Published in final edited form as:

Am Stat. 2019 ; 73(Suppl 1): 82–90. doi:10.1080/00031305.2018.1537891.

Before $p < 0.05$ to Beyond $p < 0.05$: Using History to Contextualize p-Values and Significance Testing

Lee Kennedy-Shaffer[*]

Department of Biostatistics, Harvard University

Abstract

As statisticians and scientists consider a world beyond $p < 0.05$, it is important to not lose sight of how we got to this point. Although significance testing and p-values are often presented as prescriptive procedures, they came about through a process of refinement and extension to other disciplines. Ronald A. Fisher and his contemporaries formalized these methods in the early twentieth century and Fisher's 1925 *Statistical Methods for Research Workers* brought the techniques to experimentalists in a variety of disciplines. Understanding how these methods arose, spread, and were argued over since then illuminates how $p < 0.05$ came to be a standard for scientific inference, the advantage it offered at the time, and how it was interpreted. This historical perspective can inform the work of statisticians today by encouraging thoughtful consideration of how their work, including proposed alternatives to the p-value, will be perceived and used by scientists. And it can engage students more fully and encourage critical thinking rather than rote applications of formulae. Incorporating history enables students, practitioners, and statisticians to treat the discipline as an ongoing endeavor, crafted by fallible humans, and provides a deeper understanding of the subject and its consequences for science and society.

Keywords

Education; Foundational Issues; Hypothesis Testing; Inference; Probability

1 Introduction

With new journal policies, conferences, and special issues, it is easy to view the debate around p-values and hypothesis testing as a modern invention. For many scientists whose primary connection to statistics is through these methods, the debate may seem like a challenge to the received wisdom of their profession, a rebuke to the way they have been using statistics for decades. For students learning the field, it can seem bewildering, and they might be tempted to replace one decontextualized methodology with another. Indeed, as Gerd Gigerenzer (2004, p. 589) writes, the anonymizing of the roots of the p-value and hypothesis testing has contributed to the idea that "they were given truths" and encouraged the "mindless" use of these procedures, to the point of misuse and abuse. But for those who have studied statistics, and, in particular, studied the progression of statistical theory, the

[*] Lee Kennedy-Shaffer is a doctoral candidate in the Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115 (lee_kennedyshaffer@g.harvard.edu).

debates are not a sudden attack on a completely accepted paradigm, and the statistics themselves did not arise wholly formed to be prescriptively applied. Rather, the statistics arose through the ongoing process of scientific discovery, with contributions by many along the way.

In order to properly understand the challenges that face statistics and its applications in science, medicine, and policy today, and to meet those challenges in the future, we must consider the history of the discipline and its most prominent methods. It is a history that is too poorly known, even among statisticians, but it is rich in characters, personal grudges, and academic debates. Gigerenzer (2004, pp. 587–588) laments this lack of focus on the history and controversy when he relates the story of a psychological statistics textbook author who removed all mention of Thomas Bayes, Ronald A. Fisher, Jerzy Neyman, and Egon Pearson. In a similar vein, Stephen Ziliak and Deirdre McCloskey (2008, p. 232) argue that the conscious erasure of William S. Gosset from the history contributed to the dominance of Fisher's paradigm and reduced the prominence of competing ideas.

In Section 2, I trace the use of statistical reasoning similar to the modern *p*-value before 1900, demonstrating that the statistic and the use of thresholds did not arise from Karl Pearson and Fisher alone. In Section 3, I briefly describe the contributions of Pearson, Gosset, and Fisher, covering both the similarities among them and highlighting some of the debates that occurred as early as the 1920s when Fisher's *Statistical Methods for Research Workers* began to put the *p*-value in the hands of experimenters. In Section 4, I point out some of the challenges that emerged in response to Fisher's paradigm, focusing especially on those arising from Gosset, Neyman, and Egon Pearson, and from the Bayesian paradigm. These sections are far from comprehensive; rather, they seek to provide an overview of the history that can spur thought and encourage further research. In Section 5, I present resources that can be used for that research and as teaching tools. I also discuss how the historical debates relate to modern arguments surrounding the *p*-value and how that can encourage statisticians to craft a more useful and durable response to this controversy. I further describe the role this history can play in education in formal classroom settings and in research and collaboration settings.

Understanding how *p*-values and 0.05 came to occupy their prominent role in 20th century statistics reminds us that these "arbitrary" thresholds came about through work to make mathematical statistics more practical and useful for experimentalists. But these efforts were never without controversy. Learning this history will help statisticians better appreciate the translational challenges of their own work by improving understanding of the fact that, since its inception, the modern field of statistics has grappled with the balance between mathematical rigor and practical use to scientists. Those pushing the boundaries of knowledge in the discipline will surely face this balance in their own work. They will have to consider, like the statisticians of the early twentieth century did, how others will use their theories.

Learning this history will help practitioners understand that no method is sacred and that all methods are products of the era in which they were born and the functions to which they have been applied. As technology, mathematics, and science develop, new methods or adjustments to old methods will be needed as the underlying assumptions no longer apply, whether in a world of early electronic computing devices or a world of big data.

Learning this history will help students access the discipline by learning of the faults, personal and professional, of those who came up with today's commonly used statistics and help them understand statistics as a living discipline rich with ongoing debate and new understandings. Indeed, one can find many parallels between today's debate and the controversies that arose with the development of *p*-values and significance testing, framing the ASA statement and subsequent discussion as another step in the ongoing evolution of the discipline of statistics.

2 A World Before Fisher

The *p*-value is generally credited to Karl Pearson's 1900 article in his journal *Biometrika*; Ronald A. Fisher's 1925 *Statistical Methods for Research Workers* then formalized the concept and expanded its reach to experimenters (Hubbard 2016, p. 14). But statistics similar to *p*-values and probabilistic reasoning akin to hypothesis tests existed well before then. Both Stigler (1986) and David and Edwards (2001) point to John Arbuthnott's 1710 "An Argument for Divine Providence" as perhaps the earliest use of probabilistic reasoning that matches that of a modern null hypothesis test. Using birth data from London, Arbuthnott (1710) notes that births of males exceeded births of females for 82 years in a row. Supposing that the probability of males exceeding females in a year is 50%, and implicitly assuming independence across the years, Arbuthnott calculates the miniscule probability of this 82-year pattern. "From whence it follows," Arbuthnott (1710, p. 189) confidently concludes, "that it is Art, not Chance, that governs." Any modern student who has run a test of proportions would notice the reasoning, see Arbuthnott's calculation of a *p*-value of 2.07×10^{-25}, and confirm his rejection of the null hypothesis that each year has an independent probability of 50%. The mathematically-inclined physician's goal in this endeavor was to demonstrate the work of "Divine Providence" in the sex distribution (Arbuthnott 1710, p. 186). Many statisticians would recognize the flaw in this reasoning: the lack of a clearly stated alternative hypothesis that would be logically implied by a rejection of the null hypothesis. Gigerenzer (2004, p. 588) decries this "null ritual" used by experimentalists who often fail to properly specify "alternative substantive hypotheses."

In the nineteenth century, French mathematicians used similar methods to analyze a wide variety of data. In celestial mechanics, Pierre-Simon Laplace (1827, p. S.30) found a small value for a statistic closely related to the modern *p*-value and concluded that it indicated with a high likelihood that the discrepancy in the measurements was thus "not due solely to the anomalies of chance." Stigler (1986, p. 151) notes that Laplace himself appealed to a 0.01 significance level in his work. Stigler (1986, pp. 151–153) further highlights several errors implicit in Laplace's analysis, errors that would be familiar to students and critics of modern hypothesis testing: improper assumptions of independence and improper estimation of variance.

Not long after, Siméon-Denis Poisson used a quantity equal to one minus a modern *p*-value in describing patterns in the outcomes of French jury trials. Two comparisons he makes are particularly instructive. In one, he finds a *p*-value of 0.0897, a value not large enough for him to conclude that there has been a change in causes (Poisson 1837, p. 373). Shortly thereafter, a *p*-value of 0.00468 leads Poisson to believe that in that case there is a “real anomaly in the votes of juries” (Poisson 1837, pp. 376–77). Poisson’s conclusions in these two cases, nearly a century before Fisher’s work, would comport with a 0.05 (or 0.01) significance threshold, but do not specify a threshold he used. Poisson (1837, p. 375) also refused to make a causal statement from his identified associations, noting that “the calculation cannot teach us” this answer.

Antoine Augustin Cournot formulated the *p*-value in fairly explicit terms, noting that as a measure of the importance of some discrepancy it combines the size of the effect and the sample size (Cournot 1843, p. 196). Cournot (1843, pp. 196–197) also issues a warning about the narrow-minded use of probabilistic statements, noting that this *p*-value does not fully capture the importance of the effect size and “does not at all measure the chance of truth or of error pertaining to a given judgment.” With a little modernization of language, Cournot could have written principles 2, 5, and 6 of the ASA Statement (Wasserstein and Lazar 2016).

In 1885, Francis Ysidro Edgeworth provided a more formal mathematical underpinning for the significance test and gave a simple example of how to use the standard deviation (he used the “modulus”, equal to the standard deviation multiplied by the square root of two) to perform a significance test on a given parameter (Edgeworth 1885, pp. 184–185). Using a threshold of twice the “modulus,” Edgeworth (1885) constructed a test that would be equivalent to a modern two-sided $\alpha = 0.005$. Stigler (1986, p. 311) notes that this “was a rather exacting test” and that Edgeworth also considered smaller differences as “worthy of notice, although he admitted the evidence was then weaker.”

The existence of these tests of significance and *p*-value-like quantities long before the twentieth century demonstrate that this method of inference had an alluring rationale for practitioners in a variety of fields. Their errors in interpretations and words of caution, however, presage the controversies that would follow. Throughout the twentieth century, many of the technical probability results needed for modern significance testing arose through the theory of errors, by which astronomers and other physical scientists combined measurements and discarded outliers (Gigerenzer et al. 1989, pp. 80–84). These developments allowed Pearson, Gosset, and Fisher to make key contributions that formalized, shaped, and popularized the modern form of significance tests.

3 R.A. Fisher: the Experimentalist Statistician

In the early twentieth century, the forerunners of modern statistics began to determine the properties of various useful distributions. Karl Pearson (1900) described the χ^2 distribution and uses of the χ^2 statistic, including its use in tests of independence for proportions. Pearson (1900, pp. 157–158) here denoted by P the “chances of a system of errors with as great or greater frequency than that denoted by χ.” In an example involving dice throws, Pearson (1900, pp. 167–168) finds $P = 0.000016$ on a null distribution of equal probability of each face appearing and claims that “it would be reasonable to conclude that dice exhibit bias towards the higher points." The combination of this type of probabilistic reasoning and a distribution with many practical uses made the *p*-value more approachable and brought it more or less to its modern formulation. W. Palin Elderton built on Pearson's work and produced tables of values for this distribution that would enable investigators to test the goodness of fit. His article, published in *Biometrika* in 1902, devoted roughly half of its space to these tables (Elderton 1902). Ziliak and McCloskey (2008, pp. 199–202) note that Pearson was soon teaching his students, and enforcing as a rule for authors seeking publication in *Biometrika*, that three probable errors, or two standard errors, represented "certain significance."

William Sealy Gosset, the head experimental brewer at Guinness publishing under the pseudonym "Student" (1908, p. 25), found a curve "representing the frequency distribution of values of the means of such samples," i.e., samples from a normal or "not strictly normal" distribution, "when these values are measured from the mean of the population in terms of the standard deviation of the sample." This so-called Student's *t* distribution is now taught in introductory and applied statistics courses, as it forms the basis for a substantial number of inferential procedures. Gosset's initial paper focused as much on illustrating examples of the utility of this curve as on the mathematical justification for its use, and he produced numerous tables to enable others to use it. He calculated statistics akin to the *p*-value and drew conclusions from extreme values of these. For one drug trial, he regarded a statistic equivalent to $p = 0.0015$ as "such a high probability," it would be in practical matters "considered as a certainty" (Student 1908, p. 21). For Gosset, however, whether an effect existed or not was less important than its impact, and he saw the use of the tests more in determining the "pecuniary advantage" of one decision versus another (Ziliak and McCloskey 2008, pp. 18–19). That is, any conclusion must rest on effect size and the relative loss and gain of any potential decision; this will be a recurring theme in the debate between competing frameworks for testing discussed below.

Ronald A. Fisher, who had corresponded with Pearson and Gosset at various points, was well aware of these advances and thus of the use of significance tests. His work, especially a series of three monographs published in the 1920s and 1930s, would expand the reach of significance tests, promote their use (and the use of statistically rigorous experimental design and analysis more broadly) to researchers, and provide tables that enabled investigators to conduct such tests.

Fisher, employed at the time at Rothamsted Experimental Station, an agricultural research institution, "extended the range of tests of significance" using the theory of maximum likelihood commonly used today and conceived of tests for small sample problems (Box 1978, p. 254). In 1922, he published three key manuscripts which covered the theoretical foundations of maximum likelihood estimation and the concept of the likelihood (Fisher 1922c), the use of Pearson's χ^2 distribution to calculate *p*-values from contingency tables (Fisher 1922b), and the use of Student's *t* distribution to conduct significance tests on regression coefficients (Fisher 1922a). In 1925, he published the first edition of *Statistical Methods for Research Workers,* which sought, in his words, "to put into the hands of research workers, and especially of biologists, the means of applying statistical tests accurately to numerical data" (Fisher 1925, p. 16). The book discusses in detail the meaning and practical implications of "P", the statistic now known as the *p*-value, and suggests 0.05 as a useful cutoff:

> The value for which P = .05, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant. Using this criterion we should be led to follow up a negative result only once in 22 trials, even if the statistics are the only guide available. Small effects would still escape notice if the data were insufficiently numerous to bring them out, but no lowering of the standard of significance would meet this difficulty (Fisher 1925, p. 47).

This simple paragraph demonstrates the probability-based definition of the *p*-value that is commonly misunderstood: that it is the probability of a result as or more extreme than the observed result given that the null hypothesis is true (Greenland et al. 2016). Additionally, it makes immediately apparent why 0.05 is convenient: it is roughly equivalent to the probability of being more than two standard deviations away from the mean of a normally distributed random variable. In this way, 0.05 can be seen not as a number Fisher plucked from the sky, but as a value that resulted from the need for ease of calculation at a time before computers rendered tables and approximations largely obsolete. This particular value had the added bonus of corresponding to three "probable errors," a measure of spread of the normal distribution used commonly in early statistics but now largely forgotten (Stigler 1986, p. 230). So a useful rule of thumb could be given to researchers on either of two scales of measuring the spread of the distribution. Later, in applying the statistic to the χ^2 distribution, Fisher (1925, p. 79) remarks that "[w]e shall not often be astray if we draw a conventional line at .05, and consider that higher values of χ^2 indicate a real discrepancy."

Statistical Methods was most valuable in the hands of experimentalists due to its explanations of tests and estimation procedures, illustrative examples, and a wealth of user-friendly tables. The tables further entrenched the use of Fisher's preferred *p*-value cutoffs by displaying the calculated figures so that an investigator looked up a desired probability level for the distribution and found the quantile of the statistic that corresponded to it. Among the levels presented one almost always found 0.05 and 0.01 (Fisher 1925). As Fisher's biographer and daughter, Joan Fisher Box (1978, p. 246), states, "[b]y this means he produced concise and convenient tabulations of the desired quantities" and presented values "that were of immediate interest to the experimenter." It is this accessibility that made the book popular among practicing experimentalists who "had not a hive of staff humming at their desk calculators," but it did not endear him to more rigorous mathematicians (Box 1978, pp. 242–246). And through these presentations, which Fisher (1935; Fisher and Yates 1938) continued and expanded in *The Design of Experiments* and the 1938 compilation of tables with Francis Yates entitled *Statistical Tables for Biological, Agricultural, and Medical Research,* he set the standard for the use of *p*-values and statistical inference in a variety of forms of research. One may note, however, that Fisher's tables show that he did not think 0.05 was one size fits all; if 0.05 worked in every setting, there would have been only one column in each table.

The history of the tables presented in *Statistical Methods* is interesting in itself and further demonstrates how these values came to be presented; it also foreshadows forthcoming schisms regarding these tests. Hubbard (2004, p. 311) notes that Pearson's *Biometrika* denied Fisher permission to use Elderton's table of χ^2 probabilities in his monograph. When he created his own version, according to Egon Pearson (Pearson et al. 1990, p. 52), Fisher "gave the values of χ^2 for selected values of P… and thus introduced the concept of nominal levels of significance." Because of this change from Elderton's table to Fisher's, for users of the table in *Statistical Methods* and its successors in *Statistical Tables*, it would be easier to compare a calculated χ^2 value to a set threshold of significance rather than find the precise *p*-value. For tables of the *t* statistic, as Ziliak and McCloskey (2008, p. 229) note, "Fisher himself *copyrighted again Gosset's tables in his own name*" in *Statistical Methods* (emphasis in original). Through this action, which left Gosset's name out of the book except in the phrase "Student's *t*", Fisher removed Gosset from the history of his own statistic, hid his contributions, and, more importantly, hid his competing philosophy on how the statistic should be used (Ziliak and McCloskey 2008, pp. 230–232). Reprinters of the table and those who used it in applied research would encounter only Fisher's versions and his interpretations.

Following Fisher, the use of *p*-values grew among experimentalists. In the United States, they were particularly encouraged by Harold Hotelling of Stanford University, who called some of the tables in *Statistical Methods* "indispensable for the worker with moderate-sized samples" (quoted in Ziliak and McCloskey 2008, p. 234). George Snedecor of Iowa State University played a crucial role as well, continuing to develop the methods and promoting their use in scientific fields (Hubbard 2016, p. 21). Psychologists, sociologists, political scientists, and economists all found the innovations useful (Hubbard 2016, pp. 22–27). Thus the *p*-value spread not only across oceans but beyond the natural sciences to the social sciences, echoing its use by Poisson a century earlier.

The use of 0.05 as a cutoff became customary, though not all-encompassing. Fisher's student L. H. C. Tippett (1931, p. 48), wrote in *The Method of Statistics* that the 0.05 threshold was "quite arbitrary" but "in common use." Lancelot Hogben (1957, p. 495), two decades later, wrote that Fisher's claim that the cutoff was in usual practice was "true only of those who rely on the many rule of thumb manuals expounding Fisher's own test prescriptions." For scientists and students today, perhaps the prominence of this admittedly arbitrary cutoff is difficult to comprehend. However, they need only consider a time before computers and compare the calculation of a *p*-value by hand from one of Fisher's or Gosset's or Pearson's formulae to the ease by which one can determine whether a statistic meets a threshold by reference to one of Fisher's tables. It will immediately become clear how Fisher's standard became the gold standard. Fisher led other tables to adopt his format through his role as secretary of the Tables Committee of the British Association (Box 1978, p. 247), ensuring that future statisticians who sought to reach experimentalists would need to reconcile their methods to this framework. Thus "$p < 0.05$" could grow to the prominence it holds today.

4 Challenges to Fisher's View

The other piece of history often lost in the presentation of *p*-values is that statisticians brought many challenges to Fisher's framework as soon as it was presented. As Fisher was writing his manuscripts, Jerzy Neyman and Egon Pearson (1933) were preparing their own framework for hypothesis testing. Rather than focusing on falsifying a null hypothesis, Neyman and Pearson presented two competing hypotheses, a null hypothesis and an alternative hypothesis, and framed testing as a means of choosing between them. The decision then must balance two types of error, one made by incorrectly rejecting the null hypothesis when it is true (Type I Error) and one made by incorrectly accepting the null hypothesis when it is false (Type II Error). More generally, one can consider the class of "admissible alternative hypotheses" of which the null hypothesis is a member (Neyman and Pearson 1933, p. 294); the goal is then to compare the null hypothesis to the alternative that imparts the highest likelihood on the observed data. They propose a class of tests that, for a given limit of Type I Error, minimize the risk of Type II Error, the so-called most powerful tests. The Type I Error risk, often called the significance level and denoted α, is commonly set at 0.05 (or 0.01), as the pair noted in their paper. The Type II Error risk, often denoted β, is equal to one minus what we now call the power of the test.

This procedure has many similarities to Fisher's framework that uses the *p*-value as a continuous measure of evidence against the null hypothesis; indeed, in many cases, Fisher's choice of test statistic corresponds to a reasonable choice of alternative hypothesis in a Neyman-Pearson most powerful test (Lehmann 1993, pp. 1243, 1246). In those cases, $p < \alpha$ if and only if the most powerful α-level test would reject the null hypothesis. Nonetheless, the two factions debated fiercely the merits of each version. In one sense, the controversy can be regarded as a debate over the role of the statistician and of the test itself: should the test be considered as a step along the way to deeper understanding, a piece of evidence among many to be considered in crafting and supporting a scientific theory? Or should it be considered as a guide to decision-making, a way to choose the next behavior, whether in a practical or experimental setting? Fisher's writings generally support the former view, taking the test and the *p*-value as a piece of evidence in the scientific process, one that he wrote "is based on a fact communicable to, and verifiable by, other rational minds" (Fisher 1956, p. 43). For Neyman and Pearson, on the other hand, to accept a hypothesis means "to act as if it were true" and thus the hypotheses and error probabilities should be chosen in light of the consequences of making either decision (Gigerenzer et al. 1989, p. 101).

In a practical way, the Neyman-Pearson view also meant considering the reasonability of alternative hypotheses. Berkson (1938, p. 531) provided an application of this question, discussing how someone familiar with the data would only truly reject the null hypothesis "on the willingness to embrace an alternative one." The debate took on a variety of aspects, however, including being somewhat representative of a larger controversy over the role of mathematical rigor in statistics, with Fisher assailing Neyman and Pearson as mathematicians whose work failed to reflect the nuances of scientific inference (Gigerenzer et al. 1989, p. 98). It also covered differences in the role assigned to a statistical model of data and decision-making, which in turn relate to fundamental probability questions about defining populations and samples (Lenhard 2006). All of these differences were heightened and perhaps even exaggerated by “the ferocity of the rhetoric” (Lehmann 1993, p. 1242).

While this debate raged in the halls of academic statisticians for decades (and, even today, attempts are made to clearly define the differences or reconcile the two theories), experimentalists began to follow a third way, an *“anonymous* hybrid consisting of the union of the ideas developed by” Fisher and Neyman-Pearson (Hubbard 2004, p. 296, emphasis in original). Often, reporting of results will include a comparison of the *p*-value to a threshold level (e.g., 0.05) to claim existence of an effect, reporting of the *p*-value itself, and relative measure of evidence terms such as “highly significant,” “marginally significant,” and “nearly significant.” This leads to what Hubbard (2004, p. 297) calls a “confusion between *p*’s and α’s” among applied researchers, as seen in textbooks, journal articles, and even publication manuals. This confusion undermines the rigorous Neyman-Pearson interpretation of limiting error to a pre-specified level α. And the role of the value of *p* as a quantitative piece of ongoing scientific investigation (including using null hypotheses that are not a hypothesis of zero effect) favored by Fisher is lost to the decision-making encouraged by a statement of significance or lack thereof. Neither Fisher nor Neyman and Pearson would approve of this hybrid, though it has been institutionalized by textbooks and curricula, especially in applied settings. Its popularity owes a great deal to its simplicity and the ability of applied researchers to perform this “ritual” of testing in a more mechanized fashion (Gigerenzer 2004).

While this debate was ongoing, a revival of another paradigm of probability gained steam. Based on a crucial theorem by Thomas Bayes that was published in 1763, the “inverse probability” or Bayesian viewpoint embraced the subjectivity of statistical analysis (Weisberg 2014, §10.2). With regard to testing, the Bayesian approach allows a researcher to calculate the probability of a specific hypothesis given the observed data, rather than the converse, which is what the Fisher and Neyman-Pearson approaches do. These views gained considerable traction after Leonard J. (Jimmie) Savage’s 1954 publication of *The Foundations of Statistics,* which also replied to anticipated objections to the paradigm. His work builds on that of Bruno de Finetti (1937) and Harold Jeffreys (1939).

Bayesian ideas were present before then, however, as Fisher (1922c, pp. 325–330) included in his article on maximum likelihood a rejection of Bayesian approaches. Fisher (1922c, p. 326) even notes that the works of Laplace and Poisson, discussed above, “introduced into their discussions of this subject ideas of a similar character” to inverse probability. While this article is far too short to cover the debate between the various Bayesian approaches and the frequentist approaches of Fisher, Gosset, and Neyman-Pearson, Savage’s book is a useful starting point, and a higher-level summary can be found in Weisberg (2014). Sharon McGrayne (2011) provides a very accessible overview of the Bayesian approach, its history, and the common use of Bayesian methods in practical research even while it was philosophically rejected by statisticians. These debates, too, are ongoing, with Bayesians or frequentists holding more sway in different scientific fields (Gigerenzer et al. 1989, pp. 91, 105), and Bayesian approaches are often suggested as alternatives to *p*-values, as discussed below.

In addition to these broad philosophical challenges, statisticians and scientists objected to Fisher's *p*-value on practical grounds. Gosset wrote to Fisher and to Karl Pearson of the importance of considering effect sizes and, indeed, arranging experiments "so that the correlation should be as high as possible" (quoted in Ziliak and McCloskey 2008, p. 224). Fisher's own co-author, Francis Yates, wrote in 1951 (p. 33) of his concern that experimenters were regarding "the execution of a test of significance as the ultimate objective." Fisher (1956, p. 42) himself later wrote that "no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses." His book even included a chapter entitled "Some misapprehensions about tests of significance" (Fisher 1956, p. 75). His writings on the matter, however, are sometimes contradictory and admit several interpretations (Gigerenzer et al. 1989, p. 97). Medical statistician Joseph Berkson (1942, p. 326) feared a disconnect between significance testing and "ordinary rational discourse," especially in applying a rule to these tests without regard to "the circumstances in which it is applied" (p. 329). The International Biometric Society's British Regional President warned in 1969 that significance tests might "exercise their own unintentional brand of tyranny over other ways of thinking" (Skellam 1969, p. 474). Psychologist William Rozeboom (1960) wrote of the failings of *p*-values and significance testing, including the uncritical appeals to 0.05, in 1960. Other psychologists and social scientists soon followed (Bakan 1966; Meehl 1967; Skipper et al. 1967).

These arguments, which began as soon as the paradigm-defining works were published, would all be familiar to those following the modern debates and resonate in the ASA statement (Wasserstein and Lazar 2016). Discussing and teaching the modern debate without acknowledging its historical roots does a disservice not only to those thinkers who engaged in the debate, but to the statistics profession as a whole.

5 History as Context to Inform the Present Debate

In this article, I have endeavored to recount only a small part of the history of *p*-values and significance testing, which itself forms only a small part of the history of probability and statistics. Much more can be found on these subjects. David Salsburg's 2001 *The Lady Tasting Tea* provides a highly accessible treatment. Stephen Stigler's 1986 The *History of Statistics: The Measurement of Uncertainty Before 1900* covers the early history of the *p*-value and how it fit into notions of reasoning about uncertainty; Theodore Porter's book *The Rise of Statistical Thinking, 1820-1900,* also published in 1986, covers the latter end of this pre-Pearson history. H. A. David and A. W. F. Edwards's 2001 *Annotated Readings in the History of Statistics* highlights primary source material relevant to this history. In books published twenty-five years apart, Gerd Gigerenzer et al. (1989) and Herbert Weisberg (2014) describe the rise of the dominant modern mathematical conception of probability and how that influenced and was influenced by the rise of these statistics and of data-driven sciences. Stephen Ziliak and Deirdre McCloskey (2008) describe the rise of these statistics in the early twentieth century in great detail, focusing on reviving the forgotten role of Gosset through presentation and interpretation of his archival materials; they also describe the spread of the Fisherian paradigm in economics, psychology, and law, and the consequences of that spread. Finally, Donald MacKenzie's 1981 *Statistics in Britain, 1865–1930* discusses Fisher and his immediate predecessors in detail, focusing especially on the effects of the British social context on the work of Francis Galton, Karl Pearson, and Fisher and on how eugenics shaped the statistical work of the three men and the rise of statistics in Britain.

Articles and books exploring the use of statistical inference, especially hypothesis testing, in specific fields can be informative of this history as well: Morrison and Henkel (1970) write of the controversies in the social sciences; Hubbard (2016) discusses the use of statistics in the management sciences as well as the social sciences; Hubbard (2004) and Chambers (2017) describe the controversies in psychology; Kadane, Fienberg, and DeGroot (1986) discuss the use of statistics in the field of law with several case studies; I (Kennedy-Shaffer 2017) cover some of this history with a focus on significance testing at the United States Food and Drug Administration. These various detailed accounts, among others, clarify the lessons that statisticians and practitioners can take from this history and provide ample material for statistics educators to incorporate this history into their formal and informal teaching.

5.1 Lessons for Statisticians and Practitioners

The history of the *p*-value and significance testing is useful for statisticians and scientists who use statistical methods today for a variety of reasons. The history helps clarify today's debates, adding a long-term dimension to modern discussions. In this way, it illuminates the factors that drive the creation of statistical theory and methods and what enables them to catch on in the broader community. Understanding these factors will help statisticians respond to today's debates and consider how proposed solutions to problems that have arisen will play out in the scientific community today and in the future.

First of all, the history clarifies the debates that are occuring today; in particular, many of the objections raised to *p*-values by modern scientists (and in Wasserstein and Lazar (2016) and the accompanying *Online Discussion*) were raised by contemporaries of Fisher. One particular aspect, the importance of considering effect size rather than simply statistical significance, was the crux of the difference between Fisher's framework and Gosset's (Ziliak and McCloskey 2008). Ziliak (2016) reiterates this connection in an article in the *Online Discussion,* demonstrating the relevance of historical debates to today's discussion. A thread of argument from Fisher's earliest critics (and indeed Cournot and Edgeworth before him) to Rothman (2016) indicates that the de-emphasizing of effect size in favor of the *p*-value is an easy mistake to make and one that needs to be addressed. Similarly, debates have continued over the conflating of Fisher's paradigm with the Neyman-Pearson approach, as discussed above. Lew (2016) describes how these different inferential questions have become hybridized. Discussions of power and the role of statisticians in the design of experiments arise in the commentaries by Berry (2016) and Gelman (2016). While their approaches are quite different, Fisher certainly understood that argument, writing an entire book on how to properly design experiments (Fisher 1935); Gosset, too, participated in this discussion, disagreeing with Fisher on key aspects (Ziliak and McCloskey 2008, p. 218). And the Bayesian-frequentist debate continues today, unresolved after decades of discussion. Among others, Benjamin and Berger (2016) and Chambers (2017, pp. 68–73) promote the potential use of Bayesian hypothesis testing as an alternative to *p*-values and significance testing.

It would be easy to be disheartened by this history. If we have been debating these ideas, raising similar arguments for a century, what hope do we have of solving them now? And, as Goodman (2016) puts it, "what will prevent us from dusting this same statement off 100 years hence, to remind the community yet again of how to do things right?" The history may provide the answer here. In particular, a closer look at how Fisher's ideas spread and how the hybridization of the Fisher and Neyman-Pearson paradigms occurred, processes discussed here only briefly, can inform us of what makes statistical methods catch hold in the broader scientific, policymaking, and public communities. Berry (2016) notes that statisticians should not seek to "excuse ourselves by blaming non-statisticians for their failure to understand or heed what we tell them." But we can understand why they fail to heed us. Benjamini (2016) notes that the *p*-value was so successful in science because it "offers a first-line defense against being fooled by randomness." That is, it was useful to non-mathematicians in giving them a quantitative basis for addressing uncertainty. Additionally, it has some intuitive meaning, as can be seen by the fact that methods similar to the *p*-value arose repeatedly in various fields even before Fisher. And it had passionate advocates who put the tools into the hands of scientists in a way that was easy to use, like through Fisher and Yates's *Statistical Tables.* Finally, it was responsive to conditions of the time. These approaches addressed questions about variance and experimental design that were frequently raised at the time (Gigerenzer et al. 1989, pp. 73–74). Considering these virtues, Abelson (1997) suggests in a tongue-in-cheek piece that significance tests would be re-invented if they were banned and forgotten.

A response that gains traction outside of academic statisticians and that is durable, I argue, must meet these same criteria, summarized in Table 1. And moreover, to remain valuable, it must be able to adjust to changing conditions. For example, as many authors, including Weisberg (2014, §12.3), have discussed, our computational and data-gathering capabilities have changed enormously over the last several years, to say nothing of changes since 1925. We have seen how the lack of computing power at the time rendered Fisher's tables so valuable and thus so influential to practitioners. And the limited computer capabilities of the 1950s may have limited the ability of Bayesian methods to catch on with a wider audience (Weisberg 2014, §8.4). The ease of computation is one cause of the multiplicity issues that are commonly discussed (Ionnaidis 2005; Benjamini 2016). However, there is no reason to believe that computing capabilities have plateaued, and so an appropriate response would take into account not only today's conditions, but also those likely to occur in the future. Moreover, as we have seen, statistical methods are not always used with fidelity to the original intents and assumptions, especially decades after their initial formulation. Several of the responses to *p*-values, as Benjamini (2016) notes, would be susceptible to misuse as well.

Certainly, these are high demands to make of any statistical method, or indeed of any scientific methodology at all. And the sheer variety of alternatives proposed indicate that even the statistical community has not coalesced around one. To take one example, consider the proposal to lower the significance threshold to 0.005 (Benjamin et al. 2018). Table 1 summarizes whether and how $p < 0.005$ addresses the criteria for a lasting framework, not to argue for or against it, but to suggest the utility of this framework in assessing responses beyond $p < 0.05$. This proposal has several advantages: it maintains the ease of use and familiarity that scientists prize and can be viewed as in line with the approaches of Fisher (who often wrote of different thresholds in different settings) or of Neyman and Pearson (if it represents some true cost of a Type I Error and is paired with Type II Error control). It also addresses some of the multiplicity issues that have arisen from changing conditions and the reduced computational burden. This is not even the first time it has been proposed; as discussed, Edgeworth implicitly used this threshold at times, and threshold proposals varied greatly before and even after Fisher. However, it is, as Benjamin et al. (2018) acknowledge, just as arbitrary as current thresholds and just as susceptible to misinterpretation. And the benefits in addressing multiplicity may fade as data sets get bigger and tests are run even more frequently. Little (2016) also notes that lowering the threshold fails to address the longstanding debate between statistical significance and substantive significance. But differing thresholds have worked in other fields and this proposal may have a great deal of value in certain settings. And with tables of significance thresholds no longer necessary thanks to modern computing power, it is quite easy for researchers to use different thresholds at different times. This suggests that no one method and no one response to the controversy will be sufficient.

A multitude of responses, tailored to scientific purposes and fields of study, will be much more likely to be able to address all of these needs. Indeed, one can see this as an extension of arguments made at various points by both Fisher and Neyman-Pearson that different experimenters, working in different contexts, will use different thresholds of significance or set different α and β parameters. As Fisher's work focused on agriculture and biology, perhaps his advice still holds sway there, while other fields face different needs. Beyond just significance thresholds, different scientific questions can be approached with the variety of tools available, from Bayesian approaches to confidence intervals to machine learning, to suit their context. Such an approach, however, relies on a great deal of statistical sophistication among those who use statistical methods. Fortunately, this history can help improve statistics education and guide changes that would enhance that sophistication.

5.2 The Role of History in Statistics Education

The rise in popularity of statistics books aimed at general audiences, including some listed above, demonstrates the desires of many people to learn both the practical uses of the discipline and the way in which it came to be. Statistics educators broadly defined, whether course instructors, statistical collaborators, or writers of articles aimed at non-statisticians, can benefit from this interest and use history as a teaching tool within this moment of debate in the discipline. The British mathematician John Fauvel (1991, pp. 4–5) presented a variety of reasons for incorporating history into mathematics education, including to "increase motivation for learning," "explain the role of mathematics in society", and "contextualise mathematical studies." A decade later, the Taiwanese educator Po-Hong Liu expounded these ideas. He noted specifically that "[h]istory reveals the humanistic facets of mathematical knowledge" and can challenge students' perceptions "that mathematics is fixed, rather than flexible, relative, and humanistic" (Liu 2003, p. 418).

These reasons all hold for statistics, especially as the discipline faces great change, not just in the use of conventional inferential methods but also with the rise of computing power and big data. As Goodman (2016, p. 1) notes: “that statisticians do not all accept at face value what most scientists are routinely taught as uncontroversial truisms will be a shock to many.” To meet Millar’s (2016) and Stangl’s (2016) challenge of improving statistical education, teachers and collaborators should consider the introduction of this history into their discussions of significance testing. Presenting these controversies requires educators to present other approaches and thus also serves to, as Millar (2016) suggests, “make our students aware that *p*-values are not the ‘only way.’”

The topics covered here can be introduced alongside the presentation of the tables of values of the normal, Student’s *t*, and χ^2 distributions, which still hold a place as early as the Advanced Placement Statistics curriculum (AP 2010). Inviting students to consider how the lack of computers affected the development of statistics may further appreciation for these tables (or, more likely, further appreciation for the computer software that has rendered them obsolete). This in turn will help students appreciate what has changed since 1925 and how methods may need to change to reflect that.

Presenting the debate between Fisher, Gosset, Neyman-Pearson, and the Bayesians, and how that debate has evolved into the current discussion, highlights the human aspect of statisticians and the constantly changing, challenging nature of the field. As discussed above, many of the specific points made in that debate are ongoing points of contention today. In-depth analysis of Fisher’s rationale for using the 0.05 standard can highlight how, though arbitrary, it is not without context, and how it responded to the needs of experimentalists at a certain point of history. This understanding will allow students and practitioners to form their own assessment of, for example, the proposal to lower the standard to 0.005. In this way it becomes harder to dismiss the *p*-value without providing a substitute that is similarly usable by those who perform statistical analyses today. This teaching will also give students and practitioners the ability to critique the next statistical method that comes along, and to consider alternatives to the *p*-value in the context of statistical history and the role of statistics in modern science and society.

6 Conclusion

As we consider a world “beyond $p < 0.05$,” I invite statisticians and scientists alike to consider the world before $p < 0.05$, a world where statistical analysis was less common and far more difficult an undertaking. It is then easier to see how *p*-values came to such prominence throughout science, despite the immediate disagreements among statisticians. Statistics is an evolving discipline, but it is in the difficult position of needing to evolve alongside the various disciplines that make use of its tools. In Fisher’s teaching and manuscripts, writes Box (1978, p. 242), “he aimed to give workers a chance to familiarize themselves with tools of statistical craft as he had become familiar with them, and to evolve better ways of using them.” This approach helped make statistics a fundamental tool in many disciplines, but has led to the challenges discussed in the ASA statement and elsewhere. Presenting this history as context for these discussions provides appropriate recognition of the rich debates that define statistics. It encourages statisticians to consider how their work will be used by practitioners and encourages practitioners to consider whether they are using statistical methodologies as they were intended. Through ongoing discussions and by encouraging this critical thinking, statistics can continue to be a field that helps push forward the boundaries of knowledge.

Acknowledgments

The author's studies are supported by Grant No. 5T32AI007358-28, U.S. National Institute of Allergy and Infectious Diseases, National Institutes of Health. The author thanks the editors and referees for their helpful comments.

References

Abelson RP (1997), A retrospective on the significance test ban of 1999 (if there were no significance tests, they would be invented), *in* Harlow LL, Mulaik SA & Steiger JH, eds, 'What If There Were No Significance Tests?', Multivariate Applications, Lawrence Erlbaum Associates, Mahwah, NJ, pp. 117–141.

AP (2010), 'Statistics course description', https://secure-media.collegeboard.org/ap-student/course/ap-statistics-2010-course-exam-description.pdf. Accessed: 2018-3-4.

Arbuthnott J (1710), 'An argument for divine providence, taken from the constant regularity observ'd in the births of both sexes', Philosophical Transactions (1683–1775) 27, 186–190.

Bakan D (1966), 'The test of significance in psychological research', Psychol. Bull. 66(6), 423–437. [PubMed: 5974619]

Benjamin DJ & Berger JO (2016), 'Comment: A simple alternative to *p*-values', Am. Stat., Online Discussion 70(2).

Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R , Bollen KA, Brembs B, Brown L, Camerer C, Cesarini D, Chambers CD, Clyde M, Cook TD, De Boeck P, Dienes Z, Dreber A, Easwaran K, Efferson C, Fehr E, Fidler F, Field AP, Forster M, George EI, Gonzalez R, Goodman S , Green E, Green DP, Greenwald AG, Hadfield JD, Hedges LV, Held L, Ho TH, Hoijtink H, Hruschka DJ, Imai K, Imbens G, Ioannidis JPA, Jeon M, Jones JH, Kirchler M, Laibson D, List J, Little R, Lupia A, Machery E, Maxwell SE, McCarthy M, Moore DA, Morgan SL, Munafó M, Nakagawa S, Nyhan B, Parker TH, Pericchi L, Perugini M, Rouder J, Rousseau J, Savalei V, Schönbrodt FD, Sellke T, Sinclair B, Tingley D, Van Zandt T, Vazire S, Watts DJ, Winship C, Wolpert RL, Xie Y, Young C, Zinman J & Johnson VE (2018), 'Redefine statistical significance', Nature Human Behaviour 2(1), 6–10.

Benjamini Y (2016), 'It's not the *p*-values' fault', Am. Stat., Online Discussion 70(2).

Berkson J (1938), 'Some difficulties of interpretation encountered in the application of the chi-square test', J. Am. Stat. Assoc 33(203), 526–536.

Berkson J (1942), 'Tests of significance considered as evidence', J. Am. Stat. Assoc 37(219), 325–335.

Berry DA (2016), '*p*-values are not what theyre cracked up to be', Am. Stat., Online Discussion 70(2).

Box JF (1978), R. A. Fisher, the Life of a Scientist, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, New York.

Chambers C (2017), The Seven Deadly Sins of Psychology: A Manifesto for Reforming the Culture of Scientific Practice, Princeton University Press, Princeton, NJ.

Cournot AA (1843), Exposition de la Théorie des Chances et des Probabilités, L. Hachette, Paris.

David HA & Edwards AWF (2001), *Annotated Readings in the History of Statistics,* Springer Series in Statistics: Perspectives in Statistics, Springer, New York.

De Finetti B (1937), 'La prévision: ses lois logiques, ses sources subjectives', Annales de l'Institut Henri Poincaré 7(1), 1–68.

Edgeworth FY (1885), 'Methods of statistics', J. Stat. Soc. London Jubilee Volume, 181–217.

Elderton WP (1902), 'Tables for testing the goodness of fit of theory to observation', Biometrika 1(2), 155–163.

Fauvel J (1991), ‘Using history in mathematics education’, For the Learning of Mathematics 11(2), 3–6.

Fisher RA (1922a), ‘The goodness of fit of regression formulae, and the distribution of regression coefficients’, J. R. Stat. Soc 85(4), 597–612.

Fisher RA (1922b), ‘On the interpretation of χ^2 from contingency tables, and the calculation of *P*’, J. R. Stat. Soc. 85(1), 87–94.

Fisher RA (1922c), ‘On the mathematical foundations of theoretical statistics’, Philos. Trans. R. Soc. Lond. A 222, 309–368.

Fisher RA (1925), Statistical Methods for Research Workers, Oliver and Boyd, Edinburgh.

Fisher RA (1935), The Design of Experiments, Oliver and Boyd, Edinburgh.

Fisher RA (1956), Statistical Methods and Scientific Inference, Oliver and Boyd, Edinburgh.

Fisher RA & Yates F (1938), Statistical Tables for Biological, Agricultural and Medical Research, Oliver and Boyd, Edinburgh.

Gelman A (2016), ‘The problems with *p*-values are not just with *p*-values’, Am. Stat., Online Discussion 70(2).

Gigerenzer G (2004), ‘Mindless statistics’, J. Socio Econ 33(5), 587–606.

Gigerenzer G, Swijtink Z & Daston L (1989), The Empire of Chance: How Probability Changed Science and Everyday Life, Cambridge University Press, New York.

Goodman SN (2016), ‘The next questions: Who, what, when, where, and why?’, Am. Stat., Online Discussion 70(2).

Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN & Altman DG (2016), ‘Statistical tests, *p*-values, confidence intervals, and power: A guide to misinterpretations’, Am. Stat., Online Discussion 70(2).

Hogben LT (1957), Statistical Theory: The Relationship of Probability, Credibility, and Error, Allen & Unwin, London.

Hubbard R (2004), ‘Alphabet soup: Blurring the distinctions between *p*’s and α’s in psychological research’, Theory Psychol 14(3), 295–327.

Hubbard R (2016), Corrupt Research: The Case for Reconceptualizing Empirical Management and Social Science, SAGE Publications, Thousand Oaks, CA.

Ioannidis JPA (2005), ‘Why most published research findings are false’, PLoS Med 2(8), e124. [PubMed: 16060722]

Jeffreys H (1939), The Theory of Probability, Oxford University Press, Oxford, UK.

Kadane JB, Fienberg SE & DeGroot MH (1986), Statistics and the Law, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, New York.

Kennedy-Shaffer L (2017), ‘When the alpha is the omega: *p*-values, substantial evidence, and the 0.05 standard at FDA’, Food & Drug LJ 72(4), 595–635.

Laplace PS (1827), Traité de Mécanique Céleste, Supplément, Duprat, Paris.

Lehmann EL (1993), ‘The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two?’, J. Am. Stat. Assoc 88(424), 1242–1249.

Lenhard J (2006), ‘Models and statistical inference: The controversy between Fisher and Neyman-Pearson’, Br. J. Philos. Sci 57(1), 69–91.

Lew MJ (2016), ‘Three inferential questions, two types of *p*-value’, Am. Stat., Online Discussion 70(2).

Little RJ (2016), ‘Comment’, Am. Stat., Online Discussion 70(2).

Liu P-H (2003), ‘Do teachers need to incorporate the history of mathematics in their teaching?’, The Mathematics Teacher 96(6), 416–421.

MacKenzie DA (1981), Statistics in Britain, 1865–1930; The Social Construction of Scientific Knowledge, Edinburgh University Press, Edinburgh.

McGrayne SB (2011), The Theory that Would Not Die: How Bayes’ Rule Cracked the Enigma Code, Hunted Down Russian Submarines, & Emerged Triumphant from Two Centuries of Controversy, Yale University Press, New Haven, CT.

Meehl P (1967), 'Theory-testing in psychology and physics: A methodological paradox', Philos. Sci 34(2), 103–115.

Millar AM (2016), 'ASA statement on *p*-values: Some implications for education', Am. Stat., Online Discussion 70(2).

Morrison DE & Henkel RE (1970), The Significance Test Controversy: A Reader, Methodological Perspectives, Aldine Publishing, Chicago.

Neyman J & Pearson ES (1933), 'The testing of statistical hypotheses in relation to probabilities a priori', Math. Proc. Cambridge Philos. Soc 29(4), 492–510.

Pearson ES, Plackett RL & Barnard GA (1990), 'Student': A Statistical Biography of William Sealy Gosset, Clarendon Press, Oxford, UK.

Pearson K (1900), 'X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling', The London, Edinburgh, and Dublin Philos. Mag. J. Sci 50(302), 157–175.

Poisson S-D (1837), Recherches sur la Probabilité des Jugements en Matière Criminelle et en Matière Civile: Précédées des Règles Générales du Calcul des Probabilités, Bachelier, Paris.

Porter TM (1986), The Rise of Statistical Thinking, 1820-1900, Princeton University Press, Princeton, NJ.

Rothman KJ (2016), 'Disengaging from statistical significance', Am. Stat., Online Discussion 70(2).

Rozeboom WW (1960), 'The fallacy of the null-hypothesis significance test', Psychol. Bull. 57(5), 416–428. [PubMed: 13744252]

Salsburg D (2001), The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century, W.H. Freeman and Co., New York.

Savage LJ (1954), The Foundations of Statistics, John Wiley & Sons, New York.

Skellam JG (1969), 'Models, inference, and strategy', Biometrics 25(3), 457–475. [PubMed: 5824399]

Skipper JK, Guenther AL & Nass G (1967), 'The sacredness of .05: A note concerning the uses of statistical levels of significance in social science', Am. Sociol 2(1), 16–18.

Stangl D (2016), 'Comment', Am. Stat., Online Discussion 70(2).

Stigler SM (1986), The History of Statistics: The Measurement of Uncertainty Before 1900, Belknap Press of Harvard University Press, Cambridge, Mass.

Student (1908), 'The probable error of a mean', Biometrika 6(1), 1–25.

Tippett LHC (1931), The Methods of Statistics: An Introduction Mainly for Workers in the Biological Sciences, Williams and Norgate, London.

Wasserstein RL & Lazar NA (2016), 'The ASA's statement on *p*-values: Context, process, and purpose', Am. Stat 70(2), 129–133.

Weisberg HI (2014), Willful Ignorance: The Mismeasure of Uncertainty, John Wiley & Sons, Hoboken, NJ.

Yates F (1951), 'The influence of *Statistical Methods for Research Workers* on the development of the science of statistics', J. Am. Stat. Assoc 46(253), 19–34.

Ziliak ST (2016), 'The significance of the ASA statement on statistical significance and *p*-values', Am. Stat., Online Discussion 70(2).

Ziliak ST & McCloskey DN (2008), The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives, University of Michigan Press, Ann Arbor, MI.

Table 1:

Criteria for a Lasting Framework for Inference Beyond $p < 0.05$

Criterion	$p < 0.05$ Meeting Criterion	Does $p < 0.005$ Meet Criterion?
Provides quantitative basis to address uncertainty	p-value varies with precision of estimator	Same as $p < 0.05$
Rewards increased precision	Significance achieved more easily with higher sample size	Same as $p < 0.05$
Matches intuitive understanding	Statistic akin to p-value developed several times	Stricter threshold easily understood, but requires departure from current intuition
Advocated by statisticians and non-statisticians	Promoted by Fisher, his students, and scientists in a variety of fields	Supported by some statisticians and practitioners, but value still disputed
Computationally feasible for non-statisticians	Fisher and Yates made tables user-friendly and accessible to scientists and practitioners	Any threshold feasible with modern software
Responsive to changing conditions	$p < 0.05$ met needs of a time when few tests were conducted while varying thresholds allowed responses to multiple testing	Addresses current preponderance of tests, but viewed by advocates as a stopgap measure