Preprint

A New Standard Area Diagram Set for Assessment of Severity of Soybean Rust Improves Accuracy of Estimates and Optimizes Resource Use

Vincius T. Franceschi[1], Kaique S. Alves[2], Sergio M. Mazaro[1], Claudia V. Godoy[3], Henrique S. S. Duarte[4] and Emerson M. Del Ponte[2,*]

Abstract

Soybean rust (SBR), caused by *Phakopsora pachyrhizi*, is the most important yield-damaging fungal disease of soybean due to severe reduction in healthy leaf area and acceleration of leaf fall. In experimental research, SBR severity is estimated visually aided/trained by a standard area diagram (SAD) developed and validated during the mid-2000s (Old SAD). In this study, we propose a new SAD set for SBR with six true-colour diagrams following linear increments (c.15% increments) amended with four additional diagrams at low (<10%) severities, totaling 10 diagrams (0.2%, 1%, 3%, 5%, 10%, 25%, 40%, 55%, 70%, and 84%). For evaluation, 37 raters were split into two groups. Each assessed severity in a 50-image sample (0.25% to 84%), first unaided and then using either the Old SAD or the New SAD. Accuracy, precision, and reliability of estimates improved significantly relative to unaided estimates only when aided by the New SAD (accuracy >0.95). Low precision (<0.78) and a trend of underestimation with an increase in severity were the main issues with the Old SAD, which did not differ from unaided estimates. Simulation to evaluate the impact of the errors by different methods on hypothesis tests, showed that the new SAD was more powerful for detecting the smallest difference in mean control (e.g., 70% vs. 65% disease reduction) than the Old SAD; the latter required a 2-fold increase in sample size to achieve the same power. There is a need to improve some SADs, taking advantage of new knowledge and technology to increase accuracy of the estimates, and to optimize both resource use efficiency and management decisions.

Keywords: *Glycine max*, *Phakopsora pachyrhizi*, phytopathometry

---

[1] *Universidade Tecnológica Federal do Paraná, 85660-000, Dois Vizinhos, PR, Brazil*

[2] *Departamento de Fitopatologia, Universidade Federal de Viçosa, 36570-000, Viçosa, MG, Brazil*

[3] *Empresa Brasileira de Pesquisa Agropecuária, 86001-970, Londrina, PR, Brazil*

[4] *Departamento de Fitotecnia e Fitossanidade, Universidade Federal do Paraná, 80035-050, Curitiba, PR, Brazil*

*Email: delponte@ufv.br

Vincius T. Franceschi and Kaique S. Alves contributed equally to this study.

Introduction

Soybean rust, caused by the fungus *Phakopsora pachyrhizi*, is one of the most damaging diseases of soybean (*Glycine max*; Hartmann *et al.*, 2015). In Brazil, since its discovery in 2002, the disease has caused tremendous losses due to yield reduction and increased control costs (Godoy *et al.*, 2016), as well as indirect losses due to the development of resistance to fungicides in the fungal population (Dalla Lana *et al.*, 2018). Regional to local management practices such as mandatory soybean-free periods and early sowing of short-cycle cultivars help to reduce inoculum levels, but an effective control is only achieved using sequential fungicide sprays during the season (Scherm *et al.*, 2009; Dalla Lana *et al.*, 2018).

The disease reduces healthy leaf area and, at high intensity, accelerates leaf senescence and defoliation. The proportion of diseased leaf area (severity) is the main variable used to compare treatments (e.g., fungicides and cultivars) for suppressing the disease, protecting yield (Scherm *et al.*, 2009; Godoy *et al.*, 2016; Dalla Lana *et al.*, 2018), and predict yield losses (Dalla Lana *et al.*, 2015).

Symptoms of the disease appear as numerous tiny necrotic areas where individual clusters of pustules are formed, surrounded by a chlorotic halo of variable size. This pattern of symptoms is assumed to affect the precision of the estimates, especially for low severities (Bock *et al.*, 2016). For reducing errors and standardizing visual severity estimation of SBR, a standard area diagram (SAD) was developed and validated during the mid-2000s (hereafter Old SAD; Godoy *et al.*, 2006). The Old SAD is composed (as published) of six digitally enhanced two-grey colour images of soybean leaflets with severity values

incrementing logarithmically, which conformed to the current concept of SAD technology at the time (Del Ponte *et al.*, 2017). The Old SAD has been used extensively to standardize ratings across raters of the national uniform fungicide trial (UFT) network. Visual estimates of severity are provided in a 10-leaflet sample taken from each of three canopy heights per plot, which are averaged at the plot level (Dalla Lana *et al.*, 2018).

A recent review of SAD research conducted during the last 25 years highlighted methodological trends in both the development and the validation of SAD technology (Del Ponte *et al.*, 2017). These included development of sets with an increased number of diagrams in (approximately) linear increments (as opposed to log increments), digital drawings or true-colour photos, and novel statistical methods for assessing accuracy, precision, and reliability of the estimates (Del Ponte *et al.*, 2017). Moreover, the use of photographic images has increased more recently compared with black and white drawings, due to the enhanced realism of the symptoms (Schwanck and Del Ponte, 2014; Del Ponte *et al.*, 2019). Nonetheless, it is yet to be clarified for which kind of symptoms such enhancement is advantageous and whether it leads to improvements in accuracy of practical significance. Research in this topic is scarce and results are inconsistent (Buffara *et al.*, 2014; Schwanck and Del Ponte, 2014).

Historically, the use of logarithmic increments of severity was adapted from an early concept of an ordinal disease scale (H-B scale) developed based on the claim that raters were not capable of differentiating severity at the midrange (25% to 50% or 50% to 75%). Thus, linear scales would be of limited value and slow down assessments (Horsfall and Barratt, 1945). For diseases where the

maximum severity is lower than 50%, the number of diagrams has probably been affected by the H-B scale paradigm; hence, severity has been underrepresented from the midrange to upper limits (Del Ponte *et al.*, 2017). In fact, raters seem to be capable of distinguishing among severity values at the midrange at magnitudes that vary according to the kind of symptoms (Nutter and Esker, 2006). The use of linear increments has increased in recent years, but the H-B scale paradigm, claiming the Weber-Fechner law of visual acuity, persists in a few recently developed SAD studies (Camara *et al.*, 2018; Fantin *et al.*, 2018; Costa *et al.*, 2019). Depending on the patterns of symptoms, overestimation is more common at the lower end (<10% severity). To minimize those errors, it has been suggested that linear scales or SADs should be amended with a few severity values at the lower end (Bock *et al.*, 2010b; Schwanck and Del Ponte, 2014). This approximately linear scale has been referred to as “amended linear” (Chiang *et al.*, 2014; 2016b; 2019). Mitigation of overestimation in that range is critical due to its negative impact on hypothesis tests verified using simulation (Chiang *et al.*, 2016a,b). However, it is also important that underestimation is minimized at the upper limits (Chiang *et al.*, 2016a), particularly in experimental research to evaluate disease control efficacy (percentage reduction relative to the untreated), which requires estimation of severity values that rarely reach 100% (thus no error). Simulation has been used to evaluate the effect of sources and magnitude of errors, when using different assessment methods and experimental designs, on the hypothesis tests (Bock *et al.*, 2010a; Chiang *et al.*, 2016ab). The combined effect of errors of SAD-aided estimates of severity for two treatments of interest, relative to a nontreated check, remains to be explored.

In this study, we hypothesized that errors of severity estimates provided unaided or aided by an Old SAD (Godoy *et al.*, 2006) are reduced if an amended linear SAD represented by true-colour photographs is used as an aid. In addition, the impact of the errors of the estimates when using the different methods on hypothesis tests of treatment comparison was assessed using a simulation approach (Bock *et al.*, 2010a; Chiang *et al.*, 2016a). The power of the tests was compared for scenarios of incremental differences in disease control efficacy between two treatments and increasing number of samples within a replicated plot. Increased understanding of the effects of assessment methods is critical to ensure that the best options are chosen to optimize resource use.

Materials and methods

Sampling and image preparation

Diseased soybean leaves were sourced from (a) naturally infected plants of various cultivars evaluated in experimental plots at Embrapa Soja station in Londrina (PR, Brazil); and (b) artificially inoculated greenhouse plants. For the latter, soybean plants of a susceptible cultivar (NA 5909 RG) were grown in the greenhouse under high humidity. A spore suspension of *P. pachyrhizi* (10^5 spores/ml) was spray-inoculated on the plant leaves during flowering (R1) stage. A total of 200 leaves displaying incremental severity levels, from minimum to maximum as perceived visually, from both field and greenhouse conditions, were scanned using a Hewlett Packard scanner (Model 2130) at a resolution of 300 dpi. The proportion of leaf diseased area was determined using QUANT image

processing and analysis software (Vale *et al.*, 2003). Dozens of photographic images were taken from the diseased (necrotic and chlorotic) and healthy portions of tissues of several leaves as well as the background. The RGB (red, green, blue) coordinates for these samples were used to fit a discriminant function, which was further used to estimate the percentage diseased area in the whole set of images after excluding the non-leaf image background. These values were assumed to best represent the actual severity value. The minimum and maximum severity values were 0.2% and 84%, respectively.

SAD development

The New SAD was designed with 10 severity values following an amended linear scale: 3 diagrams were amended between the minimum (0.2%) and 10% severity, and 7 other diagrams were defined at approximately 15-percentage point intervals (Figure 1). The leaflets chosen to represent each value were selected from the original sample based on their proximity to the severity levels, leaflet integrity, as well as shape and distribution of the lesions typical of the most common pattern for a specific severity level. These images were slightly enhanced digitally using PhotoImpression (ArcSoft) for adjusting severity to match the predefined severity value.

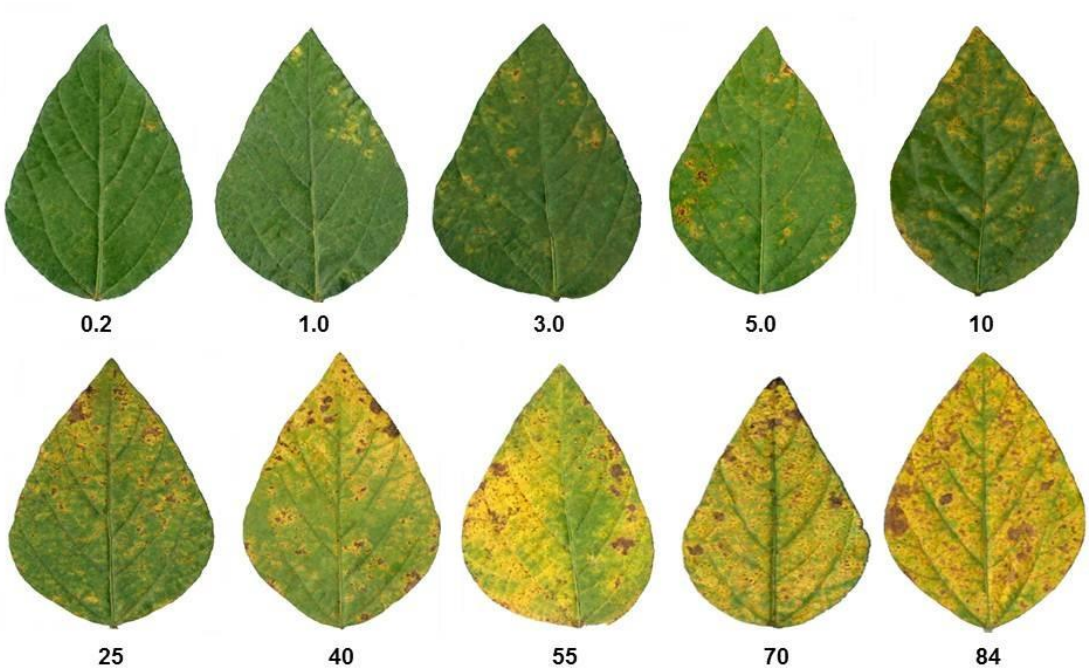

Figure 1: Standard area diagram set (SADs) for aiding visual estimates of rust (*Phakopsora pachyrhizi*) severity on soybean (*Glycine max* L.) leaves. The numbers represent percent (%) leaf area showing symptoms (necrosis and chlorosis).

SAD evaluation

A two-step process was conducted to evaluate the New SAD. First, 37 raters with no experience in quantifying plant diseases were instructed to assign a percentage value representing the relative area of the leaf depicting lesions (necrotic area + chlorotic halo) without any aid. Each image of a set of 50 photographic images of soybean leaflets, with actual severity ranging from 0.25% to 84%, embedded in PowerPoint slides, was projected on the screen for 30 s. After assessing all images, a 15-min break was taken and each rater was randomly assigned one of the two groups. One performed the visual assessments using the Old SAD (Godoy *et al.*, 2006) as an aid during the

assessment, and the other used the New SAD (Figure 1). The order of the images was randomized in the assessments.

Accuracy and reliability of the estimates

The overall accuracy (same as agreement) of the estimates, which refers to how close the severity estimates are to the actual severity (Madden *et al.*, 2007; Bock *et al.*, 2016), was determined for each rater and condition (unaided and SAD-aided) based on the Lin's concordance correlation coefficient (LCC, ρ_c) (Lin, 1989), as suggested for plant disease data (Madden *et al.*, 2007; Bock *et al.*, 2010b). In addition, the two components of overall accuracy (precision and bias correction factor) were explored to investigate the ramifications of errors.

Comparison of accuracy and reliability across methods

A generalized linear mixed model was fitted to LCC parameters data for each rater. Assessments (unaided or SAD-aided) and raters were considered fixed and random effects in the model, respectively. The least square means of each LCC parameter across the assessment methods condition were compared based on Tukey's honestly significant differences at 5% level of significance.

The inter-rater reliability, or reproducibility, was evaluated using two different methods: the intraclass correlation (ICC) (Shoukri and Pause, 1999) with decisions prior to the analysis made as described elsewhere (Schwanck and Del Ponte, 2014), and the overall concordance correlation (OCC), which is an improved LCC method for multiple raters (Barnhart *et al.*, 2002). The ICCs were compared based on the confidence interval.

To check whether the two groups show similar baseline accuracy, so that the differences in estimations were not due to one group being inherently more apt at providing accurate estimates, the mean estimate of LCC parameters for the unaided estimates were compared between the two groups of raters.

The relationship between percent point error (deviation) of the estimates (estimate minus actual) and actual severity, as well as the density distribution of errors of estimates, were depicted for each of the four groups. Finally, the functional relationship between the gain/loss (SAD aided – unaided) in overall accuracy and baseline accuracy (unaided) for each rater was also explored.

Simulations and the effect of SAD on hypothesis testing

The impact of the error of visual estimates of disease severity using either SAD was assessed via power analysis of simulated experiments. In brief, we ran virtual experiments to compare the control efficacy (percentage disease reduction relative to a nontreated check) between two treatments of interest (hereafter A and B) at different scenarios resulting from the combination of three situations: (a) four assessment methods, each using an "average rater" with no or different bias in severity estimation (details in next section); (b) incremental differences (Δ) in percentage point (p.p.) control efficacy (Δ = 0, 5, 10, or 15 p.p.) between the two fungicides; and (c) incremental number of leaflets (*n*) sampled within each replicated plot (5 to 55, by 10).

For all experiments, severity in both the nontreated check and the percentage control efficacy were fixed at 70%. These values were defined based on the fact that the leaf is more likely to fall when the severity is around 70% to

80% (Kumudini *et al.*, 2008). Finally, 70% was used to represent a reduction of disease by means of any treatment. For soybean rust, this value is within the range of efficacies determined for fungicides evaluated for soybean rust in experimental trials (Dalla Lana *et al.*, 2018). A total of 160 replicated experiments were simulated and two treatments were compared statistically with regards to control efficacy, with special interest in the power of the assessment methods, or the probability of obtaining a significant result with the correct sign.

Power analysis

The experiments were arranged in a completely randomized balanced design with five field plots as replicates. We used a *t* test, similar to previous studies (Bock *et al.* 2010a; Chiang *et al.*, 2014), to compare the control efficacy of the two treatments for each scenario that resulted from the combination of the above-mentioned situations. We were interested in detecting differences of control efficacy at incremental magnitudes (effect size), including the zero difference. A total of 2×10^4 Monte Carlo simulations were performed and the *t* test *p* value extracted from each run. The proportion of $p \leq .05$ gives the power [$p(\mathrm{reject}H_0 \vee H_1 \mathrm{istrue})$], or the probability that the test rejects the null hypothesis (H_0) when a specific alternative hypothesis (H_1) is true. The higher the power, the lower the probability of making a type II error, and thus more likely to detect an effect if it actually exists (Ellis, 2010).

Error due to the assessment method

The first source of error was due to bias in the estimates at the leaflet unit when using the different methods (except for digital analysis). The methods compared based on estimates by an "average rater" were: (a) the unaided (no SAD) situation; (b) using the Old SAD (Godoy *et al.,* 2006) as aid; (c) using the New SAD; and (d) the digital image analysis, assumed to have no error. The errors for each method based on estimates by an "average rater" were conditioned to actual severity as suggested in Figure 3a, which depicts trends in the errors over a range of actual severity. The errors were predicted from the fit of a linear regression model to the data on the relationships between error and the actual severity (Figure 2a,b,c). The parameters of the linear regression model fitted to data of each method are shown in Table 1.

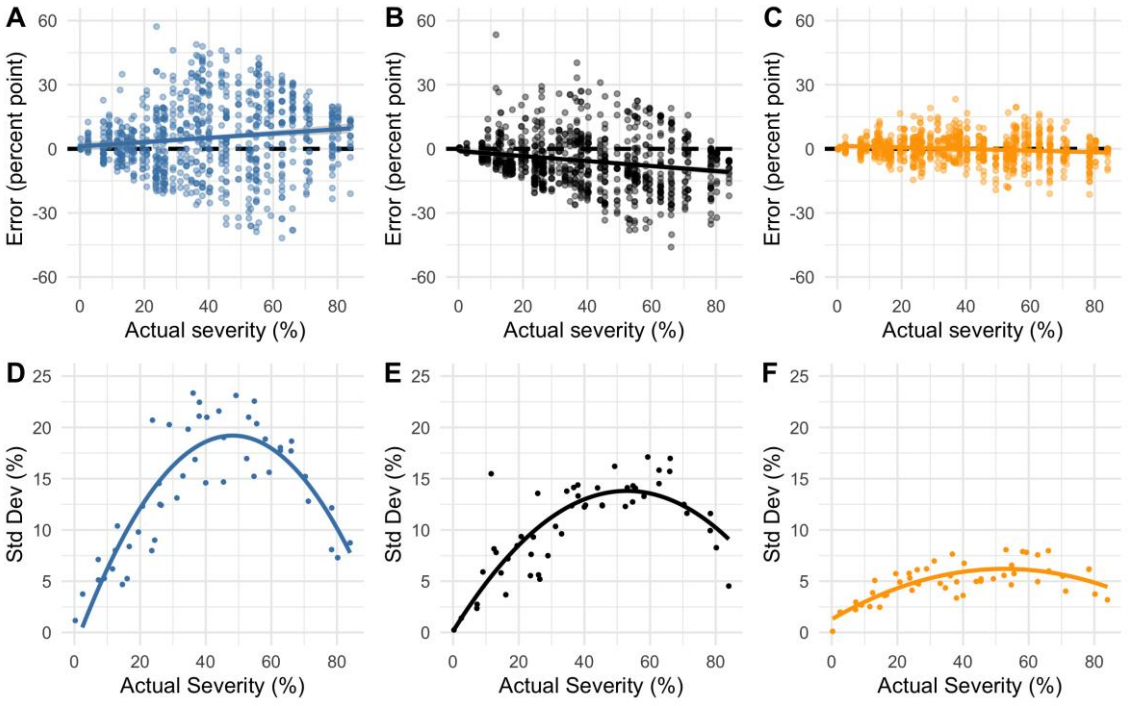

Figure 2: A-C. Relationship between the error (estimate minus actual value) of visual estimates and actual (software-based) estimates of soybean rust severity

assessed for a collection of 50 leaflets by three groups of raters where the assessments were made unaided (A), aided by the Old SAD (B) (Godoy et al. 2006) or aided by the New SAD (this study) (C). The dashed line represents the no error line and the solid line is the fit of the first-order linear model to the data, which was assumed to represent an 'average rater' (see model parameters in Table 1). D-F. Relationship between actual (software-based estimates) severity and the mean standard deviation of estimates across raters (17 to 19 raters each condition) grouped as in A to C. The colored line represents the fit of the second-order linear model to the data (see model parameters in Table 1).

Table 1: Parameters for two linear regression models: 1) a first-order model for predicting the mean (μ) error of visual severity estimates conditioned to actual severity (S), for a sample of at least 18 raters; and 2) a second-order model for predicting the standard deviation (σ) of the visual estimates, by at least 18 raters, conditioned to values of actual severity in a sample of 50 soybean leaflets. The models were fitted to estimates obtained unaided (no SAD), aided by the Old SAD (Godoy et al. 2006) and a New SAD (this study).

Method	Equation*
Unaided	$\mu = S + 1.145 + 0.10020\,S$
	$\sigma = -1.833 + 0.7794\,S - 0.007150\,S^2$
Old SAD	$\mu = S - 1.039 - 0.11694\,S$
	$\sigma = -0.114 + 0.6151\,S - 0.006718\,S^2$
New SAD	$\mu = S + 1.411 - 0.03738\,S$
	$\sigma = 1.061 + 0.1990\,S - 0.001925\,S^2$

* where S represents the actual severity value estimated by image analysis

Error due to intraplot severity heterogeneity

The second experimental error was due to the intraplot heterogeneity of soybean rust severity, which shows a vertical incremental disease gradient from the bottom to the top of the canopy (Garcés and Forcelini, 2013). We assumed

severity (S) values to be β-distributed $S_{ij} \sim B(\alpha_S, \beta_S)$ with mean μ_S and standard deviation (due to experimental error) σ_S. The shape parameters of the β distribution are given by

$$\alpha = \frac{-\mu(\sigma^2+\mu^2-\mu)}{\sigma^2} \text{ and } \beta = \frac{(\sigma^2+\mu^2-\mu)(\mu-1)}{\sigma^2},$$

where μ and σ^2 are the respective mean and standard deviation of severity. (These equations will be used for calculating the shape parameters for further variables). The standard deviation was a function of mean severity and was given by

$$\sigma_S = \mu_S 0.006(100 - \mu_S).$$

The parameter (0.006) gives a maximum $\sigma_S = 15\#$ at 50% severity, which was obtained from literature reports of severity variation at three canopy heights (Garcés and Forcelini, 2013).

Simulation of control efficacies

Control efficacy (C) values were assumed to be normally distributed $C_{ij} \sim N(\mu_C, {\sigma_C}^2)$ with mean μ_C and standard deviation σ_C ($i = 5,10,\ldots,n$, n being the number of sampled leaflets, and $j = 1,2,\ldots,R$, R being the number of replicates in the experiment). Because n leaflets are sampled in each replicate j, the total number of samples is given by their product nR.

Given the control efficacies $C_{f,ij}$ of treatment ($f = \{A, B\}$) under comparison, the percent severity in the treatment ($S_{f,ij}$) is given by

$$S_{f,ij} = \left(1 - \frac{C_{f,ij}}{100}\right) Schk_i \ .$$

The estimate of severity (s) in the treatment (f) aided by each method ($method = \{no, new, old\}$, denoted as $s_{f,ij,method}$ and severity for the check treatment, denoted as $schk_{ij,method}$ which varies with severity, due to the SAD error.

$$s_{f,ij,no} = B\left(\alpha_{ij,f,no}, \beta_{ij,f,no}\right),$$

$$s_{f,ij,old} = B\left(\alpha_{ij,f,old}, \beta_{ij,f,old}\right),$$

$$s_{f,ij,new} = B\left(\alpha_{ij,f,new}, \beta_{ij,f,new}\right),$$

$$schk_{ij,method} = B\left(\alpha chk_{ij,s}, \beta chk_{ij,method}\right).$$

The parameters of the linear models for estimating the mean error and the standard deviation, conditioned to actual severity, for each method, are shown in Table 1 and Figure 2. The mean severity of each replicate of each treatment was given by

$$\bar{s}_{f,j,method} = \frac{1}{n}\sum_{i=1}^{n} s_{f,ij,method},$$

$$\overline{schk}_{j,method} = \frac{1}{n}\sum_{i=1}^{n} schk_{ij,method}.$$

$\overline{Schk}_j$ and $\bar{S}_{f,j}$ were also calculated similarly to be used in the next step. The control efficacies for each treatment before adding the method-derived error (c^*, hereafter the unbiased estimate) and the biased control efficacy due to the method-derived errors (c) is given by

$$c_j^* = \frac{\overline{Schk}_j - \bar{S}_{f,j}}{\overline{Schk}_j} 100 \text{ and } c_{j,f,method} = \frac{\overline{schk}_{j,method} - \bar{s}_{f,j,method}}{\overline{schk}_{j,method}} 100.$$

Data processing and availability

All data processing and analyses, as well as graphical work, were performed with R v. 3.5.1 (2019-09-13). Several R packages of the tidyverse (Wickham *et al.*,

2019) were used to prepare, transform, and visualize the data. R packages for conducting some statistical analyses included lme4, emmeans, epiR, irr, car, and broom. The R scripts with text annotations were prepared as R Markdown documents (Xie *et al.*, 2018) and all files were organized as a research compendium (Gentleman and Temple Lang, 2007) structured as an RStudio project (Gandrud, 2016). To encourage and facilitate reproducibility, a website was generated to navigate through the documented code. All files, including the photographic images used for validation, are freely available and permanently stored at https://doi.org/10.17605/OSF.IO/3ZMV8.

Results

Baseline accuracy

A positive linear relationship was found between unaided estimates, using either the Old or the New SAD, and actual severity, and the pattern of the relationship was similar between these two groups of raters (Figure 3a). The measures of precision (r) and bias correction factor (C_b) of the estimates varied among raters within each group and were related in a similar linear fashion between the two groups (Figure 3b. Furthermore, the distribution of all the LCC parameters suggested that the two groups of raters were, on average, similar in their baseline accuracy (Figure 3c). This was confirmed by the non-rejection of the null hypothesis for all LCC parameters when comparing the unaided estimates of the two groups (Table 2).

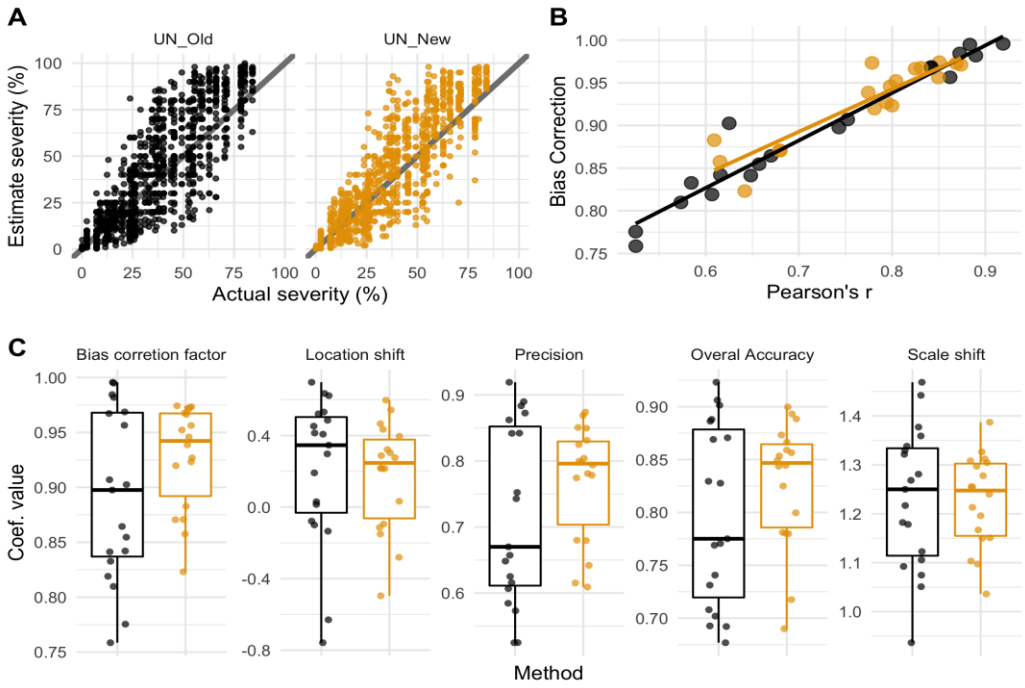

Figure 3: Relationship between actual and estimated severity of rust (*Phakopsora pachyrhizi*) severity on soybean (Glycine max) leaves without Old SAD (UN_Old) and without New SAD (UN_New) (A), relationship between bias correction factor (accuracy) and Pearson's r (precision) (B), and box plot of the LCC parameters statistics for raters's estimations of severity without Old and New SAD (C).

Table 2: Generalized linear mixed model analysis of the statistics of the Lin's concordance correlation coefficient (LCC) parameters that represent accuracy (types of bias), precision (correlation coefficient) of percent estimates of severity of soybean rust by 37 inexperienced raters during two assessments without and with the use of a new standard area diagram set (New SAD) and an Old SAD (Godoy et al. 2006) as an aid during visual assessment of disease severity.

Method	*N*	u^b	v^c	C_b^d	r^e	ρ_c^f
New SAD	18	0.01[a] a	0.99 a	0.99 a	0.96 a	0.96 a
Old SAD	19	-0.26 b	0.98 a	0.92 b	0.77 b	0.83 b
Unaided New	18	0.19 c	1.23 b	0.92 b	0.77 b	0.82 b
Unaided Old	19	0.21 c	1.23 b	0.90 b	0.72 b	0.80 b

[a] Means followed by the same letter in the column are not significantly different (Tukey's HSD, 5% level)

[b] Location shift (u, 0 = no bias relative to the concordance line).
[c] Scale shift (v, 1 = no bias relative to the concordance line).
[d] Bias correction factor (C_b) measures how far the best fitted line deviates from 45^0 and is a measure of accuracy.
[e] Correlation coefficient as a measure of precision (r).
[f]Lin's concordance correlation coefficient (LCC), that combines both measures of precision (r) and accuracy (C_b) to measure overall accuracy (agreement) with the true value.

Effect of SAD on accuracy, precision and reliability statistics

When using either SADs, the accuracy and precision of the severity estimates were improved in relation to unaided estimates only for the group of raters that used the New SAD (Table 2). The lowest absolute percentage point errors were observed for this group; most deviations were within ±15 p.p. about the actual value. For the Old SAD group, errors concentrated within ±30 p.p. about the actual severity value (Figure 4a). In general, when not using the SAD, both groups tended to slightly overestimate severity, with the largest absolute errors, either positive or negative, concentrated in the range of 20% to 60% actual severity. In general, raters of the Old SAD group tended to underestimate severity and the magnitude of the errors increased with the increase in actual severity (Figure 4a,b). The trend of underestimation was also confirmed by the negative mean value of the location-shift that represents constant bias (Table 2).

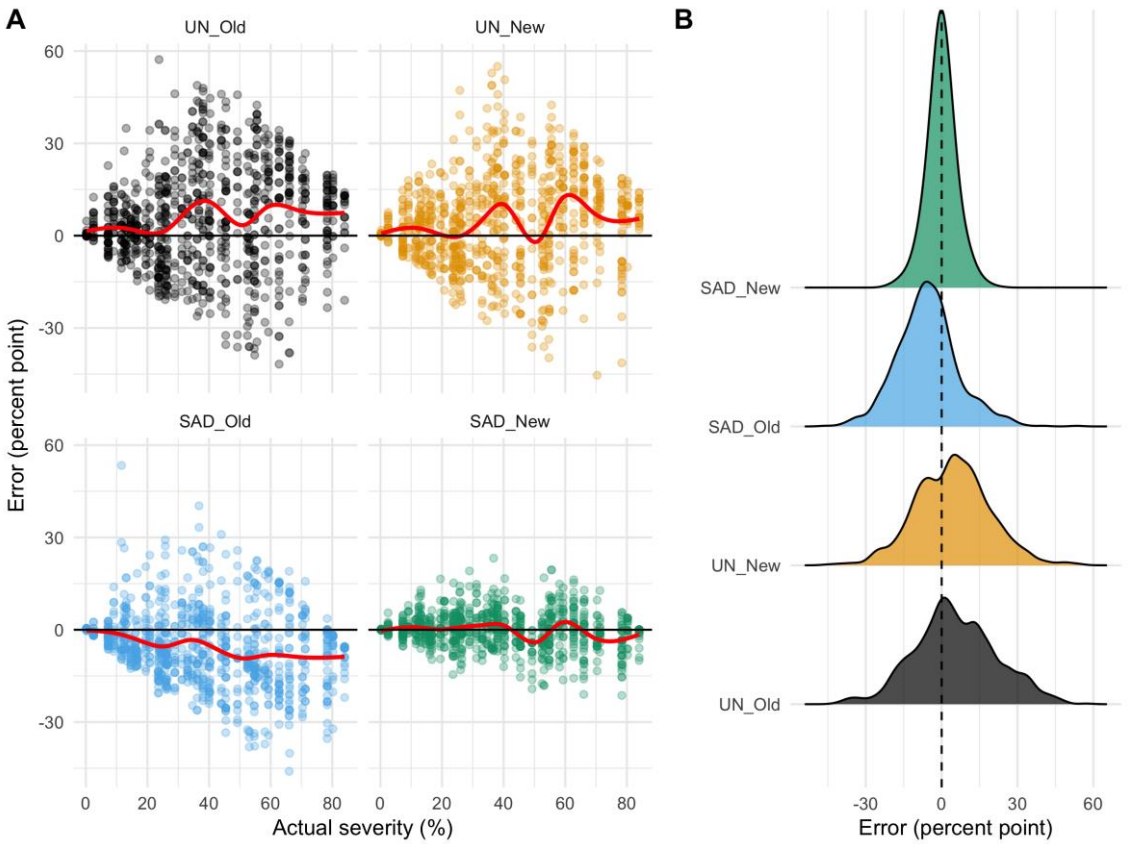

Figure 4: Scatter plot of the relationship between absolute error (estimates - actual) and actual severity (A), and their respective density plots (B), of visual estimates of soybean rust severity provided by two groups of raters. Each group, composed of inexperienced raters, firstly provided visual estimates of severity without an aid and then aided with a specific SAD, the Old SAD (Godoy et al. 2006) (19 raters) or the New SAD (this study) (18 raters). A set of 50 soybean leaflets with actual severities ranging from 0.25 to 84% were used. The smooth red line in A is the result of a local polynomial regression fitting (Loess).

Results of the ICC and OCC statistics showed an overall improvement in the inter-rater reliability for the group of raters that used the New SAD (Table 3). In fact, the between-rater variation in the concordance measures, when inspected individually, was highest for the Old SAD group, with the precision and overall accuracy of all raters being consistently below 0.9 (Figure 5a). While bias correction factor was above 0.9 for most raters, irrespective of the SAD, the precision (Pearson's *r*) varied the most among raters using the Old SAD ($r = 0.78$), which was the main reason for only a fair means of overall accuracy ($\rho_c = 0.84$),

not differing from the unaided estimates (Table 2). Moreover, the use of the Old SAD was detrimental (loss of overall accuracy up to −0.15) for around one-third of the raters who had shown good levels of baseline (unaided) accuracy (Figure 5b).

Table 3: Measures of inter-rater reliability of severity estimates by 37 inexperienced raters during two assessments without and with the use of new a standard area diagram set (New SAD) and Old SAD (Godoy et al. 2006) as an aid to assessment of disease severity

Method	Intra-class correlation coefficient (ICC) ρ (95% CI)[a]	Overall concordance correlation (OCC)[b]
New SAD	0.94 (0.92-0.96)	0.940
Unaided New	0.83 (0.77-0.88)	0.759
Old SAD	0.81 (0.74-0.87)	0.736
Unaided Old	0.83 (0.77-0.89)	0.746

[a] Calculated with decisions of ICC model described elsewhere (Schwanck and Del Ponte 2014)
[b] Overall agreement statistics based on Lin (1989) and Barnhart et al. (2002) to evaluate agreement among multiple observers

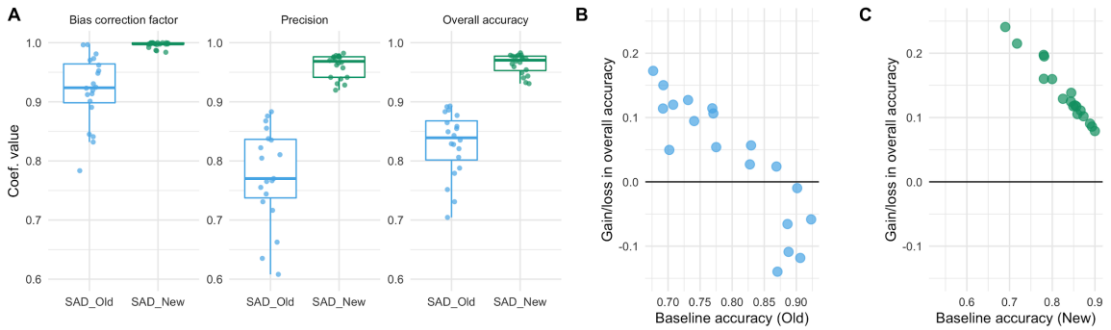

Figure 5: Distribution of bias correction factor (Cb), Pearson's correlation coefficient (r) (precision) and overall accuracy (ρc) statistics for two groups of

raters that used either an Old SAD (19 raters) (Godoy et al. 2006) a New SAD (18 raters) as an aid during visual estimation of soybean rust severity on 50 leaflets (A). The gain/loss (ρc SAD-aided - ρc unaided) in overall accuracy decreased with the increase in baseline accuracy (unaided) (B, C).

Power analysis of the assessment methods

Use of simulation to ascertain the effect of bias of the two SAD-aided estimates, considering an "average rater", the power to detect no difference (Δ = 0 p.p.) in control efficacy was not greater than the 0.05 (alpha) regardless of the method (Fig 6a). The New SAD was much more powerful to detect differences in fungicide efficacy larger than zero, especially for the lowest tested difference (Δ = 5 p.p.) in control efficacy (Figure 6b). For this same Δ, the power of the Old SAD did not differ from unaided estimates; 40 samples would be required to detect the difference with a power of 0.8. Using the New SAD and the unbiased method (image analysis) the sample size can be reduced to 20 and 10 units, respectively, for reaching the same power for Δ = 5 (Figure 6b). As expected, the larger the difference in efficacy (Δ = 10 to 15), the lower the number of samples to achieve the same power (Figure 6c,d). For those larger differences, a sample size as low as 10 using the Old SAD and only five using the New SAD is sufficient to achieve the 0.8 power. For the unbiased method, five samples would be sufficient for that range of difference.

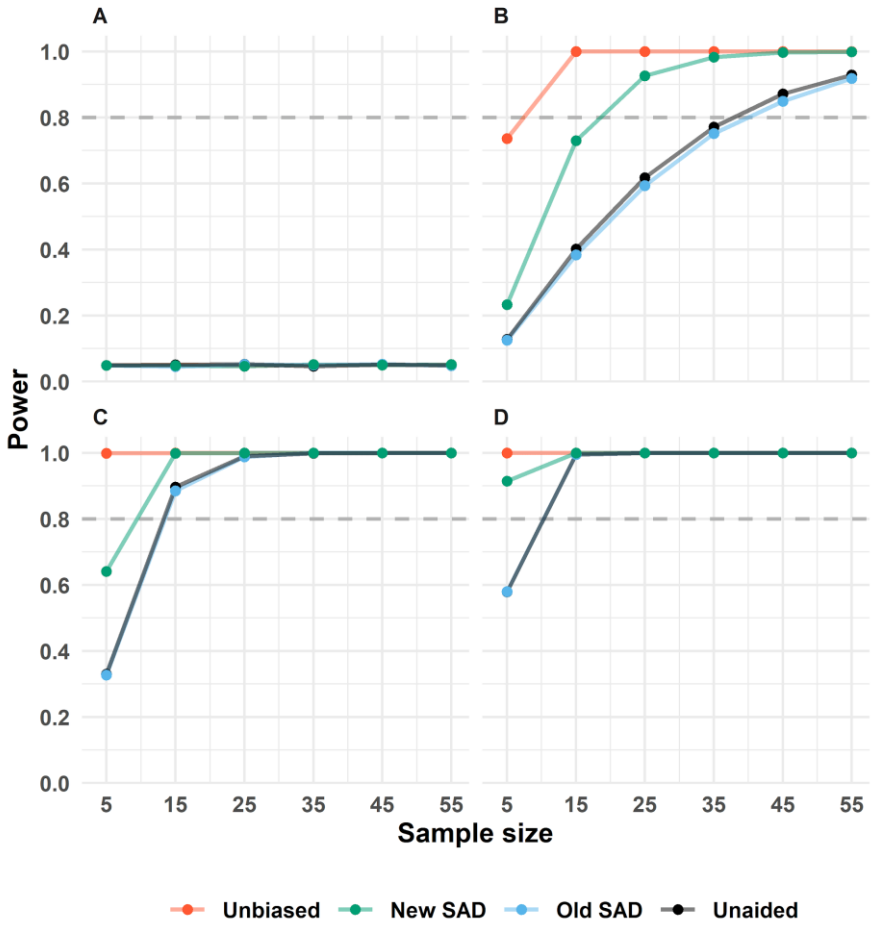

Figure 6: Statistical power of four methods for assessing soybean rust severity determined by Monte Carlo simulations of a t-test's P-value for comparing the percent control efficacy of two fungicides. Replicated (n = 5 plots) experiments were simulated for each method: unbiased estimates are based on digital analysis (no error), visual estimates with no standard area diagram (SAD) (unaided) and two estimates using a New SAD (this study) and the Old SAD (Godoy et al. 2006). Each method, excepting the unbiased, has inherent errors of the estimation by an 'average rater', for scenarios combining incremental differences (percent points): Δ = 0 (A), Δ = 5 (B), Δ = 10 (C), and Δ = 15 (D) in efficacy and incremental number of samples (leaflets) per plot (5 to 55, by 10 leaflets). While errors varied across the assessment methods, error due to intraplot heterogeneity of soybean rust severity was the same across methods.

Discussion

The development of a New SAD for aiding visual assessment of soybean rust severity was motivated by recent developments in SAD technology as well as best practices for development and statistical evaluation of the tool (Del Ponte *et*

al., 2017). Although validated and considered suitable for aiding soybean rust severity estimation, we anticipated a few issues with the Old SAD. First, besides using a rather small number of inexperienced raters (four) during validation, inferences about accuracy and precision of aided estimates were based on hypothesis tests applied separately to the linear regression coefficients at the rater rather than the group level, an approach that was deemed inappropriate in favour of concordance analysis (Madden *et al.*, 2007; Bock *et al.*, 2010b). In the Old SAD paper, the increase in accuracy was based on two results: (a) apparent decrease in the errors of estimates; and (b) lower number of raters whose null hypothesis of the slope being equal to 1 was not rejected. However, the intercepts for all raters were positive and differed from 0 irrespective of the method, evidencing a location bias (Godoy *et al.*, 2006). Moreover, although the precision of the SAD-estimates generally increased, the calculated average (r = 0.81) across inexperienced raters was very close to our estimate of precision when using the same SAD (r = 0.78). These observations help to explain why the aided estimates with the Old SAD did not differ from the unaided estimates in our study using a larger number of raters and hypothesis tests on concordance statistics at the group level. Finally, we used a much larger number of leaflets displaying severity larger than 40% in our validation dataset, while only seven were used in the previous study (Godoy *et al.*, 2006), thus confirming that the errors remained at the upper severity levels.

The general poor performance of the Old SAD compared with the New SAD may be due to several issues. First, the Old SAD was developed based on the (disproven) assumption that the increments between severities should be

based on the (nonexistent) Weber–Fechner law (Nutter and Esker, 2006; Bock *et al.*, 2010b). The Old SAD has six diagrams and only three from 18% to 78.5%, a range where the error of the aided estimates were kept at high magnitude, preventing raters from improving accuracy. There is a tendency for an increased number of diagrams in recent studies which is related to the maximum severity (Del Ponte *et al.*, 2017). The use of 10 diagrams with severity increasing following the amended linear (15% increment) concept (Chiang *et al.*, 2016a) proved sufficient to increase overall accuracy and reliability of the estimates at high levels (>0.95). Further increases in the number of diagrams are not encouraged because more choices can slow down assessment time duration (Bock *et al.*, 2016). However, new technology for delivering digital SADs, such as embedding them in tablet apps, which allows for interaction and data storage and processing, may open new possibilities and paradigms in SAD use (Del Ponte *et al.*, 2019).

The amended linear SAD displaying true-colour images significantly improved overall accuracy and interrater reliability. The use of true-colour photos may have contributed to the overall improvement but we could not tell this effect apart because the SADs have distinct colour schemes. It would be instructive to further modify the New SAD and test whether a reduction in the number of colours is of any influence in the error of the estimates. A realistic representation of lesions may be of more importance for certain patterns, such as in the present disease in which the higher the number of lesions, the higher the chlorotic area. The topic merits further investigation.

The Old SAD has been widely recommended and used in soybean rust field research, mainly to evaluate fungicide performance in field experimentation (Scherm *et al.*, 2009; Dalla Lana *et al.*, 2018). Our results show that its use should be discouraged, not only because of a clear lack of benefit, but also because of the detrimental effect on the estimates of raters who were inherently more accurate. Using simulation, we showed that the use of the New SAD may optimize resource use, especially by reducing sample size. In the disease assessment protocol of the UFTs for evaluating control efficacy, raters are instructed to collect 30 leaflets, 10 from each canopy height, and score severity at each leaflet aided or trained by Old SAD (Scherm *et al.*, 2009). Results of our power analysis suggest an increase in sample size to at least 40 leaflets if the interest is in detecting differences in efficacy of at least five percent points between treatments. With the New SAD, a reduction of sample size to 20 leaflets would result in considerable savings in time in large replicated experiments such as fungicide evaluations that usually compare 10 to 15 products. In a hypothetical situation where a rater would take five seconds, on average, to estimate and record severity on each leaf, 3.3 hr would be required to estimate severity in 2,400 leaflets for the entire experiment (15 treatments × 4 replicates × 40 leaves). With a two-fold reduction in sample size, the whole experiment could be evaluated in 1.6 hr.

Our results are in agreement with previous reports of increasing type II errors due to increased bias in severity estimates (Christ, 1991; Todd and Kommedahl, 1994; Parker *et al.*, 1995; Chiang *et al.*, 2014; 2016a,b). The simulation approach we used to evaluate the impact of bias on the power of

hypothesis test was similar to other studies (Bock *et al.*, 2010a; Chiang *et al.*, 2016a), but the difference here is that we focused on comparing control efficacy, not severity directly, thus simultaneously taking bias at both high (untreated check) and low (treated) severity into account. In conclusion, our results strongly discourage the use of the Old SAD. Because soybean rust researchers have become accustomed to estimating severity using a tool that introduces undesirable bias in the estimates and compromises resource use, it is urgent that the New SAD proposed here is adopted as a new standard and, ideally, that the soybean rust researchers are re-evaluated and retrained.

Acknowledgements

This study was financed in part by the "Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil" (CAPES) Finance Code 001. The fifth and sixth authors received research fellowships from the National Council for Scientific and Technological Development (CNPq)/Brazil.

Data availability statement

The data that support the findings of this study are openly available in Open Science Framework project at https://osf.io/3zmv8/, doi 10.17605/OSF.IO/3ZMV8.

References

Barnhart, H.X., Haber, M. and Song, J. (2002) Overall concordance correlation coefficient for evaluating agreement among multiple observers. *Biometrics*, 58, 1020–1027.

Bock, C.H., Gottwald, T.R., Parker, P.E., Ferrandino, F., Welham, S., Van Den Bosch, F. *et al.* (2010a) Some consequences of using the Horsfall-Barratt scale for hypothesis testing. *Phytopathology*,100, 1030–1041.

Bock, C.H., Poole, G., Parker, P.E. and Gottwald, T.R. (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, C.H., Chiang, K.C. and Del Ponte, E.M. (2016) Accuracy of plant specimen disease severity estimates: concepts, history, methods, ramifications and challenges for the future. *CAB Reviews*, 11, 1–21.

Buffara, C.R.S., Angelotti, F., Vieira, R.A., Bogo, A., Tessmann, D.J. and Bem, B.P. (2014) Elaboration and validation of a diagrammatic scale to assess downy mildew severity in grapevine. *Ciência Rural*, 44, 1384–1391.

Camara, G.R., Busato, L.M., Almeida, B.F. and Moraes, W.B. (2018) Elaboration and validation of diagrammatic scale for lettuce powdery mildew. *Summa Phytopathologica*, 44, 116–121.

Chiang, K.S., Liu, S.H., Bock, C.H. and Gottwald, T.R. (2014) What interval characteristics make a good disease assessment category scale? *Phytopathology*, 104, 575–585.

Chiang, K.S., Bock, C.H., El Jarroudi, M., Delfosse, P., Lee, I.H. and Liu, H.I. (2016a) Effects of rater bias and assessment method on disease severity estimation with regard to hypothesis testing. *Plant Pathology*, 65, 523–535.

Chiang, K.S., Bock, C.H., Lee, I.H., El Jarroudi, M. and Delfosse, P. (2016b) Plant disease severity assessment—how rater bias, assessment method, and experimental design affect hypothesis testing and resource use efficiency. *Phytopathology*, 106, 1451–1464.

Christ, B.J. (1991) Effect of disease assessment method on ranking potato cultivars for resistance to early blight. *Plant Disease*, 75, 353–356.

Costa, A.P., Peixoto, J.R., Blum, L.E.B. and Pires, M.C. (2019) Standard Area Diagram Set for Scab Evaluation in Fruits of sour Passion Fruit. *Journal of Agricultural Science*, 11, 298–305.

Dalla Lana, F., Ziegelmann, P.K., Maia, A.H.N., Godoy, C.V. and Del Ponte, E.M. (2015) Meta-analysis of the relationship between crop yield and soybean rust severity. *Phytopathology*, 105, 307–315.

Dalla Lana, F., Paul, P.A., Godoy, C.V., Utiamada, C.M., Silva, L.H.C.P., Siqueri, F.V. *et al.* (2018) Meta-analytic modeling of the decline in performance of fungicides for managing soybean rust after a decade of use in Brazil. *Plant Disease*, 102, 807–817.

Del Ponte, E.M., Pethybridge, S.J., Bock, C.H., Michereff, S.J., Machado, F.J.and Spolti, P. (2017) Standard area diagrams for aiding severity estimation: scientometrics, pathosystems, and methodological trends in the last 25 years. *Phytopathology*, 107, 1161–1174.

Del Ponte, E.M., Nelson, S.C. and Pethybridge, S.J. (2019) Evaluation of app-embedded disease scales for aiding visual severity estimation of cercospora leaf spot of table beet. *Plant Disease*, 103, 1347–1356.

Ellis, P.D. (2010) *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge, Cambridge University Press.

Fantin, L.H., Braga, K., Canteri, M.G., Dias, A.R. and Borges, E.P. (2018) Development and validation of diagrammatic scale to assess target spot severity in cotton. *Australasian Plant Pathology*, 47, 491–497.

Gandrud, C. (2016) *Reproducible Research with R and R studio*. Boca Raton, FL: Chapman and Hall/CRC.

Garcés, F.R. and Forcelini, C.A. (2013) Controle comparativo da ferrugem asiática da soja com fungicida triazol ou mistura de triazol+ estrobilurina. *Bioscience Journal*, 29, 805–815.

Gentleman, R. and Temple Lang, D. (2007) Statistical analyses and reproducible research. *Journal of Computational and Graphical Statistics*, 16, 1–23.

Godoy, C.V., Koga, L.J. and Canteri, M.G. (2006) Diagrammatic scale for assessment of soybean rust severity. *Fitopatologia Brasileira*, 31, 63–68.

Godoy, C.V., Seixas, C.D.S., Soares, R.M., Marcelino-Guimarães, F.C., Meyer, M.C. and Costamilan, L.M (2016). Asian soybean rust in Brazil: past, present, and future. *Pesquisa Agropecuária Brasileira*, 51, 407-421.

Hartman, G.L., Rupe, J.C., Sikora, E.J., Domier, L.L., Davis, J.A. and Steffey, K.L. (2015) *Compendium of Soybean Diseases and Pests*. St Paul, MN, USA: APS Press.

Horsfall, J.G. and Barratt, R.W.N. (1945) An improved grading system for measuring plant diseases. *Phytopathology* 35, 655.

Kumudini, S., Godoy, C.V., Board, J.E., Omielan, J. and Tollenaar, M. (2008) Mechanisms involved in soybean rust-induced yield reduction. *Crop Science*, 48, 2334–2342.

Lin, L.I. (1989) A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45, 255–268.

Madden, L.V., Hughes, G. and Van Den Bosch, F. (2007) *The Study of Plant Disease Epidemics*. St Paul, MN, USA: APS Press.

Nutter, F.W. and Esker, P.D. (2006) The role of psychophysics in phytopathology: The Weber–Fechner law revisited. *European Journal of Plant Pathology*, 114, 199–213.

Parker, S.R., Shaw, M.W. and Royle, D.J. (1995) The reliability of visual estimates of disease severity on cereal leaves. *Plant Pathology*, 43, 856–865.

Scherm, H., Christiano, R.S.C., Esker, P.D., Del Ponte, E.M. and Godoy, C.V. (2009) Quantitative review of fungicide efficacy trials for managing soybean rust in Brazil. *Crop Protection*, 28, 774–782.

Schwanck, A.A. and Del Ponte, E.M. (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, 670–682.

Shoukri, M.M. and Pause, C.A. (1999) *Statistical Methods for Health Science*. 2nd Edn. Boca Raton, FL: CRC Press.

Todd, L.A. and Kommedahl, T. (1994) Image analysis and visual estimates for evaluating disease reactions of corn to fusarium stalk rot. *Plant Disease*, 78, 876–878.

Vale, F.X.R., Fernandes Filho, E.I. and Liberato, J.R. (2003) QUANT: a software plant disease severity assessment. In: Close, R., Braithwaite, M. and Havery, I., eds. *Proceedings of the 8th International Congress of Plant Pathology, New Zealand*. Sydney, NSW, Australia: Horticulture Australia, p. 105.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R. *et al.* (2019) Welcome to the Tidyverse. *Journal of Open Source Software*, 4, 1686.

Xie, Y., Allaire, J.J. and Grolemund, G. (2018) *R Markdown: The Definitive Guide*. Boca Raton, FL: Chapman and Hall/CRC.

[dataset] Del Ponte, E.M. (2019) A new standard area diagram set for assessment of severity of soybean rust. DOI: 10.17605/OSF.IO/3ZMV8.