	0	0	0 (0)
1	0±3	3	1.5 (2.34)
2	3±6	3	4.5 (4.69)
3	6±12	6	9.0 (9.38)
4	12±25	13	18.5 (18.75)
5	25±50	25	37.5 (37.50)
6	50±75	25	62.5 (62.50)
7	75±87	13	81.5 (81.25)
8	87±94	6	91.0 (90.62)
9	94±97	3	96.5 (95.31)
10	97±100	3	98.5 (97.66)
11	100	0	100 (100)

^a The ELANCO formula is used for conversion when applying an equation developed by Redman et al. (1969). The results are similar using either direct midpoint or the ELANCO method

the arithmetic mean as averaging the category numbers leads to unacceptable bias due to the geometric nature of the scale. The H-B scale continues to be used (Mila et al. 2005; Keinath et al. 2006; Lang et al. 2007; Stevenson et al. 2007) and the claims that it provides an accurate estimate of the mean disease (Horsfall 1945; Horsfall and Cowling 1978; Redman et al. 1969) remain to be thoroughly explored.

There are two facets to the interpretation of the H-B scale. Firstly the relationship between the estimated and true disease was logarithmic—which has been established not to be the case as estimated disease is linearly related to true disease (Nutter and Esker 2006; Nita et al. 2003; Bock et al. 2008a and 2009). Secondly, variance of an estimate of disease severity increases (i.e., precision and accuracy decrease) as magnitude of true severity increases. Thus according to Horsfall (1945), a “difference” of 3% at the end of the scale was equivalent to telling the “difference” between 25% in the middle of the scale, resulting in a process that gave each grade equal weight. There is evidence that variance of direct estimates of disease is not constant with magnitude of true disease severity, but the relationship is not necessarily logarithmic (Redman and Brown 1964; Koch and Hau 1980; Forbes and Korva 1994; Forbes and Jeger 1987; Hau et al. 1989; Bock et al. 2008a, b and 2009). Furthermore, individuals vary in characteristics of assessment (Nutter and Schultz 1995; Bock et al. 2009), and in at least one study estimation error was only weakly related to true disease severity (Nita et al. 2003).

There are few studies that have ascertained effects of the H-B scale (or other scales) on the reliability, precision and accuracy of disease severity estimates in different pathosystems, and the approaches used in these studies have not been the same. Using visual estimates as the “standard”, Slopek (1989) found the H-B scale gave an accurate estimate of severity of barley diseases. However, Forbes and Korva (1994) found that both precision and accuracy were greater when estimating severity of *Phytophthora infestans* on potato directly using the percent scale compared to the H-B scale. Furthermore, Nita et al. (2003) reported the H-B scale did not improve reliability, accuracy or precision of individual estimates of *Phomopsis* on strawberry. Considering the widespread use of disease scales as tools in estimating disease severity it is desirable to thoroughly quantify their reliability, accuracy and precision, and establish advantages, or disadvantages of their use. They have not been used widely to estimate severity of citrus canker, but some characteristics of disease scales, such as the speed with which they can be applied, can be advantageous in some situations.

The objectives of this paper were to compare effects of direct estimates of the percent area infected with citrus canker to H-B scale estimates (i) on inter-rater reliability, (ii) agreement between individual raters and true disease values, (iii) on mean estimates of severity, and the characteristics of the mean estimates, (iv) on the frequency of estimates closer to the true value, and (v) assess the relationship and agreement between mean estimated severity of each leaf (averaged across raters) and true disease severity.

Materials and methods

Leaves and leaf images

Images of diseased leaves have been demonstrated to be realistic surrogates for actual leaves in various studies and are valuable and useful tools for disease assessment studies: computer-generated leaf images are used to train raters (Nutter and Schultz 1995); standard area diagrams are constructed from digitised computer images used to aid assessments (Belasque et al. 2005; Pethybridge et al. 2004); yet other studies have used computer-generated, drawn or digitised images as a basis to explore disease assessment and found them to be realistic (Amanat 1976; Forbes and Jeger 1987; Hau et al. 1989; Bock et al. 2008a). Based on these studies, and for practical reasons, images of diseased leaves, rather than actual diseased leaves, were used. The collection and processing of the images of diseased leaves was described in detail previously (Bock et al. 2008a, b, 2009). The leaves were from grapefruit trees in south Florida and displayed a natural range of severity (0 to approx. 38%). They were photographed using a 6.5 megapixel Canon EOS 10D digital camera (24–85 mm AF lens). The computer image analysis software ASSESS V1.0 (APS Press, St. Paul, MN) was used to measure the percent area with symptoms of citrus canker (necrosis and chlorosis) accurately on a leaf-by-leaf basis. Image size was standardised and printed (3.25' × 2.25') on letter-size paper and approximately reflected the dimensions of the original leaf, with six images per page. Each of the images was assessed once visually by each of twenty-eight raters (Bock et al. 2009). The data from the image analysis assessments were assumed to be the true values—the method is accurate and repeatable when leaves are assessed individually (Bock et al. 2008a). Assessments by raters were made individually, with no formal resting period although brief breaks of a few minutes could be taken.

The H-B scale was applied to each direct estimate of each leaf by each rater by placing the percent disease severity in the equivalent H-B category as described by Nita et al. (2003), assuming that raters would base the H-B category on the perceived percent area infected. The mid-point value of that category was applied to the individual leaf. For each individual, the directly estimated percent severity data was compared to the H-B scaled data to determine whether there was

any effect of scaling over the range of citrus canker severity commonly found in the field.

Data analysis

Characteristics of reliability among individual raters were estimated. The correlation coefficient (r) was calculated for each pair of raters to assess inter-rater reliability for both direct estimates and H-B scaled data. The average correlation coefficient was calculated to compare overall reliability for each method. Lin's concordance correlation (LCC, Lin 1989; Nita et al. 2003; Madden et al. 2007; Bock et al. 2008a) was used to gauge agreement of the true disease severity and disease assessments for each rater for direct estimated values and the H-B scaled data. LCC provides a quantifiable method to evaluate the characteristics of accuracy and precision (Madden et al. 2007). Accuracy is the closeness of the estimate to the true value, and precision is the variation or scatter associated with the estimates (Madden et al. 2007). Thus complete accuracy in relation to true values is only possible if there is complete precision. The LCC coefficient (ρ_c) is the product of elements of precision (the correlation coefficient, r) and accuracy (C_b). C_b in this case is a generalised bias correction factor that measures accuracy and comprises two measures that account for scale or slope shifts (coefficient ν , with a perfect relation of 1 between X and Y) and a location or elevation shift (coefficient μ , where 0 is a perfect relation between X and Y). For the individual raters, cumulative leaf sample means and variances were calculated and compared on a leaf-by-leaf basis for both direct estimates and the H-B scaled data in relation to the true values from image analysis. For these cumulative leaf means, leaves were ordered arbitrarily, but in the same sequence for each rater. The final means and the standard deviations (a measure of precision) were calculated for all 200 leaves assessed for each rater using direct estimates and H-B scaled data. Using each individual leaf estimate, and the cumulative means, the direct estimates and H-B scaled data for each rater were compared to the true values, and the total proportions that were closest to the true value calculated.

To further investigate the effects of the H-B scale on disease estimates, average severity estimates for each leaf were calculated for 3, 5, 10, 20 and 28 raters (rater-means). Raters were ordered arbitrarily and averaged

consecutively for the various analyses. Effects of H-B scaling on these more precise and/or accurate data were investigated. Mean disease is generally not based on multiple rater estimates of single leaves, but averaging several rater estimates of the same leaflet is useful in understanding and providing insight into error that might be involved in using direct estimation compared to H-B scaling, as well as other general sources of error. LCC was to gauge agreement of the rater-means and true values. Regression analysis examined the relationship between true and estimated mean-rater severity by direct estimation and H-B scaling, and that of the relationship between variance of the rater-means and true values for direct estimates and H-B scale data. The slope, intercept, coefficient of variation, and coefficient of determination were calculated, and residual plots generated. Thus residuals (d) for the i^{th} data point were calculated as $d_i = y_i - (a + (b \times x_i))$, where x is the true area infected, a is the intercept, and b is the slope parameter for the regression solution. Regression solutions for rater-means and variance of the rater-means using direct estimates or H-B scaling were compared using general linear modelling (GLM). In this analysis the regression slopes are tested to determine if there are differences in the dependent variable (ds , disease severity) due to assessment method (am) when compared to the continuous independent variable, true disease severity (tds). A test for interaction between regressor and class ($ds = tds, am, tds * am$) gauges whether there are differences due to assessment method which are reflected in slope for the two classes. If the interaction term is significant the classes are different. If the interaction is not different a further test for differences in the elevation of the regression lines (a test of parallelism) is required, which is achieved by running GLM on the reduced model (no interaction term, $ds = tds, am$). Data were analysed using SAS (SAS Systems, Cary, NC).

Results

Effect of H-B scaling on reliability of estimates among raters

Overall, inter-rater reliability measured by the correlation coefficient r , ranged from 0.25–0.93 for the direct estimates and 0.20–0.89 for the H-B scaled data. Although not all inter-rater reliabilities were best by

direct estimation, the frequency of correlation coefficients showed that they were most frequently higher for direct estimates compared to H-B scaled data (Fig. 1). The mean inter-rater reliability (r) was 0.75 for the direct estimates and 0.71 for the H-B scaled data.

Effect of H-B scaling on agreement with individual estimates

As a measure of agreement comparing the individual estimate to the true value the components of the LCC coefficient most often showed loss of agreement for all raters with the H-B scale (Table 2). This was in large part due to a loss of precision shown by the correlation coefficient for all but two raters (raters 18 and 23) when the data were converted to the H-B scale. Only rater 8 precision was the same for both methods. The bias correction factor (C_b), a measure of generalised bias or accuracy, showed that both methods were similar for most raters, although overall H-B scaling might offer a very slight improvement in the C_b (the mean C_b for direct estimation=0.87, and for H-B scaled data=0.89). The components of C_b , the scale (slope, v) and location (elevation, μ) were variable among raters. Of the twenty-eight raters, the scale was improved for eighteen raters, and the

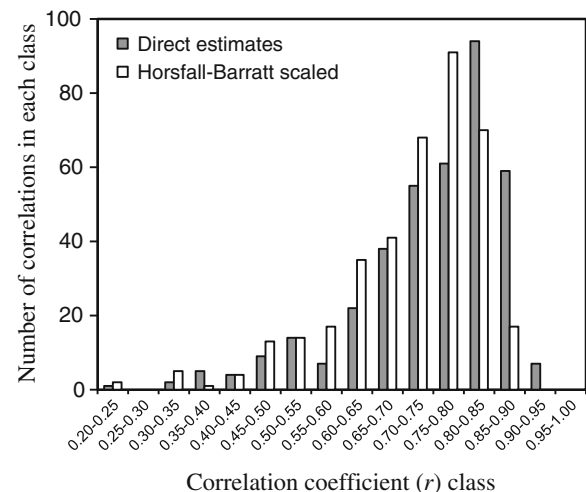


Fig. 1 Comparison of the frequency of inter-rater reliability for estimates of citrus canker severity by direct estimation or using the Horsfall-Barratt scale for twenty-eight raters assessing two-hundred canker-infected grapefruit leaves. Inter-rater reliability measured by the correlation coefficient ($r=0.0-1.0$). Total number of correlations performed=378 for direct estimates and H-B scaled data

Table 2 Correlation analysis^a and Lin's concordance correlation analysis^b of the severity of citrus canker symptoms on two-hundred leaves of grapefruit assessed by direct estimates, and by scaling the estimated percent area to the Horsfall-Barratt scale

Rater	Direct estimates					Horsfall-Barratt scaled				
	<i>r</i>	<i>v</i> (scale)	μ (location)	C_b	ρ_c	<i>r</i>	<i>v</i> (scale)	μ (location)	C_b	ρ_c
1	0.88	0.60	-0.27	0.85	0.75	0.84	0.67	-0.29	0.89	0.75
2	0.92	0.68	-0.12	0.92	0.85	0.87	0.71	-0.15	0.93	0.81
3	0.89	0.81	-0.04	0.98	0.87	0.87	0.89	-0.06	0.99	0.86
4	0.84	1.22	0.13	0.97	0.81	0.82	1.32	0.15	0.95	0.78
5	0.92	1.88	0.83	0.65	0.59	0.86	1.91	0.81	0.65	0.56
6	0.91	1.50	0.17	0.91	0.83	0.84	1.26	0.09	0.97	0.81
7	0.87	0.56	-0.37	0.81	0.71	0.76	0.60	-0.38	0.83	0.63
8	0.90	1.28	0.08	0.97	0.88	0.90	1.25	0.06	0.97	0.88
9	0.88	0.83	-0.07	0.98	0.86	0.74	0.89	-0.08	0.99	0.73
10	0.81	1.34	0.07	0.96	0.77	0.80	1.35	0.06	0.96	0.76
11	0.88	0.70	-0.21	0.92	0.81	0.79	0.71	-0.26	0.91	0.72
12	0.86	1.29	0.43	0.89	0.77	0.68	1.23	0.36	0.92	0.63
13	0.80	0.26	-1.11	0.38	0.30	0.61	0.27	-1.18	0.38	0.23
14	0.92	0.77	-0.10	0.96	0.89	0.85	0.81	-0.12	0.97	0.83
15	0.91	1.16	0.11	0.98	0.89	0.88	1.11	0.04	0.99	0.87
16	0.92	1.04	0.06	1.00	0.91	0.82	1.03	0.04	1.00	0.82
17	0.90	1.93	0.59	0.72	0.65	0.83	1.81	0.58	0.74	0.62
18	0.81	0.63	0.15	0.89	0.73	0.82	0.77	0.14	0.96	0.78
19	0.93	1.81	0.63	0.73	0.68	0.88	1.75	0.61	0.74	0.65
20	0.90	1.15	0.43	0.91	0.82	0.88	1.19	0.41	0.91	0.80
21	0.94	1.37	0.32	0.91	0.85	0.87	1.46	0.29	0.90	0.78
22	0.83	0.79	-0.40	0.90	0.75	0.80	0.79	-0.43	0.89	0.71
23	0.84	0.68	0.14	0.92	0.77	0.85	0.75	0.13	0.95	0.81
24	0.95	0.85	-0.01	0.99	0.94	0.76	0.83	-0.08	0.98	0.74
25	0.92	2.45	0.74	0.59	0.54	0.79	2.26	0.72	0.62	0.49
26	0.92	0.86	0.01	0.99	0.91	0.81	0.87	-0.02	0.99	0.80
27	0.82	0.57	-0.41	0.81	0.66	0.77	0.61	-0.43	0.82	0.63
28	0.87	0.85	0.23	0.96	0.84	0.74	0.96	0.20	0.98	0.72

^a *r*=Pearson's correlation coefficient

^b Lin's concordance correlation analysis: *v*=slope shift (1=no bias between *X* and *Y*) relative to the concordance line; μ =location, or height shift (0=no bias) relative to the concordance line; C_b =the bias correction factor, is dependent on the scale (*v*) and location (μ) shifts; ρ_c =Lin's concordance correlation is the product of precision (*r*) and accuracy (C_b)

location improved for fifteen raters by using H-B scaling, respectively. The overall result in the LCC coefficients (ρ_c) showed H-B scaled estimates generally had lower measures for most raters (except rater 18). Thus according to the components of agreement of the LCC coefficient, loss in precision (*r*) of the estimate appeared to be the major effect of H-B scaling.

The cumulative sample mean calculated on a leaf-by-leaf basis for each rater for both direct estimates and the H-B scaled data is shown for the first 100 leaves and 14 raters (Fig. 2). The true mean severity stabilised at a sample size of approximately 40, and the mean estimates of each of the raters within 15–40 consecutive leaves. The tendency to over—or under-

estimate depended on the individual rater, and was the major source of error. Rater 13 underestimated, and rater 5 overestimated disease (Bock et al. 2009). Both direct estimation and H-B scaling gave similar estimates of cumulative mean severity most of the time, but there was no benefit to using H-B scaling. The variance of these mean estimates for each rater for the first one-hundred leaves showed that magnitude of the variance was rater-dependent (Fig. 3). The variance of the true mean and that of several raters stabilised when the mean was based on a sample size of approx. fifty leaves, although some raters (e.g., rater 8) required a larger sample. The variance of some rater estimates was also higher or lower than the true sample variance. H-B scaling had neither a

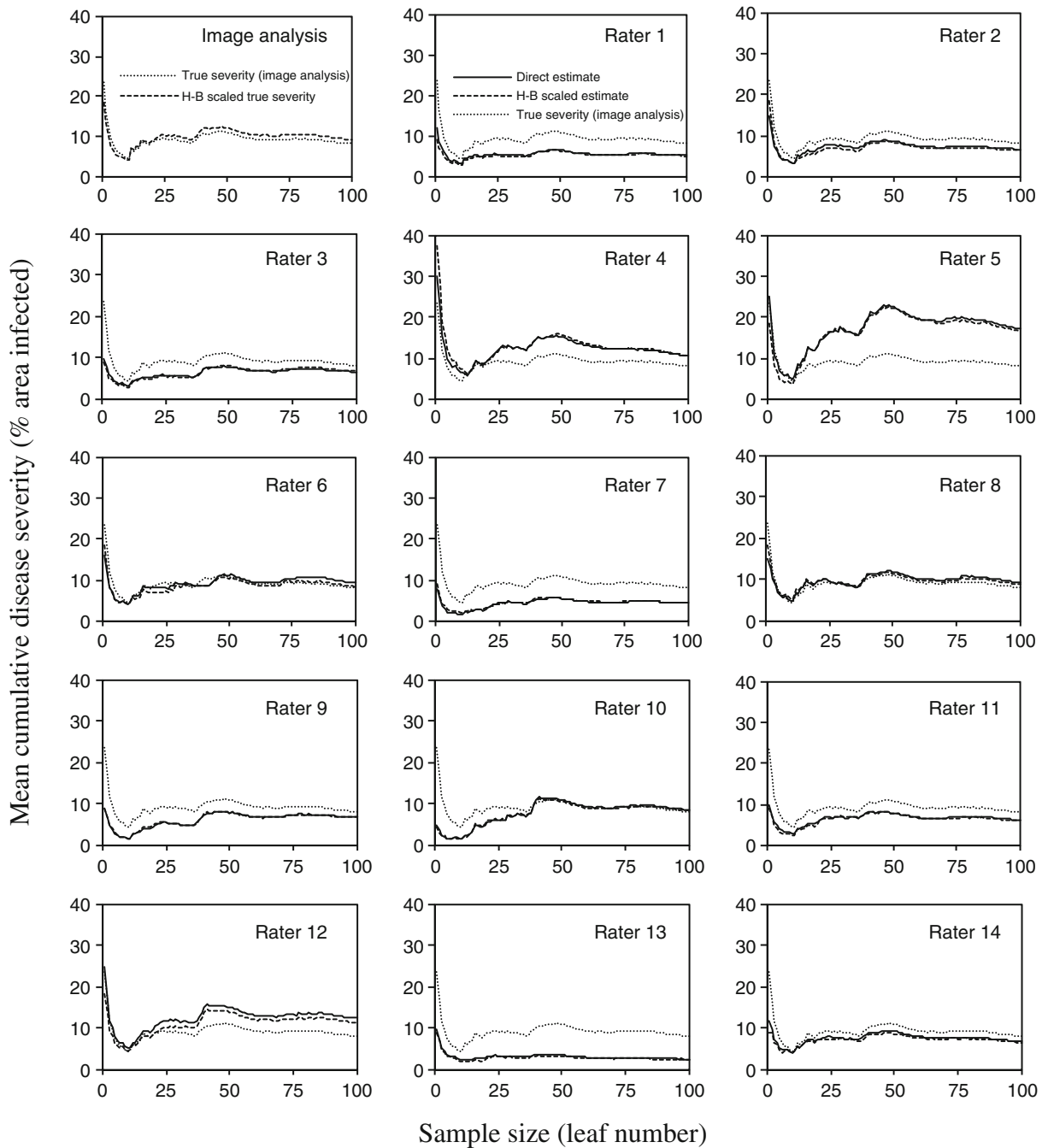


Fig. 2 The cumulative leaf-by-leaf true mean citrus canker severity by image analysis, and for direct estimation and Horsfall-Barratt (H-B) scaled data for the first fourteen raters (only leaves 1 to 100 shown)

consistent nor beneficial effect on the variance of the cumulative mean for the true value or the individual rater estimates. These results were reflected in the final sample means of all 200 leaves for the true values and each rater (Fig. 4A). With this sample of

200 leaves (true mean disease 6.56%), H-B scaling had only a slight effect on the mean estimate, and in most raters cases caused a very slight increase (<1%) in the mean estimate compared to direct estimation. The grand mean estimate was close to the true value

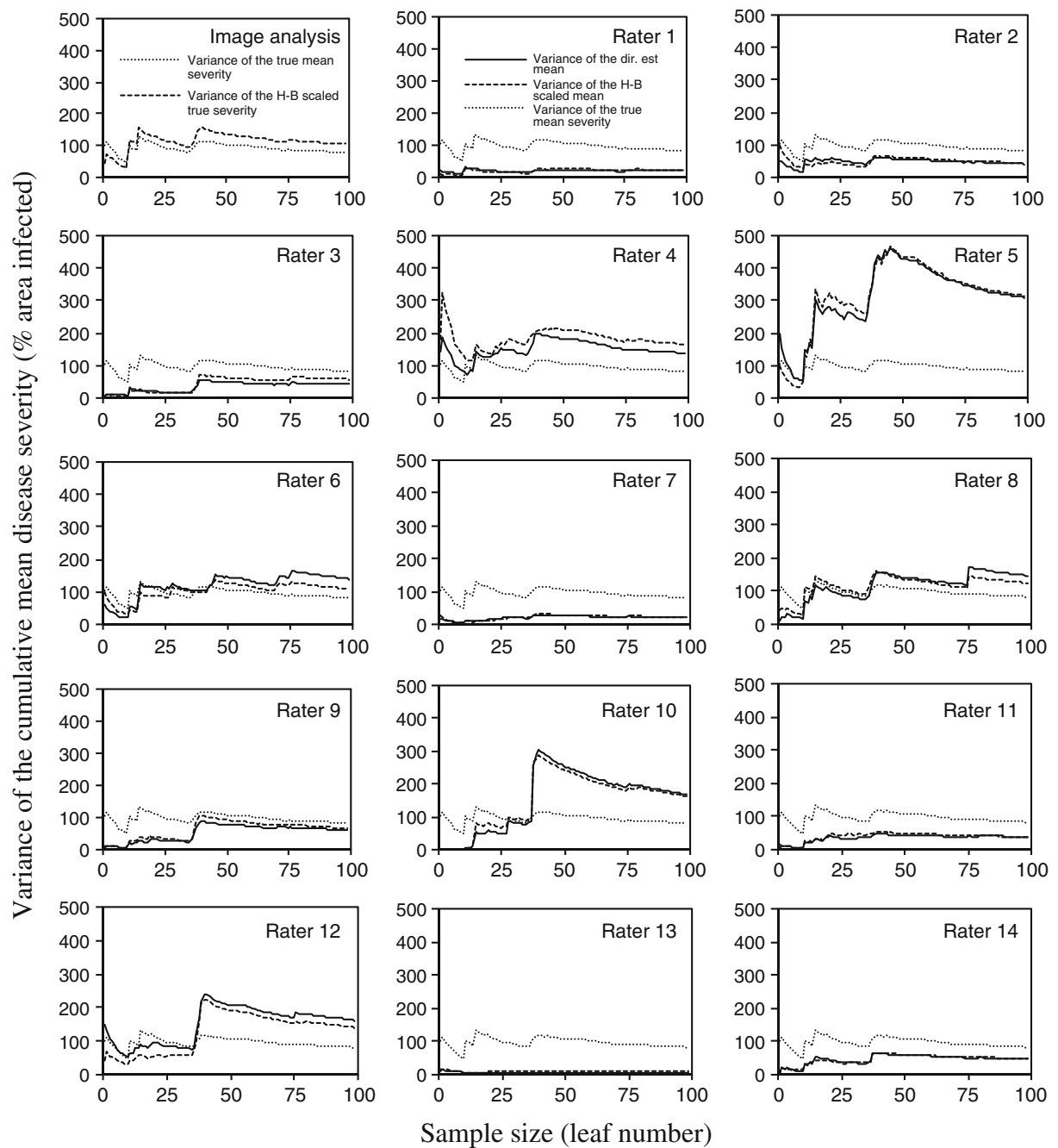
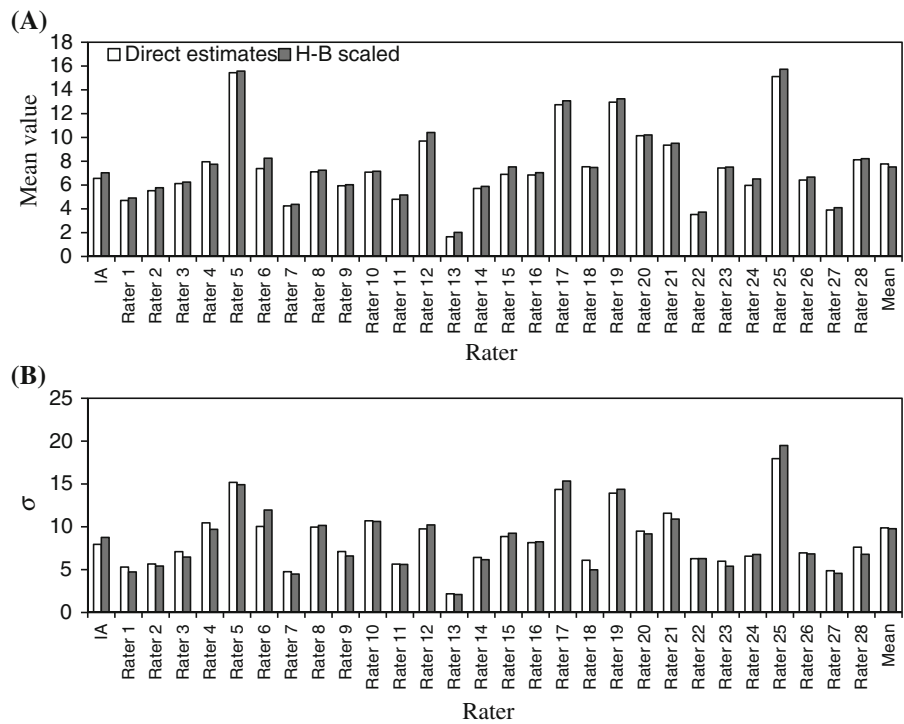


Fig. 3 The variance of the cumulative mean of the leaf sample of the true citrus canker severity for image analysis, direct estimation (dir. est.) and for Horsfall-Barratt (H-B) scaled data for the first fourteen raters (only leaves 1 to 100 shown, means shown in Fig. 2)

for both assessment methods. The standard deviation of the sample for each rater was similar for both direct estimated data and H-B scaled data (Fig. 4B), but was not consistently greater or smaller with either method. The magnitude of the standard deviation was depen-

dent on the rater ability compared to the true value (true sample standard deviation=7.95). The proportion of the total estimates that were closer to the true values was greater for direct estimation (59.5% of estimates) compared to H-B scaling, suggesting a

Fig. 4 The mean severity of a sample of two hundred canker-infected grapefruit leaves and the standard deviation ($\sigma = \sqrt{\frac{\sum x^2}{n}}$) of the mean for true values (by image analysis, IA), direct estimates and Horsfall-Barratt (H-B) scaled data for twenty-eight raters



greater discrepancy was more frequent when using the H-B scale. With cumulative mean estimates of disease for each rater, direct estimation was closer to the true disease 53.7% of the time, and the variance of these data was most often closer to the true variance (63.6% of the time).

Effect of H-B scaling on the rater-mean estimates of single diseased leaves

A linear relationship existed between all rater-mean estimates based on different numbers of raters and the true values (Table 3 and Fig. 5A–E). The coefficient

Table 3 Linear regression^a analysis between the true citrus canker severity and severity using either direct estimation or Horsfall-Barratt scaled data for two-hundred grapefruit leaves using mean severity for each leaf based on 3, 5, 10, 20 or 28 raters

Assessment method	Sample size (rater-means)	<i>a</i>	<i>b</i>	<i>r</i> ²	<i>CV</i>
Direct estimate	3	1.53***	0.63***	0.87	33.5
	5	2.01***	0.92***	0.91	29.4
	10	1.21***	0.94***	0.91	32.1
	20	1.58***	0.95***	0.92	28.8
	28	1.56***	0.95***	0.93	27.5
Horsfall-Barratt scaled	3	1.18***	0.65***	0.84	42.2
	5	1.80***	0.94***	0.88	35.0
	10	1.10***	0.93***	0.90	34.0
	20	1.47***	0.93***	0.91	30.1
	28	1.45***	0.93***	0.92	29.1

^aWhere the regression model $y=a+bx$ represents estimated % area, x is the true area infected, a is the intercept, and b the slope parameter. True severity measured with image analysis. r^2 =coefficient of determination, CV =coefficient of variation of the regression. Significance level of the slope from “1” and the intercept from “0” is indicated by ***= $P<0.001$, **= $P<0.01$, *= $P<0.05$, respectively

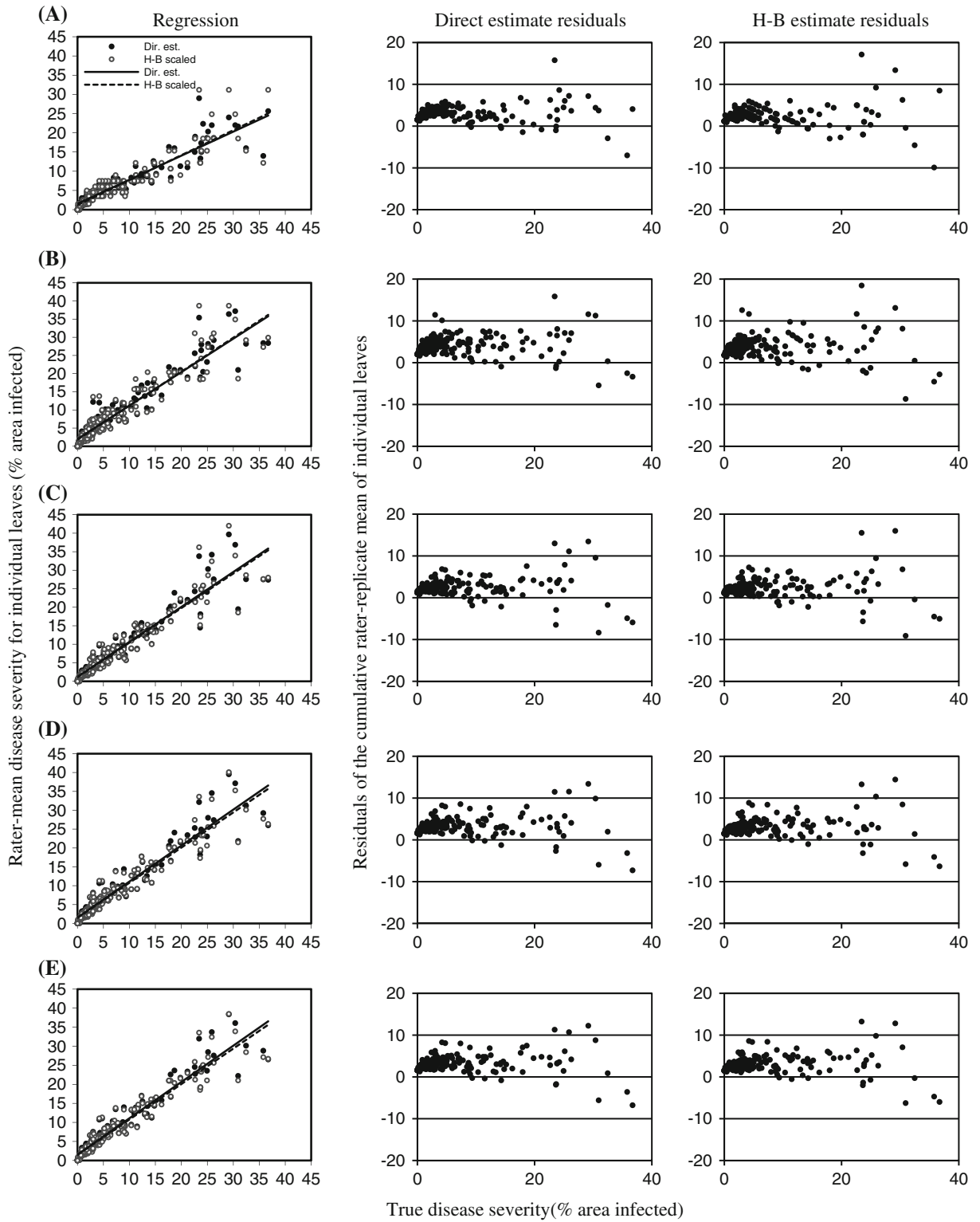


Fig. 5 The relationship and the residuals of the relationship between true citrus canker severity and the mean estimate of symptom severity by direct estimation (dir. est.) and for Horsfall-Barratt (H-B) scaled data for two-hundred leaves based on means of 3 (A), 5 (B), 10 (C), 20 (D) and 28 (E) raters. The residuals (e) for the i^{th} leaf were calculated as $e_i = y_i - (a + (b \times x_i))$, where x is the true area infected, a is the intercept, and b is the slope parameter. See regression parameters in Table 3

of determination (r^2) and the coefficient of variation (CV) suggested H-B scaled data were less precise compared to direct estimates for rater-means based on small samples sizes, but improved as individual leaf severities were based on estimates by more raters. The GLM analysis was unable to discern significant differences in the mean in slope or elevation at any sample size using either method (data not shown). An overall tendency to overestimate was pronounced at low disease severity (<10%) and is evident in the residual plots. The data for rater-means based on different numbers of raters was heteroscedastic using either rating method. Estimates of severity on leaves with >20% true area infected were not as accurate as those of less severely infected. To check the heteroscedasticity was not due to inaccurate image analysis, five of the outlying leaves with means based on 28 raters (23.4, 25.9, 30.4, 30.9 and 35.8% true area

infected) were printed on paper and the areas diseased cut out and weighed (Nita et al. 2003) and compared to the true area by image analysis. True disease by weighing was within 0.8–3.7% of the true area measured by image analysis, while the mean estimate of 28 raters compared less well to the true value by image analysis (5.7–8.7%) or by weighing (2.0–9.5%).

The LCC coefficient and its component parameters (Table 4) showed that rater-means based on more raters had better agreement with true values compared to individual estimates, but there were only minor improvements when the rater-mean was based on five or more raters. Rater-means based on a few raters (3–5 raters) were moderately more precise using direct estimated data compared to H-B scaled data (means of 3 raters, $r=0.94$ and 0.91 , respectively). Basing the rater-mean on more raters (≥ 10 raters) resulted in small improvements in precision for both assessment methods (means of 10 raters, $r=0.95$). The bias correction factor, C_b , for mean severities based on 5 or more raters using direct estimation or H-B scaling was >0.98 . Regardless of method, the scale shift (v) showed there was little deviation of the variance of the estimate compared to that of the true values for means based on ≥ 5 raters (0.97 – 1.00), but overall the location shift (μ) showed a small tendency to overestimate disease (means of ≥ 5 raters, $\mu=0.08$ – 0.19). Overall, the measure of agreement by the

Table 4 Correlation analysis^a and Lin’s concordance correlation analysis^b of the severity of citrus canker on two-hundred grapefruit leaves either estimated directly or for Horsfall-Barratt

scaled data using mean canker severity for each leaf based on 3, 5, 10, 20 or 28 raters

Assessment method	Sample size (rater-means)	r	v (scale)	μ (location)	C_b	ρ_c
Direct estimate	3	0.94	0.67	−0.14	0.92	0.86
	5	0.95	0.97	0.19	0.98	0.93
	10	0.95	0.99	0.11	0.99	0.95
	20	0.96	0.99	0.16	0.99	0.95
	28	0.96	0.99	0.16	0.99	0.95
Horsfall-Barratt scaled	3	0.91	0.71	−0.17	0.93	0.85
	5	0.94	1.00	0.17	0.99	0.92
	10	0.95	0.98	0.08	1.00	0.95
	20	0.96	0.97	0.13	0.99	0.95
	28	0.96	0.97	0.13	0.99	0.95

^a r =correlation coefficient

^b v =slope shift (1=no bias between X and Y) relative to the concordance line; μ =location, or height shift (0=no bias) relative to the concordance line; C_b =the bias correction factor, is dependent on the scale (v) and location (μ) shifts; ρ_c =Lin’s concordance correlation is the product of precision (r) and accuracy (C_b)

LCC coefficient was the same for both assessment methods with means based on ten or more raters ($\rho_c=0.95$).

The logarithm of the variance for all rater-mean estimates had a linear relationship with the logarithm of the true disease severity for both methods (Fig. 6A–E). Rater-mean estimates based on more raters had less dispersed variance for both assessment methods (Table 5). The coefficient of determination (r^2) showed that variability of mean estimates based on 3 raters was least for direct estimates ($r^2=0.63$) compared to H-B scaled data ($r^2=0.41$), but with means based on 10 or more raters, variability was similar (10-rater-means, $r^2=0.83$ and 0.82 , respectively). GLM analysis showed no difference in the slope between the two methods with severity means based on 3–10 raters (Table 6), but with means based on more raters (≥ 20) the variance based on direct estimates was significantly lower compared to H-B scaled data. Ability to discern difference in variance when rater-means are based ≥ 20 raters might be due to the more precise nature of these values (Lin 1989).

Discussion

Effect of H-B scaling on reliability and agreement of individual estimates

The inter-rater reliability of estimates between raters was diverse using both direct estimates and H-B scaled data, but the H-B scaled data was most often less reliable, which agrees with previous observations (Nita et al. 2003), suggesting that there was no gain in inter-rater reliability from using the H-B scale to assess symptoms of citrus canker on grapefruit leaves.

Compared to the true values, the H-B scale did not improve agreement of the estimate for individual raters. The H-B scale most often resulted in lower precision, which has been noted before (Forbes and Korva 1994; Nita et al. 2003). In this study generalised bias (accuracy) was not demonstrably different for direct estimation of citrus canker symptoms compared to the H-B scale—and there is no prior way of determining whether a single disease severity estimate using the H-B scale or direct estimation will be closest to the true disease value. When the true disease happens to be the same as, or very close to an H-B category mid-point (0, 1.5, 4.5,

Fig. 6 The relationship and the residuals of the relationship between the log true citrus canker severity and the log variance of the cumulative rater-mean by direct estimation (dir. est) and for Horsfall-Barratt (H-B) scaled data for two-hundred leaves based on means of 3 (A), 5 (B), 10 (C), 20 (D) and 28 (E) raters. The residuals (e) for the i^{th} leaf were calculated as $e_i=y_i-(a+(b \times x_i))$, where x is the true area infected, a is the intercept, and b is the slope parameter. See regression parameters in Table 5

9.0, 18.5, 37.5...% etc) the H-B scale will provide good agreement, but if the true disease, and direct estimate, is close to a category boundary (0, 3, 6, 12, 25, 50...% etc), then the H-B scale will be less accurate, which is more likely to occur where the intervals in the scale are wider (in the midrange of the scale) and if the estimate is in the wrong category error might be further exacerbated.

The estimate of cumulative mean disease severity on a leaf-by-leaf basis by each rater explored the effect of sample size in relation to the true values of disease using both direct estimation and H-B scaled data. The effect of rater bias on the mean estimate was much greater than any effect of the H-B scale on the mean estimate. Some raters consistently over— or underestimated the cumulative true mean, but the mean estimates (and the variance of the means) by both methods tracked each other closely, although the direct estimated mean was most often closer to the true mean, and the variance of the direct estimated mean most often closest to the variance of the true mean suggesting there was no advantage in terms of accuracy or precision in using the H-B scale.

The effect of rater-means and H-B scaling on agreement with the true values

The H-B scale did not improve agreement of the rater-mean estimates with the true values compared to direct estimates, but all rater-estimates had a linear relationship with true disease (Slopek 1989; Nutter and Schultz 1995; Nita et al. 2003). Rater-mean values were more accurate and precise when based on data from a larger sample of raters. Mean-rater estimates based on ≥ 5 raters showed that both methods had similar, good agreement with the true values, but a few rater-mean estimates of individual leaves were still inaccurate with 28 raters suggesting inherent and widespread rater bias to over— or underestimate severity on those particular leaves.

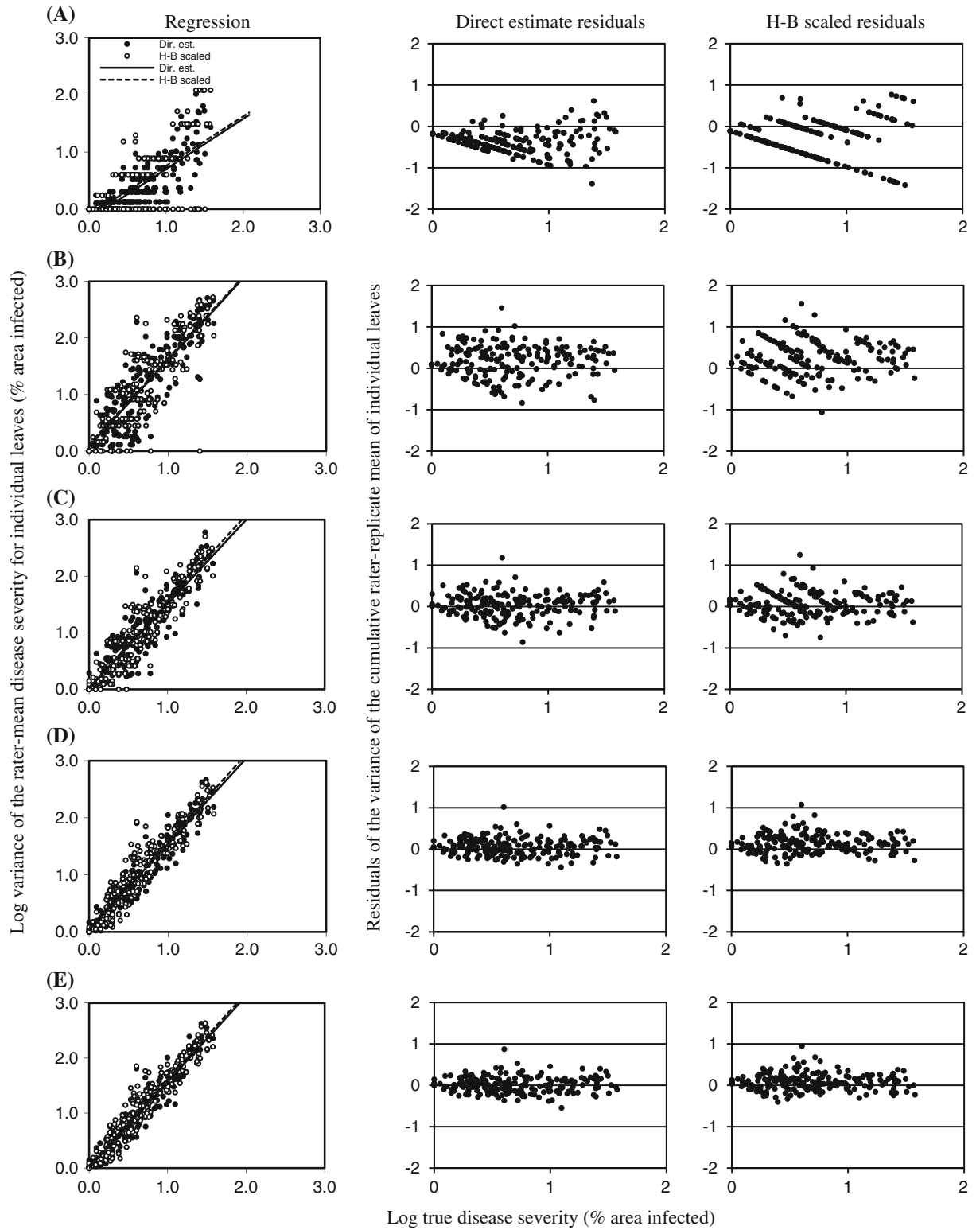


Table 5 Linear regression analysis^a of the relationship between log true disease severity and the log variance of the direct estimate and log variance of the H-B scaled data for severity of

citrus canker on two-hundred grapefruit leaves based on mean canker severity for each leaf by 3, 5, 10, 20 or 28 raters

Assessment method	Sample size (rater-means)	<i>a</i>	<i>B</i>	<i>r</i> ²	<i>CV</i>
Direct estimate	3	-0.18***	0.88***	0.63	64.5
	5	0.09	1.52***	0.73	33.5
	10	0.02	1.49***	0.83	26.9
	20	0.03	1.52***	0.90	19.4
	28	0.004	1.56***	0.92	17.8
Horsfall-Barrat scaled	3	-0.11	0.87***	0.41	87.6
	5	0.12*	1.52***	0.68	36.7
	10	0.03	1.53***	0.82	27.4
	20	0.07*	1.53***	0.89	19.9
	28	0.04	1.57***	0.91	18.3

^aWhere the regression model $y=a+bx$ represents estimated % area, x is the true area infected, a is the intercept, and b the slope parameter. True severity measured with image analysis. r^2 =coefficient of determination, CV =coefficient of variation of the regression. Significance level of the slope from 1 and the intercept from 0 is indicated by ***= $P<0.001$, **= $P<0.01$, *= $P<0.05$, respectively

Both direct estimation and the H-B scale gave similar sample means for individual raters for the 200 leaves, with H-B means being <1% different to those by direct estimation, reflecting previous data on means estimated by both methods (Nita et al. 2003). H-B scaled data appeared to have a greater effect on precision of an individual's estimates compared to the effect on the accuracy (bias), and this imprecision was reflected in larger and more scattered variance of the rater-mean estimates with H-B scaling. Variance (s^2) is a measure of the dispersion of the individual

sample estimates, and is used to determine the standard deviation ($\sigma = \sqrt{\frac{s^2}{n}}$) which describes the dispersion of the data about the sample mean value, and is thus a measure of precision of the sample mean and relevant to sample size estimation (Everitt 1998; Madden et al. 2007). Based on these data, a major effect of the H-B scale was on precision of the estimates, and the effect of less precise and/or inaccurate visual assessment data has already been demonstrated (Parker et al. 1995) to increase the probability of a Type II error falsely accepting the null

Table 6 General linear modelling analysis^a of the relationship between true citrus canker severity and the variance of the mean directly estimated (dir. est.) severity or for Horsfall-Barratt (H-

B) scaled data for two-hundred grapefruit leaves based on means of 3, 5, 10, 20 and 28 raters

Sample size (rater-means)	Assessment method	Class difference: slope (test for interaction group*regressor) Pr>F	Test of parallelism
3	Dir. est H-B	0.88 _{1,396} $P>0.05$	2.86 _{1,399} $P>0.05$
5	Dir. est H-B	0.96 _{1,396} $P>0.05$	0.44 _{1,399} $P>0.05$
10	Dir. est H-B	0.56 _{1,396} $P>0.05$	0.26 _{1,399} $P>0.05$
20	Dir. est H-B	0.71 _{1,396} $P>0.05$	4.49 _{1,399} $P=0.04$
28	Dir. est H-B	0.88 _{1,396} $P>0.05$	4.87 _{1,399} $P=0.03$

^aRegression slopes were tested to determine differences in disease severity (ds) due to assessment method (am) when compared to true disease severity (tds). A test for interaction between the regressor and class ($ds=tds, am, tds*am$) gauges whether there are differences due to assessment method which are reflected in slope for the two classes. If the interaction term is significant the classes are different. If the interaction is not different a further test for differences in the elevation of the regression lines (a test of parallelism) is required, which is achieved by running GLM on the reduced model (no interaction term, $ds=tds, am$). P =probability> F

hypothesis (Everitt 1998). Furthermore, superimposing additional imprecision on an already subjective assessment method might exacerbate error (Forbes and Korva 1994).

Variance of the rater-mean estimates was not constant with magnitude of true disease, which is in agreement with some previous observations (Forbes and Korva 1994; Forbes and Jeger 1987, Hau et al. 1989; Bock et al. 2008b, 2009). Even with rater-means based on estimates by ten or more raters, specific leaves were over— or underestimated more than others, particularly at true severities >20%. The reason for this is unknown, but might be associated with widespread bias in delineation of more complex symptoms encountered on more severely cankered grapefruit leaves, causing raters to overestimate (or underestimate) disease on those leaves. Raters are known to overestimate disease, particularly at low severity (<10%), and the number of lesions relative to area infected also influence rater estimates (Amanat 1976; Sherwood et al. 1983; Hau et al. 1989; Bock et al. 2008b; 2009). Other uncharacterised illusions or factors of symptom topography might result in rater error.

Advantages and disadvantages of the H-B scale

These results suggest the H-B scale does not improve reliability, accuracy or precision of severity estimates of citrus canker. Indeed, the estimates were generally less reliable or precise. Estimates of disease severity in plots are most often mean values based on a predetermined sample size of diseased leaves or plants (Madden et al. 2007), but the effects of sample size on the accuracy and precision of the mean value using direct estimation or H-B scaling has yet to be explored. Large sample size was an original stipulation for use of the H-B scale (Horsfall and Barratt 1945; Horsfall 1945; Horsfall and Cowling 1978) and the data presented here suggest that for individual raters the mean estimate of disease for 200 leaves was similar (within 1%) using either method, but dependent on rater accuracy (bias). There are circumstances where H-B type-scales, or other interval scales (Slopek 1989; Nita et al. 2003) might be useful or warranted for reasons of practicality. Interval scales are relatively easy to learn, and can be faster to apply in the field allowing for large sample sizes where time is a limiting factor, but if they are to be used, scales with equal increments might be preferable to loga-

rithmic-based ones (Nita et al. 2003; Nutter and Esker 2006; Madden et al. 2007).

The psychophysical basis of the perception of plant disease remains poorly defined across the full range of disease (Nutter and Esker 2006), but based on these results H-B scaling gave no predictable or consistent benefit to estimates of canker severity. The H-B scale continues to be used in plant pathology (Hagan et al. 2003; Keinath et al. 2006; Stevenson et al. 2007), and has been adopted in other disciplines (Pernezney et al. 2008; Bussotti et al. 2003; Chappelka et al. 2003). Choosing to use an H-B type disease scale will depend on a decision based on the trade-offs between time taken to perform the assessment, sample size, precision and accuracy, and reliability of the data.

References

- Amanat, P. (1976). Stimuli effecting disease assessment. *Agriculturae Conspectus Scientificus*, 39, 27–31.
- Baird, J. C., & Norma, E. (1978). *Fundamentals of scaling and psychophysics*. New York, NY: Wiley.
- Belasque Junior, J., Bassanezi, R. B., Spósito, M. B., Ribeiro, L. M., de Jesus Júnior, W. C., & Amorim, L. (2005). Escalas diagramáticas para avaliação da severidade do cancro cítrico. *Fitopatologia Brasileira*, 30, 387–393.
- Birnbaum, M.H. (1994). Psychophysics. In *Encyclopedia of Human Behavior*, San Diego: Academic.
- Bock, C. H., Parker, P. E., Cook, A. Z., & Gottwald, T. R. (2008a). Visual assessment and the use of image analysis for assessing different symptoms of citrus canker on grapefruit leaves. *Plant Disease*, 92, 530–541.
- Bock, C. H., Parker, P. E., Cook, A. Z., & Gottwald, T. R. (2008b). Characteristics of the perception of different severity measures of citrus canker and the relations between the various symptom types. *Plant Disease*, 92, 927–939.
- Bock, C.H., Parker, P.E., Cook, A.Z., Riley, T. & Gottwald, T.R. (2009). Comparison of assessment of citrus canker foliar symptoms by experienced and inexperienced raters. *Plant Disease*, 93, 412–424.
- Bussotti, F., Schaub, M., Cozzi, A., Krauchi, N., Ferretti, M., Novak, K., et al. (2003). Assessment of ozone visible symptoms in the field: perspectives of quality control. *Environmental Pollution*, 125, 81–89.
- Chappelka, A. H., Neufeld, H. S., Davison, A. W., Somers, G. L., & Renfro, J. R. (2003). Ozone injury on cutleaf coneflower (*Rudbeckia lacinata*) and crown-beard (*Verbesina occidentalis*) in Great Smoky Mountains National Park. *Environmental Pollution*, 125, 53–59.
- Everitt, B. S. (1998). *The Cambridge dictionary of statistics*, p. 360. Cambridge, UK: Cambridge University Press.
- Forbes, G. A., & Jeger, M. J. (1987). Factors affecting the estimation of disease intensity in simulated plant structures. *Zeitschrift für Pflanzenkrankheiten und Pflanzenschutz*, 94, 113–120.

- Forbes, G. A., & Korva, J. T. (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 Pathology*, *43*, 675–682.
- Gottwald, T. R., & Timmer, L. W. (1995). The efficacy of windbreaks to reduce in the spread of citrus canker caused by *Xanthomonas campestris* pv. *citri*. *Tropical Agriculture*, *72*, 194–201.
- Gottwald, T. R., Graham, J., Civerolo, J. H., Barrett, H. C., & Hearn, C. J. (1993). Differential host range reaction of citrus and citrus relatives to citrus canker and citrus bacterial spot determined by leaf mesophyll susceptibility. *Plant Disease*, *77*, 1004–1009.
- Hagan, A.K., Olive, J.W., Stephenson, J. & Rivas-Davilla, M. (2003). Impact of application rate and treatment interval on the control of *Entomosporium* leaf spot, daylily rust, and powdery mildew on container-grown nursery crops with MilStop. Timely Information - *Plant Pathology Series PP 557*. Auburn University, AL.
- Hau, B., Kranz, J., & König, R. (1989). Fehler beim Schätzen von Befallsstärken bei Pflanzenkrankheiten. *Zeitschrift für Pflanzenkrankheiten und Pflanzenschutz*, *96*, 649–674.
- Herbert, T. T. (1982). The rationale for the Horsfall-Barratt plant disease assessment scale. *Phytopathology*, *72*, 1269.
- Horsfall, J.G. (1945). Fungicides and their action. *Annales cryptogamica et phytopathologici, Vol II. Chronica Botanica*, 239 pp. Waltham, MA.
- Horsfall, J.G. (1986). This week's citation classic: Horsfall J.G. & Barratt, R.W. An improved grading system for measuring plant disease. *Phytopathology* *35*, 655. 1945. *Current Contents (Agriculture, Biology and Environmental Sciences)*, *15*, 14.
- Horsfall, J. G., & Barratt, R. W. (1945). An improved grading system for measuring plant disease. *Phytopathology*, *35*, 655. (Abstract).
- Horsfall, J. G., & Cowling, E. B. (1978). Pathometry: the measurement of plant disease. In J. G. Horsfall & E. B. Cowling (Eds.), *Plant Disease: An Advanced Treatise Vol II*, pp. 120–136. New York: Academic.
- Keinath, A. P., Dubose, V. B., May, W. H., & Wechter, W. P. (2006). Effect of Actigard on *Xanthomonas* leaf spot on rape greens, 2005. *Fungicide and Nematocide Tests*, *61*, V058.
- Koch, H., & Hau, B. (1980). Ein psychologischer aspect beim schätzen von pflanzenkrankheiten. *Zeitschrift für Pflanzenkrankheiten und Pflanzenschutz*, *87*, 587–593.
- Kranz, J. (1988). Measuring plant disease. In J. Kranz & J. Rotem (Eds.), *Experimental Techniques in Plant Disease Epidemiology*, pp. 35–50. New York: Springer-Verlag.
- Lang, J. M., Gent, D. H., & Schwartz, H. F. (2007). Management of *Xanthomonas* leaf blight of onion with bacteriophages and a plant activator. *Plant Disease*, *91*, 871–878.
- Lin, L. I.-K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, *45*, 255–268.
- Madden, L. V., Hughes, G., & van den Bosch, F. (2007). *The study of plant disease epidemics*. St. Paul, MN: APS Press.
- Mila, A. L., Driever, G. F., Morgan, D. P., & Michailides, T. J. (2005). Effects of latent infection, temperature, precipitation, and irrigation on panicle and shoot blight of pistachio in California. *Phytopathology*, *95*, 926–932.
- Nita, M., Ellis, M. A., & Madden, L. V. (2003). Reliability and accuracy of visual estimation of *Phomopsis* leaf blight of strawberry. *Phytopathology*, *93*, 995–1005.
- Nutter, F. W., Jr, & Schultz, P. M. (1995). Improving the accuracy and precision of disease assessment: selection of methods and use of computer-aided training programs. *Canadian Journal of Plant Pathology*, *17*, 174–178.
- Nutter, F. W., Jr, & Esker, P. D. (2006). The role of psychophysics in phytopathology. *European Journal of Plant Pathology*, *114*, 199–213.
- Nutter, F. W., Jr, Gleason, M. L., Jenco, J. H., & Christians, N. C. (1993). Assessing the accuracy, intra-rater repeatability, and inter-rater reliability of disease assessment system. *Phytopathology*, *83*, 806–812.
- Parker, S. R., Whelan, M. J., & Royle, D. J. (1995). Reliable measurement of disease severity. *Aspects of Applied Biology*, *43*, 205–214.
- Pernezney, K., Nuessley, G., & Stall, W. (2008). *Integrated pest management for Florida snap beans*. Gainesville, FL: Pamphlet PPP37, IFAS, University of Florida.
- Pethybridge, S. J., Hay, F. S., & Wilson, C. R. (2004). Pathogenicity of fungi commonly isolated from foliar disease in Tasmanian pyrethrum crops. *Australasian Plant Pathology*, *33*, 441–444.
- Redman, C. E., & Brown, I. F., Jr. (1964). A statistical evaluation of the Barratt and Horsfall rating system. *Phytopathology*, *54*, 904. (Abstract).
- Redman, C. E., King, E. P., & Brown Jr, I. F. (1969). *Tables for converting Barratt and Horsfall rating scores to estimated mean percentages*, p. 8. Elanco Products: Indianapolis.
- Shaw, M. W., & Royle, D. J. (1989). Estimation and validation of a function describing the rate at which *Mycosphaerella graminicola* causes yield loss in winter wheat. *Annals of Applied Biology*, *115*, 425–442.
- Sherwood, R. T., Berg, C. C., Hoover, M. R., & Zeiders, K. E. (1983). Illusions in visual assessment of *Stagonospora* Leaf Spot of orchardgrass. *Phytopathology*, *73*, 173–177.
- Slopek, S. W. (1989). An improved method for estimating percent leaf area diseased using a 1 to 5 disease assessment scale. *Canadian Journal of Plant Pathology*, *11*, 381–387.
- Stevenson, W. R., James, R. V., Inglis, D. A., Johnson, D. A., Schotzko, R. T., & Thornton, R. E. (2007). Fungicide spray programs for Defender, a new potato cultivar with resistance to late blight and early blight. *Plant Disease*, *91*, 1327–1336.