ADVANCES IN MANAGEMENT ACCOUNTING

ADVANCES IN MANAGEMENT ACCOUNTING

Series Editor: Chris Akroyd

Volumes 1–25:	Marc J. Epstein and John Y. Lee
Volumes 26 and 27:	Marc J. Epstein and Mary A. Malina
Volumes 28–30:	Mary A. Malina
Volume 31:	Laurie L. Burney and Mary A. Malina
Volume 32:	Laurie L. Burney
Volumes 33–37:	Chris Akroyd

ADVANCES IN MANAGEMENT ACCOUNTING VOLUME 37

ADVANCES IN MANAGEMENT ACCOUNTING

EDITED BY

CHRIS AKROYD
University of Canterbury, New Zealand

United Kingdom – North America – Japan
India – Malaysia – China

Emerald Publishing Limited
Emerald Publishing, Floor 5, Northspring, 21-23 Wellington Street, Leeds LS1 4DL.

First edition 2025

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN: 978-1-83662-491-2 (Print)
ISBN: 978-1-83662-490-5 (Online)
ISBN: 978-1-83662-492-9 (Epub)

ISSN: 1474-7871 (Series)

CONTENTS

SPECIAL ISSUE PAPERS

REGULAR ISSUE PAPERS

LIST OF CONTRIBUTORS

Regina F. Bento	Merrick School of Business, University of Baltimore, USA
Avishek Bhandari	Department of Accounting, University of Wisconsin Whitewater, USA
Kevin E. Dow	Department of Accounting and Information Systems, The University of Texas at El Paso, USA
Michael Falta	Department of Accounting and Information Systems, University of Canterbury, New Zealand
Joanna Golden	Fogelman College of Business and Economics, University of Memphis, USA
Frank Hartmann	Institute of Management Research, Radboud University, The Netherlands
Ned Kock	A. R. Sanchez, Jr. School of Business, Texas A&M International University, USA
R. Murray Lindsay	Dhillon School of Business, University of Lethbridge, Canada
Mahfuja Malik	Department of Accounting and Information Systems, Sacred Heart University, USA
Thomas Niederkofler	Institute of Management Research, Radboud University, The Netherlands
Eunsup Daniel Shim	Martin V. Smith School of Business & Economics, California State University Channel Islands, USA
David Trafimow	Department of Psychology, New Mexico State University, USA
Lourdes F. White	Merrick School of Business, University of Baltimore, USA

This page intentionally left blank

ASSOCIATE EDITORS

EDITORIAL BOARD

STATEMENT OF PURPOSE

Advances in Management Accounting (AIMA) is a publication of quality, theoretical, and applied research in management accounting. The journal's purpose is to publish thought-provoking articles that advance knowledge in the management accounting discipline and are of interest to both academics and practitioners. The journal seeks thoughtful, well-developed articles on a variety of current topics in management accounting, broadly defined. All research methods including survey research, field tests, case studies, experiments, meta-analyses, and modelling are welcome. Some commentaries, research notes, and critiques will be included where appropriate.

Articles may range from purely empirical to purely theoretical, from practice-based applications to speculation on the development of new techniques and frameworks. Empirical articles must present sound research designs and well-explained execution. Theoretical articles must present reasonable assumptions and logical development of ideas. All articles should include well-defined problems, concise presentations, and succinct conclusions that follow logically from the data.

REVIEW PROCEDURES

AIMA intends to provide authors with timely reviews clearly indicating the acceptance status of their manuscripts. The results of initial reviews normally will be reported to authors within two to three months from the date the manuscript is received. The author will be expected to work with the Editor and Associate Editors, who will act as a liaison between the author and the reviewers to resolve areas of concern. To ensure publication, it is the author's responsibility to make necessary revisions in a timely and satisfactory manner.

This page intentionally left blank

MANUSCRIPT FORM GUIDELINES

1. Manuscripts should include a cover page that indicates the author's name and affiliation.
2. Manuscripts should include a separate lead page with an abstract (not to exceed 250 words) and six keywords, with references in APA 6th edition style (Google Scholar APA).
3. The author's name and affiliation should not appear on the abstract.
4. Tables, figures, and exhibits should appear on a separate page. Each should be numbered and have a title.
5. To be assured of anonymous reviews, authors should not identify themselves directly or indirectly.
6. Manuscripts currently under review by other publications should not be submitted.
7. Authors should email the manuscript in two WORD files to the editor. The first attachment should include the title page with author details and the second should exclude the title page.
8. Inquiries concerning *Advances in Management Accounting* should be directed to:

Chris Akroyd
at Advances.In.MA@Gmail.com

INTRODUCTION

This volume of *Advances in Management Accounting (AIMA)* includes a special issue on the null hypothesis statistical testing (NHST) paradigm. In the lead article, Professor R. Murray Lindsay argues how the NHST undermines knowledge acquisition in management accounting research and presents reasons why it needs to be abandoned. Following this invited paper, we have five commentaries by experts in statistics from various fields who give their thoughts on Lindsay's arguments.

These special issue papers are then followed by two regular issue papers. The first by Shim and Malik examines the moderating role of corporate social responsibility on top executive compensation and provides evidence from US banks and financial institutions. The second paper by Bento and White presents an innovative experiential learning exercise dealing with ethical dilemmas in management accounting.

The eight articles in Volume 37 represent relevant, theoretically sound, and practical studies that extend our knowledge within the management accounting discipline. These articles manifest the journal's commitment to providing a high level of contribution to management accounting research and practice.

Chris Akroyd
Editor

THE NULL HYPOTHESIS STATISTICAL TESTING *PARADIGM* UNDERMINES KNOWLEDGE ACQUISITION IN MANAGEMENT ACCOUNTING RESEARCH: IT NEEDS TO BE ABANDONED

R. Murray Lindsay

University of Lethbridge, Canada

ABSTRACT

During the last decade, several areas in the biomedical and social sciences experienced a reproducibility crisis, where mounting empirical evidence indicated that many published findings could not be successfully replicated. This crisis resulted in considerable introspection within the field of statistics because the null hypothesis statistical testing (NHST) paradigm is acknowledged as one of its root causes based on widespread agreement that it is deeply flawed. However, unlike in many other areas, there has yet to be a concerted effort within the discipline of accounting to acknowledge these developments, let alone steps taken to improve practice. This essay aims to spark discussion and debate on the validity of the NHST paradigm by presenting a comprehensive case, incorporating the latest arguments and findings, that demonstrates why the paradigm needs to be abandoned, especially in fields where statistical model misspecification looms large and statistical power is low, such as in management accounting. In so doing, the analysis exposes why obtaining robust knowledge in management accounting has proved elusive. Additionally, it offers

Advances in Management Accounting, Volume 37, 1–55

ISSN: 1474-7871/doi:10.1108/S1474-787120250000037001

a new perspective on the reproducibility crisis and critical insights for improving statistical practice.

Keywords: Null hypothesis statistical testing; *P*-values; false positive risk; reproducibility crisis; model misspecification, statistical power; replication

> […] the confidence in the unlimited power of science is only too often based on a false belief that the scientific method consists in the application of a ready-made technique, or in imitating the form rather than the substance of scientific procedure, as if one needed only to follow some cooking recipes … (von Hayek, 1975, p. 438).
>
> The most important task before us in developing statistical science is to demolish the *P*-value culture … (Nelder, 1999, p. 261).

1. INTRODUCTION

The process of turning data into insight is central to the scientific enterprise. In most fields in the biomedical and social sciences, including its "… near-exclusive use" in accounting (Dyckman, 2016, p. 319), null hypothesis statistical testing (NHST) is synonymous with scientific rigor and the centerpiece of the "scientific method." It has shaped our views about science – the need to test hypotheses and how studies must be designed, analyzed, and communicated. The *P*-value has become a substitute for scientific reasoning. It is the criterion used to demarcate the importance of results and increasingly scientific truth, where statistical significance has become the methodological imprimatur for establishing facts and a gatekeeper for publication (Amrhein et al., 2019; Gigerenzer, 2004, 2018; Gigerenzer & Marewski, 2015; Goodman, 1999a; Hubbard, 2016; Lindsay, 1994; McShane et al., 2019; Stark & Saltelli, 2018; Wasserstein et al., 2019).

However, two recent developments connected to the "crisis of validity" (Ziliak, 2019) or "crisis of confidence" (Pashler & Wagenmakers, 2012) in science have introduced the possibility that significant change may be on the horizon. The first is the "reproducibility crisis" experienced in psychology, economics, and the biomedical sciences over the last decade (Baker, 2016). This crisis involved concerns regarding the credibility of scientific findings due to mounting empirical evidence, along with confirmation from analytical investigations, that many published results are false or cannot be successfully replicated (e.g., Begley & Ellis, 2012; Camerer et al., 2016, 2018; Ioannidis, 2005; Open Science Collaboration, 2015).

The second was the crisis occurring in the field of statistics arising from the acknowledgment that one of the root causes of the reproducibility crisis is the overuse, misuse, and misinterpretation of the NHST procedure (Goodman, 2019). Widespread agreement exists inside and outside the statistical community that the current process for curating knowledge claims in science, where *P*-values have played a leading role, is deeply flawed (Steel et al., 2019). In acknowledgment of this situation, the American Statistical Association (ASA)

issued an unprecedented communication in March 2016 entitled "Statement on Statistical Significance and *P*-Values" (Wasserstein & Lazar, 2016). This report denies that a *P*-value can reveal the probability of the null hypothesis, indicate the size or importance of an effect, say much about the strength of evidence for the null, be a substitute for scientific reasoning, or reflect the reproducibility of a result.

The ASA statement is unprecedented. While it contained nothing new that had not already been said in the extensive criticism levied at NHST for close to a century, the President of the ASA, Jessica Utts, writes that the Statement represented the first time the statistical community, as represented by the ASA Board of Directors, issued a formal communication on the matter (American Statistical Association, 2016). Specifically, the motivation for the Statement was to use the ASA's legitimacy as the world's largest professional association of statisticians to "… open a fresh discussion and draw renewed and vigorous attention to changing the practice of science with regards to the use of statistical inference" (Wasserstein & Lazar, 2016, p. 130). Toward this end, the ASA held a three-day conference in October 2017 entitled "Scientific Method for the 21st Century: A World Beyond $p < 0.05$." This conference was followed in 2019 by a special open access, online issue of *The American Statistician* containing over 40 articles. The ASA commissioned these two events to stimulate "a major rethinking of statistical inference, aiming to initiate a process that ultimately moves statistical science – and science itself – into a new age" (American Statistical Association, 2017).

In what Wasserstein et al. (2019) call the "perfect storm," the authoritative ASA statement, the crisis of validity, and the resultant falling public confidence in science produced newfound energy and urgency for reform so as not to squander this unprecedented opportunity given that NHST has been criticized for so many years with little effect (Goodman, 2019). Nevertheless, no consensus exists on what method should replace $P \leq 0.05$, and one is unlikely (Wasserstein et al., 2019). One reason is that the research context – the state of knowledge (theory), the soundness of measurements, and the ability to exert control or employ randomization – differs across the various sciences.[1] Another is due to the foundational controversies involving Frequentist, Likelihood, and Bayesian approaches that have permeated the statistics and philosophy literature for nearly a century and continue to simmer today (Mayo, 2018). This lack of convergence explains why the ASA statement did not provide specific recommendations, causing some commentators to worry that it will become another failed attempt at improving statistical practice (Hubbard, 2019; Matthews, 2019).

Such pessimism is bolstered by what Gigerenzer (2018) calls the "strategic-game hypothesis." The indoctrination of *P*-values as a methodological imprimatur is now part of the sociology of science. As Goodman (2019) explains, the *P*-value is a social phenomenon upon which many social rewards and penalties rest. It does not matter if it does not mean what people think it means; it becomes valuable because of what it buys within the academic community regarding publication, funding, promotion, and the pretense of scientific respectability. Consequently, change will not come easily (see Kuhn, 1970; Ravetz, 1971).

Nevertheless, while this sociologic barrier to change is formidable, it must be overcome and now is an opportune time given the widespread introspection occurring in science. While countless papers and special issues have appeared in statistics, psychology, economics, and medical-related fields – to the point where the literature is intractable – the field of accounting "… appears to be the last of the research communities to face up to the inherent problems of significance test use …" (Dyckman, 2016, p. 319). However, several recent accounting articles (all with a financial accounting orientation) suggest there is now *beginning* to be an appreciation of the validity crisis and the role NHST has played in it; consequently, the discipline may, at long last, be receptive to considering arguments espousing the need for change (see Basu, 2015; Cready et al., 2022; Dyckman, 2016; Dyckman & Zeff, 2014, 2015; Hail et al., 2020; Johnstone, 2022; Kahn & Trønnes, 2019; Kim et al., 2018; Ohlson, 2015, 2022; Stone, 2018).[2] This is an opportunity not to be missed.

This discussion leads to the motivation for this article. It aims to be a catalyst for debate and ultimately change by presenting several crucial arguments explaining why the NHST *paradigm* is, to use Ziliak and McCloskey's (2008, p. 2) phraseology, an "exceptionally bad idea" and in dire need of abandonment because of its devastating consequences for knowledge acquisition in management accounting and control (MAC) research. This paradigm reflects various practices associated with the mechanical use of *P*-values in the manner depicted in von Hayek's (1975) epigraph, including using *P*-values as a methodological curator of knowledge, dichotomizing interpretations of data ("accept" or "reject") based on the $P \leq 0.05$ anchor, ignoring the crucial importance of power (precision) in designing meaningful studies, chasing significance, and equating statistical significance with scientific or practical significance. Additionally, the paradigm underpins biases against publishing "negative" results and close replications, setting the stage for the literature to be, as Nelder (1999, p. 261) expresses, "… little more than a junkyard of false positive results" because no systematic mechanism exists to weed out false ideas.

Given the paradigm's entrenchment, it is perhaps appropriate to address at the onset whether "another" article will be more successful than earlier attempts to motivate change.[3] Several reasons exist. The first involves the current environment, where the validity crisis in science and the increasing concern for MAC's slow rate of progress (discussed later in the paper) provide fertile ground to build a case for change that previously did not exist.

Second, the last (circa) two decades have produced novel criticisms, a sharpening of older arguments, considerable empirical and simulation evidence substantiating the criticisms, and superior ways of illustrating or presenting them. These developments are incorporated in this article.

Third, where possible, the discussion is tailored to MAC's specific research context to sharpen the arguments.

Fourth, in *Nature*'s online survey polling 1,576 researchers for their views on the reproducibility crisis, 50% of respondents indicated that the number one factor required for boosting reproducibility in science is a "better understanding of statistics" (Baker, 2016; cf. Peng, 2015). Research has demonstrated that, as

humans, our intuition about probability and the consequences of variation are pervasively misleading (Gigerenzer, 2018; Tversky & Kahneman, 1974). Formal statistical training in graduate programs offers little redress because it focuses on teaching an ever-increasing number of sophisticated statistical methods at the expense of *statistical thinking* (Gelman, 2016; Stark & Saltelli, 2018; Steel et al., 2019). Statistical thinking involves synthesizing statistical knowledge, contextual knowledge of the situation under investigation, and information regarding data collection to assess concerns with model validity. It also includes understanding sample-to-sample variability and how chance alone can create wide variations in results or produce remarkable patterns where none exist. Finally, it considers how a single analysis fits into the knowledge lifecycle, which is based on following an iterative process centered on replication (both close and differentiated) within a research program rather than the focus on one-off studies (Box, 1994; Chance, 2002; Sedlmeier, 1999; Steel et al., 2019; Wild & Pfannkuch, 2007).

Without a focus on statistical thinking, any attempt to reform NHST or replace it with another method is unlikely to be successful as it too will be applied in a mechanical fashion that violates the complexities of acquiring knowledge, where statistical methods need to be considered an aid to, and not as a substitute for, scientific reasoning (Gelman, 2016; Gelman & Carlin, 2017; Gigerenzer & Marewski, 2015). In an invited article, these issues can be explored more deeply instead of by edict.

Fifth, the article examines the differences between the master statisticians' theories, whose incompatible approaches have become fused as part of NHST in a way that none would have approved. This fusion is largely unknown outside of statistical circles, and its recognition is crucial to appreciating why *P*-values have become misinterpreted into something they cannot provide. Additionally, the insights from one of these master statisticians, R. A. Fisher, are examined to reveal how *P*-values should be used and interpreted to reveal the departure of present-day practices from his prescriptions.

Finally, while it is beyond the scope of this paper to make specific recommendations regarding what a post $P \leq 0.05$ world might look like in MAC research, it will nonetheless help frame such discussions and provide essential insights on the way forward. Attention now turns to presenting the case for change.

2. NHST REPRESENTS A HYBRID OF TWO LOGICALLY INCOMPATIBLE THEORIES

Summary: NHST fuses the logically incompatible theories of R.A. Fisher, on the one hand, and Neyman and Pearson, on the other. This fusion has led to widespread confusion and serious misinterpretations of *P*-values, culminating in the *P*-value becoming a curator of knowledge. Specifically, NHST has been interpreted as capable of mechanizing inferences consistent with science's emphasis on objectivity while reflecting the probability of an erroneous rejection of the null in a specific study. None of these master statisticians would have approved of this hybrid.

NHST does not possess a sound statistical pedigree. Instead, it reflects a hybrid of Fisherian and Neyman–Pearson (NP) theories of statistical inference, whose

details appeared in leading psychology textbooks from the late 1930s to the 1960s (Gigerenzer, 1987, 2004, 2018; Huberty, 1993). As Johnstone (1986) explains, NHST follows NP formally, where there is mention of the null and alternative hypotheses, errors of the first and second kind, and the power of the test. However, epistemologically, NHST departs from NP and follows Fisher in interpreting *P*-values as a measure of evidence against the null hypothesis in the single case. The fusion is remarkable because the two theories are logically inconsistent, resulting in practices and interpretations that neither side would have approved (Gigerenzer, 2018; Goodman, 1993; Hubbard & Bayarri, 2003; Schneider, 2015).

According to Hubbard and Bayarri (2003), most applied researchers (and some statisticians!) are unaware of the conflation of these incompatible ideas, let alone the bitter and unresolved intellectual battle between Fisher and Neyman. Modern textbooks on statistical analysis typically present the subject matter as if the hybrid were a single, unified, and uncontroversial method of statistical inference. Operational similarities also played a role: both theories focused on tail-area probabilities within a relative frequency conception of probability, used the term "level of significance," and frequently adopted the 0.05 or 0.01 critical levels of significance (α) in their examples. Furthermore, both compared the *P*-value to α, which served as a conduit between them.

The discussion in this section explains and contrasts these two theories to accomplish three objectives. The first is to demonstrate that the NHST hybrid is incoherent. The second explains why this fusion has resulted in widespread confusion and serious misinterpretations of *P*-values. The third is to examine Fisher's theory in some detail. Fisher is considered the founder of the modern theory of experimentation (Kempthorne, 1983; Yates, 1964). His extensive, hands-on work at the agricultural research station at Rothamsted and his deep understanding of statistics gave him unparalleled insight into using significance tests. Specifically, he taught when to use significance tests, their purpose, and how to interpret them.

2.1. The Neyman–Pearson Theory of Hypothesis Testing

NP aimed to put classical statistical inference on a watertight mathematical foundation that was deductively valid and, therefore, "objective" (Pearson, 1962). This goal denied the possibility of an evidential interpretation from data to hypothesis in the single case – Fisher's interpretation that is so widespread today (Johnstone, 1986). NP believed that only a Bayesian posterior probability was capable of such an interpretation. They (along with Fisher) did not favor a Bayesian approach because its calculus required incorporating a subjective prior, and this was inconsistent with the early 20th-century *weltanschauungen* that prized objectivity in science (see Gigerenzer, 1987). Instead, NP adopted a strict relative frequency conception of probability and focused on limiting the number of errors over the procedure's long-run use.

NP's theory involves establishing two statistical hypotheses, H_0 and H_1, and specifying α (i.e., the Type I error) and β (i.e., Type II error) *before* conducting the experiment, based on undertaking a subjective cost–benefit consideration of incurring the two kinds of errors for a particular test. These probabilities define

the rejection region for the null and the test's power ($1 - \beta$) for rejecting H_0 when H_1 is true. The error terms refer to the mathematical idealization of data patterns for the test statistic that would arise from random sampling variation under an assumed statistical model if a study was repeated a large (infinite) number of times under constant conditions from the same population (Lehmann, 1993).

NP denied the possibility of an evidential interpretation – moving from data to hypothesis – to maintain the integrity of their theory because, under a relative frequency theory of probability, the hypothesis is either true or false (Neyman, 1937; cf. Mayo, 1985). The repeated sampling routine from a specified (and unchanging) population fixes the reference class (sample space of possibilities), resulting in the parameter of interest (as specified by H_0) being viewed as a fixed, although unknown, quantity. Instead, they counseled that a hypothesis test can only tell a person how to *act* in a specific case, not what to believe regarding a hypothesis' truth or falsity. Specifically, a hypothesis test facilitates a *decision* between two alternative courses of action – to either "accept" or "reject" the null hypothesis – following rules governing an investigator's behavior that provide mathematically determined limits for making errors over the procedure's long-run use. Citing Neyman and Pearson (1933):

> Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong. Here, for example, would be such a "rule of behaviour": to decide whether a hypothesis, H, of a given type be rejected or not, calculate a specified character, x, of the observed facts; if $x > x_0$, reject H, if $x \leq x_0$, accept H. Such a rule tells us nothing as to whether in a particular case H_0 is true when $x \leq x_0$, or false when $x > x_0$. But it may often be proved that if we behave according to such a rule, then in the long run we shall reject H when it is true not more, say, than once in a hundred times, and in addition we may have evidence that we shall reject H sufficiently often when it is false (p. 291).

2.2. R. A. Fisher's Theory of Significance Testing

Under a relative frequency conception of probability, a *P*-value formally represents the probability of getting a test statistic [$T(X)$] equal to or greater than the observed result [$T(x)$] over (hypothetical) repeated sampling, calculated on the basis that the null hypothesis is true, assuming an unbiased test. In notational terms, $P = \text{Prob}(T(X) \geq T(x)/H_0)$. Fisher eschewed NP's long-run sampling error rate interpretation because it is inconsistent with what researchers want to know (Fisher, 1929; cf. Hubbard & Bayarri, 2003). Instead, he adopted an *evidential* interpretation. Specifically, Fisher viewed the *P*-value as a continuous measure of evidence against the null hypothesis by considering it to reflect a data set's (in)compatibility with the null based on data patterns produced under hypothetical repeated sampling. Thus, lower values of *P* represent stronger evidence against the null's validity than higher values (Johnstone, 1987). As an "objective" probability reflecting the data's discordance with the null, Fisher believed that a *P*-value could be communicated to and assessed by others who could make their own (subjective) inference on what the data conveyed about a particular hypothesis.[4] However, Fisher's evidential interpretation was not derived from

mathematical logic but rather from the *psychological* condition of reluctance involving his famous disjunction argument: "*Either* an exceptionally rare chance has occurred, *or* the theory of random distribution [null hypothesis] is not true" (Fisher, 1973, p. 42; cf. Johnstone, 1987).

However, this lack of mathematical logic impacted how Fisher considered and employed *P*-values. Fisher prescribed that significance tests should only be used in contexts where the researcher's knowledge was not well grounded, resulting in the alternative hypothesis reflecting a vague lumpen hypothesis, i.e., any nonzero effect (Anscombe, 1963; Barnard, 1986; Carlson, 1976; Spielman, 1974). Their purpose was to assist in understanding the target phenomenon better by serving a heuristic function in signaling findings that were "worth another look" through further experimentation by helping to distinguish real effects from those that might arise by chance (Anscombe, 1963; Fisher, 1955, 1973, pp. 37–38, 79; Kyburg, 1985).[5] On the other hand, if one's knowledge was sufficiently advanced and a precise alternative to the null could be posited, Fisher (1973, pp. 37, 72–73) recommended calculating a likelihood ratio or a Bayesian posterior probability if objective priors existed in the form of empirical population frequencies.

Importantly, Fisher emphasized that a *P*-value is a *primitive* (pre-Bayesian) or partial measure of evidence against the null and not an inductive probability reflecting the truth of the null (Anscombe, 1963; Barnard, 1986; Fisher, 1973, pp. 37–8, 46). Efron (1998) writes that Fisher's theory reflected a compromise between (strict) frequentist and Bayesian methods. In sympathy with Bayesians, Fisher stressed the need to use all available information in making an inference (Efron, 1998), including negative results (Fisher, 1966, p. 22), which resulted in his disdain for interpreting *P*-values in terms of error rates or using a fixed critical significance level for rejecting the null (Fisher, 1955, 1973, p. 45).[6] Instead, he interpreted a *P*-value flexibly in regarding it as "… a piece of evidence that the scientist would somehow weigh, along with all other relevant pieces of evidence, in summarizing his current opinion about a hypothesis" (Fisher, 1973, p. 45, 50; cf. Cochran, 1967, p. 1461; Dempster, 1998). This postdictive assessment involved incorporating one's (subjective) priors about the null's validity, the strength and features of the study's experimental design, and prior findings (Fisher, 1966, pp. 2–3, 22; 1973, p. 45). In particular, he was clear that the infrequency of the data on the null (i.e., a low *P*-value) should not be confused with its evidential force or cogency (Fisher, 1973, p. 96). Thus, an extraordinary claim (e.g., clairvoyance) would require observing a much lower *P*-value to declare that it was worth taking another look relative to a more plausible expectation. As such, Fisher recognized that the subjectivity of inferences precluded their automaticity.

NP's concepts of "accept" and "reject" irritated Fisher because they communicated a sense of definitiveness that rarely exists in science (Cochran, 1967; Yates, 1955). Instead, Fisher wrote that significant test results must be considered provisionally within a fluid process that depends on replication as part of a cumulative research program (Fisher, 1955, 1966, p. 25; Johnstone, 1987). He stated that his personal preference for using the 0.05 cut-off represented a "low standard" (Fisher, 1926, p. 504) and that a "phenomenon is experimentally demonstratable