Does the *P* Value Have a Future in Plant Pathology?

L. V. Madden, D. A. Shah, and P. D. Esker

First author: Department of Plant Pathology, The Ohio State University, Wooster 44691; second author: Department of Plant Pathology, Kansas State University, Manhattan 66506; and third author: Center for Research in Plant Protection, School of Agronomy, University of Costa Rica, San Pedro Montes de Oca, Costa Rica.

Accepted for publication 24 August 2015.

ABSTRACT

Madden, L. V., Shah, D. A., and Esker, P. D. 2015. Does the *P* value have a future in plant pathology? Phytopathology 105:1400-1407.

The *P* value (significance level) is possibly the mostly widely used, and also misused, quantity in data analysis. *P* has been heavily criticized on philosophical and theoretical grounds, especially from a Bayesian perspective. In contrast, a properly interpreted *P* has been strongly defended as a measure of evidence against the null hypothesis, H_0. We discuss the meaning of *P* and null-hypothesis statistical testing, and present some key arguments concerning their use. *P* is the probability of observing data as extreme as, or more extreme than, the data actually observed, conditional on H_0 being true. However, *P* is often mistakenly equated with the posterior probability that H_0 is true conditional on the data, which can lead to exaggerated claims about the effect of a treatment, experimental factor or interaction. Fortunately, a lower bound for the posterior probability of H_0 can be approximated using *P* and the prior probability that H_0 is true. When one is completely uncertain about the truth of H_0 before an experiment (i.e., when the prior probability of H_0 is 0.5), the posterior probability of H_0 is much higher than *P*, which means that one needs *P* values lower than typically accepted for statistical significance (e.g., $P = 0.05$) for strong evidence against H_0. When properly interpreted, we support the continued use of *P* as one component of a data analysis that emphasizes data visualization and estimation of effect sizes (treatment effects).

P values are ubiquitous in research. Typically, investigators will do a study, calculate a test statistic (*T*) under the assumption that some null hypothesis (H_0) is true, and then either report the achieved *P* value (significance level) for *T* (e.g., $P = 0.024$), or report that *P* is less than a preassigned critical probability (e.g., $P < 0.05$). A nominally “publishable” result occurs when a “low” *P* value is found, typically less than 0.05. This form of frequentist inference, commonly called null hypothesis significance testing (NHST), grew out of the blending of concepts originating with Fisher (1925) and Neyman and Pearson (1928); see Schneider (2015) for a comprehensive discussion of the historical context. NHST was never intended to provide the final decision or culmination of a research project, and in fact, *Phytopathology* cautions authors against that fallacy in its Instructions to Authors. Put another way, NHST is a means to an end, not the end itself.

Significance testing was criticized relatively soon after its introduction, the debate continuing to the present day (Schneider 2015), with a “litany of criticisms repeatedly raised regarding statistical significance tests” (Mayo 2013). The scientific worth of *P* values continues to be heavily judged (sometimes acrimoniously), often from a Bayesian perspective (Anderson et al. 2000; Goodman 1999a, 1999b), using arguments that go back at least to Jeffreys (1961). More sweeping condemnations have been made from philosophical and logical grounds that do not rely on Bayesian arguments (Schneider 2015). Interested readers should consult Schneider (2015) for a detailed presentation on the major arguments against *P* values. In contrast, there are rigorous defenses of the frequentist viewpoint of *P* values (Barber and Ogle 2014; Hurlbert and Lombardi 2009; Spanos 2013). The debate has spilled over from statistics into the experimental and observational sciences. For example, the March 2014 issue of *Ecology* started with an article in strong support of the use of *P* values (Murtaugh 2014), followed immediately by arguments against their use (Burnham and Anderson 2014). While waiting for philosophers, mathematicians, and statisticians to iron out the shortcomings of both frequentist and Bayesian null-hypothesis testing (Grendár 2012; Mayo 2013; Robert 2014; Schneider 2015; Spanos 2013), or to offer something better, many researchers have sought a compromise by recognizing the inherent dangers of blind use of *P* values as the principal form of inference (Aarts et al. 2012; Kuss and Stang 2012; McBride et al. 2014). Fortunately, Bayesian concepts offer something in the way of a solution (Diniz et al. 2012; Sellke et al. 2001), a solution that we advocate (see below).

Notwithstanding philosophical and logical considerations, the entire debate is further clouded by serious misinterpretations of the meaning of the *P* value in experimental research (Berger and Sellke 1987; Goodman 2008; Hurlbert and Lombardi 2009; Schabenberger and Pierce 2002). Even statistical textbooks do not always agree in their definitions or explanations of *P* (Freund and Perles 1993). The *P* value “is arguably the most used and most misused quantity in all of statistical practice” (Littell et al. 2006); or as Schabenberger and Pierce (2002) put it: “The ubiquitous *p*-values are probably the most misunderstood and misinterpreted quantities in applied statistical work.” Persistent misinterpretations, abuse and misuse of *P* values have led to calls by some for their banishment from journals, though others suggested that railing against significance tests “…is not worth taking seriously” (Boruch 2007). The journal *Epidemiology* did try (unsuccessfully) banning *P* values some years ago (Lang et al. 1998). In 2010, the editors of *European Journal of Clinical Investigation* strongly discouraged significance testing in submitted papers, instead advocating Bayesian methods (Ioannidis et al. 2010). Most recently, the journal *Basic and Applied Social Psychology* banned all use of *P* values (Trafimow and Marks 2015), and most other forms of statistical analysis, sending ripples throughout the scientific community (Anonymous 2015; Leek and Peng 2015). Along with others (Leek and Peng 2015), we feel this view is far too extreme and not necessary.

Corresponding author: L. V. Madden; E-mail address: madden.1@osu.edu

*The ***e*-X**tra logo stands for “electronic extra” and indicates that one supplementary file is published online.

http://dx.doi.org/10.1094/PHYTO-07-15-0165-LE

Plant pathologists may view this latest round of P value criticism with anything from indifference to alarm. If anything, we do hope the new round of attention will increase awareness within our own discipline of the very likely misuse of the calculated P value in plant pathological science. Plant pathologists may legitimately question whether they ought to continue using classical significance (or hypothesis) testing, or if we should consider alternative methods of statistical inference. They may also wonder whether published P values ought to be interpreted in some different way. This letter attempts to address these issues. We first explain the actual meaning of the P value for significance testing, and then present an easy-to-use protocol for interpreting the magnitude of calculated values based on some straightforward Bayesian calculus (without adopting a fully Bayesian methodology for analysis). We try to show in this letter that P values—when used in conjunction with parameter estimation, confidence-interval calculations, and good statistical practices—have a valuable role in the analysis of data and the presentation of results, and that a better understanding of the P value will reduce the misinterpretations.

HYPOTHESES, STATISTICAL TESTS, AND SIGNIFICANCE LEVEL

P values exist in the framework of frequentist statistical significance testing. Going back to the notions originally developed by Fisher (1925), and following standard convention, we define a hypothesis of *no effect* (e.g., no treatment effect on a response variable, no correlation between two variables, no interaction of two or more factors, and so on), and call this the null hypothesis (H_0). The null hypothesis is, in general, not the scientific hypothesis of interest; in fact, H_0 is usually the opposite of the scientific hypothesis. For instance, we may hypothesize that a biocontrol agent has an effect on disease severity compared with the control. So, we test this by defining the null hypothesis as "the biocontrol agent has no effect on disease". We then collect data and determine (using a test statistic, which is a function of the data) the extent to which the data are consistent with H_0. Following Fisher (1925), we infer that H_0 is not supported when the test statistic is large and P is less than some small constant (O'Brien and Castelloe 2007). This is known as significance testing. We formally decide to reject H_0 if P *is less than or equal to* a specified small constant, which now moves us into the Neyman-Pearson hypothesis-testing paradigm (Box 1).

When we falsely reject H_0 (given that H_0 is actually true), we commit a so-called Type I error. The probability of rejecting H_0 when it is true is given as α; good statistical testing procedures are designed to have low α (e.g., $\alpha = 0.05$ or $\alpha = 0.01$). Therefore, in terms of hypothesis testing, we reject H_0 when $P \leq \alpha$ (Schabenberger and Pierce 2002). See Box 1 for more historical background on significance and hypothesis testing.

Example. To illustrate the above concepts, we can consider a typical and simple example that is the two-group problem, where we assume (here) that the observations have a normal distribution, with the same variance for each treatment. (We are not restricted to this distribution or experimental situation.) Define μ_1 and μ_2 as parameters, which are the expected values (means) for treatments 1 and 2, and $c = \mu_1 - \mu_2$ is a contrast of the two means. c is often called an effect size or treatment effect (when the groups specifically refer to treatments), although, depending on the problem, any parameter or combination of parameters can be labeled an effect size. Sometimes, effect size refers to c divided by the standard deviation of the observations. There are many things that the investigator can—*and should*—do other than hypothesis and significance testing, and certainly before testing (Schabenberger and Pierce 2002). Examples include graphing the data (e.g., box plots of the observations), and estimating the effect size and determining its confidence interval. Since this letter is discussing the meaning of P, we focus only on testing H_0.

Given the definition of c in the previous paragraph, we can define H_0 as follows:

$$H_0 : c = 0 \tag{1}$$

Note that c could be any type of parameter (e.g., a mean, mean difference, more complex contrast involving multiple means, a slope, correlation, or even a collection of parameters in a vector [as in a factorial design with multiple factors]). We can write an alternative to the null, known as the alternative hypothesis (H_a). The commonly used alternative is the general one where c is not equal to 0 (i.e., where $\mu_1 \neq \mu_2$):

$$H_a : c \neq 0 \tag{2}$$

We could be more precise and state that c is a specific nonzero constant (i.e., that one mean is different from the other by a fixed amount), but we do not pursue this situation here.

For this letter, we focus on versions of the simple null and alternative hypotheses given in Equations 1 and 2. Based on our

BOX 1

Amalgamation of the Fisher and Neyman-Pearson theories

Modern frequentist statistical inference is an amalgamation of two independent theories, one proposed by Fisher (1925) and the other by Neyman and Pearson (1928), that were considered to be incompatible by their respective authors. In fact, these authors were vehemently critical of each other over many years, and likely would never have accepted the ultimate hybridization of their approaches (Berger 2003; Schneider 2015). Fisher's approach, known as significance testing, involved specifying H_0 and calculating P, the latter indicating a subjective measure of evidence against H_0. In Fisher's formulation, there is no H_a, which may come as a surprise to many, and no formal decision point (e.g., $\alpha = 0.05$) for rejecting or not rejecting H_0, although later Fisher did consider 0.05 an important value for P.

The Neyman-Pearson approach, often known as hypothesis testing, was explicitly decision-analytic, requiring an acceptance of one of two competing hypotheses (which can be H_0 and H_a). Emphasis was on minimizing the long-run Type II error rate (β; the probability of accepting H_0 when it is false) subject to prespecified constraints on the Type I error rate (probability of rejecting H_0 when it is true). Decision thresholds were chosen based on critical values of the statistical distribution of the test statistic. The frequentist notion of hypothetical (infinite) repetitions of the same experiment is manifested here, as well as the idea of statistical power (probability of rejecting H_0 when it is false).

There are very deep statistical and philosophical differences underlying the two approaches, as discussed by Berger (2003), Schneider (2015), and many others. It is actually not clear when exactly, who or how the theories merged in practice. However, for practical data analysis, significance and hypothesis testing coincide when one calculates the P random variable (for strength of evidence) and then formally rejects H_0 when $P \leq \alpha$ (Schabenberger and Pierce 2002). We mostly follow the amalgam here. However, we place much less emphasis on rejecting a hypothesis or not, and instead, emphasize the use of P (and related measures that can be derived, in part, from P) to assess the strength of evidence for or against H_0. When not using the realized P value as the criterion for hypothesis acceptance or rejection, we are more in tune with the ideas of Fisher.

experience, this is the most common situation in plant pathology. There are other scenarios that could be considered, such as one-sided tests (which can be expressed as a directional alternative hypothesis; e.g., $c > 0$). Even more complicated hypotheses can be defined to test for so-called superiority, noninferiority, or equivalence (Garrett 1997; McBride et al. 2014). We do not discuss these here because of their rarity in plant pathology.

We consider, for convenience, a very simple situation where we can utilize properties of the normal distribution to test H_0. With data and statistical assumptions for the distribution of the data, we can test the null hypothesis using a Student *t* test, preferably only after looking at the data, estimating (and looking at) *c*, and calculating a confidence interval. If one had the following data points (treatment A: 22.8, 23.1, 21.8, 20.5, 19.7, and 21.4; treatment B: 19.2, 15.6, 21.2, 17.2, 17.4, and 21.7), and performed a *t* test, one would obtain a test statistic of $t = 2.53$, df = 10, and $P = 0.030$ in the output (with a single pooled variance). Based on standard frequentist notions (which are not accepted by some, as discussed above and below), the small *P* is giving a *form* of evidence against H_0 with this data set (Berger 2003; Boos and Stefanski 2011; Sellke et al. 2001), an indication that the treatment means are different. Intuitively, the larger the difference between the means, the smaller the *P* value, but this is not necessarily the case. So, does *P* tell us directly about the *strength* of the evidence against H_0? Before attempting to answer this question, we need to revisit our understanding of what *P* is (and is not).

WHAT IS *P*?

Definition. There are many misconceptions about the meaning of the *P* value (Box 2; Goodman 2008). *P* is a probability. In particular, *P* is the probability of observing a test statistic *T* as extreme as, or more extreme than, the one computed from the current data [*T*(*data*)] when H_0 is true (Freund and Perles 1993; Littell et al. 2006). Equivalently, *P* can be viewed as the conditional probability of observing data as extreme as, or more extreme than, the data actually observed, when H_0 is true. This concept can be expressed compactly as $\Pr(data|H_0)$. Given that *T*(*data*) is just a single number, where does the probability notion come from? More specifically, where does the idea of "more extreme than" originate? With continuous data, *any* value of *T*(*data*) under the null hypothesis is possible (though not necessarily likely). The probability is determined from hypothetical repetitions of the experiment in which H_0 is actually true. To better understand how statistical theory works to get this probability, one can envision that the same experiment is independently repeated a very large (infinite) number of times under identical conditions (same treatments, sample sizes, and so on), all when H_0 is true; *T* is calculated for each repetition, and the proportion with an equal or more extreme value than the observed statistic *T*(*data*) is determined. This proportion is *P*. In other words, the frequency of extreme (hypothetical) values is used to determine *P*; this is why, in part, the label "frequentist" is given to this branch of statistical philosophy and methodology. The less consistent the data are with H_0, the more extreme *T*(*data*) will be, and the smaller that *P* will be.

BOX 2

What *P* is not: Corrections to some common misconceptions about *P* values (after Goodman 2008; Littell et al. 2006; Schabenberger and Pierce 2002)

- A *P* value is not the probability that the null hypothesis (H_0) is true.
- A *P* value is not the error probability of rejecting the null hypothesis.
- A *P* value is not the probability of falsely rejecting the null hypothesis.
- A *P* value is not a measure of the probability that the null hypothesis is wrong.
- A *P* value is not the probability of a Type I error.
- A small *P* value does not necessarily mean that the alternative hypothesis is true.
- A small *P* values does not necessarily mean that the result is biologically significant.

Of course, one only has the single set of experimental data at hand, not an infinite number of hypothetical replications. While this would seem to create a quandary, theoretical work based on the principles of randomization (Fisher 1922, 1925; Schabenberger and Pierce 2002), has shown that the Normal, *t*, *F*, and χ^2 distribution functions, depending on the situation, appropriately represent the test statistic over these hypothetical repetitions. Furthermore, for certain discrete data, various permutation tests can be used to determine the *P* value without reliance on a theoretical distribution under H_0.

This basic approach to significance (or hypothesis) testing is routinely criticized by Bayesians (Goodman 1999a, 1999b). As Jeffreys (1961) wrote provocatively, "a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred." Nevertheless, the *P* value has been used across all realms of science over the better part of a century, and we take it as axiomatic that science is progressing overall, so we are not prepared to reject the notion of *P* as a useful concept (when used and interpreted in a reasonable fashion). It should be pointed out that Hurlbert and Lombardi (2009) do not believe that one needs to resort to the notion of unobserved observations to justify or properly interpret *P* values.

***P* is conditional.** The *P* probability value is a random variable that is conditional on several factors. To reiterate, the *P* value is conditional on the null hypothesis being true [$\Pr(data|H_0)$]. In fact, under H_0, *P* has a uniform, continuous distribution bounded by 0 and 1 (Boos and Stefanski 2011). *P* is implicitly dependent on the statistical model used for the data. For instance, one would get different *P* values if one analyzed disease incidence data using no transformation, the angular transformation (arcsine-square-root), or the logit transformation, if one used a binomial or beta-binomial distribution for the dependent variable, or if one used parametric or nonparametric analytical methods. *P* is also dependent on the experimental design (e.g., completely randomized or randomized block) and the treatment design [number of treatments (factors), crossing or nesting of treatment levels, and so on]. The so-called sampling space (the region over which the response variable varies) also affects *P*. One needs to appreciate the conditioning of *P* on all of the above factors, because when "the conditions change, the probability itself will also very likely change, often substantially" (Littell et al. 2006).

Quite importantly, the sample size (number of replications, *n*) can have a huge impact on *P*. All other things being equal, *P* declines as *n* increases. With the above example ($\hat{\mu}_1 = 21.55$, $\hat{\mu}_2 = 18.72$, $\hat{c} = 2.83$, $P = 0.03$, $n = 6$ [per group]), one would find a smaller *P* if *n* equaled 8 instead of 6 (assuming all other things, such as the standard deviation and means, remain the same). Because of this sample-size conditioning, any H_0 involving a point value (e.g., $c = 0$) will be rejected ($P \leq \alpha$) with sufficiently large sample sizes (Spanos 2013). Thus, investigators always should distinguish between statistical significance and biological (or practical) significance. For instance, with (much) larger *n* in the simple two-treatment example, an estimated *c* very close to 0 could result in $P < 0.05$, even though such a small difference in means could have no biological significance or relevance.

The online supplementary file contains annotated R code and output to visualize and explore some of these properties of P given here and below.

EVIDENCE, LIKELIHOOD RATIOS, PRIOR AND POSTERIOR PROBABILITIES

Evidence for H_0? The rationale for P values as a *measure* of evidence (in the sense that "measure" conveys the notion of quantifying something that is estimable) is certainly debatable (Hubbard and Lindsay 2008; Schervish 1996; see literature review in Schneider 2015). However, within the frequentist paradigm, many do consider P values as a measures of evidence: "P values are the most commonly used tool to measure evidence against a hypothesis or hypothesized model" (Sellke et al. 2001); and "P-values are useful statistical measures of evidence against a null hypothesis" (Boos and Stefanski 2011). Intuitively, the smaller the P value, the further away the data (but not necessarily the summary effect size) are from H_0, and the "stronger" the evidence against H_0. However, because of the conditioning of P on so many factors (as discussed above), and the uncertainty at which the random variable P is determined (Boos and Stefanski 2011), one cannot quantitatively compare P values among studies (e.g., the same effect size will give two different P values with two different sample sizes).

Putting debate aside in order to move forward, the crux is that researchers often would like some idea of the probability that H_0 is true given the observed data, that is, $\Pr(H_0|data)$, which is the posterior probability of H_0 given the data. However, as discussed above, the P value gives $\Pr(data|H_0)$. The heart of the problem is when investigators mistakenly equate these two quantities and then incorrectly interpret the P value as the probability that H_0 is true (Berger 2003; Goodman 2008; Hubbard and Lindsay 2008; Schneider 2015). See Nuzzo (2015) for a very readable explanation of this all-too-common fallacy. Returning to the idea that $\Pr(H_0|data)$ is of interest from a research perspective, we *can* approximate this probability, but we need additional information or additional assumptions. Our exposition below is motivated by Littell et al. (2006, chapter 13). Similar approaches to addressing the P value topic have been taken by Nuzzo (2015) and Held (2010).

Bayes factor. To get to $\Pr(H_0|data)$, we shall make use of some Bayesian calculus, known sometimes as Bayesian updating, without utilizing a full Bayesian analysis (Berger and Sellke 1987; Good 1992). The Bayes Factor (BF) is the ratio of the likelihoods of the data under any two models for the data, such as H_0 and H_a:

$$\mathrm{BF} = \frac{\Pr(data|H_0)}{\Pr(data|H_a)} \qquad (3)$$

The numerator is the probability (likelihood) of the data given the null-hypothesis model, and the denominator is the probability (likelihood) of the data given the alternative-hypothesis model (Goodman 1999b; Sellke et al. 2001). The BF is a generalized likelihood ratio. Based on the estimated BF, the strength of evidence against H_0 can be classified as: weak (BF = 0.2 = 2/10), moderate (BF = 0.1 = 1/10), strong (BF = 0.033 = 1/30), and very strong (BF = 0.01 = 1/100) (Jeffreys 1961; Littell et al. 2006). These are subjective categories in the same sense that 0.05 and 0.01 are subjective categories for P. With a BF, however, the investigator typically is not attempting to outright accept or reject a hypothesis. The BF is very challenging to calculate, especially the denominator (Christensen et al. 2011; Gelman et al. 2014; Lesaffre and Lawson 2012). Under H_a, an indefinite (infinite) number of values for μ_1 and μ_2 are possible that satisfy the nonzero difference of means in Equation 2, although some values are more likely than others (but which ones?). The maximum-likelihood parameter estimates, for instance, are just one set of many possible values. Although Bayesians often advocate the BF (Good 1992; Goodman 1999b), it is not commonly calculated in Bayesian analyses. Rather, Bayesians usually place emphasis on the posterior distributions of parameters in a given model, and do not deal with the general alternative hypothesis.

There is a remarkable link between the BF and the P value that was shown by Sellke et al. (2001). An *approximate* lower bound for the BF (BF*) can be calculated as

$$\mathrm{BF}^* = -e \cdot P \cdot \ln(P) \qquad (4)$$

for $P < 1/e$ (~0.368), where e is the base of the natural log system. When $P > 1/e$, BF* = 1. One can consider BF* as a measure of surprise of the outcome given that H_0 was true: the smaller the BF*, the greater the surprise (Bayarri and Berger 1998; Good 1988). Figure 1 shows the relationship between BF* and P. For example, at $P = 0.01$ the BF* = 0.12, which indicates moderate evidence against the null hypothesis (and moderate surprise) based on the Jeffreys' subjective scale given above. For $P = 0.05$, there is only weak evidence, at best, against the null hypothesis, given that BF* = 0.41. We reiterate that these are lower-bound limits, and that the actual BF for any P is likely greater, although we cannot know the actual BF without more specific information about H_a (Bayarri and Berger 1998). Very strong evidence against H_0, using Jeffreys' criterion (BF = 0.01), therefore requires P values of less than 0.001 (Fig. 1), a sobering result given the fact that a majority of scientists appear to use $P = 0.05$ as the decision point (Cowles and Davis 1982).

The BF is a useful alternative to the P value, and Equation 4 shows that P actually provides a substantial amount of evidence for or against H_0 (at least as an approximation), despite criticisms of P in the literature. That is, if BF is a useful quantity, then so is P. However, as BF* is a (deterministic) nonlinear transformation of P, it still does not directly indicate $\Pr(H_0|data)$. The latter requires a consideration of posterior probabilities. As Lambdin (2012) put it: "Demonstration that [$\Pr(data|H_0)$] is low may indeed reduce [$\Pr(H_0|data)$], but it does not demonstrate that [$\Pr(H_0|data)$] *is also low*, which is what (as scientists) we would be interested in seeing…".

Posterior probability of H_0. Note that terms comprising the BF in Equation 3 represent the probability of the data given a particular hypothesis (or given a particular model). What we want here is the probability of H_0 given the data [$\Pr(H_0|data)$]. In other words we need to turn the conditional probabilities around [from $\Pr(data|H_0)$ to $\Pr(H_0|data)$]; see Madden (2006) and Madden et al. (2007) section 11.7 for examples of applying this concept with disease prediction decision tools. To accomplish this "turning around", we require the *prior probability* that the null hypothesis is true, $\Pr(H_0)$, which is an *unconditional* probability that exists *before* the experiment is performed and data collected. It could be

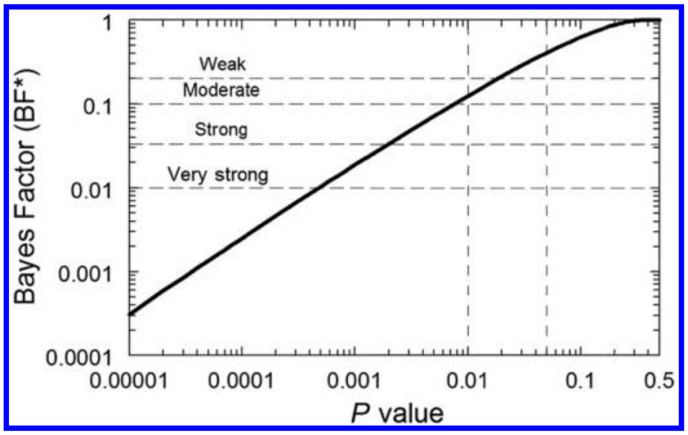

Fig. 1. Relationship between the lower bound of the Bayes Factor (BF*) and the P value (Equation 4). Commonly used thresholds of the BF* for evidence against the null hypothesis are given as horizontal lines. Vertical lines indicate two commonly used critical values of P, 0.05 and 0.01. See Sellke et al. (2001) for details.

objectively determined from previous studies that dealt with the same (or similar) set of treatments; or, it could be subjectively determined based on the expertise of the investigator. A full Bayesian approach to the problem would be to specify a distribution for the prior probability; we take the simpler, and more practical, approach of Littell et al. (2006) of assigning a point value to the probability and using Bayesian calculus (updating) to derive a point estimate of the posterior probability. Readers should see Lesaffre and Lawson (2012; section 1.3) and O'Brien and Castelloe (2007) for other uses of point values for the prior probability and Bayesian updating. Lesaffre and Lawson (2012) is also a good source for learning about fully Bayesian approaches to data analysis.

If π is a probability (of an event), then *odds* = π/(1-π). The *prior odds* of H_0 are therefore given by

$$odds(H_0) = \frac{\Pr(H_0)}{1-\Pr(H_0)}$$

The *posterior odds* of H_0, given the data are

$$odds(H_0|data) = \text{BF} \cdot odds(H_0) \tag{5}$$

Equation 5 is the fundamental step in the Bayesian updating. We substitute BF* (Equation 4) for BF in this odds equation. By inverting *odds* = π/(1-π) we obtain π = *odds*/(1+*odds*), and hence

$$\Pr(H_0|data) = \frac{odds(H_0|data)}{1+odds(H_0|data)} \tag{6}$$

One obtains the lower bound for the posterior probability by substituting Equation 5 into Equation 6, and then using BF* for BF:

$$\Pr(H_0|data) = \frac{\text{BF*} \cdot odds(H_0)}{1+\text{BF*} \cdot odds(H_0)} \tag{7}$$

The key to Equation 7 is utilization of prior knowledge about H_0. For some situations, $\Pr(H_0)$ will be quite low (e.g., <0.1), such as for the comparison of a control with a very effective treatment, when we expect the treatment to be effective. Often a generally effective treatment is included in a study in order to judge the experimental methods; the corresponding treatment effect would then have a low $\Pr(H_0)$. On the other hand, when testing a novel (potential) disease-control product, such as a possible biocontrol agent, or in screening for gene expressions (for hundreds or thousands of genes) in response to a treatment, $\Pr(H_0)$ could be quite high (e.g., >0.75). For example, one may be screening microbes isolated from the soil in order to detect possible biocontrol agents. Based on past work, we know that most agents will not provide effective control; here, $\Pr(H_0)$ for a treatment effect (product versus the control) would be high (e.g., 0.9). Ioannidis (2005), writing primarily about medical studies, has argued strongly—based on the number of failures of drugs and other medical treatments after earlier "positive" results—that $\Pr(H_0)$ is far above 1/2 for these explorative (and other) studies.

TABLE 1. Posterior probability of the null hypothesis being true (conditional on the data), $\Pr(H_0|data)$, in relation to the *P* value (significance level) and the prior probability of the null hypothesis being true, $\Pr(H_0)$[a]

	$\Pr(H_0)$						
P value	0.99	0.90	0.75	0.50	0.25	0.10	0.01
0.150	0.987	0.874	0.699	0.436	0.205	0.079	0.008
0.100	0.984	0.849	0.652	0.385	0.173	0.065	0.006
0.075	0.981	0.826	0.613	0.346	0.150	0.055	0.005
0.050	0.976	0.786	0.550	0.289	0.120	0.043	0.004
0.025	0.961	0.693	0.429	0.200	0.077	0.027	0.0025
0.010	0.925	0.530	0.273	0.111	0.040	0.014	0.0013
0.001	0.650	0.145	0.053	0.018	0.006	0.002	0.0002
0.0001	0.199	0.022	0.007	0.003	0.001	0.0003	0.00003

[a] See Equations 4 to 7 and Sellke et al. (2001) for details.

Table 1 presents the lower bounds for posterior probabilities $\Pr(H_0|data)$, for a range of prior probabilities [$\Pr(H_0)$] and *P* values. We emphasize that the calculation (using Equation 7) depends on the validity of the calculated *P* value. That is, if an inappropriate model was used for the analysis, then the *P* value would also be invalid. Likewise, if the study was poorly conducted, possibly with unsuitable experimental methods, then *P* would also be invalid. Researchers are encouraged to actively work with a statistician so that their analysis is appropriate for the distribution of the data and experimental design. These calculations are directly applicable when H_a is two-sided (as in "not equal") and when H_0 is for a specific constant (e.g., $c = 0$; Equation 1). Numerical simulations show that the general approach is also valid for the "negligible" small-interval null hypothesis (e.g., H_0: $|c| < \varepsilon$, where ε is a very small constant) (Sellke et al. 2001).

Consider when the prior probability is 0.5, so that we are ambivalent about the truth of H_0 before the experiment (this is equivalent to both hypotheses having equal prior probability of being true). Suppose that after analyzing the data we obtained $P = 0.048$. We can either use Equation 7 directly, or assume that *P* is close enough to 0.05 and use Table 1; we do the latter here. For this situation, despite the significant result by the classical NHST analysis, $\Pr(H_0 \mid data)$ is 0.289. Therefore there is nearly a 30% chance after the experiment that H_0 is actually true despite this "significant" result (with this ambivalent prior). Although this result still favors the alternative hypothesis of nonequality of the two means, in the sense that $\Pr(H_0|data)$ is less than 0.5, evidence is clearly not as strong as one might suspect based on the naïve (incorrect) interpretation of *P*.

Returning to Bayesian arguments and moving down the $\Pr(H_0) = 0.50$ column in Table 1 (where H_0 and H_a have equal prior probability), we see there is still about an 11% chance that H_0 is true when $P = 0.01$ (a *P* value often considered "highly significant" in frequentist analysis); *P* needs to be less than 0.01 before the chance of H_0 being true drops below 5% (specifically, $P \approx 0.0035$ based on Equation 7). Suppose now that one is testing a new treatment (or treatments) for disease control. Assume we have no specific prior knowledge about the treatment; but that past screenings of similar products found that about three-quarters of those had no effect. Then, it is reasonable to define $\Pr(H_0) = 0.75$ as a point estimate of the prior. If a data analysis results in $P \approx 0.05$, the posterior probability of H_0 is 0.55 based on the Bayesian updating of Equation 7; that is, there is greater than a 50% chance that the treatment truly has no effect when $P = 0.05$ (Table 1). A reasonably small posterior probability [say, $\Pr(H_0|data) = 0.05$] is only found when $P \leq 0.001$ when $\Pr(H_0) = 0.75$. If the prior was even higher, say, $\Pr(H_0) = 0.90$, so that one clearly does not think that a treatment can be effective, then the posterior probability of the null hypothesis being true is a high value of 0.79 when the frequentist test is "just" significant ($P \approx 0.05$). This is why Ioannidis (2005) claimed that most published research findings in medical science are wrong. Others do not accept Ioannidis' claim (e.g., Samsa 2015), but he does have a good point, in that "positive" frequentist results (i.e., $P \leq 0.05$) of individual studies can be quite misleading, especially when $\Pr(H_0)$ is high.

Now consider situations where $\Pr(H_0)$ is low, where we expect to find an effect (a nonzero effect size, treatment effect, or interaction). Suppose we are considering the correlation between incidence and severity measurements of disease intensity. Based on prior epidemiological theory and practice (Madden et al. 2007), we believe the prior probability of no correlation (after transformation to a linear scale) is low; we shall use $\Pr(H_0) = 0.10$. After analyzing a data set, we obtain $P = 0.059$ (which we round down to 0.05 to use Table 1). The posterior probability of H_0 (no relationship) is 0.043, giving reasonably strong evidence against H_0 (in favor of H_a). Even with *P* values between 0.075 and 0.10, the posterior probability of H_0 remains reasonably low (between 0.055 and 0.065). When the prior probability of the null hypothesis is even lower

[$\Pr(H_0) = 0.01$], all P values result in small posterior probabilities (i.e., strong evidence against the null).

We now return to our original two-treatment simple example ($P = 0.03$) and apply Equation 7 directly (rather than using the table). With the ambivalent $\Pr(H_0) = 0.50$, the posterior probability of the null hypothesis, $\Pr(H_0|data)$, is 0.22, indicating only a little evidence against the null hypothesis, even though P is less than 0.05. Even when $\Pr(H_0) = 0.25$, the posterior probability is still above 0.05 [specifically, $\Pr(H_0|data) = 0.087$].

To summarize, Table 1 is a very useful tool for helping interpret analytical results. Figure 2 plots the posterior probabilities for a wider range of P values between 0.00001 and 0.6, which may help readers quickly calibrate between the two probabilities. The online supplemental file gives the R code to produce the posterior probabilities for selected P and $\Pr(H_0)$. There are several other interrelated, and more complex, ways of addressing inference regarding the null hypothesis that we do not cover here. Sellke et al. (2001), Berger (2003), Berger and Sellke (1987), and Bayarri and Berger (1998) thoroughly describe some of the methods, giving frequentist and Bayesian arguments. Note that some approaches are concerned with calculating $\Pr(H_0 \mid P \leq 0.05)$, which is *not* the same thing as in Equation 7 (O'Brien and Castelloe 2007; Sellke et al. 2001). $\Pr(H_0 \mid P \leq 0.05)$ indicates the posterior probability of H_0 conditional on the decision that the null hypothesis is false (i.e., conditional on finding *any* significant result [using 0.05 as the threshold]). The actual value of P could be 0.049, 0.01, 0.00025, 10^{-5}, or any other value less than 0.05. A weighted result over all possible P values less than or equal to 0.05 has to be determined. But, all possible low P values do not occur within a single experiment, only one is realized for a given hypothesis. As such, in this letter we are interested in the posterior probability conditional just on a specific P value (i.e., conditional on the specific data being analyzed).

Use of Table 1 and Figures 1 and 2 requires that one is following good statistical practices (Kirk 2001). For instance, it is improper to try several different statistical models or procedures in a search for the one that gives the *desired* result (e.g., small P for one's favorite treatment). This is one specific aspect of "p-hacking". Likewise, it would be inappropriate to choose the prior probability that gives the *desired* posterior probability of H_0. Rather, Table 1 and the figures are meant to inform the investigator and the reader, within the context of the experiment and system being studied.

DISCUSSION

> "The thing that's unusual about good scientists is that they're not so sure of themselves as others usually are. They can live with steady doubt, think 'maybe it's so' and act on that, all the time knowing it's only 'maybe'." (Feynman 1999).

This quote from Richard Feynman, borrowed from O'Brien and Castelloe (2007), describes the situation we explore in this letter. Based on the body of knowledge and observation, plant pathologists conduct experiments (or surveys) to confirm past work, gain insights, or make new discoveries. As scientists we appreciate there being the chance that the inference made from a single study will be incorrect, and hope that inferences are mostly right in the long run, as earlier errors are corrected by the total collection of conducted studies. Data analysis is an important part of this process, and the calculation of P values is one very common component of data analysis that can increase our knowledge of a system or process based on the results (Mudge 2013). Other approaches, such as Bayesian analysis or direct use of likelihood ratios, are also possible and useful.

Nevertheless, P values (and NHST) have been harshly criticized over many decades (e.g., Anderson et al. 2000; Goodman 1999a,b; Schneider 2015), from their philosophical and theoretical foundations to their abuse and misinterpretation by applied practitioners. Others are quite vigorous in their defense of P values, especially when they are not used strictly for NHST, but for assessing evidence against the null hypothesis (Hagen 1997; Hurlbert and Lombardi 2009; Mulaik et al. 1997; Murtaugh 2014). There is no shortage of (sometimes acrimonious) back-and-forth debate over P values and their place in science. Yet, statisticians around the world continue to develop or expand on theory and methodology that rely on the principles of frequentist statistics (and resulting P values); these statisticians clearly have not been swayed by the recurring arguments in the literature. Plant pathologists as applied practitioners may (rightly so) feel overwhelmed. However, we strongly feel that use of P is not likely to disappear, despite viewpoints to the contrary (Trafimow and Marks 2015).

P values clearly are widely misunderstood across all disciplines (Berger 2003; Hupe 2015; Lambdin 2012; McBride et al. 2014; Murtaugh 2014), but this does not need to be the case going forward. Scientists often want to know $\Pr(H_0|data)$; that is, an estimate of the probability of the null hypothesis being true, given the data, is often the desired outcome. However, through the use of frequentist-based statistical testing, the opposite conditional probability, $\Pr(data|H_0)$, the probability of the data given that the null hypothesis is true, is estimated after the collection and analysis of data (Stang and Poole 2013). Then, unfortunately, the latter is misinterpreted as $\Pr(H_0|data)$ (Lambdin 2012; Nuzzo 2015) after the results are in. It can be easily argued that this has led to many false claims in the literature. Fortunately, researchers are not at an impasse with these conditional probabilities. The research by Sellke et al. (2001) provides an approximate lower bound for the Bayes Factor (BF*) for the strength of evidence for (or against) H_0 and the resulting posterior probability for H_0 (Equations 3 and 7; Table 1; Fig. 2). Investigators can remain ambivalent about the prior probability of the null by choosing $\Pr(H_0) = 0.5$, or they can use their professional expertise to assign more informative priors, as demonstrated here. Clearly, use of $P = 0.05$ as the decision point provides only weak evidence against the null when $\Pr(H_0) = 0.5$ (i.e., null and alternative hypotheses are equally likely before the experiment), but it provides much stronger evidence when $\Pr(H_0)$ is small (i.e., the alternative hypothesis is more likely before the experiment). The approach also provides an intuitive framework to judge suspicious or surprising results in the literature.

To benefit from Table 1 or Equation 7 and the methodology of Sellke et al. (2001), appropriate statistical practices for the data and experimental design must be practiced so that estimated P values are valid, at least in terms of their magnitude (Boos and Stefanski 2011). Improperly analyzed data could yield unreliable P values. This could occur, for instance, if one failed to account for the correlation structure of repeated measures, or used an inappropriate

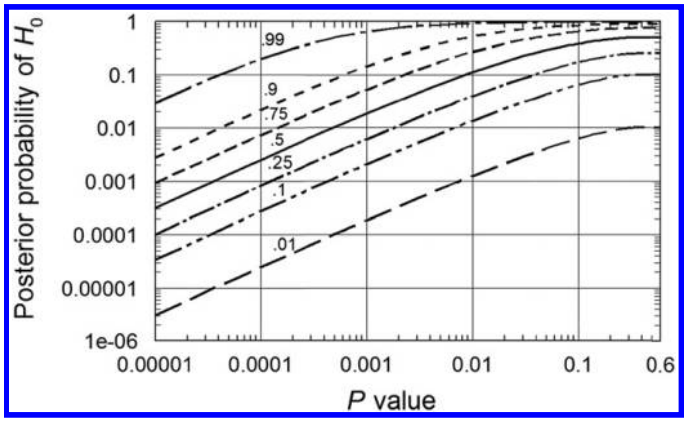

Fig. 2. Lower bound of the posterior probability of the null hypothesis conditional on the data, $\Pr(H_0|data)$, in relation to the significance level (P) and the prior probability of the null hypothesis, $\Pr(H_0)$. Curves based on Equation 7. Numbers next to lines indicate specific values of $\Pr(H_0)$. Based on Sellke et al. (2001) and Littell et al. (2006).

distribution for the data, or did not take into account the random effects, or analyzed an ordinal response variable as if it were a continuous one (Madden et al. 2007; Schabenberger and Pierce 2002; Shah and Madden 2004). Of course, all statistical analyses involve some level of approximation, and in the real world, one can never be certain that an ideal analysis has been carried out, because one does not have full information on reality. However, there are many ways to judge when an improper analysis has been conducted, or, conversely, when a *reasonable* analysis has been performed, based on statistical theory, data, and model diagnostics (Schabenberger and Pierce 2002). This requires advice from an applied statistician, which we highly recommend.

The weak evidence provided by $P = 0.05$ (Fig. 1), in general, might suggest that the critical significance level for decision making should be uniformly lower. Because the 0.05 threshold is so ingrained throughout science, it is unlikely that such a change will happen. Rather, we feel it is much more likely that investigators can become better able to judge the actual strength of evidence for the analyses conducted by them and others, based on reported P values (and the information given in Table 1 and Figures 1 and 2). In fact, there is no single critical P value (or BF or posterior probability) that one should use for all objectives. One should have very strong evidence [low P, BF*, and $\Pr(H_0|data)$] for late-stage testing of a disease-control product, where a well-supported conclusion is required for efficacy. Sample sizes (replications, blocks) typically would need to be large to achieve a low P (or low BF*) unless the effect size or treatment effect is very large (O'Brien and Castelloe 2007). Alternatively, in early exploratory studies screening to find *possible* efficacious treatments (cultivars, genes, etc.), one could use a large P value (even larger than 0.2). Here, only treatments with "favorable" results in one study are kept for later confirmatory studies. Most of these "significant" results with high P values would be incorrect, but one would be less likely to be discarding prematurely possible viable treatments (or cultivars, genes, etc.).

There is clearly an overemphasis on P values and binary decision-making in the scientific literature (Murtaugh 2014). With low-power studies, one generally finds only weak evidence against H_0, at best, when H_a is true. This occurs if one is using P directly, or BF* or $\Pr(H_0|data)$ for the measure of evidence. With high-power studies (typically, with many blocks or replications), P will often be lower, providing stronger evidence against H_0, even when the effect size (treatment effect) is small. As discussed by O'Brien and Castelloe (2007), increasing the power of a study is the best way to increase the evidence against the null hypothesis (when the null is false). However, even biologically nonimportant effects can be found to be statistically significant (or have very low BF*) with high-power studies. Thus, as stated above, one must always distinguish biological from statistical significance, and place importance on estimation of effect sizes (including their confidence intervals), rather than simply relying on NHST (Murtaugh 2014; Schabenberger and Pierce 2002). The question is no longer "Does treatment have an effect?", but "What is the magnitude of the treatment effect?". When effect sizes are available for multiple studies dealing with the same topic, it is also advisable to use meta-analysis to combine the results (Madden and Paul 2011).

In conclusion, we support the continued use of P values in plant pathology as part of the (frequentist-based) data analytical process, but also strongly support the general trend of moving away from the arbitrary "significant/not significant" decision-point cutoffs that are carryovers from the Neyman-Pearson hypothesis-testing paradigm. Instead, by using the magnitude of the calculated P value, the lower bound of the posterior probability of the null hypothesis being true, given the data, can be approximated in order to interpret the evidence against (or for) the null hypothesis in the context of the study.

ACKNOWLEDGMENTS

Work by the first author was supported by state and federal funds appropriated to the Ohio Agricultural Research and Development Center (OARDC), The Ohio State University.

LITERATURE CITED

Aarts, S., Winkens, B., and van den Akker, M. 2012. The insignificance of statistical significance. Eur. J. Gen. Pract. 18:50-52.

Anderson, D. R., Burnham, K. P., and Thompson, W. L. 2000. Null hypothesis testing: Problems, prevalence, and an alternative. J. Wildl. Manage. 64: 912-923.

Anonymous. 2015. A psychology journal bans p-values. Significance 12:6.

Barber, J. J., and Ogle, K. 2014. To *P* or not to *P?* Ecology 95:621-626.

Bayarri, M. J., and Berger, J. O. 1998. Quantifying surprise in the data and model verification. Pages 53-82 in: Bayesian Statistics 6. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, eds. Oxford University Press, Oxford.

Berger, J. O. 2003. Could Fisher, Jeffrey and Neyman have agreed on testing? Stat. Sci. 18:1-32.

Berger, J. O., and Sellke, T. 1987. Testing a point null hypothesis: the irreconcilability of the p-values and evidence (with discussion). J. Am. Stat. Assoc. 82:112-122.

Boos, D. D., and Stefanski, A. 2011. P-value precision and reproducibility. Am. Stat. 65:213-221.

Boruch, R. 2007. The null hypothesis is not called that for nothing: Statistical tests in randomized trials. J. Exp. Criminol. 3:1-20.

Burnham, K. P., and Anderson, D. R. 2014. *P* values are only an index to evidence: 20th- vs 21st-century statistical science. Ecology 95:627-630.

Christensen, R., Johnson, W., Branscum, A., and Hanson, T. E. 2011. Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians. CRC Press, Boca Raton, FL.

Cowles, M., and Davis, C. 1982. On the origins of the .05 level of statistical significance. Am. Psychol. 37:553-558.

Diniz, M., Pereira, C. A. B., Polpo, A., Stern, J. M., and Wechsler, S. 2012. Relationships between Bayesian and frequentist significance indices. Inter. J. Uncert. Quant. 2:161-172.

Feynman, R. P. 1999. The Pleasure of Finding Things Out. Perseus Books, Cambridge, MA.

Fisher, R. A. 1922. On the mathematical foundations of theoretical statistics. Phil. Trans. Royal. Soc. A. 222:309-368.

Fisher, R. A. 1925. Statistical Methods for Research Workers. 1st ed. Oliver & Boyd, London.

Freund, J. E., and Perles, B. M. 1993. Observations on the definition of P-values. Teach. Stat. 15:8-9.

Garrett, K. A. 1997. Use of statistical tests of equivalence (bioequivalence tests) in plant pathology. Phytopathology 87:372-374.

Gelman, A., Carlin, J. B., Stern, S. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. 2014. Bayesian Data Analysis. 3rd ed. CRC Press, Boca Raton, FL.

Good, I. J. 1988. Surprise index. Pages 104-109 in: Encyclopedia of Statistical Sciences. S. Kotz, N. L. Johnson, and C. B. Reid, eds. John Wiley & Sons, New York.

Good, I. J. 1992. The Bayes/Non-Bayes compromise: A brief review. J. Am. Stat. Assoc. 87:597-606.

Goodman, S. N. 1999a. Toward evidence-based medical statistics. 1: The *P* value fallacy. Ann. Intern. Med. 130:995-1004.

Goodman, S. N. 1999b. Toward evidence-based medical statistics. 2: The Bayes Factor. Ann. Intern. Med. 130:1005-1013.

Goodman, S. N. 2008. A dirty dozen: Twelve P-value misconceptions. Semin. Hematol. 45:135-140.

Grendár, M. 2012. Is the p-value a good measure of evidence? Asymptotic consistency criteria. Stat. Probab. Lett. 82:1116-1119.

Hagen, R. L. 1997. In praise of the null hypothesis statistical test. Am. Psychol. 52:15-24.

Held, L. 2010. A nomogram for P values. BMC Med. Res. Methodol. 10:21.

Hubbard, R., and Lindsay, R. M. 2008. Why P values are not a useful measure of evidence in statistical significance testing. Theory Psychol. 18:69-88.

Hupe, J.-M. 2015. Statistical inferences under the null hypothesis: common mistakes and pitfalls in neuroimaging studies. Front. Neurosci. 9:19.

Hurlbert, S. H., and Lombardi, C. 2009. Final collapse of the Neyman-Pearson decision theoretic framework and the rise of the neoFisherian. Ann. Zool. Fenn. 46:311-349.

Ioannidis, J. P. A. 2005. Why most published research findings are false. PLoS Med. 2:e124.

Ioannidis, J. P. A., Tatsioni, A., and Karassa, F. B. 2010. A vision for the European Journal of Clinical Investigation: Note from the editors. Eur. J. Clin. Invest. 40:1-3.

Jeffreys, H. 1961. Theory of Probability. 3rd ed. Oxford University Press, Oxford, UK.

Kirk, R. E. 2001. Promoting good statistical practices: Some suggestions. Educ. Psychol. Meas. 61:213-218.

Kuss, O., and Stang, A. 2012. The p-value—A well-understood and properly used statistical concept? Contact Dermat. 66:1-3.

Lambdin, C. 2012. Significance tests as sorcery: Science is empirical—significance tests are not. Theory Psychol. 22:67-90.

Lang, J. M., Rothman, K. J., and Cann, C. I. 1998. The confounded *P*-value. Epidemiology 9:7-8.

Leek, J. T., and Peng, R. D. 2015. *P* values are just the tip of the iceberg. Nature 520:612.

Lesaffre, E., and Lawson, A. B. 2012. Bayesian Biostatistics. John Wiley & Sons, Chichester, UK.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., and Schabenberger, O. 2006. SAS for Mixed Models. 2nd ed. SAS Press, Cary, NC.

Madden, L. V. 2006. Botanical epidemiology: Some key advances and its continuing role in disease management. Eur. J. Plant Pathol. 115:3-23.

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

Madden, L. V., and Paul, P. A. 2011. Meta-analysis for evidence synthesis in plant pathology: An overview. Phytopathology 101:16-30.

Mayo, D. G. 2013. Discussion: Bayesian methods: Applied? Yes. Philosophical defense? In flux. Am. Stat. 67:11-14.

McBride, G., Cole, R. G., Westbrooke, I., and Jowett, I. 2014. Assessing environmentally significant effects: A better strength-of-evidence than a single *P* value? Environ. Monit. Assess. 186:2729-2740.

Mudge, J. F. 2013. Explicit consideration of critical effect sizes and costs of errors can improve decision-making in plant science. New Phytol. 199:876-878.

Mulaik, S. A., Raju, N. S., and Harshman, R. A. 1997. There is a time and place for significance testing. Pages 65-115 in: What if There Were No Significance Tests? L. L. Harlow, S. A. Mulaik, and J. H. Steger, eds. Erlbaum, London.

Murtaugh, P. A. 2014. In defense of *P* values. Ecology 95:611-617.

Neyman, J., and Pearson, E. S. 1928. On the use and interpretation of certain test criteria of statistical inference. Part I. Biometrika 20A:175-240.

Nuzzo, R. L. 2015. The inverse fallacy and interpreting P values. Phys. Med. Rehabil. (PM&R) 7:311-314.

O'Brien, R. G., and Castelloe, J. 2007. Sample-size analysis for traditional hypothesis testing: Concepts and issues. Pages 237-271 in: Pharmaceutical Statistics Using SAS; A Practical Guide. SAS Press, Cary, NC.

Robert, C. P. 2014. On the Jeffreys-Lindley paradox. Philos. Sci. 81:216-232.

Samsa, G. P. 2015. Has it really been demonstrated that most genomic research findings are false? Am. Stat. 69:1-4.

Schabenberger, O., and Pierce, F. J. 2002. Contemporary Statistical Models for the Plant and Soil Sciences. CRC Press, Boca Raton, FL.

Schervish, M. J. 1996. *P* values: What they are and what they are not. Am. Stat. 50:203-206.

Schneider, J. W. 2015. Null hypothesis significance tests. A mix-up of two different theories: The basis for widespread confusion and numerous misinterpretations. Scientometrics 102:411-432.

Sellke, T., Bayarri, M. J., and Berger, J. O. 2001. Calibration of *p* values for testing precise null hypotheses. Am. Stat. 55:62-71.

Shah, D. A., and Madden, L. V. 2004. Nonparametric analysis of ordinal data in designed factorial experiments. Phytopathology 94:33-43.

Spanos, A. 2013. Who should be afraid of the Jeffreys-Lindley paradox? Philos. Sci. 80:73-93.

Stang, A., and Poole, C. 2013. The researcher and the consultant: A dialogue on null hypothesis significance testing. Eur. J. Epidemiol. 28:939-944.

Trafimow, D., and Marks, M. 2015. Editorial. Basic Appl. Soc. Psych. 37:1-2.