Reproducibility of the Development and Validation Process of Standard Area Diagram by Two Laboratories: an Example Using the *Botrytis cinerea*/*Gerbera jamesonii* Pathosystem

Vilma Pereira de Melo[1], Ana Claudia da Silva Mendonça[2]; Hudson Sergio de Souza[2], Lorrant Cavanha Gabriel[3], Clive H. Bock[4], Mahogani J. Eaton[5], Kátia Regina Freitas Schwan-Estrada[1.3], William Mário de Carvalho Nunes[2,3]

[1] Programa de Pós-Graduação em Agroecologia, Departamento de Agronomia, Universidade Estadual de Maringá;

[2] Núcleo de Pesquisa em Biotecnologia Aplicada, Universidade Estadual de Maringá;

[3] Programa de Pós-Graduação em Agronomia, Departamento de Agronomia, Universidade Estadual de Maringá;

[4] United States Department of Agriculture–Agricultural Research Service Southeastern Fruit & Tree Nut Research Lab, Byron, GA 31008;

[5] Fort Valley State University, Fort Valley, GA 31030

Corresponding author: William Mário de Carvalho Nunes

E-mail: wmcnunes@uem.br

Abstract

Standard area diagrams (SADs) are plant disease severity assessment aids demonstrated to improve the accuracy and reliability of visual estimates of severity. Knowledge of the sources of variation, including those specific to a lab such as raters, specific procedures followed including instruction, image analysis software, image viewing time, etc., that affect the outcome of development and validation of SADs can help improve standard operating practice of these assessment aids. As reproducibility has not previously been explored in development of SADs, we aimed to explore the overarching question of whether the lab in which the measurement and validation of a SADs was performed affected the outcome of the process. Two different labs (Lab 1 and Lab 2) measured severity on the individual diagrams in a SADs and validated them independently for severity of gray mold (caused by *Botrytis cinerea*) on Gerbera daisy. Severity measurements of the 30 test images were performed independently at the two labs as well. A different group of 18 raters at each lab assessed the test images first without, and secondly with SADs under independent instruction at both Lab 1 and 2. Results showed that actual severity on the SADs as measured at each lab varied by up to 5.18%. Furthermore, measurement of the test image actual values varied from 0 to up to 24.29%, depending on image. Whereas at Lab 1 an equivalence test indicated no significant improvement in any measure of agreement with use of the SADs, at Lab 2, scale shift, generalized bias and agreement were significantly improved with use of the SADs ($P \leq 0.05$). An analysis of variance indicated differences existed between labs, use of the SADs aid, and the interaction, depending on the agreement statistic. Based on an equivalence test, the inter-rater reliability was significantly ($P \leq 0.05$) improved at both Lab 1 and Lab 2 as a result of using SADs as an aid to severity estimation. Gain in measures of agreement and reliability tended to be greatest for the least able raters at both Lab 1 and Lab 2. Absolute error was reduced at both labs when raters used SADs. The results confirm that SADs are a useful tool; but the results demonstrated that aspects of the development and validation process in different labs may affect the outcome.

Key words: Reproducibility, disease evaluation, assessment, diagrammatic scales, Gerbera, *Gerbera jamesonii*, gray mold, *Botrytis cinerea*

Gerbera (*Gerbera jamesonii* H. Bolus ex. Hooker), is an important nursery plant for both cut flower production and as a container-grown plant. It is among the three most important container-grown flowers produced in Brazil (Ferronato et al., 2008; Andrade, 2016) and is an important crop in the U.S.A. (Anonymous, 2009). Gerberas are most often cultivated under protected environments which provides a favorable place for development of many diseases (Brisco-McCann and Hausbeck, 2016). Among the diseases common on foliage of Gerbera is gray mold, caused by the fungus *Botritys cinerea* Pers. Although common on foliage causing spotting and blighting, *Botrytis* can also cause damping-off, crown rot and infection of flowers (Daughtery et al., 2000; Töfoli et al, 2011). Leaves develop gray-brown zonate lesions of variable size and shape; in some situations, the disease may cause drying and necrosis of leaf tips and edges. Flower petals show tan spots and tip necrosis or are entirely blighted. The disease may be seed borne (Daughtery et al., 2000). The infection reduces the profitability of gerbera production. Although endeavors are underway to develop *Botritys*-resistant gerbera (Fu et al., 2015), this will take time and screening of progeny for disease resistance based on severity of symptoms can be a requirement.

Accuracy and reliability of visually acquired disease estimates are important for several aspects of plant pathology and related disciplines (Madden et al., 2007; Bock et al., 2016). Inaccurate individual estimates, and the resulting imprecision and unreliability can result in incorrect conclusions (Parker et al., 1995; Chiang et al., 2016a). Standard area diagrams (SADs, otherwise called diagrammatic scales) are important tools to aid in the accuracy and reliability of estimates of the severity of plant diseases (Bock et al., 2010; Del Ponte et al., 2017). SADs have been developed for over 100 pathosystems, and are habitually used in the field by many researchers as an aid to improve the accuracy and reliability of an individual's disease severity estimates. Although SADs are well established, there remain many facets that have yet to be understood regarding their development, usage and benefit (Del Ponte et al., 2017). Very recently the first 'best practices' or standard operating procedures (SOPs) were developed for SADs, but these do not provide definitive detail regarding specific instructions, image analysis processing, number of images in a SADs, validation, rater selection, etc. (Del Ponte et al., 2017), partly because information is lacking on the impact of these factors. One aspect that has not been explored is whether the laboratory in which the development and validation of a SADs affects the overall outcome of the process. Sources of variation specific to a laboratory may include raters, SOPs used, image analysis software, viewing time for images, and amount or quality instruction provided to raters. Ideally, the recommended SOP for development and validation process should be sufficiently robust to prevent unwanted variability among labs. We aim to explore the overall effect of lab in which measurement and validation of a SADs is performed.

Furthermore, development and validation of SADs that demonstrably improve accuracy and reliability of disease estimates is valuable as they become more widely available for use on hand held devices for application in the field in real time (Pethybridge and Nelson, 2018). There are challenges to how these device-based SADs may be implemented (Del Ponte et al., 2019), but they need to be based on SADs that are effective at improving accuracy and reliability of estimates for the disease in question.

As noted, SADs have been instrumental in improving accuracy and precision of disease severity assessments. Unfortunately, unaided severity estimates of individual diseased specimens are known to be subjective and variable among raters, with estimates deviating from the actual value to differing degrees (Nutter et al., 1993; Bock et al., 2010 and 2016). Thus, SADs are useful and fundamental tools to assist the evaluator and reduce subjectivity and error (Sposito et al., 2004; Barbosa et al., 2006; Barguil et al, 2008; Sussel et al., 2009; Lens et al, 2009; Mesquini et al, 2009; Spolti et al., 2011; Braido et al., 2014). Various considerations and stages in the development of a SADs include: a) the upper and lower limits of the scale, which should correspond, respectively, to the maximum and minimum intensity of the disease observed in the field (ensure an adequate sample); b) if diagrammatic (rather than photographic), the symptoms represented on the SADs should be sufficiently representative of those observed on living plants; c) the number of SADs should be appropriate for the range of severity and to reflect the frequency characteristics of the symptoms; d) measurements of disease severity on the SADs and the unknown test images should be as accurate as possible using image analysis or an alternative method; e) selection of sufficient numbers of test images for the validation process to represent the range and characteristics of the disease; f) clear instructions should be provided to the raters so they can recognize the symptoms, delineate the edges of diseased tissue, and be aware of how to estimate a percentage area (proportionally to represent the diseased part); g) ensure the conditions for assessments are consistent and constant; and h) use appropriate statistical analysis to demonstrate if there is an effect of the SADs improving accuracy and precision. How these factors taken as a whole can vary when interpreted or applied in different studies is unknown. As noted above, a new SOPs exists (del Ponte et al., 2017), but the ramifications of how overall differences in the SOPs between labs in the SAD measurement and validation process have not been explored. Ideally, when two labs measure and validate a SADs, the results should be the same.

The objectives of this study were i) to determine whether the interpretation and application of SOPs for SAD measurement and validation by two labs affects the overall outcome of the process, and ii) to develop and validate a SAD set as an assessment aid for the estimation of the severity of gray mold symptoms on leaves of gerbera.

MATERIALS AND METHODS

Laboratories. The studies were conducted at the Departamento de Agronomia, Universidade Estadual de Maringá (Paraná State, Brazil), designated Lab 1, and at the USDA–ARS-SEFTNRL (Byron, GA, USA), designated Lab 2. As outlined below all preliminary aspects of the study were prepared at Lab 1.

Inoculation of plants and collection of leaves. Gerbera daisy plants (cultivar Revolution Yellow DC, Ball Seeds, Toledo, Paraná State, Brazil) were grown in a compost of pine bark, vermiculite and macro nutrients (MecPlant Agricola, Telemaco Borba, Paraná State, Brazil) in containers under greenhouse conditions with mean temperature of ~27°C, natural photoperiod, and daily watering. The plants were inoculated with a suspension of *Botrytis* conidia prepared from cultures in Petri dishes (90 × 15 mm) grown on potato dextrose agar at 23°C with a 12-hour photoperiod. Conidia were collected by flooding the culture with sterile distilled water and scraping the surface using a glass bar. The conidia concentration was adjusted to 2 × 10^5 per mL using a hemocytometer. The plants were inoculated when they were 37 days old using the suspension of *Botrytis* conidia. Inoculation was by hand held sprayer (Pulverizador Sanremo Boulevard 580 mL, Sanremo, Esteio, Rio Grande do Sol, Brazil), the inoculum sprayed on the leaves to run-off. After inoculation, plants were placed in a humid chamber and held at 90-100% relative humidity for 48 hours. Spray inoculation, as opposed to wounding, was used to emulate natural infection. Plants were returned to the greenhouse, where disease developed under conditions already noted. When plants were 60 days old and 23 days after inoculation, 126 leaves with symptoms of *Botrytis* infection were arbitrarily collected.

The leaves had a range of severity and were photographed individually against a blue background immediately after collection using a digital camera (Sony CyberShot 5.1MP, Tokyo, Japan). For image capture the leaves were illuminated using a 40-W light bulb (Fluorescent Lights, Taschibra 6400K, Encano do Norte, Santa Catarina, Brazil) placed 30 cm over the leaves using a support – images were captured from the same distance overhead to ensure uniform light conditions. All images were captured at Lab 1.

Image analysis. A trained individual measured the severity of *Botrytis* on all 126 leaves at Lab 1 using the image analysis program Quant V1.0.2 (Vale et al., 2001). The percentage diseased area in relation to the total surface area of the leaf was calculated. The minimum and maximum percent severity measured on the 126 images of the leaves were 0.2 and 68.0%, respectively (Fig. 1). The majority of leaves (69%) had severity <20%, demonstrating the need to focus the diagrams at severities of <20%.

Selection of images and measurement of disease on SADs. We specifically wanted to compare laboratories holistically and account for any differences that might occur due to the entirety of different approaches taken by independent groups subsequent to sample collection and identification of specimen leaves for use as SADs. Thus, using a selected sub-sample of 6 leaves representing the range of severity in the greenhouse, a common set of SADs were prepared at Lab 1 based on the results from the image analysis of all 126 leaves collected. The leaves were recolored in Quant V1.0.2 to generate a color SAD set with brown (diseased area) and green (healthy area). Thus, the SAD set was structured to have six diagrams of leaves with upper and lower limits based on the image analysis-measured minimum and maximum disease severity in the sample of 126 leaves as noted in the previous section, and was performed at Lab 1.

Once generated, the resulting six images of the SADs were subject to independent image analysis by a test administrator to measure the diseased area in each leaf diagram using Quant V1.0.2 at Lab 1, and using APS Assess V2.0 (Lamari, 2002) at Lab 2. As noted above, the same SAD set was used at both labs to maintain a common starting point, but independent measurements and approaches taken thereafter to explore the effect of lab on the downstream process of SAD development and validation.

Validation of the SADs. To maintain common images for testing in the two labs, a subset of 30 images from the remaining 120 images on which actual severity had been measured by image analysis were selected at Lab 1 for the rater-validation process (leaves with measured actual values are required for validation). A sample size of 30 is deemed adequate for mean disease severity estimation based on prior studies if taking two estimates per specimen (Chiang et al., 2016b); here we were taking 18 estimates per specimen at each Lab. These 30 images had been independently subject to image analysis by the test administrator at Lab 1 (using Quant

V1.0.2), and were again subject to image analysis by the test administrator at Lab 2 where the diseased area was measured using APS Assess V2.0. The subsequent approach to validate the SADs was intentionally independently selected in each lab. Thus in Lab 1, 18 raters were instructed to estimate the severity of gray mold symptoms on each of the selected subset of 30 images of the diseased leaves using a MS PowerPoint (Microsoft Inc., Redmond, WA) slide presentation, projecting each leaf image at random on a screen using an LCD Epson projector (Model H855A, Seiko Epson Corp., Japan) with evaluation programmed to last 30 seconds per image**.** The raters had a range of experience with disease assessment and familiarity with disease symptoms. At Lab 1, prior to the first assessment, all raters received the same instructions describing the symptoms of the disease and instructions in use of the SAD set. Initially, each rater estimated the severity of gray mold symptoms without the aid of the SAD set. After a 30-min break, each rater again estimated the severity of symptoms on the same 30 leaves, again shown at random but with the aid of the 6-diagram SAD set to guide estimation. In Lab 2, 18 raters were independently but similarly instructed to estimate the severity of gray mold symptoms on each of the selected subset of 30 images of the diseased leaves, but using approximately life-sized images of the leaves on sheets of paper that were randomized (1 per sheet). No time limit was imposed at Lab 2. Similar to Lab 1, the raters had a range of experience with disease assessment and familiarity with disease symptoms. As for Lab 1, all raters in Lab 2 received the same instructions describing the symptoms of the disease and instructions in use of the SAD set. Initially each rater estimated the severity of gray mold symptoms without the aid of the SAD set. After up to a two-week break (minimum 1 day), each rater again estimated the severity of symptoms on the same 30 leaves which were randomized again, but using the six-diagram SAD set as an assessment aid.

Data analysis. The visual estimates of severity of gray mold symptoms on the 30 leaves without and with SADs at Lab 1 and Lab 2 were compared to the actual values measured by image analysis from each Lab 1 and Lab 2, respectively. Lin's concordance correlation (LCC, Lin, 1989; Nita *et al.*, 2003) analysis was used to evaluate the degree to which the estimates fell on the line of concordance (45°, where slope =1, intercept =0). When there is perfect concordance between the estimates and the true values, then the LCC statistics of systematic bias, $\upsilon = 1$, constant bias, $\mu = 0$, overall bias or accuracy, $C_b = 1$, precision, $r = 1$, and agreement, $\rho_c = 1$. Deviation from these values indicates bias, loss of precision and loss of agreement. Analyses were performed in MS Excel following the standard calculations for calculating the LCC statistics (Lin, 1989). The difference in each of these statistics when estimated without using SADs and using SADs was calculated for each rater. An equivalence test (Yi et al., 2008; Yadav et al., 2013; Bardsley and Ngugi, 2013) was used to calculate 95% confidence intervals (CIs) for the difference between the means for υ, μ, C_b, r, ρ_c by 1000 balanced bootstrap samples using the percentile method. The equivalence test assumes groups are different, and was performed independently for each statistic from each lab. If the resulting CIs span zero, there is no significant difference between the means. The equivalence test was performed using SAS V9.4 using PROC SURVEYSELECT and PROC UNIVARIATE (SAS Institute Cary, NC).

In addition to the equivalence test, an analysis of variance (ANOVA) using a generalized linear model (PROC GLIMMIX) was performed to explore fixed effects of SADs and Lab, and the SADs × Lab interaction on each of the dependent variables for υ, μ, C_b, r and ρ_c. In contrast to the equivalence test, an ANOVA tests the null hypothesis (H_0) that there is no difference between groups. A Tukey's means separation was performed to compare the means for the two fixed effects and the interaction ($\alpha = 0.05$).

The inter-rater reliability with and without SADs at each lab was measured using two methods. Firstly, the coefficient of determination (R^2) for each pairwise combination of rater-based estimates without or with SADs was calculated for the data at each lab. The R^2 reflects the proportion of variation explained by the linear relationship (PROC REG), and indicates how closely one measurement predicts the other. The R^2 was calculated for all pairwise combinations in each lab with and without SADs using SAS V9.4. The within lab SAD effect on the R^2 was explored using an equivalence test. The R^2 was also subject to a GLIMMIX analysis as described in the previous paragraph. Secondly, the intra-class correlation coefficient (ICC, ρ) was determined for estimates by raters at each lab with and without SADs. The ICC compares between-subject and within-subject variance and thus accounts for chance correspondence of the variance between the two measurements. The ICC and its confidence limits were calculated step by step in MS Excel using a two-way ANOVA as described by Nita et al. (2003). The 95% CIs were calculated.

The relationship between the change in rater ability based on all LCC statistics (υ, μ, C_b, r, ρ_c) and inter-rater reliability (R^2) for estimates made without SADs and those made using the SADs (with SADs assessment – No SADs assessment) was regressed against the assessment statistics without SADs. Because υ and μ are centered on 1 and 0, respectively, we standardized the values by transforming υ using 1- υ, while μ was converted to absolute values prior to calculating the mean difference between assessments. Linear regression analysis was performed to examine the relationship between the change in the statistics without and with SADs, and the statistic (υ, μ, C_b, r, ρ_c or R^2) without SADs. The regression solution was assessed using the F and P values for the model (significant if $P<0.05$), the R^2, and the coefficient of variation (CV), a unit-less measure of variation, calculated as [(Mean Square Error/Mean) × 100]. Regression was also used to explore the relationships between measurements of the actual values by Lab 1 and Lab 2.

Finally, absolute error (the visual estimate made with or without SADs – actual disease severity) was calculated for all estimates.

RESULTS

Actual values. The SADs consisted of six images (Fig. 2). The measurements of actual values on the SAD images varied between the two labs. The differences were not large, ranging from 0.2 to 5.18%. The measurements of the SADs diseased areas at Lab 1 were consistently lower compared to those at Lab 2. The actual values measured on the 30 'unknown' images for the tests at Lab 1 and Lab 2 also differed (Fig. 3). The relationship indicated moderate to strong agreement ($R^2 = 0.88$). Only one image had an identical measurement. The differences in measured diseased area ranged from 0.22 to 24.29%. Of the thirty measurements at each lab, 18 at Lab 1 had a lower measurement.

Bias, precision and agreement. Each of the 36 raters from the two labs showed a unique profile when estimating severity without or with SADs. Despite instructions, one rater from Lab 1 used the SADs as categories into which the unknowns were binned (data not shown). Based on the test of equivalence, the two labs differed: when the SADs were used by raters at Lab 1, they failed to significantly improve any measure of bias (systematic bias, constant bias or generalized bias), precision or agreement (Table 1). There was no significant effect on location bias, systematic bias, generalized bias, precision or agreement. Overall, the tendency to underestimate severity of *Botrytis* of leaves of Gerbera daisy was greater with SADs. In contrast, the raters at

Lab 2 showed significant reductions in systematic bias, generalized bias, and agreement, but not in constant bias and precision. The mean % change in accuracy of the overall mean estimate of severity also confirmed these trends: the actual mean severity of gray mold on the 30 leaves measured at Lab 1 was 19.43%; without SADs the mean rater estimated severity was 18.69% (underestimate of 0.75%), and with SADs it was 15.47% (underestimate of 3.97%). In contrast, the actual mean severity of gray mold on the 30 leaves measured at Lab 2 was 20.49%; without SADs the mean rater estimated severity was 27.08% (overestimate of 6.59%), and with SADs it was 20.17% (underestimate of 0.32%).

Raters varied in their responses to using SADs. The diversity of rater response to SADs can be ascertained from the gain or loss for each of the statistics defining bias, precision and agreement (Fig. 4A-E). For all statistics (υ, μ, C_b, r, and ρ_c) there were individual raters who responded in unexpected and in extreme ways and as a result are outliers in gain or loss. The phenomenon was true for both Lab 1 and Lab 2. There are outliers among these data, which were included in the analysis. Despite these outliers, the trends for most raters are clear and consistent in these figures. The majority of rater's response to the use of SADs was for small to large gains in each statistic, with similar trends. The extreme rater exceptions caused the regression to behave contrary to the trend in the majority of data points for both systematic bias (Fig. 4A) and constant bias (Fig. 4B), particularly for data from Lab 1. For the majority of raters for each statistic the response confirms that less accurate and less precise raters tended to improve the most when using SADs (Table 2).

The analysis of variance revealed effects of Lab and SADs on the LCC statistics (Table 3). Thus there were significant effects of Lab only for constant bias (F=6.2, P=0.02), with raters from Lab 2 being slightly less biased on average. Overall, there were significant effect of SADs for generalized bias (F=5.8, P=0.02), and agreement (F=6.9, P=0.01). Overall, SADs resulted in less biased estimates that had greater agreement with the actual values. There was no significant interaction effect for any of the LCC statistics.

Inter-rater reliability. Whereas Lab had no discernible effect (Table 3), use of SADs significantly improved inter-rater reliability (F=33.6, P<0.0001). There was a significant Lab × SADs interaction (F=3.9, P=0.05) with both labs showing an improvement in inter-rater reliability with use of SADs although the improvement when using SADs was greater for Lab 1.

These results were borne out by the test of equivalence using all pairwise coefficients of determination for the raters (Table 4). Use of the SADs resulted in improvement in inter-rater reliability by raters at Lab 1 and Lab 2. This was mirrored in improvements in the intra-class correlation coefficient at both labs. It should be noted that the confidence intervals for the ICC do not represent differences between the means based on a hypothesis test, but represent the confidence intervals of each population (no SADs and SADs for each lab).

The overall frequency of the levels of the coefficients of determination for the two labs with and without SADs indicates that the raters at Lab 2 tended to have slightly higher inter-rater reliability values with and without SADs (Fig. 5A). The gain or loss of inter-rater reliability showed that most pairwise comparisons of raters showed improved inter-rater reliability with use of SADs at both Lab 1 and Lab 2. However, as with agreement statistics, there were raters at both Labs who did not show typical gains in inter-rater reliability (Fig. 5B; Table 2).

Absolute error. Raters at Lab 1 tended to underestimate disease when not using SADs, but at Lab 2 the tendency was for raters to overestimate disease, particularly at low disease severities (<40%) (Fig. 6). Using SADs reduced the absolute error of raters at both labs. Estimates of zero (or almost zero) disease acted as a barrier to more extreme underestimates at both labs, but even with SADs individual disease severities were underestimated up to 60.0% and overestimated up to 40.0% at Lab 1, and underestimated up to 42.5% and overestimated up to 64.0 % at Lab 1, respectively.

Discussion

The results of our study demonstrate that the SAD experiments are not necessarily reproducible among different laboratories, even when the same SADs and test images are used for disease assessment. Although this study did not explore the reasons for the lack of reproducibility between labs, it forms the basis for exploring sources of variation in future studies. Our study was observational in that we observed the effect of independently developed SAD measurement and validation processes on the outcome of using SADs. Thus, our study relates directly to an ongoing discussion about reproducibility of research in science in general (Baker, 2016) and specifically within the microbiology and plant pathology community (Schloss, 2018; https://openplantpathology.org/tags/reproducibility/).

Different approaches have been used to develop and validate SADs (del Ponte et al., 2017). The image analysis process of measuring diseased area on the SADs and on the test images is a potential source of some error. Image analysis systems may rely on different algorithms and is inevitably prone to error as two individuals may not delineate the disease the same way; thus pixels may be included in the healthy or diseased grouping depending at what point in the color grade the differentiation is made by the individual performing the measurement. Indeed, due to these subjectivities even the same individual measuring actual severity using image analysis may have different results performing the measurement a second time (Bock et al., 2008). Accurate segmentation of diseased areas is more challenging for symptom with unclear boundaries, perhaps with a gradation of chlorosis from necrotic to healthy. Symptoms of gray mold on gerbera has these characteristics that may lend themselves to error due to subjectivity of delineation. No formal analysis has yet been done to determine whether symptoms with poorly defined boundaries are more difficult to estimate severity accurately. But SADs do exist for diseases where chlorosis or other factors make symptom delineation a little more challenging (Spolti et al., 2011; Correa et al., 2017; Domiciano et al., 2014), and for those diseases with relatively clear-cut symptoms (Lima et al., 2011; Schwanck and Del Ponte, 2014; González-Domínguez et al., 2014). Thus, the agreement (ρ_c) with and without SADs for estimates of severity of Phomopsis leaf blight of eggplant that has variable chlorosis associated with lesions (similar to *Botrytis* on *Gerbera*) was 0.73 and 0.92, respectively (Correa et al., 2017), while a pathosystem with a very clear-cut symptom like brown spot of rice was 0.53 and 0.87, respectively (Schwanck and Del Ponte, 2014). Thus, symptoms that are poorly defined do not necessarily preclude a significant improvement and accuracy of estimation at least equivalent to those with more clearly defined symptoms. Much research remains to be done to understand these factors in SAD development and validation.

In Lab 1 the images were presented as a MS PowerPoint presentation with timed, 30 sec viewings for rater estimation, while in Lab 2, the images were printed on paper and there was no time limit for the rater to estimate severity. The raters selected can also impact the overall outcome of the study. Raters are diverse in ability (Bock et al., 2009) and, although a minimum of 15 raters is recommended (del Ponte et al., 2017), the characteristics of the raters will likely impact the outcome of the study too. Raters in both labs showed a wide range of capability and response to SADs. Also, instruction provided to raters, regardless of expertise, is critical (Ngugi and Bardsley, 2013) and how instruction is provided by a test administrator can vary between labs. It is important to ensure that raters know how to recognize symptoms of the disease, and how to delineate healthy tissue from the diseased tissue. Raters need to understand the concept and process of estimating a proportion based on the continuous percentage ratio scale. Furthermore, raters must clearly understand the SADs are an aid to help with the process of estimation by interpolation, and are not to be used as categories into which the disease estimates are binned. One rater in Lab 1 appeared not to understand this point. Del Ponte et al. (2017) provided a list of SOPs for SADs. It may be that the SOPs should be amended to further refine and standardize SAD approaches. However, before additions to the SOPs are proposed, research must be conducted to identify methods that result in accurate, and highly reproducible disease assessment data. .

We used a robust number of raters (18 at each lab) and unknown images (30), yet in Lab 1 there was only some numeric evidence of improvements in accuracy, agreement or precision of rater estimates, while in Lab 2, there was a significant improvement in accuracy and agreement of the estimates when using SADs. Both labs demonstrated significant gains in inter-rater reliability when using the SADs, confirming that the SADs increased the uniformity of rater estimates, in of itself a very valuable improvement where multiple raters might be assessing disease on different samples in a study. Interestingly, there was no significant difference in the precision of estimates between the two labs, although both did show numeric improvements. The results indicating improvements in agreement and reliability reflect those reported for many other SADs (Sposito et al., 2004; Barbosa et al., 2006; Barguil et al, 2008; Sussel et al., 2009; Lens et al, 2009; Mesquini et al, 2009; Spolti et al., 2011; Braido et al., 2014). Why the tendency of raters in Lab 1 was to underestimate disease severity, and those in Lab 2 to overestimate disease severity is not understood. Use of SADs generally reduced the tendency to overestimate.

Bias is an important source of error in disease severity estimation and can affect the outcome of hypothesis testing (Chiang et al., 2016a), so it is important to understand and minimize. Raters from different areas may have small, inherent differences in characteristics of estimation. Although not considered here, differences in individuals' personality types might also affect the accuracy of estimates.

Overall our study reaffirms that the use of SADs is a useful method to improve accuracy and reliability of disease assessment by at least some raters – although most often the gain in a particular statistic as a result of using SADs is greatest for those least capable raters. We observed this phenomenon in the current study, as has been observed and commented on previously (Yadav et al., 2013; Braido et al., 2014). Thus, it would be advantageous to use these newly developed SADs in future studies where more accurate and reliable estimates of severity of *Botrytis* on Gerbera are sought. Furthermore, these SADs to aid estimation of severity of symptoms of gray mold on leaves of Gerbera has additional utility too. It may also be useful for other diseases of Gerbera with similar symptoms. A recently described disease of Gerbera in Brazil is caused by *Pseudomonas cichorii* (Marques et al., 2016) and has symptoms that are reminiscent of gray mold infection. The SADs described here may be useful as an aid to estimate severity of symptoms caused by *P. cichorii*.

To conclude, this study provides evidence that labs may vary in the outcome of the SAD development and validation process; in one lab they may result in statistically different improvements in measures of accuracy, yet not in another. This is useful to know. In this case both showed a significant increase in reliability using the SADs. Various factors in the process of SAD development and validation may affect outcome including components unrelated to the raters involved in the test, who themselves are a source of potential discrepancy. However, given suitable sample size, a test to ascertain SADs utility should provide the same outcome regardless of lab. These results suggest that we need more rigorous SOPs for developing and using SADs.

Acknowledgements

ACSM received a grant from CNPq and LCG received a grant from CAPES. KRFSE and WMCN received CNPQ productivity grants. We thank the students of the Graduate Program in Agronomy (PGA) at UEM who helped with evaluation of the severity of the disease. CHB is supported by the USDA-ARS National Programs through CRIS project 6042-21220-012-00. We appreciate the technical and logistical support of Wanda Evans, and temporary hire Susan Burrell, and assistance from several student hires other colleagues in assessing disease severity at the ARS location.

This article reports the results of research only. Mention of a trademark or proprietary product is solely for the purpose of providing specific information and does not constitute a guarantee or warranty of the product by the U.S. Department of Agriculture and does not imply its approval to the exclusion of other products that may also be suitable.

Literature Cited

Andrade, P.F.S. 2016. Análise da conjuntura agropecuária safra 2015/16 – Floricultura. Secretaria da Agricultura e do Abastecimento - Departamento de Economia Rural (SEAB-DERAL). 19 p.

Anonymous. 2009. Floriculture crops: 2008 summary. USDA, NASS ISSN: 1949-0917, p 59. Available at https://www.nass.usda.gov/Publications/Todays_Reports/reports/floran09.pdf. Accessed 17 Feb 2019.

Baker, M. 2016. Is there a reproducibility crisis? Nature 533: 452-454.

Barbosa, M. A. G., Michereff, S. J. and Mora-Aguilera, G. 2006. Elaboração e validação de escala diagramática para avaliação da severidade da ferrugem branca do crisântemo Summa Phytopathologica 32: 57-62.

Bardsley, S.J. and Ngugi, H.K. 2013. Reliability and accuracy of visual methods to quantify severity of foliar bacterial spot symptoms on peach and nectarine. Plant Pathol. 62: 460–474.

Barguil, B.M., Albert, I. C. L. and Oliveira, S. M. A. de. 2008. Escala diagramática para avaliação da severidade da antracnose em bastão do imperador. Ciência Rural 38: 807-810.

Bock, C.H., Parker, P.E., Cook, A.Z., and Gottwald, T.R. 2008. Visual assessment and the use of image analysis for assessing different symptoms of citrus canker on grapefruit leaves. Plant Dis. 92: 530–541.

Bock, C.H., Parker, P.E., Cook, A.Z., Riley, T. and Gottwald, T.R. 2009. Comparison of assessment of citrus canker foliar symptoms by experienced and inexperienced raters. Plant Dis. 93: 412-424.

Bock, C.H., Poole, G., Parker, P.E., and Gottwald, T.R. 2010. Plant disease severity estimated visually, by digital photography and image analysis, and by hyperspectral imaging. Crit. Rev. Plant Sci. 29: 59-107.

Bock, C.H., Chiang, K.-S. and Del Ponte, E.M. 2016. Accuracy of plant specimen disease severity estimates: concepts, history, methods, ramifications and challenges for the future. CAB Reviews: Perspectives in Agriculture, Veterinary Science, Nutrition and Natural Resources 11, 039: 1-13.

Braido, R., Goncalves-Zuliani, A.M.O., Janeiro, V., Carvalho, S.A., Belasque Junior, J., Bock, C.H. and Nunes, W.M.C. 2014. Development and validation of standard area diagrams as assessment aids for estimating the severity of citrus canker on unripe oranges. Plant Dis. 98: 1543-1550.

Brisco-McCann E.I. and Hausbeck M.K. 2016. Diseases of *Gerbera*. In: McGovern R., Elmer W. (eds) Handbook of Florists' Crops Diseases. Handbook of Plant Disease Management. Springer, Cham.

Chiang, K.-S., Bock, C.H., El Jarroudi, M., Delfosse, P., Lee, I.H., and Liu, S.-H. 2016a. The effects of rater bias and assessment method used to estimate disease severity on hypothesis testing. Plant Pathol. 65: 523–535.

Chiang, K.-S., Bock, C.H., Lee, I.-H., El Jarroudi, E. 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.

Correia, K.C., de Queiroz, J.V.J., Martins, R.B., Nicoli, A., Del Ponte, E.M. and Michereff, S.J. 2017. Development and evaluation of a standard area diagram set for the severity of phomopsis leaf blight on eggplant. Eur J Plant Pathol 149: 269–276 (2017) doi:10.1007/s10658-017-1184-y

Daughtrey, M.L., Wick, R.L. and Peterson, J.L. 2000. Botrytis blight of flowering potted plants. Plant Health Progress. Doi:10.1094/PHP-2000-0605-01-HM.

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 98: 1543-1550.

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.

Domiciano, G. P., Duarte, H. S. S., Moreira, E. N., and Rodrigues, F. A. 2014. Development and validation of a set of standard area diagrams to aid in estimation of spot blotch severity on wheat leaves. Plant Pathol. 63:922-928. https://doi.org/10.1111/ppa.12150

Ferronato, M. de Lurdes; Lima-Neto, V. de Costa and Tomaz, R. 2008. Gerbera diseases in the state of Paraná, Brazil. Scientia Agraria, Curitiba 9: 481-489.

Fu, Y., Chen, M., van Tuyl, J., Visser, R.G.F. and Arens, P. 2015. The use of a candidate gene approach to arrive at botrytis resistance in gerbera. ActaHort. 1087: 461-466.

González-Domínguez, E., Martins, R.B., Del Ponte, E.M., Michereff, S.J., Garcia-Jimenez, J. and Armengol, J. 2014. Development and validation of a standard area diagram set to aid assessment of severity of loquat scab on fruit. Eur J Plant Pathol 139: 419–428. doi:10.1007/s10658-014-0400-2

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

Lenz, G., Costa, I. D. da, Balardin, R. S., Stefanelo, M. S., Marques, L. N., Arrué, A. 2009. Escala diagramática para avaliação de severidade de mancha-de-septoria em girassol. Ciência Rural 39: 2527-2530.

Lima, G., Assunção, I. P., Martins, R. B., Santos, H. V., and Michereff, S. M. 2011. Development and validation of a standard area diagram set for assessment of Alternaria spot on the cladodes of the prickly pear cactus. J. Plant Pathol. 93:691-695.

Lin, L.I. 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics 45: 255–68.

Madden, L.V., Hughes, G. and van den Bosch, F. 2007. The Study of Plant Disease Epidemics. APS Press, St. Paul, MN.

Marques, E., Borges, R.C.F. and Uesugi, C.H. 2016. Identification and pathogenicity of *Pseudomonas cichorii* associated with a bacterial blight of gerbera in the Federal District. *Horticultura Brasileira* 34: 244-248. DOI - http://dx.doi.org/10.1590/S0102-053620160000200015

Mesquini, R.M., Schwan-Estrada, K.R.F., Godoy, C.V., Vieira, R.A., Zarate, N.A.H., Vieira, M. do C. 2009. Escala diagramática para a quantificação de *Septoria apiicola* e *Cercospora arracacina* em mandioquinha-salsa. Tropical Plant Pathology 34: 250-255.

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

Nutter, F.W. Jr., Gleason, M.L., Jenco, J.H., and Christians, N.L. 1993. Accuracy, intra-rater repeatability, and inter-rater reliability of disease assessment systems. Phytopathology 83: 806–812

Parker, S.R., Shaw, M.W. and Royle, D.J. 1995. Reliable measurement of disease severity. Aspects of Applied Biology 43: 205–214.

Pethybridge, S.J., and Nelson, S.C. 2018. Estimate, a new ipad application for assessment of plant disease severity using photographic standard area diagrams. Plant Disease 102: 276-281.

Schloss, P.D. 2018. Identifying and overcoming threats to reproducibility, replicability, robustness, and generalizability in microbiome research. mBio: 9 (3) e00525-18. DOI: 10.1128/mBio.00525-18.

Spolti, P., Schneider, L., Sanhueza, R.M.V., Batzer, J.C., Gleason, M.L. and Medeiros Del Ponte, E. 2011. Improving sooty blotch and flyspeck severity estimation on apple fruit with the aid of standard area diagrams. European Journal of Plant Pathology 129: 21–9.

Spósito, M. B., Amorim, L., Belasque, J., Jr., Bassanezi, R. B., and Aquino, R. 2004. Elaboração e validação de escala diagramática para avaliação da severidade da mancha preta em frutos cítricos. Fitopatologia Brasileira 29: 81–85.

Sussel, A.A.B., Pozza, E. A. and Castro, H. A. 2009. Elaboração e validação de escala diagramática para avaliação da severidade do mofo cinzento em mamoneira. Tropical Plant Pathology, vol.34, no.3, p. 186-191.

Töfoli, J.G., Ferrari, J.T., Domingues, R.J. and Nogueira, E.M.C. Mofo cinzento em plantas oleráceas, frutíferas e ornamentais. 2011. Artigo em Hypertexto. Disponível em: http://www.infobibos.com/Artigos/2011_2/MofoCinzento/index.htm. Acessed on 21/11/2017.

Vale, F.X.R., Fernandes Filho, E.I. and Liberato J.R. 2003. QUANT – A software for plant disease severity assessment. 8th International Congress of Plant Pathology. Christchurch, New Zealand, 2-7. Pp 105, abstract 8.18.

Yadav, N.V.S., de Vos, S.M., Bock, C.H. and Wood, B.W. 2013. Development and validation of standard area diagrams to aide assessment of pecan scab symptoms on pecan fruit. Plant Pathology 62: 325-335.

Yi Q, Wang PP, He Y, 2008. Reliability analysis for continuous measurements: Equivalence test for agreement. Statistics in Medicine 27: 2816-2815.

Table 1. Mean concordance statistics (Lin's concordance correlation, LCC - bias, precision and agreement) with bootstrap analysis of the differences between means for two groups (Lab 1 and Lab 2) of 18 raters estimates of severity of symptoms of gray mold on a set of 30 images of leaves of *Gerbera jamesonii* without and with a standard area diagram set (SADs) assessment aid.

Lab	**LCC statistic**	**Mean**		**Mean diff**[a]	**95% CIs**[b] **(upper and lower)**
		No SAD	**SAD set**		
1	υ[c]	0.948	0.926	0.048	-0.033 to 0.158[h]
	μ[d]	-0.264	-0.370	0.096	-0.117 to 0.389
	C_b[e]	0.856	0.891	0.037	-0.038 to 0.139
	r[f]	0.825	0.857	0.032	-0.006 to 0.080
	ρ_c[g]	0.736	0.787	0.052	-0.015 to 0.143
2	υ[c]	1.138	1.052	**0.092**	**0.008 to 0.186**
	μ[d]	0.288	0.022	0.096	-0.117 to 0.389
	C_b[e]	0.860	0.967	**0.107**	**0.046 to 0.175**
	r[f]	0.853	0.861	0.008	-0.048 to 0.060
	ρ_c[g]	0.744	0.833	**0.089**	**0.033 to 0.154**

[a] Mean of the difference between each rating.

[b] Confidence intervals (CIs) were based on 1000 bootstrap samples. If the CIs embrace zero, the difference is not significant ($\alpha = 0.05$).

[c] Systematic bias, or scale shift (υ, 1 = no bias relative to the concordance line).

[d] Constant bias, or height shift (μ, 0 = no bias relative to the concordance line).

[e] Generalized bias (C_b) measures how far the best-fit line deviates from the line of concordance.

[f] The correlation coefficient (r) measures precision.

[g] Lin's Concordance Correlation Coefficient (ρ_c) combines both measures of precision (r) and accuracy (C_b) to measure the degree of agreement with the true value.

[h] Bold text indicates a significant difference.

Table 2. The regression solutions for the relationship between bias, precision, agreement and inter-rater reliability without the use of standard area diagrams (SADs) and the difference (no SADs – SADs) for the two groups (Lab 1 and Lab 2) of 18 raters estimating severity of symptoms of gray mold on a set of 30 images of leaves of *Gerbera jamesonii*. See Fig. 4.

LCC statistic	**Lab**	**Intercept**	**Slope**	**F-value (P-value)**	**CV**[a]	**R^2**[b]
υ[c]	Lab 1	0.19	-0.15	1.1 (0.3)	470.4	0.07
	Lab 2	-0.53	0.55	18.0 (0.0006)	156.0	0.53
μ[d]	Lab 1	0.01	-0.30	5.6 (0.03)	560.3	0.26
	Lab 2	0.11	0.60	51.1 (<0.0001)	65.5	0.76
C_b[e]	Lab 1	0.46	-0.49	14.1 (0.002)	427.6	0.47
	Lab 2	0.94	-1.00	458.5 (<0.0001)	26.6	0.97
r[f]	Lab 1	0.26	-0.28	3.0 (0.1)	276.3	0.16
	Lab 2	0.29	-0.33	1.4 (0.3)	1426.5	0.08
ρ_c[g]	Lab 1	0.38	-0.44	10.2 (0.006)	288.0	0.39
	Lab 2	0.45	-0.49	16.1 (0.001)	111.1	0.50
R^2	Lab 1	0.28	-0.32	33.9 (<0.0001)	173.0	0.18
	Lab 2	0.34	-0.47	24.9 (<0.0001)	481.6	0.14

[a] The coefficient of variation (*CV*) is a unit-less measure of variation, and is calculated as [(Mean Square Error/Mean) x 100].

[b] The coefficient of determination (R^2) is the proportion of the variation explained by the association between two sets of measurements.

[c] Systematic bias (scale or slope shift, υ, 1 = no bias relative to the concordance line) can be less than or greater than 1 so it was necessary to obtain standardized (as 1-υ) absolute data prior to calculating the mean difference.

[d] Constant bias (location or height shift, μ, 0 = no bias relative to the concordance line) can be less than or greater than 0, so it was necessary to obtain absolute data prior to calculating the mean difference.

[e] Generalized bias (C_b) measures how far the best-fit line deviates from 45° and is thus a measure of accuracy.

[f] The correlation coefficient (r) measures precision.

[g] Lin's Concordance Correlation Coefficient (ρ_c) combines both measures of precision (r) and accuracy (C_b) to measure the degree of agreement with the true value.

Table 3. General linear mixed model analysis and lsmeans separation of measures of accuracy, precision and agreement for two groups (Lab 1 and Lab 2) of 18 raters estimates of severity of symptoms of gray mold on a set of 30 images of leaves of *Gerbera jamesonii* without and with a standard area diagram set (SADs) assessment aid. For each statitic, numbers in comparison groups (‘Lab’, ‘SADs’ and ‘Interaction (Lab × SADs)’) followed by different letters are significantly different (Tukey’s HSD, α = 0.05).

Statistic	**Main effects**				**Interaction (Lab × SADs)**			
	Lab		**SADs**		**Lab 1**		**Lab 2**	
	1	**2**	**No SADs**	**SADs**	**No SADs**	**SADs**	**No SADs**	**SADs**
υ[a]	0.937 a	1.095 a	1.043 a	0.989 a	0.948 a	0.926 a	1.138 a	1.052 a
F (P)[g]	3.9 (0.06)		1.7 (0.2)		0.6 (0.4)			
μ[b]	-0.317 b	0.155 a	0.012 a	-0.174 a	-0.264 ab	-0.370 b	0.288 a	0.022 ab
F (P)	6.2 (0.02)		3.8 (0.06)		0.7 (0.4)			
C_b[c]	0.874 a	0.914 a	0.858 b	0.929 a	0.856 a	0.891 a	0.861 a	0.967 a
F (P)	0.5 (0.5)		5.8 (0.02)		1.5 (0.2)			
r[d]	0.841 a	0.857 a	0.839 a	0.859 a	0.825 a	0.857 a	0.853 a	0.861 a
F (P)	0.2 (0.7)		1.2 (0.3)		0.4 (0.5)			
ρ_c[e]	0.762 a	0.789 a	0.740 b	0.810 a	0.736 a	0.787 a	0.744 a	0.833 a
F (P)	0.2 (0.7)		6.9 (0.01)		0.5 (0.5)			
R^2[f]	0.622 a	0.661 a	0.608 b	0.675 a	0.577 c	0.667 ab	0.639 bc	0.683 a
F (P)	3.2 (0.07)		33.6 (<0.0001)		3.9 (0.05)			

[a] Systematic bias (υ, 1 = no bias relative to the concordance line).

[b] Constant bias (μ, 0 = no bias relative to the concordance line).

[c] Generalized bias (C_b) measures how far the best-fit line deviates from 45° (Madden *et al.*, 2007).

[d] The correlation coefficient (*r*) measures precision.

[e] Lin’s Concordance Correlation Coefficient (ρ_c) combines both measures of precision (*r*) and generalized bias (C_b) to measure accuracy.

[f] R^2= the coefficient of determination, is a quantitative measure of inter-rater reliability - the degree to which the X-data explain the Y-data.

[g] *F*-value and P-values indicate a significant effect where P<0.05.

Table 4. The inter-rater reliability for two groups (Lab 1 and Lab 2) of 18 raters estimating severity of symptoms of gray mold on a set of 30 images of leaves of *Gerbera jamesonii* without and with a standard area diagram set (SADs) assessment aid. Inter-rater reliability was measured using either the coefficient of determination (R^2)[a] or the intra-correlation coefficient (ρ)[b].

Lab	**Statistic**	**Variable**	**Value**	**Mean diff**[c]	**95% CIs**
1	Coefficient of determination (R^2)	No SADs SADs	0.578 0.667	**0.089**	**0.062 to 0.116**[d]
	Intra-class correlation coefficient (ICC, ρ)	No SADs SADs	0.575 0.730	0.155	0.451 to 0.705 0.620 to 0.825
2	Coefficient of determination (R^2)	No SADs SADs	0.639 0.683	**0.043**	**0.009 to 0.079**
	Intra-class correlation coefficient (ICC, ρ)	No SADs SADs	0.575 0.757	0.182	0.452 to 0.706 0.651 to 0.844

[a] The coefficient of determination (R^2) is the proportion of the variation explained by the association between two sets of measurements.

[b] The ICC (ρ) compares the between-subject variance with the within-subject variance and is the relative amount of variation from the combined mean of the two test sessions explained by differences between the subjects.

[c] Mean of the difference between each rating (i.e., without and with SADs).

[d] [b] Confidence intervals (CIs) were based on 1000 bootstrap samples. If the CIs embrace zero, the difference is not significant ($\alpha = 0.05$). Bold text indicates a significant difference.

[e] The intra-class correlation coefficient and confidence intervals (CIs) were calculated in MS Excel©.

Fig. 1. The frequency of severity (percentage area diseased) of symptoms caused by infection with *Botrytis cinerea* on 126 diseased leaves of *Gerbera jamesonii*. Severity measured using image analysis program Quant V1.0.2 (Vale et al., 2001).

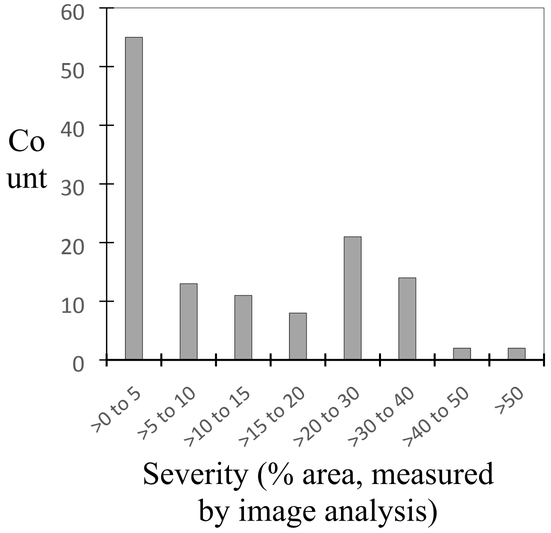
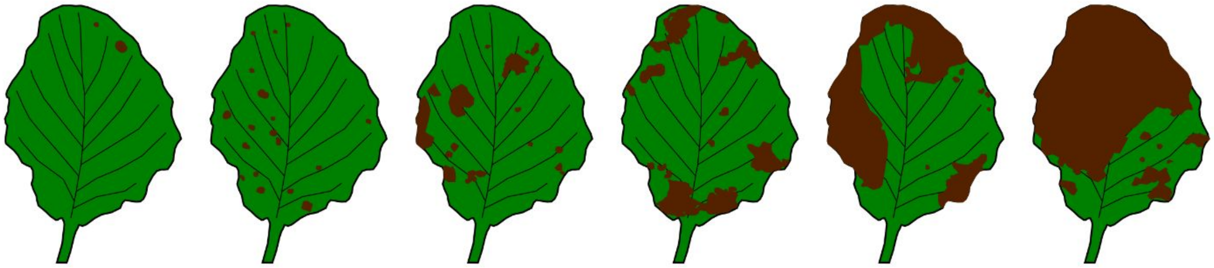
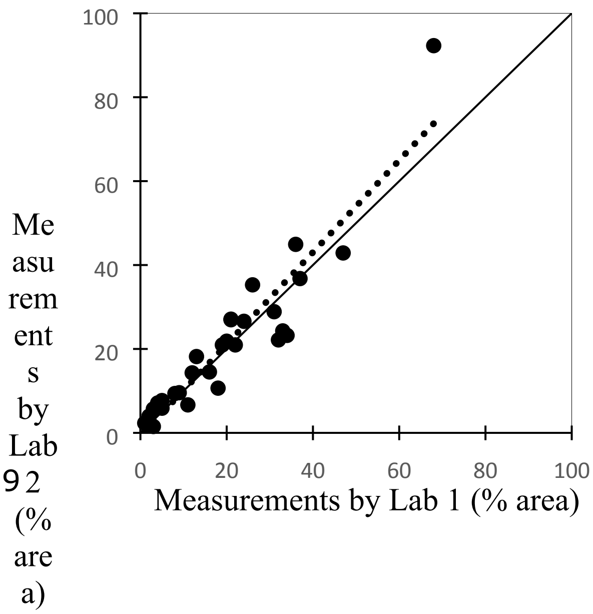


Fig. 2. Standard area diagrams developed and independently measured for diseased area using image analysis by two the administrator of the test for two groups at Lab 1 and Lab 2, respectively. The test groups comprised 18 raters who estimated severity of symptoms of *Botrytis cinerea* on a set of 30 images of leaves of *Gerbera jamesonii* without and with a standard area diagram set (SADs).



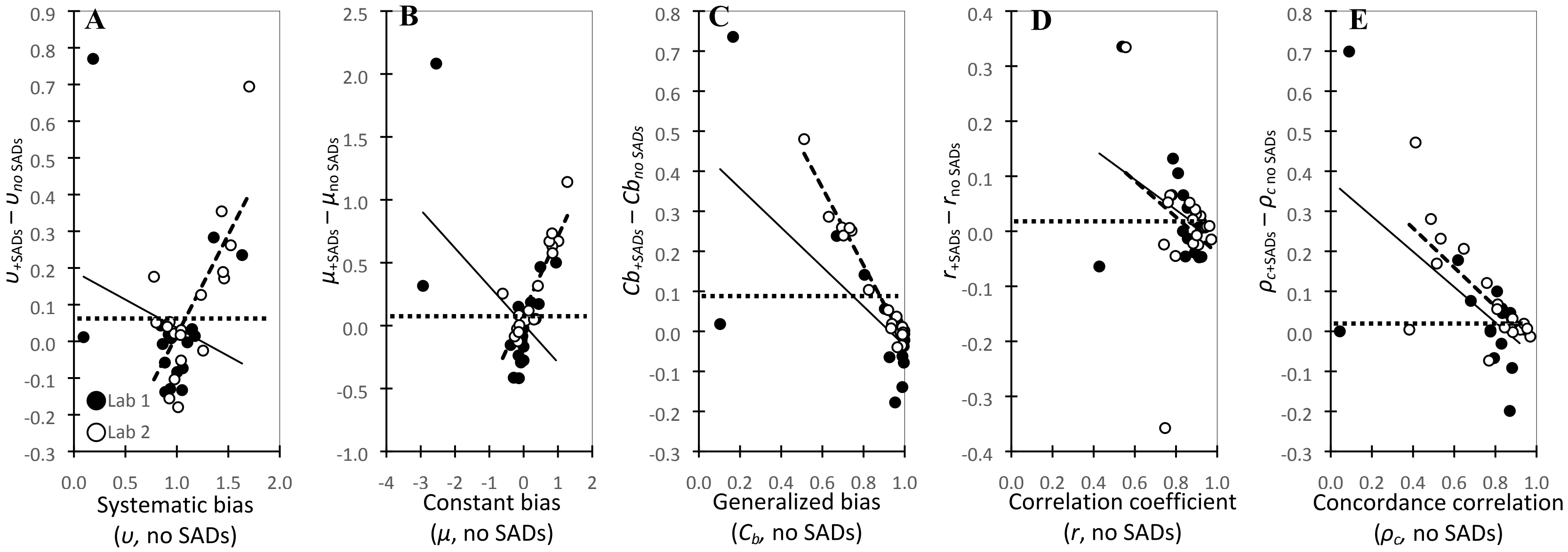
Lab 1	0.2	1.0	3.0	11.0	31.0	64.0
Lab 2	0.4	1.8	7.4	15.1	36.2	65.2

Fig. 3. The relationship between measurements of actual values of severity of symptoms of *Botrytis cinerea* on a set of 30 images of leaves of *Gerbera jamesonii* as made by two administrators of two test groups (Lab 1 and Lab 2) of 18 raters who estimated the severity on the images without and with the use of a standard area diagram set. Solid line is the line of concordance; dashed line is the line fit to the data (regression solution: Lab 2 = Lab 1*1.096 - 0.819 [F=197.7 (P<0.0001), R^2 = 0.88]).



10

Fig. 4. The relationship between bias, precision and agreement without the use of standard area diagrams (SADs) assessment aides and the difference (+SADs – no SADs) demonstrating raters with the least good scores most often benefitted the most for all variables. **A**) Systematic bias, **B**) Constant bias, **C**) Generalized bias, **D**) Correlation coefficient, and **E**) Lin's concordance correlation coefficient. Disease was assessed on a set of thirty images of symptoms of *Botrytis cinerea* on leaves of *Gerbera jamesonii* by 18 raters in two different labs (Lab 1 and Lab 2). Solid line is fitted to data from Lab 1, the dashed line to data from Lab 2. Raters above the horizontal dotted line improved in score relative to the first rating; below the dotted line, raters ability declined compared to the first rating. Regression solutions are presented in Table 2.



13

Fig. 5. The A) frequency of the inter-rater reliability of two groups of 18 raters in different labs (Lab 1 and Lab 2) who assessed thirty images of leaves of *Gerbera jamesonii* with symptoms of *Botrytis cinerea* measured by the coefficient of determination (R^2) without and with use of a standard area diagram set (SAD), and B) relationship between the gain or loss in inter-rater reliability by the two groups when using the standard area diagrams (SADs, the difference (+SADs – no SADs)). Raters above the horizontal dashed line improved in score relative to the first rating; below the dashed line, raters ability declined compared to the first rating.

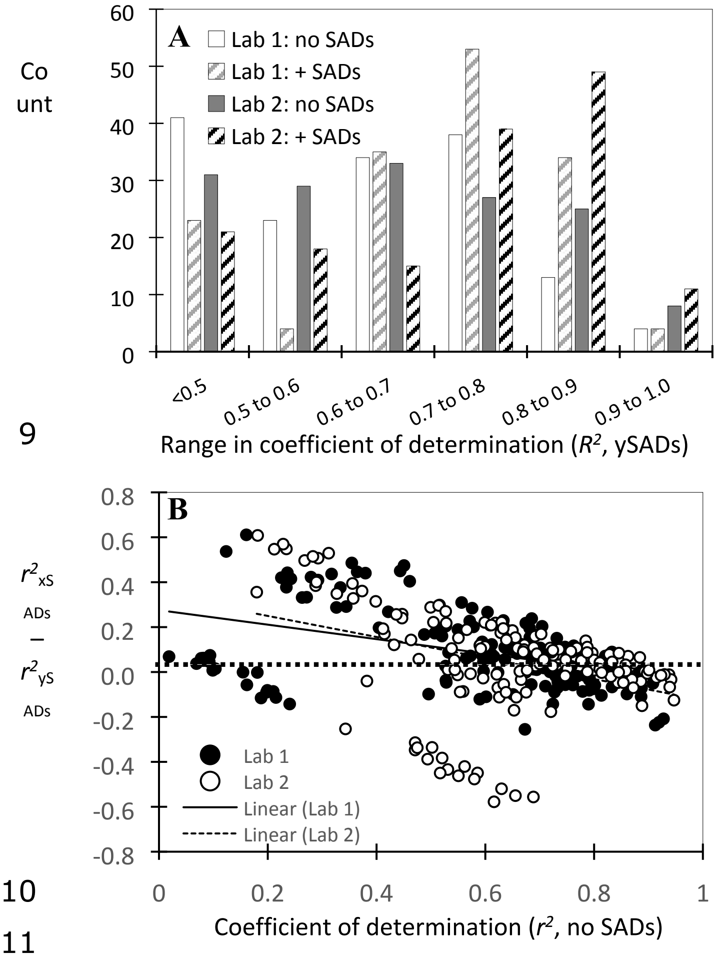
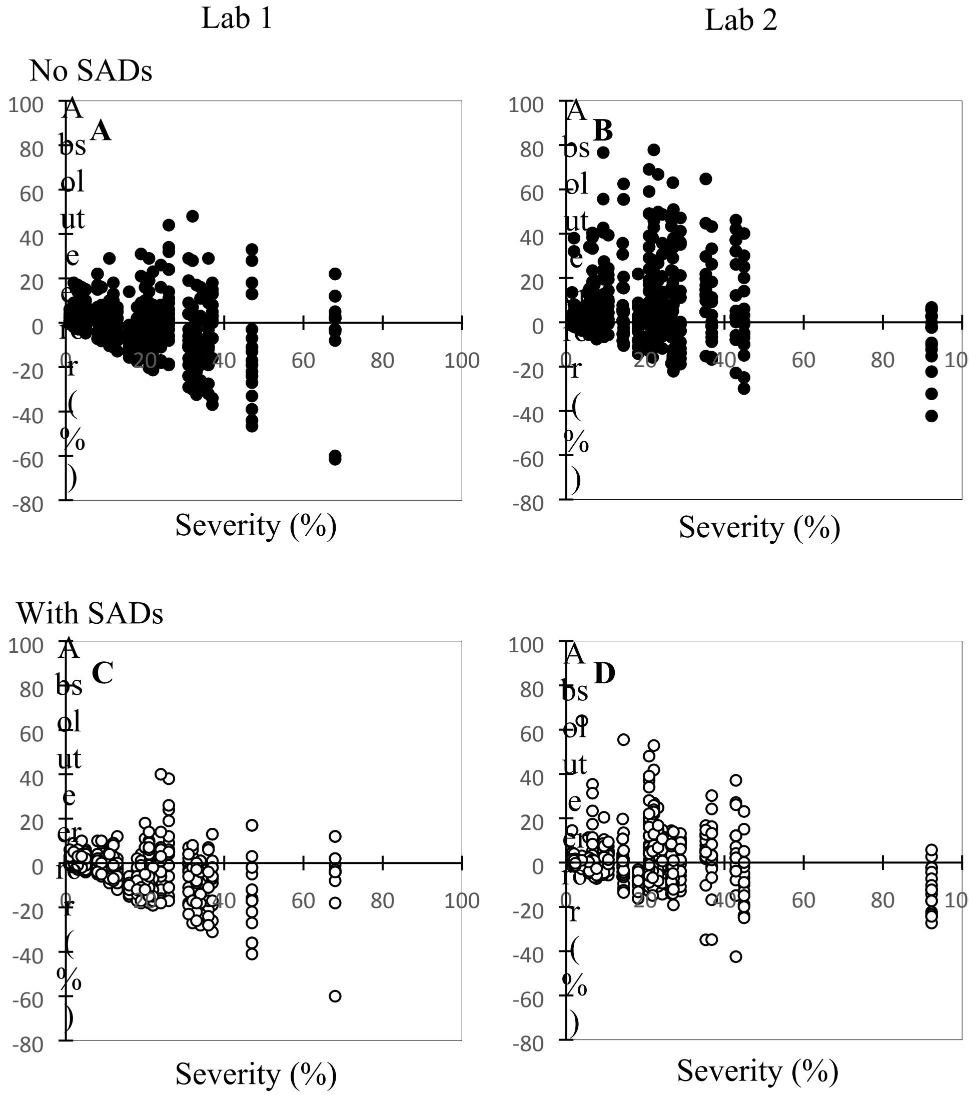


Fig. 6. The absolute error (estimate minus true disease) of estimates of severity of symptoms of *Botrytis cinerea* on 30 images of leaves of *Gerbera jamesoni* by two groups (Lab 1 and Lab 2) of 18 raters without use of standard area diagram sets (No SADs) as assessment aides (**A, B**), or using a SADs (**C, D**).


Lab 1
Lab 2
No SADs
With SADs
A
B
C
D
Absolute error (%)
Severity (%)