Effects of rater bias and assessment method on disease severity estimation with regard to hypothesis testing

K. S. Chiang[a], C. H. Bock[b]*, M. El Jarroudi[c], P. Delfosse[d], I. H. Lee[a] and H. I. Liu[a]

[a]*Division of Biometrics, Department of Agronomy, National Chung Hsing University, Taichung, Taiwan 402;* [b]*USDA-ARS-SEFTNRL, 21 Dunbar Road, Byron, GA 31008, USA;* [c]*Department of Environmental Sciences and Management, Université de Liège, 185 Avenue de Longwy, 6700 Arlon, Belgium; and* [d]*Luxembourg Institute of Science and Technology, 41 Rue du Brill, 4422 Belvaux, Luxembourg*

The effects of bias (over- and underestimates) in estimates of disease severity on hypothesis testing using different assessment methods was explored. Nearest percentage estimates (NPE), the Horsfall–Barratt (H-B) scale, and two linear category scales (10% increments, with and without additional grades at low severity) were compared using simulation modelling to assess effects of bias. Type I and type II error rates were used to compare two treatment differences. The power of the H-B scale and the 10% scale were least for correctly testing a hypothesis compared with the other methods, and the effects of rater bias on type II errors were greater over specific severity ranges. Apart from NPEs, the amended 10% category scale was most often superior to other methods at all severities tested for reducing the risk of type II errors. It should thus be a preferred method for raters who must use a category scale for disease assessments. Rater bias and assessment method had little effect on type I error rates. The power of the hypothesis test using unbiased estimates was most often greater compared with biased estimates, regardless of assessment method. An unanticipated observation was the greater impact of rater bias compared with assessment method on type II errors. Knowledge of the effects of rater bias and scale type on hypothesis testing can be used to improve accuracy and reliability of disease severity estimates, and can provide a logical framework for improving aids to estimate severity visually, including standard area diagrams and rater training software.

Keywords: phytopathometry, plant disease quantification, rating scales

Introduction

The assessment of disease severity (the proportion of a plant unit diseased) is a basic need in many studies in plant disease epidemiology (Madden *et al.*, 2007). Moreover, accurate estimates of disease severity are important for predicting yield loss, for testing treatment (e.g. chemical, biological) efficacy, for assessing crop germplasm for disease resistance, and for understanding fundamental biological processes including coevolution (Bock *et al.*, 2010b). Estimates of disease severity are most often made visually, but must be accurate and reliable. Here 'accuracy' is defined as it is used in measurement science as the closeness of the estimate to the true value, and 'reliability' is the extent to which the same estimate obtained under different conditions yields similar results (Madden *et al.*, 2007). If inaccurate or unreliable disease assessments are obtained, this might lead to faulty conclusions being drawn from the data, which in turn might lead to incorrect actions being taken in disease management (Bock *et al.*, 2010b).

There is a growing body of literature comparing different methods for estimating disease severity based on empirical studies (Forbes & Korva, 1994; Nita *et al.*, 2003; Bock *et al.*, 2008a,b, 2009a, 2013a,b; Bardsley & Ngugi, 2013). Although Forbes & Korva (1994) were the first to use simulation to study aspects of disease assessment, only recently has simulation modelling of disease assessment been used to understand the impacts on hypothesis testing, thereby providing a basis to compare the assessment methods quantitatively (Bock *et al.*, 2010a; Chiang *et al.*, 2014). Hypothesis testing requires that the collected data be sufficiently accurate to reject the null hypothesis (H_0) when H_0 is false, or conversely, to accept H_0 when there are no treatment differences. Failure to reject H_0 when H_0 is false results in commission of a type II error, while rejection of H_0 when H_0 is true results in commission of a type I error. Type II errors have been reported where different methods have been used to compare treatments (Todd & Kommedahl, 1994; Bock *et al.*, 2010b; Chiang *et al.*, 2014), resulting in discrepancies in means separation ranking (Christ, 1991; Todd & Kommedahl, 1994; Parker *et al.*, 1995), and in analysis of the relationship of disease to yield (Vereijssen *et al.*, 2003; Danielsen & Munk, 2004). In addition, it is important to note that there is a theoretical possibility that raters could inadvertently commit a type I error. Type II and type I errors might arise from inaccurate data.

The Bock *et al.* (2010a) and Chiang *et al.* (2014) studies simulated and compared the performance of nearest

*E-mail: clive.bock@ars.usda.gov

© 2015 British Society for Plant Pathology

percentage estimates (NPEs) with the performance of different category scales based on a known disease assessment distribution developed from estimates of actual values over a range of known disease severities. However, the issue of bias (over- and underestimation) and how it might impact hypothesis testing was not addressed in either study. Furthermore, the data in Bock *et al.* (2008a,b) used in the two simulation-study articles are based on a sample of 210 citrus canker-diseased grapefruit leaves with a range of disease severity from 0 to 60%, thus precluding investigation of the effects of error at severities >60%. In many plants, foliar diseases are most often present at severities <50–60% (Kranz, 1977) (leaves often abscise if disease becomes too severe, making it difficult to obtain samples with severity >50%), so the data presented is of great value to the range of disease most often observed in the field for many pathosystems. However, there are several important pathogens that regularly cause severity of disease >50% (e.g. late blight of potato and tomato, cereal rusts; Peterson *et al.*, 1948; Forbes & Korva, 1994; Corrêa *et al.*, 2009; Duarte *et al.*, 2013). Thus it behooves phytopathologists to explore the effects of rater error over the full range of disease severity from 0 to 100%.

There is a widespread tendency to overestimate disease severity at low actual severities (<10%) (Sherwood *et al.*, 1983; Bock *et al.*, 2010b). Furthermore, Forbes & Jeger (1987) found that estimation of severity when due to fewer larger lesions was less error prone compared to estimation of severity when disease was due to numerous small, random or uniformly distributed lesions. Overestimation of disease severity has tremendous ramifications for epidemiological studies when projecting yield loss, where disease progress is measured based on estimates of disease severity. Furthermore, rater bias resulting in overestimates of actual disease severity may adversely affect advances in plant breeding programmes as well (Sherwood *et al.*, 1983). Bock *et al.* (2010b) hypothesized a possible additional reason for rater overestimation at low disease severity: it is not possible to estimate a disease severity <0%, thus all estimates at low disease severity (at least up to 5–10%) experience an invisible ‘barrier’ (i.e. 0%) to underestimation, yet no ‘barrier’ to overestimation. However, the cause of overestimation at any severity does not appear to have been fully explored, although the effect is now well recognized (Amanat, 1976; Sherwood *et al.*, 1983; Beresford & Royle, 1991; Forbes & Korva, 1994; Bock *et al.*, 2008b, 2009b). Also, estimates close to 100% are constrained by a possible maximum, so it follows that there is likely to be a tendency to underestimate at severities approaching 100% (El Jarroudi *et al.*, 2015), although this has not been explored previously. Thus, considering these inherent tendencies, it is particularly relevant to explore the effects of bias on hypothesis testing when using different assessment methods to estimate disease severity.

The occurrence of disease symptoms in some pathosystems at severities in excess of 50% is a compelling reason to investigate the effects of assessment methods over the full severity range from 0 to 100%. Thus, in this study, a data set of rater estimates and actual measured values of septoria leaf blotch (SLB, caused by *Zymoseptoria tritici*) severity on winter wheat spanning the range 0 to 100% was used. The data set was previously described in the development of a decision support system for fungicide application (El Jarroudi *et al.*, 2009, 2012a,b, 2015). To the best of the authors’ knowledge, no previous study has comprehensively defined the characteristics of assessment from 0 to 100%, although some limited information is available on estimation in this range (Hau *et al.*, 1989; Forbes & Korva, 1994). Bock *et al.* (2010a) and Chiang *et al.* (2014) considered a range of disease severity from 0 to 50%. In each of these studies the relationship between the standard deviation of the rater mean NPEs and actual disease was described as a hyperbolic function. However, over the range of disease severity from 50 to 100%, this relationship is likely to be different, as the standard deviation almost certainly declines toward zero at 100% severity.

A better understanding of the disease assessment process, the ramifications of error, and a basis on which to determine improved methods to assess disease severity are required. Such knowledge will help provide a framework to develop improved scales, aids for visual severity estimation, including optimizing the number and range of severities for standard area diagrams (SADs), and improved methods of using severity assessment training software. The purpose of this study was to investigate the effects of rater bias and assessment scale method on hypothesis testing over the full range of disease severity from 0 to 100%.

Materials and methods

Leaves of winter wheat with symptoms of SLB

The assessed leaves were sampled from plants in both fungicide-treated and control (no fungicide spray) plots in field experiments in the Grand-Duchy of Luxembourg that were previously described in detail (El Jarroudi *et al.*, 2009, 2015; Bock *et al.*, 2015).

Estimates of disease severity

For each assessed leaf, visual estimates (NPEs) of SLB and associated senescence were made on the flag leaf (F1), and on the two leaves below the flag leaf (F2 and F3). Five wheat stems per plot were assessed by each of four raters on both the fungicide-treated and control plots. The characteristics of the rating abilities of these four raters has been described previously (Bock *et al.*, 2015; El Jarroudi *et al.*, 2015); they covered a spectrum of ability to overestimate (raters 3 and 4), underestimate (rater 2) and show relative accuracy of estimation (rater 1), thereby providing a fair representation of the rater population. A total of 20 plants from each treatment were assessed on each assessment date (El Jarroudi *et al.*, 2015). In both 2006 and 2007, assessments were made weekly from the end of May to the beginning of July, with final observations at GS 73 to GS 77 (Zadoks *et al.*, 1974), assuring a wide range of disease severities. There were six assessments in 2006 and four in 2007. Images of the same leaves used for the visual assessments were taken with a digital camera (Powershot A620, 7.1 megapixels; Canon Inc.), and the area with symptoms of SLB and associated senescence was measured using Assess v. 2.0 (APS Press; Lamari, 2002). Due to advanced leaf senescence and death, some sample sizes were <20 with the later sample dates for leaves F2 and F3, in both years. In the data set, for non-treated plots, a total of 345 leaves in 2006 and 201 leaves in 2007 were photographed, image analysed and assessed; for fungicide-treated plots, 240 leaves in 2006 and 171 leaves in 2007 were subjected to the same procedure (a grand total of 957 leaves).

The absolute errors (visual estimate minus digital image measurements) were plotted against the assumed actual severity (digital image measurements) for the SLB data set for each rater (Fig. 1). Rater 1 was exceptionally accurate at estimating the diseased area; rater 2 tended to underestimate SLB severity; but raters 3 and 4 tended to overestimate SLB severity.

Simulations and hypothesis testing

First, the lognormal distribution has the advantage that the tails do not tend to infinity, which is realistic for estimation of disease severity using the percentage scale, with a minimum of 0 and a maximum of 100% (Bock *et al.*, 2010a; Chiang *et al.*, 2014); thus, the frequency of NPEs of specific actual disease severities by raters was assumed to follow a lognormal distribution. That is,

$$y_i \sim \text{Lognormal}\ (\mu, \rho^2) \qquad (1)$$

$$\text{where } \mu = \ln(\mu_{\text{rater}}) - \frac{1}{2}\ln\left[1 + \left(\frac{\sigma_{\text{rater}}}{\mu_{\text{rater}}}\right)^2\right] \qquad (2)$$

$$\text{and } \rho = \sqrt{\ln\left[1 + \left(\frac{\sigma_{\text{rater}}}{\mu_{\text{rater}}}\right)^2\right]} \qquad (3)$$

Here, the mean rater-estimated severity, μ_{rater}, was regarded as a linear function of the actual severity (Y_{actual}); σ_{rater} was regarded as a function of Y_{actual} determined by the rater estimates of severity of SLB on leaves of winter wheat. The relationships are as follows:

$$\mu_{\text{rater}} = \theta Y_{\text{actual}} \qquad (4)$$

$$\text{and } \sigma_{\text{rater}} = f(Y_{\text{actual}}) \qquad (5)$$

For the unbiased situation, θ for Eqn (4) is constant and equals 1. However, for situations of over- or underestimation, θ is not constant because the effects of rater bias are different for individual raters. Also, if the same rater assesses the same leaf several times, the results should be variable. Therefore, θ is originally assumed to be 1. But to account for the bias, further simulation values were produced in order to account for the uncertainty of bias (both over- and underestimation).

When a value from a treatment was simulated, it represented the actual severity (Y_{actual}). Subsequently, the rater-estimated severity (μ_{rater}) and standard deviation (σ_{rater}) were acquired through Eqns (4) and (5). The parameters of the lognormal distribution were obtained using Eqns (2) and (3). Finally, a simu-

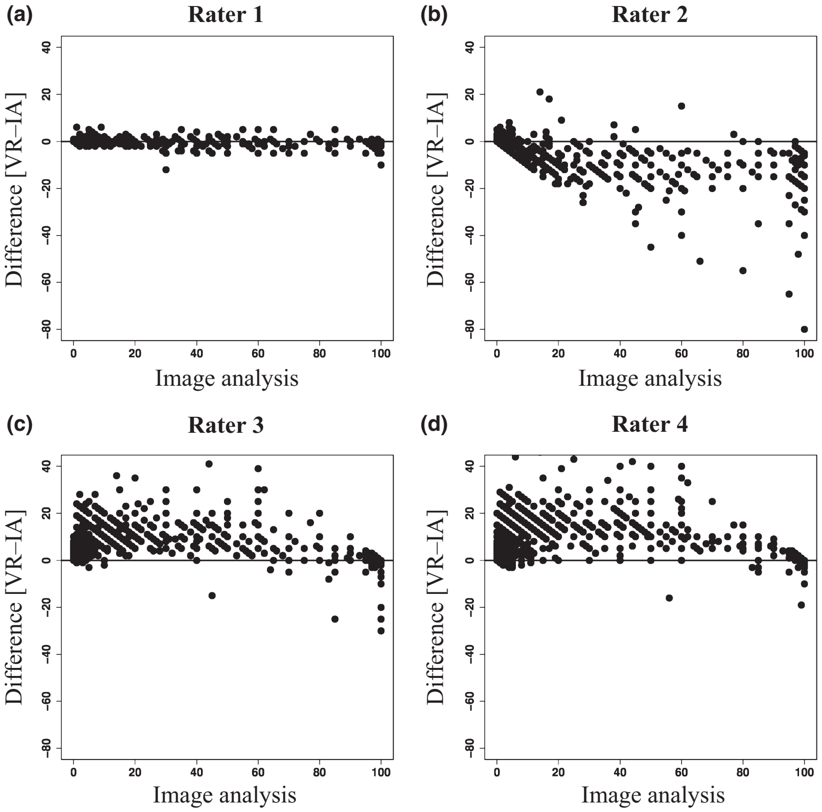

Figure 1 Absolute errors (visual estimate minus digital image analysis measurements, VR-IA) plotted against the actual severity (digital image measurements) for estimates of septoria leaf blotch severity on leaves of winter wheat as estimated by four different raters (raters 1–4).

lated value based on the distribution of rater-estimated disease severities was obtained using Eqn (1).

In order to quantify the effect of rater bias (over- and underestimation), a simulation approach was employed, as outlined in Figure 2. The lognormal distribution is a positively skewed distribution. As usual, there is a lognormal distribution for disease at low severities, and a 1 minus lognormal distribution (a negatively skewed distribution) for high disease severities. The SLB data set used in this study confirms these characteristics. If the effect of rater bias was overestimation, only random samples with values greater than the mean of both distributions were drawn, in order to represent the effect of the overestimation. On the other hand, where rater ability was characterized by underestimation, only random samples with values less than the mean of both distributions were drawn. However, the right shaded area of Figure 2b was not used because overestimation of high severity is of less interest.

The performance of different assessment methods was compared. Assuming that two treatments, A and B, affect epidemics, the disease severity distribution of treatment A has mean μ_A and treatment B has mean $\mu_B = \mu_A + \mu_\Delta$, where μ_Δ represents the difference between the means of the two severity distributions (Bock *et al.*, 2010a). The standard deviations (φ) of the disease severity distributions of treatments A and B are assumed to be equal. A truncated-normal distribution, rather than a normal distribution, was assumed for each treatment because the actual severities cannot be negative values (Bock *et al.*, 2010b; Chiang *et al.*, 2014).

Comparing assessment methods

Four different disease assessment scales were compared as described in a previous study (Chiang *et al.*, 2014). The characteristics of the assessment scales were: (i) NPEs (disease estimated by the rater to the nearest 1%); (ii) the logarithmic Horsfall–Barratt (H-B) scale (Horsfall & Barratt, 1945); (iii) a linear category scale (10% increments); and (iv) an amended 10% linear category scale with additional grades at low severities (0·1, 0·5, 1, 2, 5, 10, 20, 30…100%).

To simulate observations using the H-B scale and the two different linear category scales, NPEs were converted to the appropriate grade for assessment methods (ii)–(iv). These scale data were subsequently converted to the appropriate midpoint value of each grade for analysis (as is a standard practice for these H-B data; Madden *et al.*, 2007). Subsequently, a two-tailed *t*-test was used to determine whether an observed difference between the means of the two severity distributions could be attributed to chance. A parametric test was considered appropriate for these data as they were midpoint values on a ratio scale rather than categories with uneven intervals (Madden *et al.*, 2007). To mimic actual hypothesis testing, an overestimation (or underestimation) in treatment A was compared to an overestimation (or underestimation) in treatment B at different severities from 0 to 100%. Here, the criteria used were type I and type II errors when comparing the treatment means based on the simulated rater estimates relative to the means based on the ‘actual’ data.

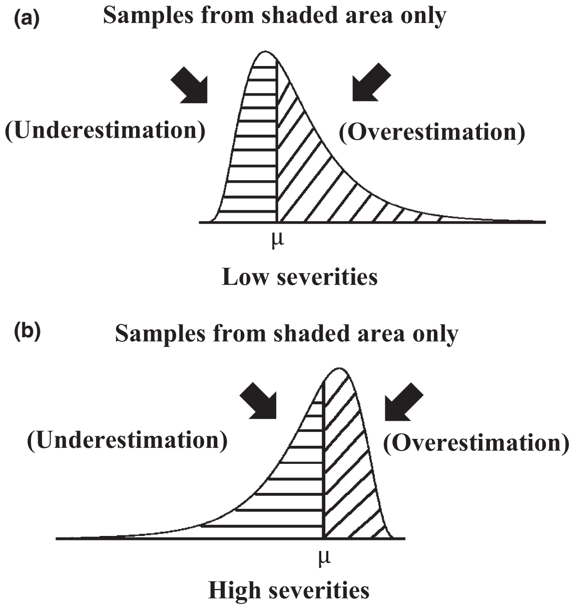

Figure 2 The lognormal distribution is a positively skewed distribution. There is a lognormal distribution for disease at low severities (a) and a 1 minus lognormal distribution (a negatively skewed distribution) at high disease severities (b). Where rater bias resulted in overestimation, random samples with values greater than the mean of both distributions were drawn (right shaded area). Where rater bias resulted in underestimation, random samples with values less than the mean of both distributions were drawn (left shaded area). The right shaded area in (b) was not used as overestimation of high severity is of less interest. Mean severity = μ.

Characteristics of the simulated data

For investigating the effects of bias (over- and underestimations), the following approaches were used:

Seven actual severities were chosen: low (1, 5, 20%); mid-range (50%); and high (80, 90 and 95%). Based on Figure 1, overestimation, underestimation and unbiased estimates were tested at low and mid-range severities. However, for high severities, only the underestimation and unbiased estimates were tested because overestimation of high severity is of less interest.

In order to quantify the relationship between the standard deviation of the mean estimated severity and the actual severity, four different situations were considered (Fig. 3). For an example of an unbiased situation, the relationship was obtained based on estimates by rater 1. For the situation where raters overestimated severity, the relationship was based on estimates by raters 3 and 4. With respect to underestimation of severity, the relationship was based on estimates by rater 2. However, raters 3 and 4 also tended to underestimate at high severities; thus, the estimates by raters 3 and 4 were combined with those of rater 2 to quantify the effects of underestimation in this range (for example, see the left shaded area of Fig. 2b).

To establish the relationship between the standard deviation of the rater mean NPE and the actual disease severity for non-biased estimates, overestimates and underestimates (Fig. 3), the rater estimates from 0 to 100% were divided into 16 consecutive groupings with an approximately equivalent number of estimates in each interval. For each of the scenarios in Figure 3, the data were subject to polynomial curve fitting, and a parabolic curve was found to be best suited to describe the relationship between the standard deviation of the rater mean NPE (σ_{rater}) and the actual disease severity:

$$\sigma_{\text{rater}} = aY_{\text{actual}}^2 + bY_{\text{actual}} + c \tag{6}$$

The parameters, the corresponding standard error, and the coefficient of determination (R^2) for each of the scenarios were used to evaluate the appropriateness of the parabolic model (Fig. 3). The analyses were performed using SAS v. 9.3 (SAS Institute).

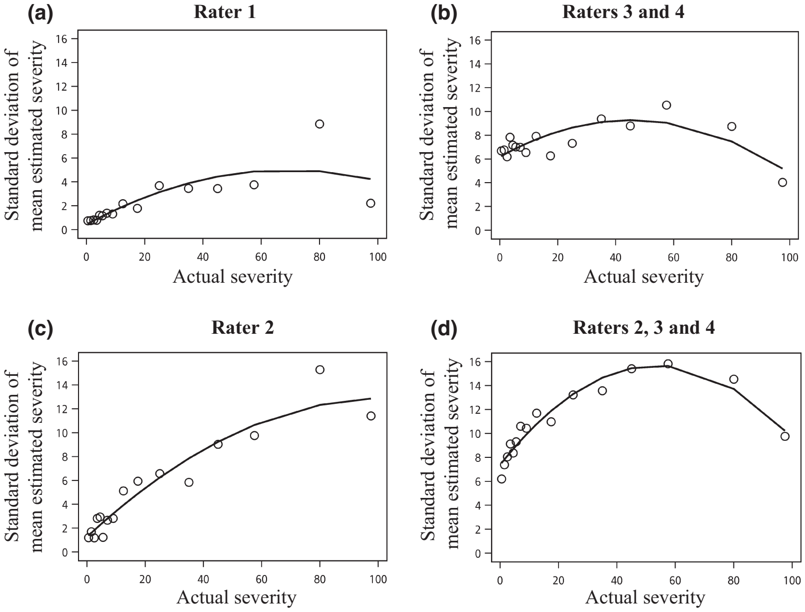

Figure 3 The parabolic relationship ($\sigma_{\text{rater}} = aY^2_{\text{actual}} + bY_{\text{actual}} + c$) between the standard deviation (σ_{rater}) of the rater nearest percentage estimates and the actual disease severity for estimates of severity of septoria leaf blotch for four different raters showing either accurate or biased estimates. (a) Accurate estimates (rater 1) parameters (standard error) are *a*, *b* and *c* = −0·0010 (0·0004), 0·1322 (0·0406) and 0·4268 (0·5477), respectively, R^2 (coefficient of determination) = 0·63. (b) Overestimates (raters 3 and 4) parameters (standard error) are *a*, *b* and *c* = −0·0015 (0·0003), 0·1366 (0·0307) and 6·1634 (0·4143), respectively, R^2 = 0·61. (c) Underestimates (rater 2) parameters (standard error) are *a*, *b* and *c* = −0·0011 (0·0004), 0·2268 (0·0392) and 1·2611 (0·5286), respectively, R^2 = 0·92. (d) The parabolic curve, using the standard deviations of estimates of raters 2, 3 and 4; the parameters (standard error) of *a*, *b* and *c* = −0·0029 (0·0003), 0·3085 (0·0242) and 7·3640 (0·3266), respectively, R^2 = 0·93.

Type II error

Different disease severities (μ_A) of 1, 5 and 20% (low severity), 50% (mid-range severity), and 80, 90 and 95% (high severity) were used to explore effects of bias on type II error rates. The threshold for rejection of the null hypothesis was set at $P = 0{\cdot}05$ (Bock *et al.*, 2010a). Simulations were computed and results were plotted. For the specified mean population disease severities, three relationships were presented.

First, the relationship was presented between the probability of rejecting H_0 when this hypothesis was false (i.e. the power of the assessment method) and sample sizes (5–50, step size = 5) for the different assessment scales. The standard deviation (φ) of the mean disease severity of the population was assumed equal to 5%. Secondly, the relationship was plotted between the power of the assessment method and the standard deviation (φ = 2–20%) of the mean disease severity of the population, with the number of samples (n) set at 20. For the two relationships above, the difference between the population means ($\mu_\triangle$) was assumed to be 5%. Thirdly, in order to take the magnitude of the population difference into account, the relationship between the differences in population means ($\mu_\triangle$ = 2–20%) and the power of the assessment method at a sample size (n) of 20 was explored. Assumed fixed values were $\varphi = 5\%$. To calculate the probability that H_0 is rejected, the simulation procedure outlined above was repeated 10 000 times and a t-test performed on each simulated data set.

Type I error

Type I error rates were investigated in a similar manner to the type II error rates described above. The relationships between the probability of rejecting H_0 (when this hypothesis is true) at different sample sizes and population standard deviations for the different assessment methods were calculated as described, but with the assumption that there was no difference between the means ($\mu_\triangle = 0$).

Results

Type II error

Effect of sample size at disease severities ≤20%

To establish the ramifications of bias (over- and underestimation) at the low disease severities, biased estimates were compared to non-biased estimates at 1, 5 and 20% severity (Fig. 4). The power of the hypothesis test increased for all assessment methods with larger sample size (n) (as expected). However, the power of the hypothesis test was greater with unbiased estimates at low disease severities compared to biased estimates, particularly overestimates. Interestingly, little difference was found in the power of the test between the unbiased estimates and underestimates.

As previously noted (Bock *et al.*, 2010a; Chiang *et al.*, 2014), if estimates are unbiased, the H-B scale has the lowest power compared with all the other assessment methods tested at $\mu_A = 20\%$ (Fig. 4). Regardless of severity (μ_A = 20, 5 or 1%) tested, there are no apparent differences in the power of the hypothesis test among the assessment methods due to overestimates. But compared with the unbiased situation, there is a reduction in the power of the hypothesis test due to overestimates with

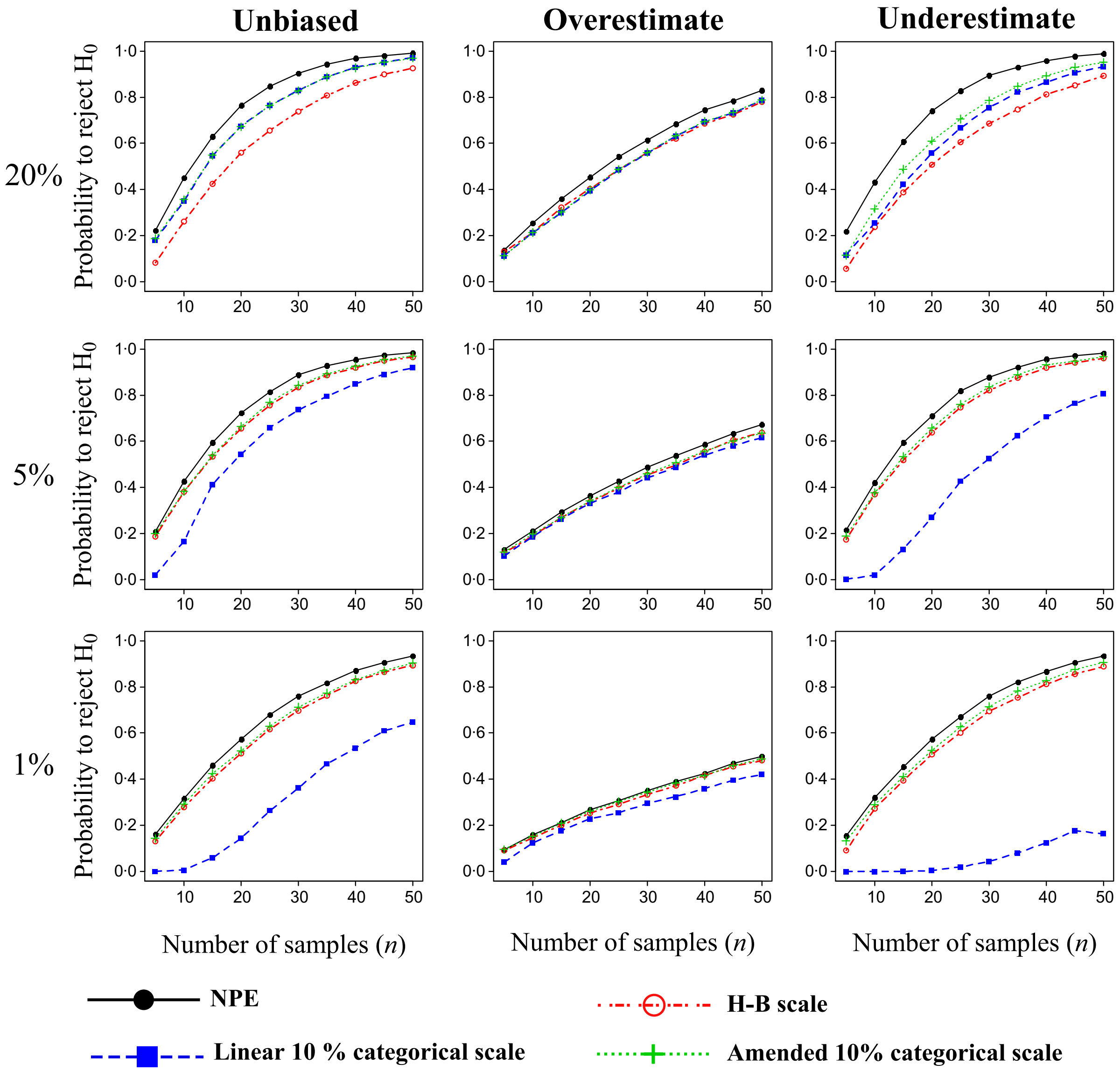

Figure 4 The relationships between the probability to reject H_0 (when this hypothesis is false) and samples size (n = 5–50) for the different assessment scales, and the effects of rater bias on that probability at mean disease severity (μ_A) of 20, 5 and 1%, respectively. The assumed difference between the population means (μ_Δ) = 5%, the standard deviation (φ) = 5%, with significance at P = 0·05. Assessment methods: (i) nearest percentage estimates (NPEs); (ii) Horsfall-Barratt (H-B) scale; (iii) linear scale (10% categories); and (iv) amended linear scale (10% categories with additional grades at severities <10%).

all the different assessment methods. When μ_A = 5% or 1%, the power was least for the linear scale (10%) over a wide range of sample sizes (n), and particularly at μ_A = 1% (Fig. 4). There was little difference among the other methods at 5 and 1% severity, although NPEs consistently had a slightly higher power compared to the other methods.

Effect of sample standard deviation at disease severities ≤20%

At disease severities ≤20%, bias affects the power of the hypothesis test depending on the standard deviations (φ) of the mean severities and the method used (Fig. 5). Regardless of assessment method, when the standard deviation of the severity distribution is large, the hypothesis test has lower power. At standard deviations ~ ≤10%, overestimates consistently reduced the power of the hypothesis test compared to unbiased estimates, while underestimates (with the exception of the H-B scale and the linear 10% scale) had little effect on the power of the hypothesis test. With unbiased estimates it should be noted that at small standard deviations, the power of the H-B scale at μ_A = 20%, and of the linear 10% scale at $\mu_A \leq$ 5%, were lower compared to the other methods.

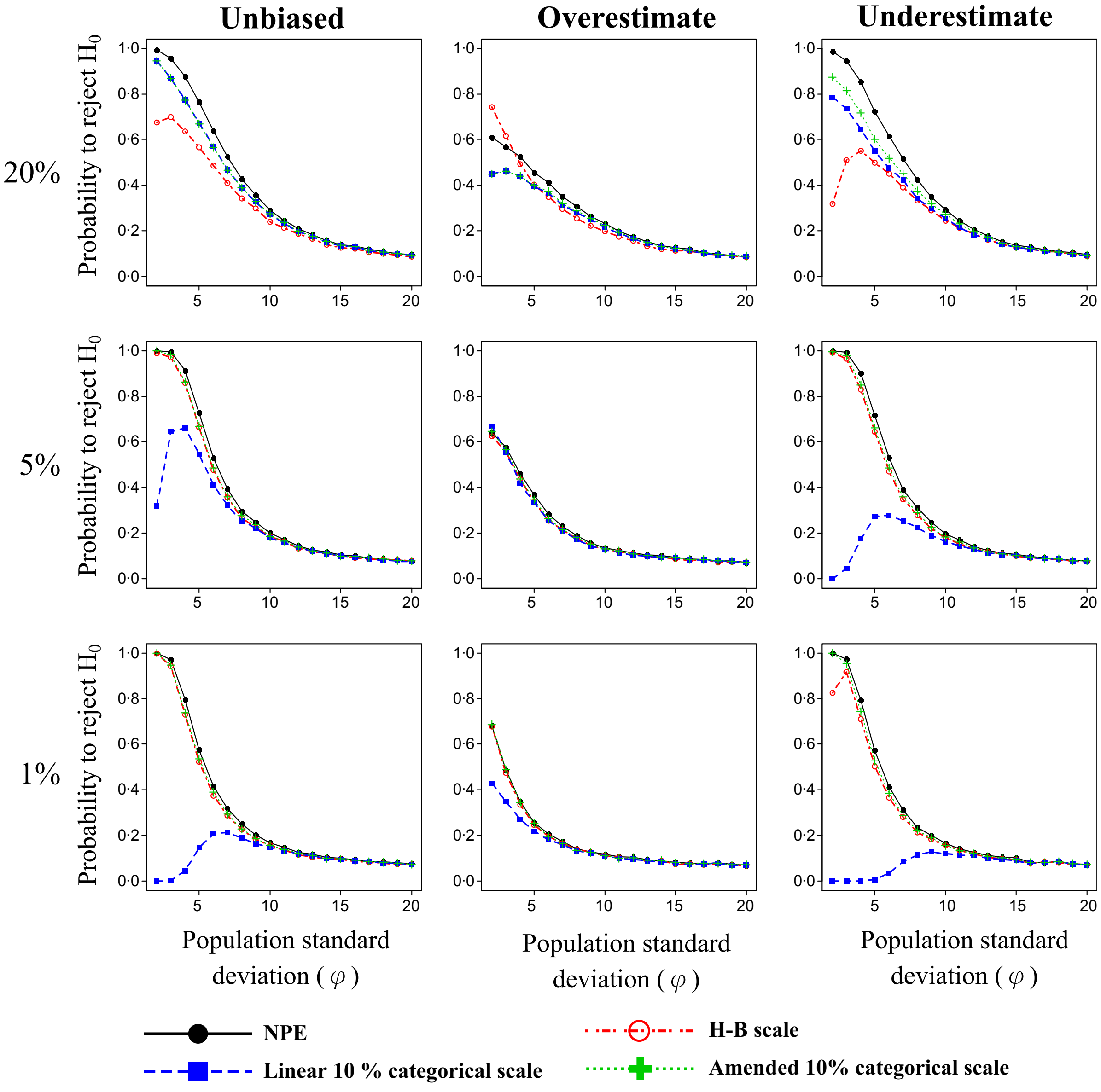

Figure 5 The relationship between the probability to reject H_0 (when this hypothesis is false) and magnitude of the population disease severity standard deviation (φ = 2 to 20%) for the different assessment scales, and the effects of rater bias on that probability at mean disease severity (μ_A) of 20, 5 and 1%, respectively. The assumed sample size (n) = 20; the assumed difference between the populations means ($\mu_\triangle$) = 5% with significance at P = 0·05. Assessment methods: (i) nearest percentage estimates (NPEs); (ii) Horsfall-Barratt (H-B) scale; (iii) linear scale (10% grades); and (iv) amended linear scale (10% categories with additional intervals at severities <10%).

Effect of the difference between sample means at disease severities ≤20%

Increasing the difference ($\mu_\triangle$) between the population means increased the power of the hypothesis test for all assessment methods (Fig. 6), regardless of whether they were biased. When $\mu_\triangle$ is ≥10% (and n = 20 and φ = 5), the power is near 1 for all methods at low severities. However, overestimates tended to have a slightly lower power when $\mu_\triangle \leq 10\%$. Furthermore, for unbiased estimates and underestimates, when $\mu_A \leq 5\%$, the 10% scale has a lower power for hypothesis testing (which is similar to the effect observed in Fig. 4).

Effects of sample size, sample standard deviation and the difference between sample means at disease severities ≥80%

Raters 2, 3 and 4 tended to underestimate severity at 80–100% (high disease severity) actual severity in this data set. Thus the effect of rater bias in relation only to underestimation and assessment method was explored at μ_A = 80, 90 and 95%.

The results of increasing sample size (Fig. 7) at μ_A = 80, 90 and 95% are comparable to those at severities of 20, 5 and 1%, respectively. This is not unexpected as the comparisons of the unbiased estimates versus

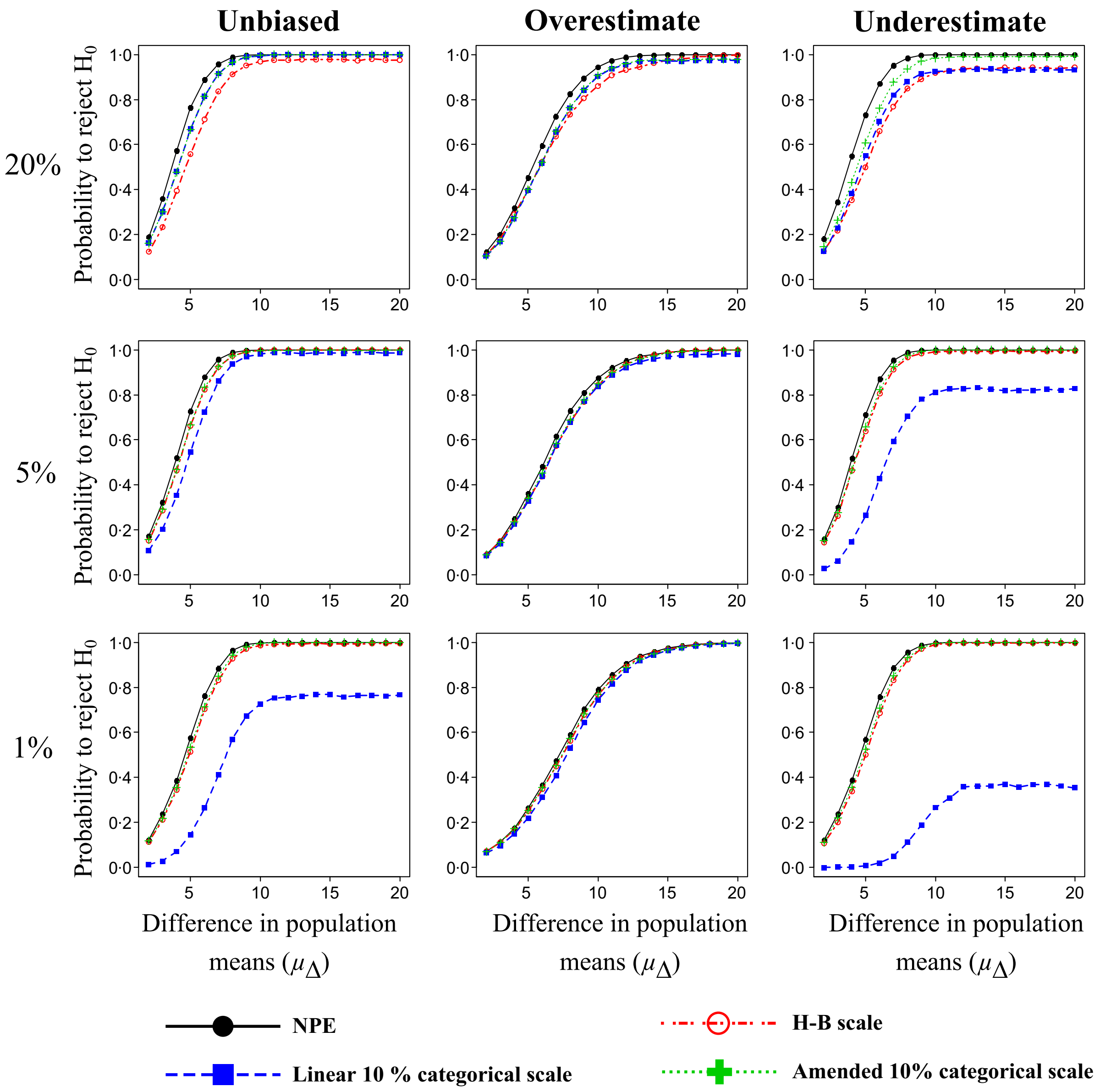

Figure 6 The relationship between the probability to reject H_0 (when this hypothesis is false) and the difference between the population means (μ_Δ = 2–20%) for the different assessment scales, and the effects of rater bias on that probability at mean disease severity (μ_A) of 20, 5 and 1%, respectively. The assumed sample size (*n*) = 20; the population standard deviations (φ) = 5% with significance at *P* = 0·05. Assessment methods: (i) nearest percentage estimates (NPEs); (ii) Horsfall-Barratt (H-B) scale; (iii) linear scale (10% categories); and (iv) amended linear scale (10% categories with additional grades at severities <10%).

underestimates at high severities are basically the inverse of those of the unbiased estimates versus overestimates at low severities. Briefly, the power of the unbiased estimates was greater compared with that of the underestimates for each assessment method when μ_A = 80, 90 or 95% (Fig. 7). Compared with NPEs, all other methods reduced the power of the hypothesis test at μ_A = 80, 90 or 95% when based on unbiased estimates. There was little difference among assessment methods when based on underestimates (Fig. 7).

At small population standard deviations (φ), the unbiased estimates of each method had more power at μ_A = 80, 90 and 95% compared with the underestimates (Fig. S1). With unbiased estimates, the linear 10% scale and the amended 10% scale reduced the power of the hypothesis test at μ_A = 80, 90 or 95%, and the H-B scale reduced the power of the hypothesis test at μ_A = 80% compared with the power using NPEs. With underestimates, only at μ_A = 95% was the power of the hypothesis test reduced by the linear 10% scale and the amended 10% scale when compared with the other methods.

With increasing magnitude of the difference between population means (μ_Δ), the power of the test increased (Fig. S2). The unbiased estimates tended to have higher

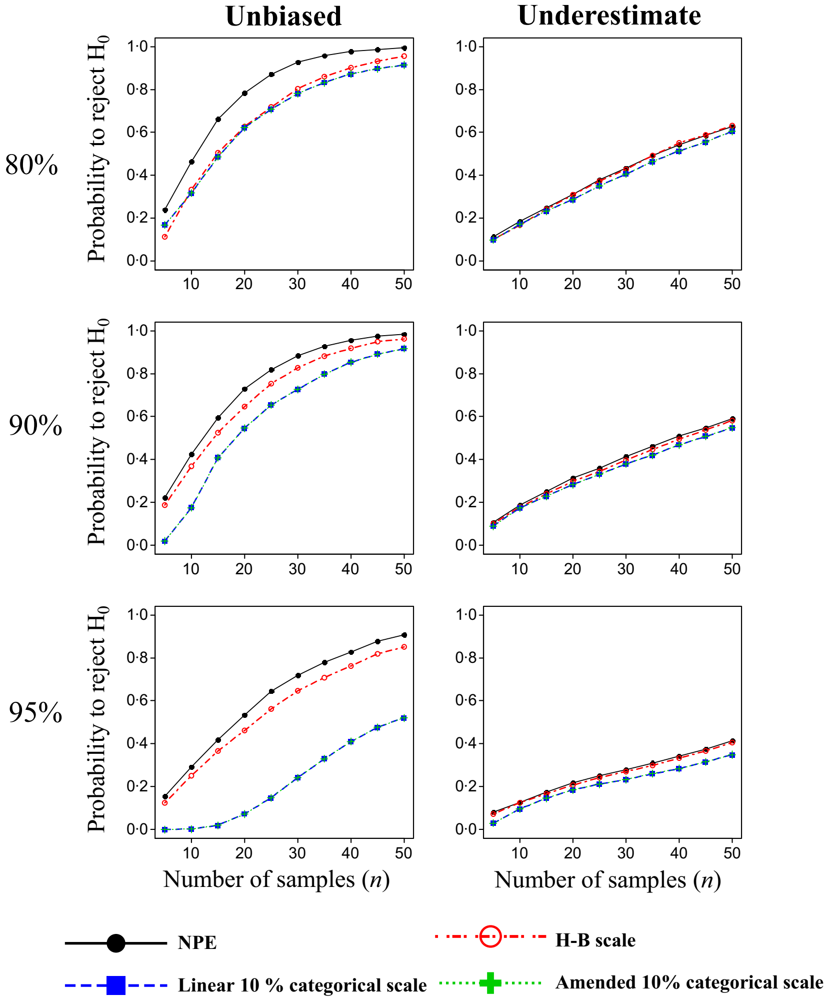

Figure 7 The relationships between the probability to reject H_0 (when this hypothesis is false) and sample size (n = 5–50) for the different assessment scales and the effects of rater bias on that probability at mean disease severity (μ_A) of 80, 90 and 95%, respectively. The assumed difference between the population means ($\mu_\triangle$) = 5%; the standard deviation (φ) = 5%, with significance at P = 0·05. Assessment methods: (i) nearest percentage estimates (NPEs); (ii) Horsfall-Barratt (H-B) scale; (iii) linear scale (10% categories); and (iv) amended linear scale (10% categories with additional grades at severities <10%).

power at μ_A = 80, 90 and 95% (up to $\mu_\triangle$ = 12, above which they were equal).

Effects of sample size, sample standard deviation and the difference between sample means at disease severities of 50%

At mean disease severity (μ_A) of 50% (mid-range disease severity), the H-B scale had consistently lower power for hypothesis testing compared with the other methods, at all sample sizes (n) tested, with differences between population means ($\mu_\triangle$) ≤12, and with population standard deviations (φ) ≤10 (Fig. S3). The unbiased estimates for all assessment methods tested tended to have slightly higher power to reject the null hypothesis compared with biased estimates. Both the 10% category scales reduced the power of the hypothesis test only marginally (at some sample sizes, differences in population means and population standard deviations tested), when compared with NPEs.

The results in Figure 4 demonstrate that a greater sample size of biased estimates can be taken to achieve the same probability as that based on unbiased estimates for rejecting H_0, when H_0 is false. Thus, the additional number of samples needed to obtain the same probability was calculated using biased estimates as compared with the probability based on unbiased estimates for the effects of both overestimation and underestimation when using NPEs. For example, if two treatments are compared with a difference between the population means ($\mu_\triangle$) of 5%, standard deviations (φ) of 5%, with the significance set at P = 0·05, and a sample size of 50, the corresponding sample sizes required for the same power based on overestimates are 107 and 71 at disease severities of 20 and 50%, respectively (based on the data in Figs 4 & S3, respectively). Moreover, for severities of 50 and 80%, the corresponding sample sizes when using underestimates are 59 and 185, respectively (Figs S3 & 7, respectively).

Type I error

Similar to testing type II error rates, the relationships between the probability of rejecting H_0 (when this hypothesis is true) at different sample sizes (Fig. 8), and

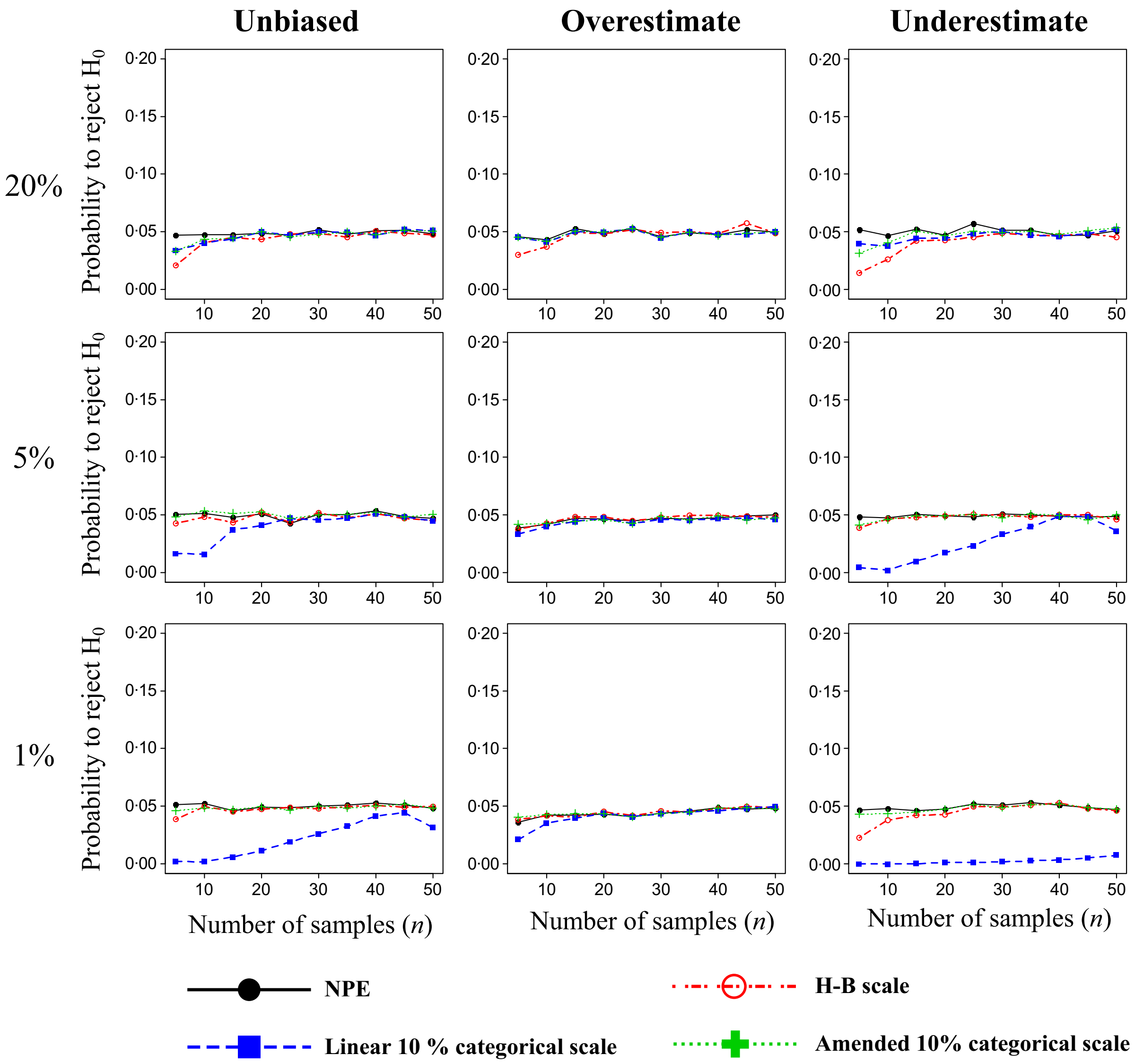

Figure 8 The relationships between the probability to reject H_0 (when this hypothesis is true – a type I error) and sample size (n = 5–50) for the different assessment scales and the effects of rater bias on that probability at mean disease severities (μ_A) of 20, 5 and 1%, respectively. The assumed difference between the population means (μ_Δ) = 0, the sample standard deviation (φ) = 5%, with significance at P = 0·05. Assessment methods (i) nearest percentage estimates (NPEs); (ii) Horsfall-Barratt (H-B) scale; (iii) linear scale (10% categories); and (iv) amended linear scale (10% categories with additional grades at severities <10%).

population standard deviations (data not shown) for the different assessment methods were calculated. There was almost no effect of rater method or bias on type I error rate at low severities (Fig. 8), although at μ_A = 20% and small sample sizes the H-B scale had a lower type I error (especially for underestimates), and at μ_A = 5 or 1% the linear 10% scale had a lower type I error compared with the other methods (especially with underestimates and at low sample sizes). With higher severities (μ_A = 80, 90 and 95%), neither assessment method nor bias had much effect on the probability to reject H_0 when this hypothesis is true (data not shown). At disease severity (μ_A) of 50% there was no apparent effect of assessment method or bias on type I error rates (data not shown). In short, rater bias and assessment method do not inflate type I errors.

Discussion

The main objective of this study was to explore the effects of rater bias and assessment method on disease severity estimation over the full range of disease severities (from 0 to 100%). The power of the hypothesis test was used to ascertain the effects. The comparison of bias (overestimates and underestimates) with unbiased estimates of disease severity at different mean severities, sample sizes, standard deviations, and difference between population means showed that the power of the hypothesis test is greatest when estimates are unbiased. This was true at all the severities tested and for all assessment methods used. These results reinforce how essential it is to provide some way of reducing the bias in rater estimates in order to improve accuracy and reliability of visual estimates of disease severity, and thus avoid type II errors in analysis of disease severity data, which have been noted before (Christ, 1991; Todd & Kommedahl, 1994; Parker *et al.*, 1995).

There is a widespread tendency to overestimate disease severity at low actual severities (<10%) (Sherwood *et al.*, 1983; Bock *et al.*, 2010b). This tendency was observed for two of the four raters in the data set used in this study (raters 3 and 4). Although there might sometimes be underestimation of low disease severities (as demonstrated with rater 2 at low severities of SLB in this data set), it has not been a widely reported bias among raters. As severities of <10% are often observed in the field on annual crops, it is worthwhile considering ways of reducing error in this range. Standard area diagrams (SADs) (Rios *et al.*, 2013; Yadav *et al.*, 2013; Schwanck & Del Ponte, 2014) and computer-aided training (Nutter & Schultz, 1995) can assist in reducing the absolute error overall, but have not been investigated in detail for their value at guiding accurate estimates specifically at these low disease severities. Based on the results of the present study, it seems valuable to investigate whether methods (perhaps better designed SADs) and training software can be specifically developed to reduce rater error in this range.

Thus SAD design combined with computer-aided training tools are likely to be an important remedial measure for reducing error. Based on results of a previous study, Chiang *et al.* (2014) concluded that a 10% category scale with additional divisions at low severity provided good estimates of disease severity that minimized type II errors, and was comparable with NPEs, at least for raters of average accuracy (very accurate raters would likely still benefit from using NPEs (Bock *et al.*, 2010a)). Based on the results of the Chiang *et al.* (2014) study, a recently published SADs developed by Rios *et al.* (2013) provides the characteristics that should help minimize rater bias at low severity, and thus improve accuracy and reliability of the estimates. In order to improve accuracy and reliability of estimates using SADs, further discussion and research measuring the impact of SADs and computer-aided tools on rater accuracy is needed.

Assessment methods used to estimate disease severity affected the power of the hypothesis test. If the severity of disease was 50%, the H-B scale had the lowest probability to reject H_0 when this hypothesis was false compared with all the other assessment methods tested for both unbiased and biased estimates. Also, the H-B scale tended to have a slightly greater risk of type II error at severities of 20 and 80% regardless of whether raters were biased or not, which might be expected. The probable reason for this is that the H-B scale intervals of 12–25% and 25–50% are so wide (13 and 25%, respectively) that when mid-points are taken for analysis they inevitably result in less accurate estimates compared with what is achieved based on NPEs. At severities close to 20%, if a rater overestimates (or underestimates) severity, the estimate will have a high probability of falling into the 25–50% interval (or the 12–25% interval, if underestimated). Subsequently, having been placed in this range, the value is transformed to a mid-point 37·5% (or 18·5%, if underestimated) for analysis. Thus depending on the specific disease severity means, mean differences, variances and bias in the range of disease severity from 12 to 88%, using the H-B scale is more likely to lead to a type II error situation where it is more difficult to establish the difference between the means of two severity distributions. It should be noted that increasing sample size helps mitigate the effect of the H-B scale in elevating the type II error rate (Bock *et al.*, 2010a; Chiang *et al.*, 2014), but a disadvantage is that more samples take more time to observe in the field or collect for later assessment, under sometimes demanding field conditions.

Assessment method (iii) (the 10% category scale) tended to have a greater risk of type II error at <10% and >85% severities. Although the power of method (iv) (the amended 10% category scale with additional grades at low severities) is the same as that of assessment method (iii) at high severities (both scales have the same structure at >10%), this should not be detrimental to most cases requiring hypothesis testing. Except in the case of very high severities, the power of the amended 10% category scale is almost equivalent to that of NPEs. One reason for choosing a category scale may be for convenience and speed of rating (Madden *et al.*, 2007), and the amended 10% category scale is reasoned to be superior to other category-scale methods for researchers who want to base their severity estimation on a disease category scale for hypothesis testing. In particular, the additional low grades are sensitive to the range of severity at which many diseases frequently occur. Nonetheless, NPEs provided the consistently greatest power to reject H_0 when H_0 was false, and thus the 0–100% scale is recommended whenever possible (and particularly so in the case of very accurate raters (Bock *et al.*, 2010a)). The conclusions here based on severity estimates of SLB are consistent with those of a previous study based on estimates of the severity of citrus canker by different raters (Chiang *et al.*, 2014). That is, a disease assessment category scale should be sensitive to low disease severity (1–10%) by incorporating additional categories to account for disease severity ≤5%. Category intervals in the mid-range should not exceed 10%. This consideration should be taken into account when designing SADs and computer-aided tools to train raters in these sensitive disease severity ranges.

Nita *et al.* (2003) had tested a 5% category scale and found it to be more accurate and reliable compared to the H-B scale, and in a previous study, Chiang *et al.* (2014) included a 5% linear scale for comparison with other assessment scales. The 5% scale was not included in the current study as the authors believe it has too many divisions negating the assumed advantages of simplicity offered by category scales with divisions ≥10%. However, the 5% linear scale does have advantages over other methods (with the exception of NPEs) for reducing the risk of type II errors (Chiang *et al.*, 2014). In the present study, increasing the difference ($\mu_{\triangle}$) between the population means increased the probability of rejecting H_0 for all assessment methods. As $\mu_{\triangle}$ increased to more than 10%, the power of the test was close to 1. This observation agrees with the work of Bock *et al.* (2015), which found there were no type II errors observed between fungicides treated and control plots when using estimates of SLB severity based on either NPEs or the H-B scale. There were consistent significant differences in mean estimates of disease severity for control and fungicide-treated plot for each rater and for the actual values based on image analysis (Bock *et al.*, 2015), with differences between treatments (>50% disease severity) providing a robust basis for correctly rejecting H_0.

In conclusion, this research indicated that rater bias has a greater effect on type II errors compared with the effect of assessment methods. It is noteworthy that neither rater bias nor assessment method had any notable effect on type I error rates. The H-B scale (and other lower resolution, non-linear category scales) and the linear 10% category scale had the lowest power for hypothesis testing compared to the other methods tested. The results of the study should contribute to developing improved disease assessment category scales and understanding the effects of rater bias in disease estimation. This information helps focus research on the improvement and optimization of category scales (in situations where they must be the assessment method of choice) and for developing SADs as aids for estimating disease severity both accurately and reliably, taking particular account of low disease severities.

Acknowledgements

The authors appreciate many useful discussions with several colleagues on the topic of plant disease severity assessment based on visual estimates. K. S. Chiang gratefully acknowledges financial support from the Ministry of Science and Technology of Taiwan, R.O.C. (NSC103-2313-B-005-016). C. H. Bock acknowledges the research was partially supported by the USDA national program for Crop Production (project number 6042-21220-012-00D).

References

Amanat P, 1976. Stimuli effecting disease assessment. *Agriculturae Conspectus Scientificus* **39**, 27–31.

Bardsley SJ, Ngugi HK, 2013. Reliability and accuracy of visual methods to quantify severity of foliar bacterial spot symptoms on peach and nectarine. *Plant Pathology* **62**, 460–74.

Beresford RM, Royle DJ, 1991. The assessment of infectious disease for brown rust (*Puccinia hordei*) of barley. *Plant Pathology* **40**, 374–81.

Bock CH, Parker PE, Cook AZ, Gottwald TR, 2008a. Visual rating and the use of image analysis for assessing different symptoms of citrus canker on grapefruit leaves. *Plant Disease* **92**, 530–41.

Bock CH, Parker PE, Cook AZ, Gottwald TR, 2008b. Characteristics of the perception of different severity measures of citrus canker and the relationships between the various symptom types. *Plant Disease* **92**, 927–39.

Bock CH, Gottwald TR, Parker PE *et al.*, 2009a. The Horsfall-Barratt scale and severity estimates of citrus canker. *European Journal of Plant Pathology* **125**, 23–38.

Bock CH, Parker PE, Cook AZ, Gottwald TR, 2009b. Automated image analysis of the severity of foliar citrus canker symptoms. *Plant Disease* **93**, 660–5.

Bock CH, Gottwald TR, Parker PE *et al.*, 2010a. Some consequences of using the Horsfall-Barratt scale for hypothesis testing. *Phytopathology* **100**, 1031–41.

Bock CH, Poole GH, Parker PE, Gottwald TR, 2010b. Plant disease severity estimated visually, by digital photography and image analysis, and by hyperspectral imaging. *Critical Reviews in Plant Sciences* **29**, 59–107.

Bock CH, Wood BW, Gottwald TR, 2013a. Pecan scab severity – effects of assessment methods. *Plant Disease* **97**, 675–84.

Bock CH, Wood BW, van den Bosch F, Parnell S, Gottwald TR, 2013b. The effect of Horsfall-Barratt category size on the accuracy and reliability of estimates of pecan scab severity. *Plant Disease* **97**, 797–806.

Bock CH, El Jarroudi M, Kouadio AL, Mackels C, Chiang KS, Delfosse P, 2015. Disease severity estimates – effects of rater accuracy and assessment methods for comparing treatments. *Plant Disease*. doi: 10.1094/PDIS-09-14-0925-RE.

Chiang KS, Liu SH, Bock CH, Gottwald TR, 2014. What interval characteristics make a good disease assessment category scale? *Phytopathology* **104**, 575–85.

Christ BJ, 1991. Effect of disease assessment method on ranking potato cultivars for resistance to early blight. *Plant Disease* **75**, 353–6.

Corrêa FM, Bueno Filho JSS, Carmo MGF, 2009. Comparison of three diagrammatic keys for the quantification of late blight in tomato leaves. *Plant Pathology* **58**, 1128–33.

Danielsen S, Munk L, 2004. Evaluation of disease assessment methods in quinoa for their ability to predict yield loss caused by downy mildew. *Crop Protection* **23**, 219–28.

Duarte HSS, Zambolin L, Capucho AS *et al.*, 2013. Development and validation of a set of standard area diagrams to estimate severity of potato early blight. *European Journal of Plant Pathology* **137**, 249–57.

El Jarroudi M, Delfosse P, Maraite H, Hoffmann L, Tychon B, 2009. Assessing the accuracy of simulation model for Septoria leaf blotch disease progress on winter wheat. *Plant Disease* **93**, 983–92.

El Jarroudi M, Kouadio L, Bertrand M *et al.*, 2012a. Integrating the impact of wheat fungal diseases in the Belgian crop yield forecasting system (B-CYFS). *European Journal of Agronomy* **40**, 8–17.

El Jarroudi M, Kouadio L, Delfosse P *et al.*, 2012b. Typology of the main fungal diseases affecting winter wheat in the Grand Duchy of Luxembourg. *Journal of Agricultural Science and Technology* **A2**, 1386–99.

El Jarroudi M, Kouadio AL, Mackels C, Tychon B, Delfosse P, Bock CH, 2015. A comparison between visual estimates and image analysis measurements to determine septoria leaf blotch severity in winter wheat. *Plant Pathology* **64**, 355–64.

Forbes GA, Jeger MJ, 1987. Factors affecting the estimation of disease intensity in simulated plant structures. *Zeitschrift fur Pflanzenkrankheiten und Pflanzenschutz* **94**, 113–20.

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 Pathology* **43**, 675–82.

Hau B, Kranz J, Konig R, 1989. Fehler beim Schätzen von Befallsstärken bei Pflanzenkrankheiten. *Zeitschrift fur Pflanzenkrankheiten und Pflanzenschutz* **96**, 649–74.

Horsfall JG, Barratt RW, 1945. An improved grading system for measuring plant disease. *Phytopathology* **35**, 655.

Kranz J, 1977. A study on maximum severity in plant disease. *Travaux dédiés à G. Viennot-Bourgin* **16**, 9–73.

Lamari L, 2002. *Assess: Image Analysis Software for Plant Disease Quantification*. St Paul, MN, USA: APS Press.

Madden LV, Hughes G, van den Bosch F, 2007. *The Study of Plant Disease Epidemics*. St Paul, MN, USA: APS Press.

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 Jr, Schultz PM, 1995. Improving the accuracy and precision of disease assessments: selection of methods and use of computer-aided training programs. *Canadian Journal of Plant Pathology* **17**, 174–85.

Parker SR, Shaw MW, Royle DJ, 1995. The reliability of visual estimates of disease severity on cereal leaves. *Plant Pathology* **44**, 856–64.

Peterson RF, Campbell AB, Hannah AE, 1948. A diagrammatic scale for estimating rust intensity on leaves and stems of cereals. *Canadian Journal of Research* **26**, 496–500.

Rios JA, Debona D, Duarte HSS, Rodrigues FA, 2013. Development and validation of a standard area diagram set to assess blast severity on wheat leaves. *European Journal of Plant Pathology* **136**, 603–11.

Schwanck AA, Del Ponte EM, 2014. Accuracy and reliability of severity estimates using linear or logarithmic disease diagram sets in true colour or black and white: a study case for rice brown spot. *Journal of Phytopathology* **162**, 670–82.

Sherwood RT, Berg CC, Hoover MR, Zeiders KE, 1983. Illusions in visual assessment of *Stagonospora* leaf spot of orchardgrass. *Phytopathology* **73**, 173–7.

Todd LA, Kommedahl T, 1994. Image analysis and visual estimates for evaluating disease reactions of corn to *Fusarium* stalk rot. *Plant Disease* **78**, 876–8.

Vereijssen J, Schneider JHM, Termorshuizen AJ, Jeger MJ, 2003. Comparison of two disease assessment methods for assessing *Cercospora* leaf spot in sugar beet. *Crop Protection* **22**, 201–9.

Yadav NVS, de Vos SM, Bock CH, Wood BW, 2013. Development and validation of standard area diagrams to aid assessment of pecan scab symptoms on pecan fruit. *Plant Pathology* **62**, 325–35.

Zadoks JC, Chang TT, Konzak CF, 1974. A decimal code for the growth stages of cereals. *Weed Research* **14**, 415–21.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site.

Figure S1 The relationship between the probability to reject H_0 (when this hypothesis is false) and magnitude of the population disease severity standard deviation (φ = 2–20%) for the different assessment scales, and the effects of rater bias on that probability at mean disease severity (μ_A) of 80, 90 and 95%, respectively. Assumed sample size (n) = 20; the difference between the population means ($\mu_\triangle$) = 5% with significance at $P = 0{\cdot}05$. Assessment methods: (i) nearest percentage estimates (NPEs); (ii) Horsfall-Barratt (H-B) scale; (iii) linear scale (10% categories); and (iv) amended linear scale (10% categories with additional grades at severities <10%).

Figure S2 The relationship between the probability to reject H_0 (when this hypothesis is false) and the difference between the population means ($\mu_\triangle$ = 2–20%) for the different assessment scales, and the effects of rater bias on that probability at mean disease severity (μ_A) of 80, 90 and 95%, respectively. Assumed sample size (n) = 20; sample standard deviation (φ) = 5% with significance at $P = 0{\cdot}05$. Assessment methods: (i) nearest percentage estimates (NPEs); (ii) Horsfall-Barratt (H-B) scale; (iii) linear scale (10% categories); and (iv) amended linear scale (10% categories with additional grades at severities <10%).

Figure S3 The relationship between the probability to reject H_0 (when this hypothesis is false) for samples size (n = 5–50), the difference between the population means ($\mu_\triangle$ = 2–20%) and magnitude of the population disease severity standard deviation (φ = 2–20%), for each of the different assessment scales, and the effects of rater bias on that probability at a disease severity of 50%. Details of parameters held constant are already described in Figures 4–6. Assessment methods: (i) nearest percentage estimates (NPEs); (ii) Horsfall-Barratt (H-B) scale; (iii) linear scale (10% categories); and (iv) amended linear scale (10% categories with additional grades at severities <10%).