



Plant disease severity estimated visually: a century of research, best practices, and opportunities for improving methods and practices to maximize accuracy

Clive H. Bock¹ · Kuo-Szu Chiang² · Emerson M. Del Ponte³

Received: 10 March 2021 / Accepted: 12 May 2021 / Published online: 22 June 2021

© This is a U.S. government work and not under copyright protection in the U.S.; foreign copyright protection may apply 2021

Abstract

Plant disease quantification, mainly the intensity of disease symptoms on individual units (severity), is the basis for a plethora of research and applied purposes in plant pathology and related disciplines. These include evaluating treatment effect, monitoring epidemics, understanding yield loss, and phenotyping for host resistance. Although sensor technology has been available to measure disease severity using the visible spectrum or other spectral range imaging, it is visual sensing and perception that still dominates, especially in field research. Awareness of the importance of accuracy of visual estimates of severity began in 1892, when Cobb developed a set of diagrams as an aid to guide estimates of rust severity in wheat. Since that time, various approaches, some of them based on principles of psychophysics, have provided a foundation to understand sources of error during the estimation process as well as to develop different disease scales and disease-specific illustrations indicating the diseased area on specimens, similar to that developed by Cobb, and known as standard area diagrams (SADs). Several rater-related (experience, inherent ability, training) and technology-related (instruction, scales, and SADs) characteristics have been shown to affect accuracy. This review provides a historical perspective of visual severity assessment, accounting for concepts, tools, changing paradigms, and methods to maximize accuracy of estimates. A list of best-operating practices in plant disease quantification and future research on the topic is presented based on the current knowledge.

Keywords Plant disease assessment · Reliability · Repeatability · Reproducibility · Inter-rater reliability · Intra-rater reliability · Phytopathometry

Introduction

Quantification of plant disease intensity (amount of disease in a population, Nutter Jr et al. 1991) is required for many different purposes including monitoring epidemics in experiments or surveys, understanding yield loss, comparing phenotypes for disease resistance, and evaluating effects of

treatments (chemical, biological, agronomic, or environmental factors) on disease (James 1974; Kranz 1988; Cooke 2006; Madden et al. 2007; Bock et al. 2010a). Throughout all of these applications, visual estimates of disease are used to draw conclusions and/or take actions—and thus they should be as accurate as possible given available resource and purpose—where accuracy is operationally defined as the closeness of the visual estimate to the actual value Nutter Jr et al. (1991). The term agreement can be considered synonymous with accuracy where actual values are concerned (Madden et al. 2007). During the research process, incorrect quantification could result in a type II error (failure to reject the null hypothesis when it is false) when comparing treatments in any experiment situation, which will have ramifications for the conclusions drawn from those experiments. Decisions based on such conclusions could result in wasted resources, increased disease, yield loss, and ultimately reduced profit. Thus, accurate disease intensity estimates would appear to be vital.

✉ Clive H. Bock
clive.bock@usda.gov

¹ United States Department of Agriculture – Agriculture Research Service – Southeastern Fruit and Tree Nut Research Station, Byron, GA 31008, USA

² Division of Biometrics, Department of Agronomy, National Chung Hsing University, Taichung, Taiwan

³ Departamento de Fitopatologia, Universidade Federal de Viçosa, Viçosa, MG 36570-000, Brazil

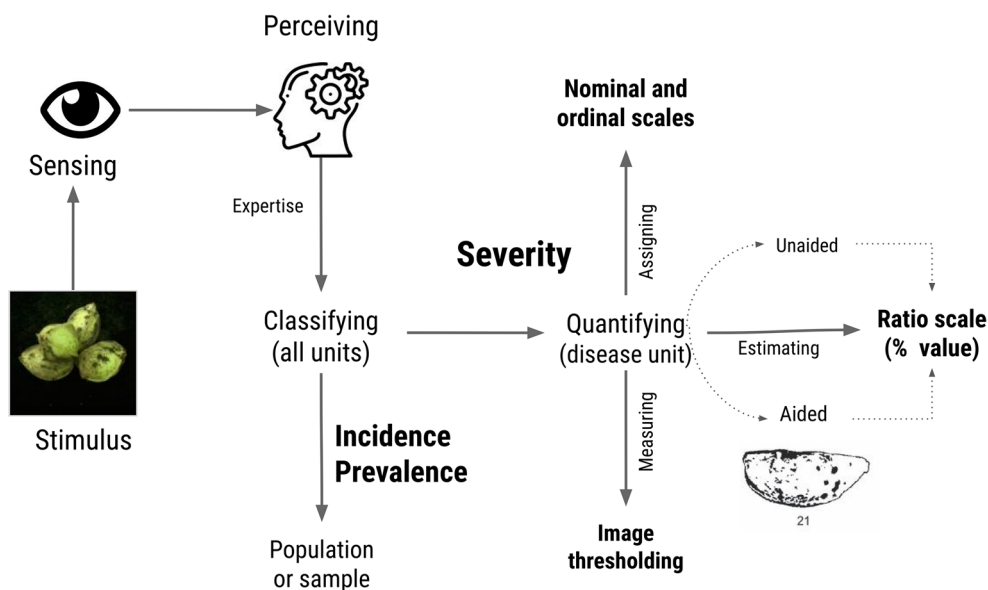
Plant disease severity, the subject of this review, is currently defined as the “area of a sampling unit (plant surface) affected by disease expressed as a percentage or proportion of the total area” Nutter Jr et al. (1991). It is worth considering this definition of disease severity as it is limited to only the increase in the magnitude of disease intensity that can be measured or estimated based on a proportion or percentage of specimen area (for example, soybean rust [Godoy et al. 2006], pecan scab [Yadav et al. 2013], and rice brown spot [Schwanck and Del Ponte 2014]). The current definition is not applicable to diseases that manifest by a progression of symptoms that do not lend themselves to area estimations (for example, huanglongbing [Gottwald et al. 1989], zucchini yellow mosaic virus [Xu et al. 2004], and cassava mosaic disease Hougue et al. (2019)) and for which severity is quantified and represented by other means. Perhaps it is time to revise the definition of plant disease severity to include those diseases that have a symptomatology that does not lend itself to proportion of specimen area diseased. However, as noted, this review focuses on those diseases that can be assessed quantitatively as a proportion based on visible symptomatic area. Full definitions of all the terms used in this review and in phytopathometry can be found in the glossary of terms and concepts in this special issue. Other terms of disease intensity germane to this review include incidence (the proportion of diseased specimens) and prevalence (the proportion of diseased plots or fields in a defined area).

The use of the term visual estimation refers to the eye sensing a stimulus (a diseased specimen, say), followed by perception of the sensation by our brains, which is in turn followed by a cognitive process based on our training, knowledge, and expertise to classify parts of the specimen as diseased (Fig. 1). Such elementary cognition is sufficient to determine incidence of disease, but more complex cognition is

required if an estimate of severity based on the proportion of area diseased is to be made. That mental process of estimation may be achieved using various scales, or can be performed using sensor-based systems and image analysis. There are three commonly used scale types for visual estimates of plant disease severity, as defined by Stevens (1946): nominal scales (where the rater uses simple descriptors to indicate degrees of severity), ordinal scales of two types, qualitative and quantitative (where the rater may use either a qualitative [descriptive] or quantitative [defined ranges of the percentage scale, respectively]), both based on rank-ordered classes [Chiang et al. 2020]), and ratio scales (where the rater bases severity estimates on the proportion or percentage of the specimen area diseased). Quantitative ordinal scales are discussed in detail in another article in this issue (Chiang and Bock 2021). The argument could be made that a fourth common type of scale, the interval scale, is not used in measurement of plant disease severity as interval scales have no defined zero—and all disease severity by definition has a defined zero when the host status is healthy. For this reason, we choose to recognize just the three aforementioned scale types, although the authors recognize some may have valid reasons for considering additional scale types.

Our understanding and knowledge of the processes, methods, and factors affecting the accuracy of severity assessment have evolved as new research results have become available, and consequently approaches to improve accuracy have been developed. This review has two purposes: firstly, to briefly chart the history of plant disease severity estimation and factors that affect those estimates (including sources of error), and secondly to outline the approaches and tools available to maximize accuracy of rater estimates. It is a synthesis of the history of visual disease severity estimation, and corals the tools that we have available at this time to maximize the

Fig. 1 The stages in plant disease severity estimation by visual raters and by image analysis *via* a sensor



accuracy of estimates made visually by different raters; the endeavor will distil a list of best-operating practices that may be used to maximize accuracy of visual disease severity estimation, and point out some avenues for future research.

A brief history of visual disease severity assessment

Various original research studies have described error in visual estimates of plant disease severity as well as novel tools and approaches (Supplementary Table 1 and Fig. 2 list and provide a timeline of some of the most significant). Review articles (Table 1) have been written over the decades that have charted practices and developments in severity estimation.

Interestingly, since 1970 with the publication by Kranz of an article on rater error and scale design (Kranz 1970), phytopathometry has fallen primarily under the purview of botanical epidemiology, perhaps due to its quantitative nature and the treatment of the subject in many reviews and books primarily by epidemiologists since 1970 (most references in Supplementary Table S1). Phytopathometry is indeed critical to epidemiology, but it is equally vital to other branches of plant pathology and in other disciplines where plant disease measurement is required (for example, horticulture, agronomy, ecology, and plant breeding). Phytopathometry is needed

in these disciplines, and the importance of accurate assessments is vital in many studies that cut across the needs of these scientific endeavors. Based on these needs, we contend that phytopathometry in its broadest sense including visual and sensor-based assessments should play a more prominent role in plant pathology than it has hitherto occupied, or at least for which it has been recognized. Indeed, we believe phytopathometry warrants a status as an independent branch of plant pathology, of importance and application in several other branches of the discipline, and in other disciplines.

The rise of phytopathometry

The history of phytopathometry as it relates to visual estimation of disease severity can be divided into two phases. A pre-1970 phase when there was no basis for assessing accuracy of severity estimates (Chester 1950; Large 1953 and 1966), and the phase since 1970 during which there have been quantitative approaches to understanding error and improving accuracy and reliability (“the extent to which the same measurements of individuals [e.g., diseased specimens] obtained under different conditions yield similar results” Everitt 1998) of estimates of severity (for example, Kranz 1970; Forbes and Jeger 1987; Nutter Jr et al. 1993; Nutter Jr and Schultz 1995; Nita et al. 2003; Nutter Jr and Esker 2006; Godoy et al. 2006;

Fig. 2 The history of phytopathometry, 1892 to the present. Significant events and articles are indicated

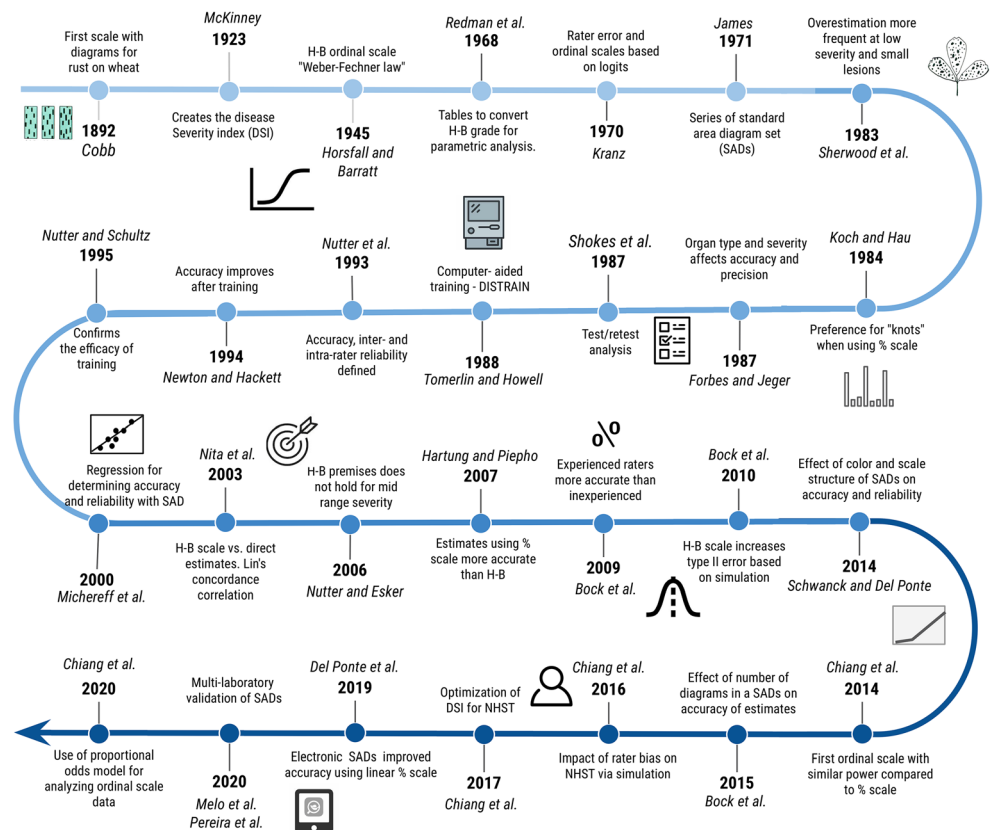


Table 1 Some review articles and book chapters (or components of book chapters) that have described the status of visual plant disease severity estimation

Year	Author	Topic
1950	Chester	Review of the first 50+ years of plant disease severity assessment. The first comprehensive treatise.
1955	Large	A review of disease severity assessment in the UK, with emphasis on use of the percentage scale.
1966	Large	A review of plant disease assessment, including severity. First use of the term “Phytopathometry.”
1970	James	A general review of plant disease assessment in relation to estimating crop losses.
1974	Kranz (Chapter 2)	The importance of disease assessment to epidemiological studies.
1978	Horsfall and Cowling (Chapter 6)	A general review on phytopathometry, current knowledge, and approaches used for plant disease assessment.
1979	Zadocks and Schein	Considers disease assessment in the context of crop loss.
1982	Hebert	In a Letter to the Editor, <i>Phytopathology</i> was first to articulate strong opposition to the basis of the Horsfall-Barratt scale (the Weber-Fechner law).
1988	Kranz (Chapter 3)	A discussion of plant disease measurement, error, and methods.
1989	Hau et al.	Combining new data and results from Kranz (1970 and 1977) and Amanat (1977), error in disease severity estimation is discussed. The authors speculate that the power function may be more appropriate than the Weber-Fechner logarithmic function for ordinal scales and diagrams.
1990	Campbell and Madden (Chapter 6)	Describe disease assessment, accuracy, precision and reliability, and sensor-based approaches.
1991	Chaube and Singh	Review of phytopathometry in relation to crop losses.
1991	Nutter et al.	Presents a comprehensive list of definitions used in plant disease assessment.
1999	Nutter	Article in “Encyclopedia of Plant Pathology” providing an overview of disease assessment, including severity.
2001	Nutter	Conference proceeding on disease assessment and approaches to improve accuracy.
2006	Nutter et al.	A description of concepts and status of accuracy and precision in plant disease severity estimation.
2006	Cooke (Chapter 2)	An updated discussion of plant disease measurement in relation to crop losses.
2007	Madden et al. (Chapter 3)	Authoritative overview with statistical treatment of plant disease severity assessment.
2010	Bock et al.	Comprehensive review visual plant disease severity estimation, image analysis, and hyperspectral imaging to measure disease.
2016	Bock et al.	A general review of accuracy in plant disease severity estimation.
2017	Del Ponte et al.	A scientometric review of SADs describing the development, validation, and impact of the technology on plant disease severity estimation.
2020	Bock et al.	A review of visual estimation of disease severity in relation to sensor-based methods.

Bardsley and Ngugi 2013; Bock et al. 2016b; Pereira et al. 2020). The pivot around which these two phases occur is the recognition of a need for unifying methods of assessment to quantify crop loss in particular, including standardized methods allowing for reproducibility (“the extent to which two or more raters obtain the same estimates of disease severity of the same specimens”, Madden et al. 2007, also known as inter-rater reliability), recognized by the United Nations Food and Agriculture Organization in the mid-1960s (Chiarappa 1970), and culminating in publication of a crop loss assessment manual (Chiarappa 1971).

Over the last 140 years, since the first tool was developed as an aid to quantify severity (Cobb 1892), various approaches have been developed in an attempt to standardize and improve accuracy of visual estimates of disease severity (Chester 1950; Large 1966; James 1974; Nutter Jr 1999; Nutter Jr 2001; Nutter Jr and Esker 2006; Madden et al. 2007; Bock et al. 2010a, 2016a, and Bock et al. 2021; Chiang et al. 2014; Del

Ponte et al. 2017). The term “Phytopathometry” was first suggested in the 1950s (Large 1953, 1966)—a term that was at that time defined as equivalent to “plant disease measurement” or “disease assessment.” Perhaps defining phytopathometry with the narrower, former meaning is most appropriate, as disease assessment is more typically the physical process of measuring disease.

Although various disease assessment terms had been used and defined previously (Chester 1950; James 1974; Berger 1980), it was only in 1991 that the first comprehensive and authoritative list of definitions and concepts used in plant disease assessment was presented in the journal *Plant Disease* Nutter Jr et al. (1991). Research on phytopathometry has since provided knowledge of sources of error, and various methods for augmenting visual estimates of disease severity, which are now a basis for recommending practices to improve the accuracy of visual estimates of disease severity. During this evolution, new terms have been coined, new technologies and

methods used, and definitions have been revised. Thus, an updated glossary of terms and concepts in phytopathometry has been developed (this issue, Bock et al. 2021).

The early years of quantification: scales, diagrams, and field keys

The first tool developed to standardize severity assessments, and which utilized a SAD set, was the “Cobb scale” published by Nathan Cobb (Cobb 1892). This ordinal scale had 5 classes which corresponded to 1, 5, 10, 20, and 50% severity (and was used for “classifying” rather than interpolation to the nearest percent estimate). The Cobb scale was modified twice. Firstly, by Melchers and Parker (1922) who labelled a maximum severity of 37% as 100% “infection level”; and secondly, by Peterson et al. (1948) who included additional infection levels. Such ordinal scales, both those that are proposed as “qualitative ordinal scales” (descriptive of symptoms) or “quantitative ordinal scales” (each class representing defined ranges on the percentage scale) (Chiang et al. 2020), with or without diagrams, proliferated during the following decades, often with diagrams designed to rate severity at the organ or plant level (Gassner 1915; Tehon and Stout 1930; Trumblower 1934; Ullstrup et al. 1945; Horsfall and Barratt 1945; Croxhall et al. 1952a, b). In contrast, field keys were developed to estimate disease severity in whole fields, which may combine characteristics of qualitative and quantitative ordinal scales. An example is the widely used 9-class key developed to assess late blight of potato (Anon. 1947). Many of the SADs developed during this period were used to group estimates in an appropriate “class” or illustrated “degree of symptoms.” Nonetheless, the value of using the continuous percentage scale was well recognized even in the 1940s (Anon. 1948). In that article, the authors point out the nearest percentage estimates have direct biological meaning and may be compared among seasons and raters, and the percentage scale provides a single, uniform method for many different diseases (compared to a diversity of ordinal scales or diagram based systems).

An early quantitative ordinal scale was that of Gassner (1915). Other linear and logarithmic scales and methods were developed to assess severity on individual plant organs, and whole plants (see Chester 1950). The usefulness of these tools to quantify severity accurately was implicit, and although considerations of “reproducibility and reliability” (*sic*) were considered important (Marsh et al. 1937) they were not addressed statistically, nor were they defined. Indeed, it was during these nascent years of phytopathometry that pre-processing of ordinal data for analysis was considered important. McKinney (1923) proposed the “infection index” (a kind of disease severity index, or DSI), which basically summarized frequency of severity class ratings on an ordinal scale. The early history

of the DSI is described by Chester (1950). Marsh et al. (1937) commented that the DSI reduced what may be non-linear data to a single expression that is continuous and amenable to statistical analysis “...although the estimates are not necessarily in direct linear relation to the amount of fungus present... they are reducible to a linear function of this amount.”

Phytopathometry encounters psychophysics

Historically, a widely-used scale for quantifying plant disease severity is the Horsfall and Barratt (H-B) scale (Horsfall and Barratt 1945). It is a quantitative ordinal scale with 12 classes that divide the percentage scale into logarithmically increasing and decreasing sized ranges below and above 50%, respectively. The rationale for the scale design was based in psychophysics. According to the authors, the scale was structured to reflect the “Weber-Fechner law,” which actually combined two independent laws (Nutter Jr and Esker 2006): (1) there is a logarithmic relationship between the intensity of the stimulus (in this case severity of disease) and the estimated value (Fechner’s law, which is false), and (2) the change in a stimulus that will be just noticeable is a constant ratio of the original stimulus (Weber’s law, which holds true). Horsfall and Barratt also presumed that the eye perceives diseased tissue at severity <50%, and healthy tissue at severity >50%, which has never been established. Redman et al. (1968) developed a set of tables based on a formula to convert multiple H-B ratings to estimated mean percentages, effectively taking the percentage midpoint values of the ranges for each class to facilitate determination of percentage mean severity. The H-B scale and its basis in psychophysics was perhaps the most dominant paradigm in phytopathometry for many decades, and received praise as late as the 1980s (Hollis 1984), and remains a tool used in modern research in the field, although not without the psychophysical basis and structure of the scale being seriously questioned (Hebert 1982; Nutter Jr et al. 2006; Bock et al. 2010b). Contrary to the claims of a logarithmic relationship between estimates and actual severity, it has now been demonstrated on many occasions that there is a linear relationship between estimates of disease severity and actual severity (Nita et al. 2003; Nutter Jr and Esker 2006; Bock et al. 2009b).

A flourish of manually prepared black and white SADs

The major contributions of W. Clive James, a researcher from the Canada Department of Agriculture, to the field of phytopathometry, began when he published an influential article in the Canadian Plant Disease Survey (James 1971). In the article, he presented and described the preparation and usage of what he defined as “assessment keys,” which in fact

were SADs for cereal, forage, and field crops, representing 19 diseases. Each key was accompanied by detailed instructions for sampling and usage. To ensure that percent affected area was accurate, a drum scanner coupled to a computer was used to measure drawings made on paper sheets. James recommended interpolation to the nearest percent estimate using the SADs. Another important unit of research was that of Dixon and Doodson (1971) who also published disease-specific SAD sets, some being recommended to be used for interpolation, and others to be used alongside ordinal scales for classification of severity. Given the extensive variation of type and intensity of symptoms across several diseases, the diagram sets in those two studies varied in purpose and number, from as few as three illustrations to 6 or more depicting different disease severities. James (1971) recognized the advantages of using the percentage scale, but also warned that, because only a few severities are shown in the SADs, the extent of interpolation was determined by the ability of the observer. Moreover, it is interesting that the rationale for defining the few diagrams and their values was convenience, rather than laws of psychophysics, as suggested decades earlier as well as, surprisingly, in more recent SADs research (Del Ponte et al. 2017). During that time, no formal quantification of the accuracy of the estimates was determined—rather, it was implicitly presumed that the SADs or scales with diagrams improved accuracy.

Exploring and understanding error

It was also during the early 1970s that a quantitative understanding of characteristics of error and accuracy of visual estimates of disease severity was established. Kranz (1970), Analytis (1973), and Amanat (1976, 1977) investigated rater error and disease scales, and determined standard deviations of multiple rater estimates of the same model leaves were non-constant with severity, a relationship demonstrated for several other diseases too (Fig. 3). Standard deviations of unaided rater estimates tended to be greatest in the range 18 to 62% severity. Kranz (1970) was first to report the minimum and maximum estimate, range, and relative errors of unaided estimates, which increased up to 50% severity, then decreased up to the maximum severity of 100% (Fig. 3). The same pattern has been confirmed more recently for other diseases, as indicated in the figure. Analytis (1973) confirmed non-homogeneity of variance with severity in the apple scab pathosystem. Various transformations of severity data were suggested to account for the non-homogeneous variance and lack of normality of these data (Kranz 1970; Analytis 1973). Amanat (1976) was first to show that training improved precision (which is the degree of variability; the greater the variability, the less precise the estimates, in these cases in relation to the actual values, which is an important point;

Madden et al. 2007). Precision was measured as the scatter of the points in a regression analysis, and in early studies it was noted that precision of estimate of severity was low where symptoms were comprised of small lesions, and raters tended to overestimate with such symptoms (Amanat 1976). Koch and Hau (1980) showed that raters preferred certain severities (“knots”) when estimating severity—generally at 5 and 10% intervals at severities >10 to 20%, which has since been observed in other pathosystems (Bock et al. 2008a). Sherwood et al. (1983) and Hock et al. (1992) also showed overestimation was greatest at low disease severities and that, given the same severity, a disease with smaller lesion size will generally be overestimated, and that, overall, visual estimates by raters were not particularly precise, confirming previous reports. Error associated with estimation due to organ types and disease severity was explored further by Forbes and Jeger (1987).

Intra-rater reliability is the closeness of repeated estimates of severity of the same specimens by the same rater (also known as “repeatability”). Inter-rater reliability is the closeness of repeated estimates of the same specimen by different raters, also known as “reproducibility” (Madden et al. 2007). Reliability does not embrace the concept of accuracy, as no actual values are involved. Statistical analyses of inter-rater and intra-rater reliability were made by Shokes et al. (1987) using analysis of variance and correlation analysis, respectively. Indeed, in regard to plant disease severity estimation, the test/retest method to gauge intra-rater reliability was first promulgated by Shokes et al. (1987) but was based on correlation—although Amanat (1976) used the same test/retest concept it was in relation to learning capacity. Hau et al. (1989) summarized much of this early quantitative work to explore accuracy, and provided further insights into the nature of the relationships between actual disease severity and rater estimates. They also questioned the nature of the logarithmic relationships espoused by Horsfall and Barratt (1945). Nutter Jr et al. (1993) used regression analysis to further establish and understand the concepts of accuracy (using image-analyzed acetate images), inter-rater and, using the test/retest method, intra-rater reliabilities in visual plant disease severity estimation compared to sensor-based methods. Others also explored accuracy and variability in rater estimates using various approaches (Beresford and Royle 1991; Newton and Hackett 1994). Several studies compared rater estimation of symptom components (Beresford and Royle 1991) and rater variability (Beresford and Royle 1991; Newton and Hackett 1994; Parker et al. 1995a, 1995b).

Arrival of personal computers and programs to aid in assessment training

Research on visual assessments of severity was impacted in the mid-1980s when personal computers and programming

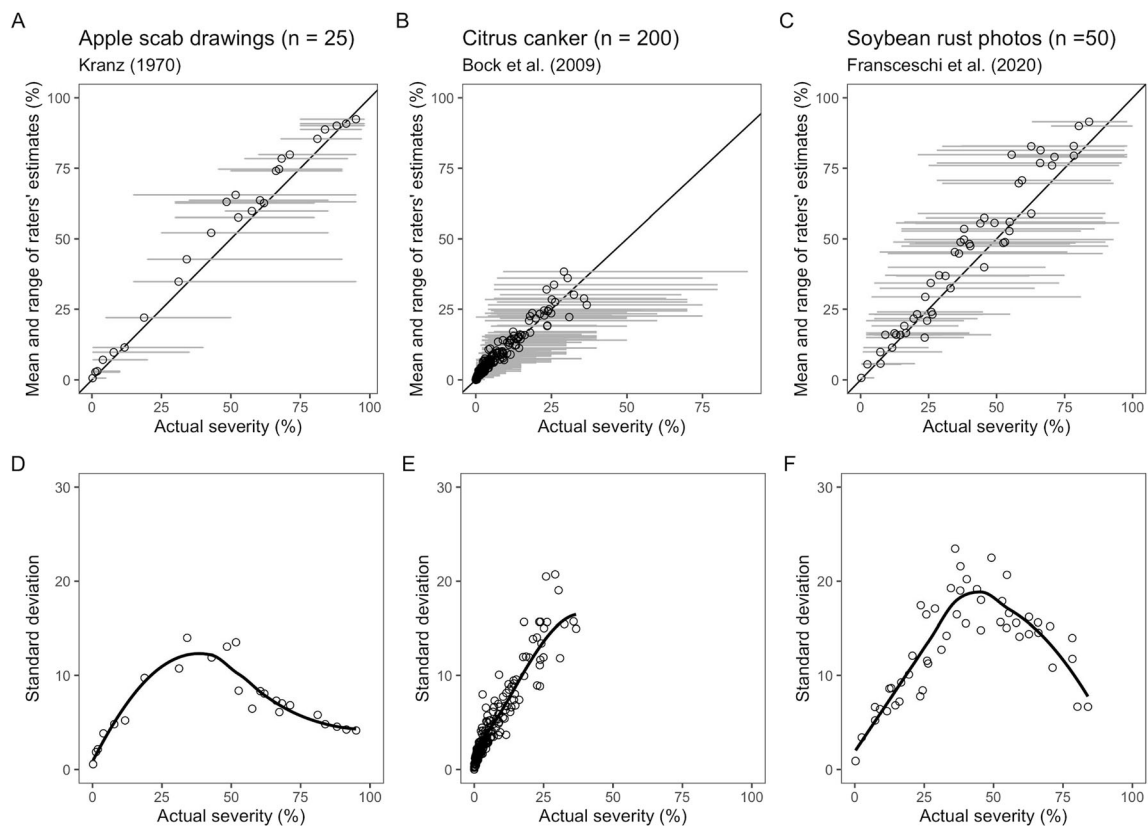


Fig. 3 The means and ranges of unaided estimates of disease severity **A** of stylized disease on 25 model leaves by 200 raters (Kranz 1970), **B** of symptoms of citrus canker on 200 leaves by 28 raters (Bock et al. 2009a),

and **C** of symptoms of soybean rust on 50 leaves by 37 raters Franceschi et al. (2020). The standard deviations of the means are indicated in **D**, **E**, and **F**, respectively

languages become more accessible. Several computer programs were developed with the purpose of improving raters' accuracy *via* training based on computer-generated images of specific and measured disease severity. The estimate could be compared to the actual value. AREAGRAM was the first, described in a university report by Shane et al. (1985) to develop a program where leaves of fixed severities (not randomly generated) were shown to raters. This was followed by other software with similar functionality, but allowing randomly generated series of diseased leaf images in a defined severity range, including DISTRAIN (Tomerlin and Howell 1988), DISEASE.PRO (Nutter Jr and Worawitlikit 1989), and ESTIMATE (Weber and Jorg 1991). Later in the decade, new software was developed for specific diseases, symptoms, and leaf types, for example, SEVERITY.PRO (Nutter Jr et al. 1998). COMBRO (Canteri and Giglioti 1998) was developed specifically for sugarcane rust and borer-rot complex. Research using these tools demonstrated statistically detectable improvement in the accuracy of estimates of disease severity after training (Newton and Hackett 1994; Nutter Jr and Schultz 1995; Parker et al. 1995b; Giglioti and Canteri 1998). A potential issue with computer training is that the benefits may be short-lived (Parker et al. 1995b), with raters requiring regular re-training.

The computer capability to quickly generate digital drawings of diseased leaves not only without the need to hand draw, scan, and measure, but also with the ability to process and analyze the data in real time, was a significant advance. Raters could also do in-house training at any time of the year. Interestingly, the development of these computer programs in the 1980s and 1990s was not immediately followed by computerized systems with greater sophistication to draw more realistic symptomatic digital leaves, despite the advances in software engineering. Indeed, there have been very few training programs developed since (Aubertot et al. 2004; Sachet et al. 2017).

The early psychophysical basis of severity perception challenged

Starting in the 1980s, the so-called Weber-Fechner law and the ideal of the H-B scale began to be challenged. Although Kranz (1970) presented results which showed that error is not symmetrical (and logits were a suitable transformation), estimates did not follow the so-called Weber-Fechner law, because the standard deviation of rater estimates was similar and greatest between ≈ 18

52% compared to other severities (Fig. 3). Hebert (1982) was first to question the presumed psychophysical basis of plant disease severity assessment. Forbes and Jeger (1987) provided the first valuable insights into a number of factors affecting estimation of severity on different plant structures, identifying rater, actual disease severity, and plant structure as factors affecting the accuracy of estimates, and unequivocally demonstrated and stated that the rater error was not compliant with the assumptions of the Weber-Fechner law. The results were reinforced by other observations that estimation error was not greatest at 50% (as had been argued by Horsfall and Barratt 1945) (Hau et al. 1989).

Forbes and Korva (1994) showed that direct use of a H-B type scale did not necessarily resolve uneven variances of estimates, and direct percent estimates were more accurate and precise. Nita et al. (2003) compared scale types using measured, actual values. In a study on Phomopsis leaf blight of strawberry comparing direct visual estimates to the H-B scale, the authors pioneered use of Lin's concordance correlation in determining accuracy in phytopathometry. Accuracy can be considered a product of bias and precision, where bias is the difference between the estimated mean severity and the actual mean severity, and precision is as previously defined. Bias has two forms. First, bias may be constant with estimates being higher, or lower on average by a constant amount when compared to the actual values, and second, bias may be systematic, where the estimates are higher (or lower) than the actual values by an amount that is proportional to the actual severity measured. Constant bias is also known as "fixed bias" or "location shift," while systematic bias is also known as "proportional bias" or "scale shift." Nita et al. (2003) demonstrated that the use of the H-B scale did not result in greater accuracy or reliability when compared to direct nearest percentage estimates, and the results of the study further questioned the basis of the Weber-Fechner law. These and other observations were confirmed experimentally by Nutter Jr and Esker (2006) who used the concept of the "just noticeable difference" to demonstrate that accuracy of raters was far greater in the mid-ranges (25 to 75%) of the H-B scale than the scale structure suggests, which is a significant argument against its use where more accurate methods can be applied (or use of scales similar to the H-B scale).

Various simulation studies have since confirmed that the H-B scale lacks the same power for hypothesis testing compared to the percentage scale (Bock et al. 2010b; Chiang et al. 2014; Chiang et al. 2016a, b). Indeed, since Hebert (1982) first articulated his concerns, it is now generally accepted that a linear relationship exists between estimated severity and actual severity (Nutter Jr and Esker 2006; Bock et al. 2009b), although the relationship between the error of those estimates and the actual values remains to be fully established.

The importance of instruction and experience

Associated with training is instruction. But this is not like computer-based training; rather it relates to written or oral descriptions of symptoms, how to delineate them, and a description of how to implement the rating scales used for assessments. Only recently has research shown that detailed instruction in a pathosystem and how to rate disease severity using the methods of choice is critical for accurate and reliable assessments (Bardsley and Ngugi 2013). Indeed, instruction on use of the rating scale is critical too, as error may result from misuse, as has been noted (Kranz 1988; Bock et al. 2013a, b; Forbes and Korva 1994). Studies demonstrating the importance of the basic procedure of instruction should be repeated with other pathosystems to confirm these results.

Over the last 10 years, several studies have demonstrated that raters' lack of experience can result in inaccuracy and unreliability (Bock et al. 2009b; Pedroso et al. 2011; Yadav et al. 2013; Lage et al. 2015). Experienced raters tend to estimate disease severity on specimens more accurately (although some novice raters may also be intrinsically accurate too). The research has demonstrated that as a group experienced raters are more accurate, but experience does not guarantee more accurate estimates.

Establishment and evolution of SADs research

The pioneering work by W Clive James was highly influential to subsequent SAD research. A selected list of SADs is presented in the chapter on disease monitoring in the Plant Disease Epidemiology book by Campbell and Madden (1990). The list shows 17 published studies by other authors from 1971 to 1988, averaging one per year, but between 1991 and 2017, 105 articles were published (averaging 6 articles per year; Del Ponte et al. 2017). A study conducted by Amorim et al. (1993) was a turning point and the first to use regression analysis to report a measure of accuracy, although the benefits from using SADs were not checked because there was no data on unaided estimates. The Amorim et al. study was used as a model in several articles that followed (Godoy et al. 1997; Michereff et al. 1998, 2000; Diaz et al. 2001; Leite and Amorim 2002). Nutter Jr and Litwiller (1998) were first to show that SADs improved rater estimates of disease severity in an abstract from a conference. Michereff et al. (2000) formally published the first comparison of accuracy of estimates without and with SADs for assessing citrus leprosis *via* comparison of linear regression coefficients. A plethora of SADs followed (see review by Del Ponte et al. 2017) with analyses demonstrating statistically detectable improvements in accuracy and reliability due to using SADs. Many SADs from 1970 to 2010 were based on the "Weber-Fechner" law. As noted earlier, the Weber-Fechner law is non-existent, although

Weber's law holds true. Consequently, the Weber-Fechner law as a principle to guide SAD design has generally been abandoned as a stated basis for scale development in more recent years (Yadav et al. 2013; Lage et al. 2015; Araújo et al. 2019). Interestingly it was not a stated basis for defining incremental interval and number of diagrams in the pioneering work of James (1971). The basis for SADs design should probably be a linear scale, but with additional diagrams at low severity (Bock et al. 2010a; Schwanck and Del Ponte 2014; Franceschi et al. 2020).

Two advances in the methodology for SADs validation were important to more appropriately understand the benefits of SADs. First, the shift from using linear regression to Lin's concordance coefficients as a measure of accuracy and its two main components (precision and bias), recommended as more appropriate for the purpose (Nita et al. 2003; Madden et al. 2007). Spolti et al. (2011) were the first to apply them to the study of SADs. Second, the use of statistical approaches other than regression analysis to explore accuracy including (ordered by first use) confidence intervals (Spolti et al. 2011) equivalence tests (Yadav et al. 2013), non-parametric tests (Schwanck and Del Ponte 2014), and generalized linear mixed models (Correia et al. 2017).

Research on the topic has demonstrated that several aspects of the SADs design and evaluation might affect accuracy (and reliability) including rater experience (Yadav et al. 2013), pathosystem (Godoy et al. 1997), number of diagrams and structure and/or color of SADs (Schwanck and Del Ponte 2014; Bock et al. 2015; Franceschi et al. 2020), and the procedures followed during SAD development and validation, and other factors (Melo et al. 2020; Pereira et al. 2020). Franceschi et al. (2020) demonstrated the substantial improvements that could be made with carefully designed SADs compared to older, basic, previously developed SADs (Fig. 4)—raters' estimates were significantly more accurate with the new SADs. Thus, there may be useful room for improving accuracy based on SADs characteristics. With SADs, research showed that those raters who are least accurate tend to benefit the most from using SADs, while raters who are already accurate remain about the same (Yadav et al. 2013; Bock et al. 2015).

The inexpensive availability of scanners and portable digital cameras in the early 2000s, and the development of plant disease-specific image analysis software facilitated development of SADs (Del Ponte et al. 2017). Image analysis software included APS Assess 2.0 (Lamari 2002) and QUANT (Vale et al. 2003). The development of empirical approaches to develop more realistic SADs, combined with accessibility of image analysis for measuring actual values of test images, made the use of SADs as a training tool a practical and easy to use option compared to computer training programs (and a less expensive one), which may have contributed to the decline of computer training systems. Only a few examples exist

linking SADs and training software either based on an ordinal scale (Aubertot et al. 2004) using Didacte-PIC (Training program: Canker-didacte. Online https://www62.dijon.inrae.fr/didactepic/choix_nombre_et_mode.php) or a percent scale (Sachet et al. 2017).

The intersection of portable devices and SADs was explored by Pethybridge and Nelson (2018). The iPad app "Estimate" has SADs for assessing the severity of *Cercospora* leaf spot in red and yellow table beets and allows direct data entry, using either different ordinal (linear or logarithmic) or continuous scale data. For the ordinal scales, a higher resolution linear scale was most accurate (Del Ponte et al. 2019).

Comparing scale types and characteristics and evaluating impact on decisions

An early study was that of O'Brein and van Bruggen (1992), which compared three quantitative ordinal scales to relate to yield loss caused by corky root of lettuce. The scales had 7, 10, and 12 (the H-B scale) classes. Although the actual values on which accuracy was based were merely "expert" visual estimates, the authors concluded that no scale was most accurate and precise overall, and depended on the specific severity ranges or lettuce growth stages. Two years later, Forbes and Korva (1994) were the first to demonstrate that direct use of H-B scale types did not necessarily resolve uneven variances of estimates, and direct percent estimates were more accurate and precise (direct use of the scale resulted in a "linearization" of unequal scale class intervals). As noted, Nita et al. (2003) compared direct visual estimates to H-B scale converted values and demonstrated the H-B scale was not more accurate or reliable compared with direct nearest percentage estimates. Similar studies on citrus canker by Bock et al. (2009b) drew similar conclusions, and Bardsley and Ngugi (2013) demonstrated that direct estimation resulted in more accurate and reliable estimates than an ordinal scale when estimating severity of foliar bacterial spot symptoms on peach and nectarine. Hartung and Piepho (2007) also showed that accuracy was greatest using the percentage scale (although they considered a 5% ordinal scale to be sufficient).

Some studies have compared treatments using different methods of assessment on the outcome of an analysis. Todd and Kommedahl (1994) compared severity of symptoms of *Fusarium* stalk rot of corn caused by three different species of *Fusarium* assessed either as a percentage by image analysis (considered objective) or visually using a 1 to 4 severity scale—means separation was dependent on assessment method. Similarly, Parker et al. (1995a, b) found that data from objectively measured severity of barley powdery mildew (using image analysis) gave different outcomes compared to visual estimates after data analysis. Bock et al. (2015) also

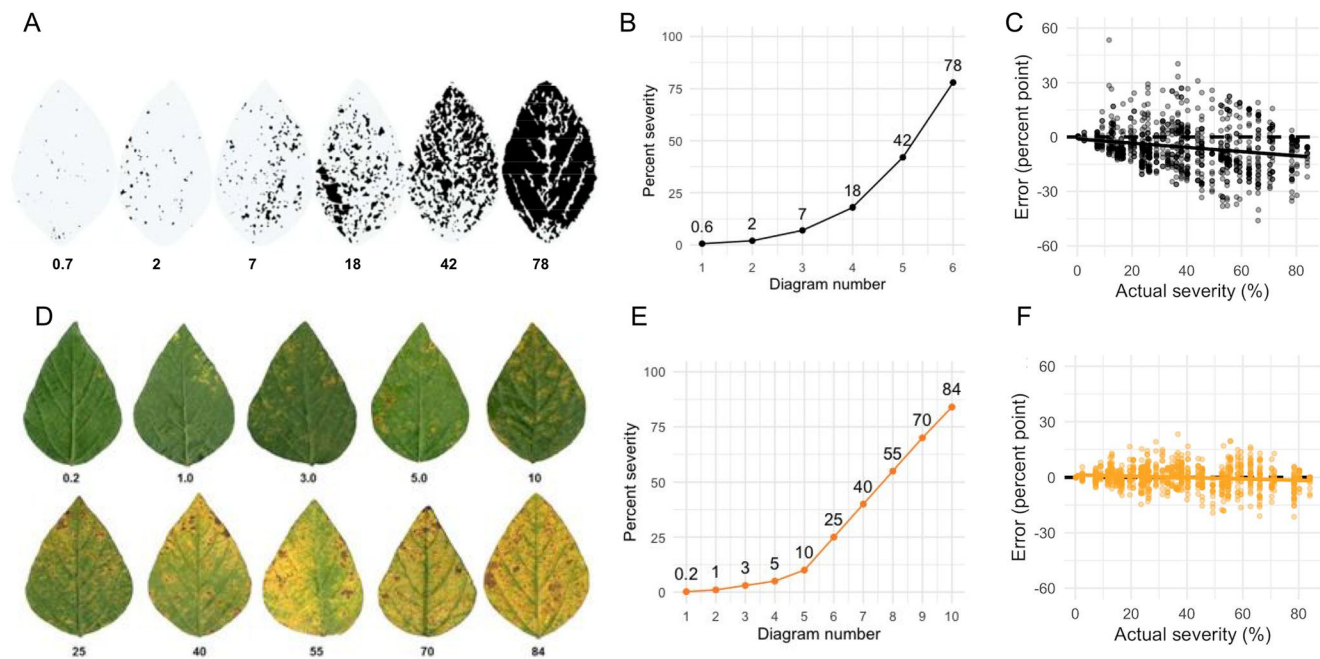


Fig. 4 Standard area diagrams (SADs) to estimate severity of rust (*Phakopsora pachyrhizi*) on soybean (*Glycine max*) leaves. **A** The original SADs (Godoy et al. 2006) **B** the relationship between the illustrated SAD severity and diagram number for the original SAD **C** the absolute errors of estimates when using the original SADs **D** the newly developed and validated SADs (Franceschi et al. 2020) that is a tool for more

accurate estimates of rust severity **E** the relationship between the illustrated SAD severity and diagram number for the newly developed SAD **F** the absolute errors of estimates when using the newly developed SADs. The numbers under each leaf represent actual percentage leaf area showing symptoms (necrosis and chlorosis)

found that use of the H-B scale could result in different means separation among treatments compared to direct percentage estimates. In a study of QTLs for oat crown rust resistance genes, a quantitative analysis found that 64% of the phenotypic variation was accounted for using q-PCR to quantify the pathogen (which also most precisely mapped the gene), 52% was accounted for using digital image analysis, but only 41% by visual assessments, respectively (Jackson et al. 2007). Although Poland and Nelson (2011) observed little difference in the QTLs identified to northern leaf blight of corn using either a 1 to 9 scale or a direct percentage estimation, the results showed the direct percentage estimates to be more precise.

During the last decade, several simulation-based studies exploring the power of the hypothesis test have demonstrated the issues associated with using quantitative ordinal scales compared to a continuous ratio scale (Bock et al. 2010b; Chiang et al. 2014). In the former, type II errors are elevated, although increasing sample size can resolve most issues. Rater bias also has problematic effects that can be magnified by quantitative ordinal scales (Chiang et al. 2016a). Several of the studies described in this paragraph, and in other works (for example, Chiang et al. 2014), indicate that the H-B scale (and similar scales) has drawbacks and can result in elevated type II errors. The research has also provided a basis for developing ordinal scales that have similar accuracy or minimized risk of type

II error compared to nearest percentage estimates (Hartung and Piepho 2007; Chiang et al. 2014) (Table 2). Furthermore, selection of scale type can affect resource use efficiency (Chiang et al. 2016b), with more replications required to achieve the same level of power in a hypothesis test (the type II error rate) when using some ordinal scales. Furthermore, percentage scale severity data estimated by very accurate raters almost always leads to the rejection of the null hypothesis (when it is false), but for accurate raters using H-B type scales is more detrimental to the probability to reject the null hypothesis compared to inaccurate raters (Bock et al. 2010b).

The impact of using a DSI on accuracy and type II error when using a quantitative ordinal scale was investigated by Chiang et al. (2017a, b). Results showed that DSIs based on ranges of the percentage scale are prone to overestimation if the midpoint values of the rating class are not considered. Rater bias can further detract from accuracy of the DSI compared to the actual mean. However, Chiang et al. (2017b) using quantitative ordinal rating grades or the midpoint conversion for the ranges of disease severity resulted in similar powers of hypothesis testing. The authors concluded that the principal factor determining the power of the hypothesis test (the complement of the type II error rate) when using a DSI is the nature of the intervals in the quantitative ordinal scale—an amended 10% interval scale provided a type II error rate close to direct estimation of disease severity. Thus, steps can be

Table 2 The “Chiang” scale: an improved 16-class quantitative ordinal scale for general assessment of plant disease severity based on the scale developed by Chiang et al. (2014) (with modification by Bock et al. 2021)

Ordinal equivalent	Midpoint	Severity (% range)
0	-	-
1	0.05	0 ⁺ to 0.1
2	0.30	0.1 ⁺ to 0.5
3	0.75	0.5 ⁺ to 1.0
4	1.50	1.0 ⁺ to 2.0
5	3.50	2.0 ⁺ to 5.0
6	7.50	5.0 ⁺ to 10.0
7	15.0	10.0 ⁺ to 20.0
8	25.0	20.0 ⁺ to 30.0
9	35.0	30.0 ⁺ to 40.0
10	45.0	40.0 ⁺ to 50.0
11	55.0	50.0 ⁺ to 60.0
12	65.0	60.0 ⁺ to 70.0
13	75.0	70.0 ⁺ to 80.0
14	85.0	80.0 ⁺ to 90.0
15	95.0	90.0 ⁺ to 100.0

taken that maximize the utility of the DSI when selecting the scale intervals on which it will be based. DSIs remain quite widely used (Hunter and Roberts 1978; Koitabashi 2005; Nsabiya et al. 2012; Gafni et al. 2015).

The previous sections have outlined the history and many of the advances in phytopathometry since Nathan Cobb developed a cereal rust scale in the 1890s. But there remain many unanswered questions, and there are further avenues to explore that may provide a basis for added improvements in accuracy and reliability of visual estimates of plant disease severity.

The need for a baseline for accuracy

So, a couple of questions may remain regarding all this progress: what is an accurate visual estimate of disease severity? How do we know when we are close enough to the actual value? Accuracy may in part be dependent on the needs of a specific study, so these are not easy questions to answer. Nonetheless, based on empirical results from rater studies over the last 10 years we can determine that raters with Lin’s concordance correlation coefficient (ρ_c) of approximately 0.90, or more may be considered accurate (Capucho et al. 2011; Spolti et al. 2011; Rios et al. 2013; Domiciano et al. 2014; Duarte et al. 2013; Yadav et al. 2013; Bardsley and Ngugi 2013; Schwanck and Del Ponte 2014; Lage et al. 2015; Araújo et al. 2019; Franceschi et al. 2020). Inevitably this is somewhat arbitrary, and the references show that it varies with the

study, and quite likely the pathosystem and several other factors. But based on the studies that have been done, and the accuracies achieved with and without instruction, training, and SADs, it is a reasonable magnitude for a ρ_c for the rater to be considered accurate on the spectrum of known rater capability. Rarely will a visual rater have a consistent $\rho_c > 0.95$. Most commonly raters with training instruction and/or SADs will have an $\rho_c \geq 0.85$ to 0.95. The SADs, individual rater, and other factors will contribute to imprecision, constant bias, and systematic bias. In a real disease assessment situation, it is quite likely that accuracy will be a little lower. But this is a lot better compared to the capability of some raters with no experience or assessment aid, and possibly just rudimentary instruction, who may have a $\rho_c = 0.60$ or less. It should be noted that accuracy too may be in the eyes of the beholder: whereas Altman (1991) considers 0.90 to be accurate, and McBride (2005) does not believe “substantial” accuracy is achieved until $\rho_c \geq 0.95$ (anything less being considered only moderately accurate or poor).

Using regression analysis, permissible accuracy has been based on a range of percentages around the actual severity (Amanat 1977; Newton and Hackett 1994). In these studies, ranges considered accurate at specific actual severities were 1% (0.5 to 1.5%); 5% (3.75 to 7.00%), 10% (7.50 to 12.50%), and 30% (25 to 35%). As observed by Newton and Hackett (1994), this gives an upper limiting regression line with an intercept of 0.28 and a slope of 1.2, and a lower limiting regression line with an intercept of -0.28 and a slope of 0.78.

As noted, many decisions are based on estimates or measurements of plant disease severity. Thus, for these decisions to be of greatest value, they must be based on data that is true to the actual severities—i.e., it must be accurate. The work done to date has explored many facets that affect accuracy and reliability, and plant pathologists have developed an understanding of sources of error to address some of the shortcomings by implementing improved tools and approaches to estimate plant disease severity. The main sources of error are briefly considered in the next section.

Factors affecting accuracy

Scale type: Several studies have demonstrated that assessment method can affect accuracy and the outcome of an analysis (Todd and Kommedahl 1994; Parker et al. 1995a, b; Nita et al. 2003; Bock et al. 2015; Bock et al. 2009a; Bock et al. 2010b; Jackson et al. 2007; Poland and Nelson 2011; Chiang et al. 2014; Chiang et al. 2016a, b).

Raters: Probably the single biggest source of error and variability in assessment. Raters have been demonstrated to be inherently variable (Hau et al. 1989; Nutter Jr et al. 1993; Bock et al. 2009b). The majority of raters tend to overestimate

disease severity, especially at low severity, while a few may also underestimate, but it is less common.

Rater preferences for particular severities: Raters tend to have a preference for certain severities, generally at 5 and 10% interval, particularly at severities >20% (Koch and Hau 1980; Bock et al. 2008b, 2009b; Schwanck and Del Ponte 2014).

Lack of experience, training, and instruction: Over the last 10 years, several studies have demonstrated that lack of experience can result in inaccuracy and unreliability (Bock et al. 2009b; Pedroso et al. 2011; Yadav et al. 2013; Lage et al. 2015). Training improves accuracy of estimates (Parker et al. 1995a; Nutter Jr and Schultz 1995; Bardsley and Ngugi 2013). Similarly, instruction in the pathosystem and/or rating methods can result in more accurate and reliable estimates (Bardsley and Ngugi 2013).

Symptoms: The characteristics of the symptoms can influence rater accuracy. Severity characterized by numerous small lesions will tend to be more seriously overestimated compared to diseases with fewer, larger lesions (Sherwood et al. 1983; Forbes and Jeger 1987). Also, that tendency to overestimate is relatively greatest at severities <20% (Sherwood et al. 1983; Bock et al. 2008b). Whether lesions are regularly or irregularly distributed may also impact error (Hock et al. 1992).

Plant structure: The organ (plant part) or whole plant being assessed can influence the accuracy of estimation (Amanat 1977; Forbes and Jeger 1987; Nita et al. 2003). Roots in particular are especially challenging to accuracy in severity estimation (Forbes and Jeger 1987).

Time: The speed with which rating is performed may affect accuracy, although not many studies have been performed. Faster raters tended to have less precise estimates of severity (Parker et al. 1995b), and by extension these estimates would be individually less accurate.

Other causes: Color blindness has been reported to be detrimental to estimation of disease severity (Nilsson 1995).

There may also be interactions among the various factors listed here. A more in-depth discussion of sources of error affecting disease severity estimation is provided by Bock et al. (2010a). Other factors not yet studied may also play a role in rater error. A chart presenting the sources of error in plant disease assessment and the tools, methods, and approaches to increase accuracy is presented in Fig. 5.

Can visual estimates be more accurate? A primer on best practices

The advantages of the percentage ratio scale for estimating those diseases amenable to such estimations were articulated in an article in the Transactions of the British Mycological Society (Anon. 1948). In addition, the authors: (i) encouraged use of pictorial diagrams of known severities to more

accurately guide estimates, (ii) commented that the percentage estimates have direct biological meaning, and (iii) stated it provides a single, unifying method for estimating severity for all those diseases where area estimates are appropriate measures for severity. As noted, the data also lend themselves to direct parametric analysis. It is also bounded by 0 and 100%, the scale can be subdivided, is universally known, and is applicable to measures of incidence as well as severity (James 1971). Large (1955) stated that wherever possible they strove to assess using percentages because it provided the “percentage of the total green area of the plant rendered inoperative by reason of the disease at the time of observation,” rather than arbitrary or subjective ordinal grading systems based on psychology of perception, and due to its objectivity allowing comparisons.

There may be reasons for selecting any one of the types of scales used in plant pathology for a specific disease assessment purpose, but the user should remember that the objectivity and statistically available information content is least with the nominal scale, and increases progressively with the ordinal and ratio type scales, respectively. Of course, there are many diseases that must be assessed using a qualitative ordinal scale, but those are not considered in this review. There are various criteria to consider, and a sequence to approaching severity assessment that can be followed that will help contribute to accuracy of rater estimates, and at the same time minimize risk of type II errors. Best-operating practices (summarized in Table 3) for consideration in a disease severity assessment activity to maximize accuracy should include:

First, **select the most appropriate scale** for the pathosystem involved, the requirements of the experiment, and the resource availability. In some cases, a pathosystem may dictate the scale to be used: thus, many systemic diseases that have relatively amorphous symptoms are more readily scored using a qualitative ordinal scale. Other pathosystems where symptoms are easily defined and quantified on an organ or whole plant lend themselves to rating using a quantitative ordinal scale or a ratio scale (the percentage scale). The percentage scale may be preferable to provide greater accuracy of individual estimates, and the ability to use parametric statistics directly with no loss in accuracy or precision (taking mid-points of quantitative scale ordinal estimates is less accurate and precise compared to direct estimates). Furthermore, a rater must learn the characteristics of the quantitative ordinal scale.

Second, **provide raters with detailed instruction** of (i) the pathosystem, (ii) the rating scale being used, and (iii) common sources of error in rating. These instructions should include a description of the disease symptoms and the stages they may go through, and any fungal structures that are relevant to assessment, and where to consider a boundary between healthy and diseased tissue. Other diseases or conditions that could be a source of misidentification, confusion, and error should also be described. Explicit instruction should be provided, even for the




Error source	Factor affecting accuracy	Approaches and tools to improve accuracy
 Specimen	Symptom (lesion size and pattern) Symptom distinctiveness Actual disease severity Plant organ type	Oral and visual Instruction Description of lesion pattern variation
 Rater	Inexperience Intrinsic ability Value preference (“knots”) Rating environment and duration Color blindness	Description of common sources or error Training in specific scale and specimens Re-training
 Rating method	Selected scale (nominal, ordinal or ratio) Lack of: instruction & training Insufficient sample size (+ uncertainty)	Appropriate scale selection Instruction in scale use Well-designed standard area diagrams Adequate sample size Appropriate analysis for scale type used

Fig. 5 Sources of error that affect rater accuracy of individual specimen disease severity estimates during the assessment process, and approaches and tools to increase accuracy

percentage ratio scale (for example, if using SADs, raters should understand to use the SADs as a guide for interpolation of their best estimate, not as a tool to classify the specimen as represented by a SAD or a preferred value—which has happened in some studies [Parker et al. 1995a, b; Melo et al. 2020]). Raters should be instructed on common sources of error including avoiding the common tendency to overestimate (especially at low severity) and to avoid rating in “knots”—specific values at 5 and 10% intervals. The importance of instruction is demonstrated (Bardsley and Ngugi 2013).

Third, **raters should be tested and trained** for two reasons: (i) to ascertain their native ability and (ii) to ascertain whether they can improve with experience, training, and/or the use of SADs. A rater who is consistently very inaccurate should probably be replaced. Most raters respond favorably to training and it ensures that their estimation accuracy is sufficient. This can be done using computer training programs (not easy to obtain now) or through the use of SADs and sets of image-analyzed, diseased specimens of known actual value that raters can gain experience by using. The value of training for improving accuracy is demonstrated (Nutter Jr and Schultz 1995; Bardsley and Ngugi 2013).

Fourth, related to the previous two criteria is experience. Wherever possible **raters should be experienced** (perhaps through instruction and training) so that they are comfortable rating disease severity. Thus, raters should be selected based on demonstrated experience wherever possible. Again, experience has been shown in several studies to be an important gauge of accuracy (Pedroso et al. 2011; Yadav et al. 2013; Lage et al. 2015).

Fifth, wherever possible, **raters should use SADs as an aid**, especially if not highly experienced and demonstrated to be accurate. There are now well over 100 studies that show SADs improve accuracy, particularly for those less experienced or less accurate raters (Pedroso et al. 2011; Yadav et al. 2013; Lage et al. 2015; Del Ponte et al. 2017). The SADs also improve inter- and intra-rater reliability, which is most likely a result of the increase in accuracy of individual raters.

Sixth, where possible the **minimum number of raters should be used** in any particular experiment, and if different raters are used, ideally they should be allocated randomly across the experimental units. This provides a further method to isolate rater-related error in a way that can be accounted for in the analysis. If raters vary and have assessed across statistical units, the error will detract from the power of the analysis. Although peripheral to disease assessment, resource use efficiency and sample size may be critical considerations. Chiang et al. (2016b) demonstrated the need for a minimum sample size to minimize the risk of type II errors. Ideally this should be at least 30 samples. Subsequent analysis should be appropriate for the data type (Shah and Madden 2004; Chiang et al. 2020).

The future of visual plant disease severity estimation

Based on applying these methods, and implementing appropriate tools, the potential improvement in accuracy of direct visual severity estimates, especially for inherently “accurate” raters, is likely approaching its limit. Further gains, of variable

Table 3 Best-operating procedures for maximizing the accuracy of plant disease severity estimates (adapted from Bock et al. 2021)

Factor	Procedures
Scale selection	<p>Know the pathosystem – are there appropriate, previously published scales? Select the most appropriate scale for measuring disease severity for that pathosystem and for the purpose of the experiment. Scale choice is:</p> <ul style="list-style-type: none"> i) Nominal scales ii) Field keys iii) Qualitative ordinal scales iv) Quantitative ordinal scales v) Ratio scales (most often the % scale) <p>The scales have an increasing information content from (i) to (v), and only some quantitative ordinal scales and ratio scales offer themselves to parametric analysis without the need of transformation of the original data (e.g., index or ranks).</p>
Experience	<p>Where possible ensure that raters have at least some prior experience and appreciation of their own ability to estimate severity:</p> <ul style="list-style-type: none"> i) Provide raters with some experience rating the disease. Check rater results against known values and check accuracy. This can be repeated as needed. ii) Where possible, gain knowledge of rater ability and select the most accurate raters for assessing experiments or conducting surveys.
Instruction	<p>Instruction on the specificities of pathosystem and scale used to rate the disease:</p> <ul style="list-style-type: none"> i) Describe characteristics of healthy and diseased foliage, and where the boundary should be drawn. ii) Describe characteristics, and variability, of the symptoms of the target disease iii) Describe other conditions, most likely to be encountered, that might be confused with the target symptoms, and how to differentiate them. iv) Explain common sources of error in estimating disease severity v) Particularly the tendency for overestimation at low severity, and use of preferred values at 5 and 10% intervals. vi) Explain how to use the selected scale.
Training	<p>Train raters on how to assess the particular disease. If available, use computer- or internet-based software or diagrams (printed or in digital format) of specimens with symptoms of known severity to train raters prior to assessment of experiment specimens.</p>
Use of SADs	<p>SADs can be developed for any scale type, but are most commonly used for the percent ratio scale. But they have been developed and used as an aid to visual classifications for qualitative and quantitative ordinal scales. Wherever possible validated SADs should be used for rating to increase accuracy and reliability, regardless of rater ability.</p> <ul style="list-style-type: none"> i) If no SAD exists that match the target disease, consider developing one. If a SAD has been developed, it is important to confirm it is adequate. ii) Train raters in use of SADs – how to interpolate, and if being used for a quantitative ordinal scale how to assign class appropriately for an interval.
Deployment of raters	<p>Although not directly related to accuracy, if using multiple raters deploy them in a way that can ensure the rater variability is accounted for in the experiment design.</p>
Sample size and analysis	<p>Also not directly related to accuracy of estimates, employ a sufficient sample size, ideally determined by a power analysis, and choose the most appropriate methods for descriptive and inferential statistics and data visualization.</p>

magnitude, may be made from results of additional studies understanding aspects of rater error, and from the optimization of SADs design and their deployment, in particular. But there remain many areas for future research.

For example, how does rater accuracy really vary over the full range of disease severity? Nutter Jr and Esker (2006) provided valuable information over the mid-range of disease severity (25 to 75%) using the just noticeable difference. But what about severity estimates <25% or >75%? A comprehensive study will lay to rest the question of the relationship between ability to estimate and actual disease severity.

How do symptom types and characteristics, and likely range of disease severity for a pathosystem affect an optimum selection of SAD severities and the range and individual severities illustrated? Further research is needed to determine how many diagrams are really needed in a SAD set. And at what point are there too many? What are critical aspects of SAD development and validation that should be followed in

all labs developing this tool to ensure that differences due to approach are not a source or error in design or validation?

What aspects of rater instruction are most important to accurate assessments? How do personality types and other psychological or gender factors affect rater accuracy?

Similarly, with ordinal scales: what further improvements might be made to scale structure that will improve accuracy of estimation? Does training or instruction impact quantitative ordinal scale use (this has never been explored)? Do SADs aid accuracy of classification using ordinal scales (both quantitative and qualitative)?

Do the same raters need to be used for all stages in disease assessment studies of accuracy and reliability? Or can a random “sample” of raters be used to represent the population? If so, how many raters should be used in any given study to encompass likely variability?

There are several other methods used for assessing disease severity which are sensor-based. These may incorporate

artificial intelligence (AI) and have the potential for an eventual capability to provide accurate estimates of disease severity Bock et al. (2021). Nonetheless, in most pathosystems, visual disease estimation is and will be a standard for many years to come, underlining the importance of accuracy in visual estimation of disease severity.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s40858-021-00439-z>.

Data availability statement Where applicable, the datasets analyzed during the current study are available from the corresponding author on reasonable request.

Author contribution Conceptualization: CHB; methodology: CHB, EMDP; resources: CHB, KSC, EMDP; writing - original draft preparation: CHB; writing - review and editing: CHB, KSC, EMDP; visualization: CHB, EMDP; project administration, CHB; funding acquisition, CHB, KSC, EMDP. All authors have read and agreed to the published version of the manuscript.

Funding Clive H. Bock was supported by the USDA-ARS National Programs through CRIS project 6042-21220-012-00. Kuo-Szu Chiang was supported by Ministry of Science and Technology of Taiwan, R.O.C. (MOST 109-2313-B-005-036). Emerson M. Del Ponte was supported by the National Council for Scientific and Technological Development (CNPq) through a Productivity Research Fellowship (PQ) project 310208/2019-0.

Declarations

Conflict of interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Altman DG (1991) Practical statistics for medical research. Chapman and Hall, London
- Amanat P (1976) Stimuli effecting disease assessment. *Agric Conspectus Scientificus* 39:27–31
- Amanat P (1977) Modellversuche zur Ermittlung individueller und objektabhangeriger schatzfehler bei pflanzenkrankheiten. Diss. Universitat Gieen. Cited in: Hau, B., Kranz, J., and Konig, R. 1989. Fehler beim Schatzen von Befallsstarken bei Pflanzenkrankheiten. *Zeitschrift fur Pflanzenkrankheiten und Pflanzenschutz* 96:649–674
- Amorim L, Bergamin Filho A, Palazzo D, Bassanezi RB, Godoy CV, Torres GAM (1993) Clorose variegada dos citros: uma escala diagramatica para avaliaao da severidade da doena. *Fitopatol Bras* 18:174–180
- Analytis S (1973) Zur methodik der analyse von epidemien dargestellt am apfelschorf. *Acta Phytomedica* 1:1–72
- Anon. (1947) The measurement of potato blight. *Transactions of the British Mycological Society* 31:140–141
- Anon. (1948) Disease measurement in plant pathology. *Transactions of the British Mycological Society* 31:343–345
- Araujo ER, Resende RS, Krezanoski CE, Duarte HSS (2019) A standard area diagram set for severity assessment of botrytis leaf blight of onion. *European Journal of Plant Pathology* 153:273–277
- Aubertot J-N, Sohbi Y, Brun H, Penaud A, Nutter FW (2004) Phomadacte: a computer-aided training program for the severity assessment of phoma stem canker of oilseed rape. *Integrated Control in Oilseed Crops*:247–254
- Bardsley SJ, Ngugi HK (2013) Reliability and accuracy of visual methods to quantify severity of foliar bacterial spot symptoms on peach and nectarine. *Plant Pathology* 62:460–474
- Beresford RM, Royle DJ (1991) The assessment of infectious disease for brown rust (*Puccinia hordei*) of barley. *Plant Pathol* 40:374–381
- Berger RD (1980) Measuring disease intensity. In: *Proc. E.C. Stakman Commemorative Symposium on Crop Loss Assessment*. University of Minnesota. Miscellaneous Publications, St Paul 7:28–31
- Bock CH, Parker PE, Cook AZ, Gottwald TR (2008a) Visual rating and the use of image analysis for assessing different symptoms of citrus canker on grapefruit leaves. *Plant Disease* 92:530–541
- Bock CH, Parker PE, Cook AZ, Gottwald TR (2008b) Characteristics of the perception of different severity measures of citrus canker and the relations between the various symptom types. *Plant Disease* 92:927–939
- Bock CH, Gottwald TR, Parker PE, Cook AZ, Ferrandino F, Parnell S, van den Bosch F (2009a) The Horsfall-Barratt scale and severity estimates of citrus canker. *European Journal of Plant Pathology* 125:23–38
- Bock CH, Parker PE, Cook AZ, Riley T, Gottwald TR (2009b) Comparison of assessment of citrus canker foliar symptoms by experienced and inexperienced raters. *Plant Disease* 93:412–424
- Bock CH, Poole GH, Parker PE, Gottwald TR (2010a) Plant disease severity estimated visually, by digital photography and image analysis, and by hyperspectral imaging. *Critical Reviews in Plant Science Sci* 29:59–107
- Bock CH, Gottwald TR, Parker PE, Ferrandino F, Welham S, van den Bosch F, Parnell S (2010b) Some consequences of using the Horsfall-Barratt scale for hypothesis testing. *Phytopathology* 100:1031–1041
- Bock CH, Wood BW, Gottwald TR (2013a) Pecan scab severity – effects of assessment methods. *Plant Disease* 97:675–684
- Bock CH, Wood BW, van den Bosch F, Parnell S, Gottwald TR (2013b) The effect of Horsfall-Barratt category size on the accuracy and reliability of estimates of pecan scab severity. *Plant Disease* 97:797–806
- Bock CH, El Jarroudi M, Kouadio AL, Mackels C, Chiang K-S, Delfosse P (2015) Disease severity estimates – effects of rater accuracy and assessment methods for comparing treatments. *Plant Disease* 99:1104–1112
- Bock CH, Chiang K-S, del Ponte EM (2016a) Accuracy of plant specimen disease severity estimates: concepts, history, methods, ramifications and challenges for the future. *CAB Reviews* 11:1–21
- Bock CH, Hotchkiss MW, Wood BW (2016b) Assessing disease severity: accuracy and reliability of rater estimates in relation to number of diagrams in a standard area diagram set. *Plant Pathology* 65:261–272
- Bock CH, Pethybridge SJ, Barbedo JGA, Esker PD, Mahlein A-K, and Del Ponte EM (2021) A phytopathometry glossary for the 21st

- century: towards consistency and precision in intra- and inter-disciplinary dialogues. *Tropical Plant Pathology*. This issue
- Campbell CL, Madden LV (1990) Introduction to plant disease epidemiology. Wiley, New York, 532 p
- Canteri MG, Giglioti EA (1998) COMBRO: um software para seleção e treinamento de avaliadores de ferrugem e do complexo broca-podridões em cana-de-açúcar. *Summa Phytopathol* 24:190–192
- Capucho AS, Zambolim L, Duarte HSS, Vaz GRO (2011) Development and validation of a standard area diagram set to estimate severity of leaf rust in *Coffea arabica* and *C. canephora*. *Plant Pathology* 60: 1144–1150
- Chaube HS, Singh US (1991) Pathometry-assessment of disease incidence and loss (chapter 9). In: *Plant disease management: principles and practices*. CRC Press, Boca Raton, Florida, pp 119–131
- Chester KS (1950) Plant disease losses: their appraisal and interpretation. *Plant Disease Reporter* 193:190–362
- Chiang K-S, Bock CH (2021) Understanding the ramifications of quantitative ordinal scales on accuracy of estimates of disease severity and data analysis in plant pathology. *Tropical Plant Pathology*. This issue
- Chiang K-S, Liu SC, Bock CH, Gottwald TR (2014) What interval characteristics make a good categorical disease assessment scale? *Phytopathology* 104:575–585
- Chiang K-S, Bock CH, El Jarroudi M, Delfosse P, Lee IH, Liu HI (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 CH, Lee IH, El Jarroudi M, 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
- Chiang K-S, Liu HI, Bock CH (2017a) A discussion on disease severity index values: I. Warning on inherent errors and suggestions to maximize accuracy. *Annals of Applied Biology* 171:139–154
- Chiang K-S, Liu HI, Tsai JW, Tsai JR, Bock CH (2017b) A discussion on disease severity index values. Part II: using the disease severity index for null hypothesis testing. *Annals of Applied Biology* 171: 490–505
- Chiang K-S, Liu HI, Chen YL, El Jarroudi M, Bock CH (2020) Quantitative ordinal scale estimates of plant disease severity: comparing treatments using a proportional odds model. *Phytopathology* 110:734–743
- Chiarappa L (1970) FAO international collaborative programme for the development of reproducible methods for the assessment of crop losses. *PANS Pest Articles and News Summaries* 16:733–734
- Chiarappa L (1971) Crop loss assessment methods: FAO manual on the evaluation and prevention of losses by pests, diseases and weeds. Farnham Royal, UK: Commonwealth Agricultural Bureaux for FAO. (Loose leaf, plus supplements)
- Cobb NA (1892) Contribution to an economic knowledge of the Australian rusts (Uredinae). *Agricultural Gazette (NSW)* 3:60
- Cooke BM (2006) Disease assessment and yield loss. In: Cooke BM, Gareth Jones D, Kaye B (eds) *The epidemiology of plant diseases*. Springer, Second edition. The Netherlands
- Correia KC, de Queiroz JVJ, Martins RB, Nicoli A, Del Ponte EM, Michereff SJ (2017) Development and evaluation of a standard area diagram set for the severity of phomopsis leaf blight on eggplant. *European Journal of Plant Pathology* 149:269–276.
- Croxhall HE, Gwynne DC, Jenkins JEE (1952a) The rapid assessment of apple scab on leaves. *Plant Pathology* 1:39–41
- Croxhall HE, Gwynne DC, Jenkins JEE (1952b) The rapid assessment of apple scab on fruit. *Plant Pathology* 1:89–92
- Del Ponte EM, Pethybridge SJ, Bock CH, Michereff SJ, Machado FJ, 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 EM, Nelson SC, Pethybridge SJ (2019) Evaluation of app-embedded disease scales for aiding visual severity estimation of *Cercospora* leaf spot of table beet. *Plant Disease* 103:1347–1356
- Diaz CG, Bassanezi RB, Bergamin Filho A (2001) Desenvolvimento e validação de uma escala diagramática para *Xanthomonas axonopodis* pv. *phaseoli* em feijoeiro. *Summa Phytopathol* 27:35–39
- Dixon GR, Doodson JK (1971) Assessment keys for some diseases of vegetable, fodder and forage crops. *Journal of National Institute of Agricultural Botany (GB)* 23:299–307
- Domiciano GP, Duarte HSS, Moreira EN, Rodrigues FA (2014) Development and validation of a set of standard area diagrams to aid in estimation of spot blotch severity on wheat leaves. *Plant Pathology* 63:922–928
- Duarte HSS, Zambolim L, Capucho AS, Nogueira Junior AF, Rosado AWC, Cardoso CR, Paul PA, Mizubuti ESG (2013) Development and validation of a set of standard area diagrams to estimate severity of potato early blight. *European Journal Of Plant Pathology* 137: 249–257
- Everitt BS (1998) *The Cambridge dictionary of statistics*. pp 360. Cambridge University Press. Cambridge, UK
- Forbes GA, Jeger MJ (1987) Factors affecting the estimation of disease intensity in simulated plant structures. *Journal of Plant Disease Protection* 94:113–120
- Forbes GA, Korva JT (1994) The effect of using a Horsfall-Barratt scale on precision and accuracy of visual estimation of potato late blight severity in the field. *Plant Pathology* 43:675–682
- Franceschi VT, Alves KS, Mazaro SM, Godoy CV, Duarte HS, Del Ponte EM (2020) A new standard area diagram set for assessment of severity of soybean rust improves accuracy of estimates and optimizes resource use. *Plant Pathology* 69:495–505
- Gafni A, Calderon CE, Harris R, Buxdorf K, Dafa-Berger A, Zeilinger-Reichert E, Levy M (2015) Biological control of the cucurbit powdery mildew pathogen *Podosphaera xanthii* by means of the epiphytic fungus *Pseudozyma aphidis* and parasitism as a mode of action. *Frontiers in Plant Science* 6:132
- Gassner G (1915) Die getreideroste und ihr auftreten im subtropischen ostlichen sudamerika. *Ctbl Bakt* 44:305–381
- Giglioti EA, Canteri MG (1998) Desenvolvimento de software e escala diagramática para seleção e treinamento de avaliadores da severidade do complexo broca-podridões em cana-de-açúcar. *Fitopatologica Brasileira* 23:359–363
- Godoy CV, Carneiro SM, Iamauri MT, Dalla-Pria M, Amorim L, Berger RD, Bergamin Filho A (1997) Diagrammatic scales for bean diseases: development and validation. *Journal of Plant Disease Protection* 104:336–345
- Godoy CV, Koga LJ, Canteri MG (2006) Diagrammatic scale for assessment of soybean rust severity. *Fitopatologica Brasiliensis*. 31:63–68
- Gottwald TR, Aubert B, Xue-Yaun Z (1989) Preliminary analysis of citrus greening (Huanglungbin) epidemics in the People's Republic of China and French Reunion Island. *Phytopathology* 79: 687–693
- Hartung K, Piepho H-P (2007) Are ordinal rating scales better than percent ratings? – a statistical and “psychological” view. *Euphytica* 155:15–26
- Hau B, Kranz J, König R (1989) Fehler beim Schätzen von Befallsstärken bei Pflanzenkrankheiten. *Zeitschrift für Pflanzenkrankheiten und Pflanzenschutz* 96:649–674
- Hebert TT (1982) The rationale for the Horsfall-Barratt plant disease assessment scale. *Phytopathology* 72:1269
- Hock J, Kranz J, Renfro BL (1992) Tests of standard diagrams for field use in assessing the tar spot disease complex of maize. *Tropical Pest Management* 38:314–318
- Hollis JP (1984) The Horsfall-Barratt grading system. *Plant Pathology* 33:145–146

- Horsfall JG, Barratt RW (1945) An improved grading system for measuring plant disease. *Phytopathology* 35:655 (Abstract)
- Horsfall JG, and Cowling EB (1978) Pathometry: the measurement of plant disease (pp 120–136). In: *Plant disease: an advanced treatise*. Vol II. J. G. Horsfall and E. B. Cowling, (eds.). Academic Press, New York
- Houngue JA, Zandjanakou-Tachin M, Ngalle HB, Pita JS, Cacaï G, Ngatat SE, Bell JM, Ahanhanzo C (2019) Evaluation of resistance to cassava mosaic disease in selected African cassava cultivars using combined molecular and greenhouse grafting tools. *Physiological and Molecular Plant Pathology* 105:47–53
- Hunter RE, Roberts DD (1978) A disease grading system for pecan scab [*Fusicladium effusum*]. *Pecan Quarterly* 12:3–6
- Jackson EW, Obert DE, Menz M, Hu G, Avant JB, Chong J, Bonman JM (2007) Characterization and mapping oat crown rust resistance using three assessment methods. *Phytopathology* 97:1063–1070
- James WC (1971) An illustrated series of assessment keys for plant diseases, their preparation and usage. *Canadian Plant Disease Survey* 51:39–65
- James WC (1974) Assessment of plant disease losses. *Annual Review of Phytopathology* 12:27–48
- Koch H, Hau B (1980) Ein psychologischer aspect beim schätzen von pflanzenkrankheiten. *Zeitschrift für Pflanzenkrankheiten und Pflanzenschutz* 87:587–593
- Koitaishi M (2005) New biocontrol method for parsley powdery mildew by the antifungal volatiles-producing fungus Kyu-W63. *Journal of General Plant Pathology* 71:280–284
- Kranz J (1970) Schätzklassen für Krankheitsbefall. *Phytopathology Z* 69:131–139
- Kranz J (1977) A study on maximum severity in plant disease. *Travaux dédiés à G. Viennot-Bourgin, Société Française de Phytopathologie, Paris 1977*:169–173
- Kranz J (1988) Measuring plant disease. Pages 35–50 In: *Experimental techniques in plant disease epidemiology* (J.Kranz and J. Rotem, eds.), Springer-Verlag, New York
- Lage DAC, Marouelli WA, Duarte HSS, Café-Filho AC (2015) Standard area diagrams for assessment of powdery mildew severity on tomato leaves and leaflets. *Crop Protection* 67:26–34
- Large EC (1953) Some recent developments in fungus disease survey work in England and Wales. *Annals of Applied Biology* 40:594–599
- Large EC (1955) Methods of plant disease measurement and forecasting in Great Britain. *Annals of Applied Biology* 42:344–354
- Large EC (1966) Measuring plant disease. *Annual Review Phytopathology* 4:9–26
- Leite MVBCL, Amorim L (2002) Elaboração e validação de escala diagramática para mancha de *Alternaria* em girassol. *Summa Phytopathology* 28:14–19
- Madden LV, Hughes G, van den Bosch F (2007) *The study of plant disease epidemics*. APS Press, St. Paul, MN
- Marsh RW, Martin H, Munson RG (1937) Studies upon the copper fungicides. III. The distribution of fungicidal properties among certain copper compounds. *Annals of Applied Biology* 24:853–866
- McBride GB (2005) A proposal for strength-of-agreement criteria for Lin's concordance correlation coefficient. *NIWA Client ReportHAM 62* <https://www.medcalc.org/download/pdf/McBride2005.pdf>
- McKinney HH (1923) Influence of soil temperature and moisture on infection of wheat seedlings by *Helminthosporium sativum*. *Journal of Agricultural Research* 26:195–218
- Melchers LE, Parker JH (1922) Rust resistance in winter wheat varieties. U.S. Department of Agriculture, Washington DC. *Bulletin* 1046. 32 p
- Melo VP, Mendonça ACS, Souza HS, Gabriel LC, Bock CH, Eaton MJ, Schwan-Estrada KR, Nunes WMC (2020) Reproducibility of the development and validation process of standard area diagram by two laboratories: an example using the *Botrytis cinerea*/*Gerbera jamesonii* pathosystem. *Plant Disease* 104:2440–2448
- Michereff SM, Pedrosa RA, Noronha MDA, Martins RB, Silva FV (1998) Escala diagramática e tamanho de amostras para avaliação da severidade da mancha parda da mandioca (*Cercosporidium henningsii*). *Agrotropica*. 10:143–148
- Michereff SM, Maffia LA, Noronha MDA (2000) Escala diagramática para avaliação da severidade da queima das folhas do inhame. *Fitopatologica Brasiliensis* 25:612–619
- Newton AC, Hackett CA (1994) Subjective components of mildew assessment on spring barley. *European Journal of Plant Pathology* 100:395–412
- Nita M, Ellis MA, Madden LV (2003) Reliability and accuracy of visual estimation of Phomopsis leaf blight of strawberry. *Phytopathology* 93:995–1005
- Nsabiya V, Ochwo-Ssemakula M, Sseruwagi P (2012) Hot pepper reaction to field diseases. *African Crop Science Journal* 20:77–97
- Nutter FW Jr (1999) Disease assessment theory and practice: “what we think we see is what we get”. *Fitopatologica Brasiliensis (Supplement)* 24:229–231
- Nutter FW, Jr (2001) Disease assessment. Pages 312–323. In: *Encyclopedia of plant pathology*, O. C. Maloy and T. D. Murray, eds. John Wiley and Sons, Inc., New York, NY. Pages 229–326
- Nutter FW Jr, Esker PD (2006) The role of psychophysics in phytopathology. *European Journal of Plant Pathology* 114:199–213
- Nutter FW Jr, Litwiller D (1998) A computer program to generate standard area diagrams to aid raters in assessing disease severity. *Phytopathology* 88:S117
- Nutter FW Jr, Schultz PM (1995) Improving the accuracy and precision of disease assessments: selection of methods and use of computer-aided training programs. *Canadian Journal of Plant Pathology* 17:174–185
- Nutter FW Jr, and Worawitlikit O (1989) Disease.Pro: a computer program for evaluating and improving a person ability to assess disease proportion. *Phytopathology* 79: 1135 (Abstract)
- Nutter FW Jr, Teng PS, Shokes FM (1991) Disease assessment terms and concepts. *Plant Disease* 75:1187–1188
- Nutter FW Jr, Gleason ML, Jenco JH, Christians NL (1993) Accuracy, intrarater repeatability, and interrater reliability of disease assessment systems. *Phytopathology* 83:806–812
- Nutter FW Jr, Miller DL, Wegulo SN (1998) Do standard diagrams improve the accuracy and precision of disease assessment. 7th International Congress of Plant Pathology Edinburgh v 2:2.1.16
- Nutter FW Jr, Esker PD, Coelho Netto RA (2006) Disease assessment concepts in plant pathology. *European Journal of Plant Pathology* 115:95–103
- O’Brein RD, van Bruggen AHC (1992) Accuracy, precision, and correlation to yield loss of disease severity scales for corky root of lettuce. *Phytopathology* 82:91–96
- Parker SR, Shaw MW, Royle DJ (1995a) The reliability of visual estimates of disease severity on cereal leaves. *Plant Pathol* 44:856–864
- Parker SR, Shaw MW, Royle DJ (1995b) Reliable measurement of disease severity. *Aspect of Applied Biology* 43:205–214
- Pedrosa C, Lage DAC, Henz GP, Café-Filho AC (2011) Development and validation of a diagrammatic scale for estimation of anthracnose on sweet pepper fruits for epidemiological studies. *Journal of Plant Pathology* 93:219–225
- Pereira WEL, de Andrade SMP, Del Ponte EM et al. (2020) Severity assessment in the *Nicotiana tabacum*-*Xylella fastidiosa* subsp. *pauca* pathosystem: design and interlaboratory validation of a standard area diagram set. *Tropical Plant Pathology* 45:710–722.
- Peterson RF, Campbell AB, Hannah AE (1948) A diagrammatic scale for estimating rust severity on leaves and stems of cereals. *Canada Journal of Research C* 26:496–500

- Pethybridge SJ, Nelson SC (2018) Estimate, a new iPad application for assessment of plant disease severity using photographic standard area diagrams. *Plant Disease* 102:276–281
- Poland JA, Nelson RJ (2011) In the eye of the beholder: the effect of rater variability and different rating scales on QTL mapping. *Phytopathology* 101:290–298
- Redman CE, King EP, Brown IF Jr (1968) Tables for converting Barratt and Horsfall rating scores to estimated mean percentages. Elanco Products, Indianapolis, IN, 8 p
- Rios JA, Debona D, Duarte HSS, Rodrigues FA (2013) Development and validation of a standard area diagram set to assess blast severity on wheat leaves. *European Journal of Plant Pathology* 136:603–611
- Sachet MR, Citadin I, Danner MA, Guerresi MT, Pertille RH (2017) DiseasePlan - a spreadsheet application for training people to assess disease severity and to assist with standard area diagram development. *Ciência Rural* 47. <https://doi.org/10.1590/0103-8478cr20160924>
- Schwanck AA, Del Ponte EM (2014) Accuracy and reliability of severity estimates using linear or logarithmic disease diagram sets in true colour or black and white: a study case for rice brown spot. *Journal of Phytopathology* 162:670–682
- Shah DA, Madden LV (2004) Nonparametric analysis of ordinal data in designed factorial experiments. *Phytopathology* 94:33–43
- Shane WW, Thomson C, and Teng PS (1985) AREAGRAM – a statistical area diagram computer program. Epidemiology Report No. 3, Dept. of Plant Pathology, University of Minnesota, St. Paul, 34
- Sherwood RT, Berg CC, Hoover MR, Zeiders KE (1983) Illusions in visual assessment of *Stagonospora* leaf spot of orchardgrass. *Phytopathology* 73:173–177
- Shokes FM, Berger RD, Smith DH, Rasp JM (1987) Reliability of disease assessment procedures: a case study with late spot of peanut. *Oleagineux* 42:245–251
- Spolti P, Schneider L, Sanhueza RMV, Batzer JC, Gleason ML, Del Ponte EM (2011) Improving sooty blotch and flyspeck severity estimation on apple fruit with the aid of standard area diagrams. *European Journal of Plant Pathology* 129:21–29
- Stevens SS (1946) On the theory of scales of measurement. *Science* 103: 677–680
- Tehon LR, Stout GL (1930) Epidemic diseases of fruit trees in Illinois., 1922–1928. Illinois Natural History Survey, Bulletin 18:415–502
- Todd LA, Kommedahl T (1994) Image analysis and visual estimates for evaluating disease reactions of corn to *Fusarium* stalk rot. *Plant Disease* 78:876–878
- Tomerlin JR, Howell TA (1988) DISTRAIN: a computer program for training people to estimate disease severity on cereal leaves. *Plant Disease* 72:455–459
- Trumblower JA (1934) Control of elm leaf spots in nurseries. *Phytopathology* 24:62–73
- Ullstrup AJ, Elliott C, Hoppe PE (1945) Report of the committee on methods for reporting corn disease ratings. Mimeographed. Unnumbered Publication of the USDA, Division of Cereal Crops and Diseases, p 5
- Weber GE, Jorg E (1991) Errors in disease assessment – a survey. *Phytopathology* 81:S1238
- Xu Y, Kang D, Shi Z, Shen H, Wehner T (2004) Inheritance of resistance to zucchini yellow mosaic virus and watermelon mosaic virus in watermelon. *Journal of Heredity* 95:498–502
- Yadav NVS, de Vos SM, Bock CH, Wood BW (2013) Development and validation of standard area diagrams to aid assessment of pecan scab symptoms on fruit. *Plant Pathology* 62:325–335
- Zadoks JC, Schein RD (1979) *Epidemiology and plant disease management*. Oxford University Press, New York, p 429

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.