Dean&Francis
ISSN 2959-6157

Statistical Significance vs. Practical Significance: Analyzing the Roots of the Replicability Crisis in Modern Scientific Research

Jiacheng Mao

Abstract:

Modern scientific research is facing a severe crisis of reproducibility, with numerous published findings failing to replicate in subsequent independent verification, severely undermining the reliability of scientific knowledge. This paper systematically argues that the methodological imbalance between statistical significance and practical significance constitutes the core root cause of this crisis. We delve into how the misuse of p-values in current research practices—through p-value manipulation and selective reporting—spawns fragile findings. We also reveal the deep-seated impacts of academic incentive structures, publication bias, cognitive biases, and insufficient scientific literacy. Building on this analysis, we propose multidimensional solutions: at the research culture level, reforming incentive mechanisms to prioritize replicable studies and open science practices; and at the educational level, strengthening the cultivation of statistical thinking. The scientific community must achieve a paradigm shift from pursuing statistical significance to evaluating evidence strength and practical relevance, thereby building a more resilient and credible research ecosystem.

Keywords: Reproducibility crisis; Statistical significance; Practical significance; P-value manipulation

1. Introduction

In recent years, the reproducibility crisis has posed a fundamental challenge to the evidence-based scientific knowledge system. This crisis manifests primarily across multiple disciplines—including psychology, medicine, biomedicine, and the social sciences—where published research findings cannot be successfully replicated in subsequent independent, rigorous validation studies. This reveals deep structural flaws within the modern research paradigm at the methodological, institutional, and cultural levels. In the social sciences, the "Psychology Reproducibility Project" initiated by the Center for Open Sci-

ISSN 2959-6157

ence sought to replicate 100 key studies. Its final report, published in Science, revealed that only about 39% of the original findings could be successfully replicated, with the replicated effects showing significantly lower average strength than the initial reports. More alarmingly, in the biomedical field, independent internal studies by Amgen and Bayer Pharmaceuticals both indicate that 70% to 90% of cutting-edge cancer biology research published in top journals—considered cornerstones of drug development—cannot be effectively replicated under standardized industrial laboratory protocols. This signals that science's self-correcting mechanism has, to some extent, failed and produced a dual destructive effect: First, the authority of scientific knowledge is rooted in its objectivity, testability, and cumulative nature. When a large number of established "facts" are proven to be based on unreliable evidence, it undermines public and policymakers' trust in science and demoralizes the scientific community itself. Second, subsequent research, clinical trials, and technology translation based on irreproducible preliminary findings effectively channel precious intellectual and financial resources into illusory avenues. This constitutes a massive waste of opportunity costs and may delay the exploration of genuine scientific questions and effective therapies.

Facing this crisis of deep structural flaws, academic attribution analyses have identified multiple factors, including publication bias, p-value manipulation, low statistical power, and selective reporting. However, this paper argues that beneath the surface lies a more fundamental and critical methodological root cause: the blind worship of "statistical significance" within research practices and evaluation systems, coupled with the systematic neglect of "practical significance." Together, these form the epistemological core of the reproducibility crisis. The prevailing research paradigm has largely transformed "statistical significance" ($p < 0.05$) into the primary criterion for scientific discovery. This has replaced comprehensive evaluations of a scientific question's depth, research design robustness, and the true value of findings with the pursuit of a p-value below a specific threshold. Consequently, we explore how the methodological imbalance between statistical and practical significance constitutes and exacerbates the reproducibility crisis in modern scientific research. This paper delves into the dialectical relationship between statistical significance and practical significance, clarifying how their imbalance leads to the alienation of scientific practice. It provides a systematic theoretical framework for understanding and alleviating the reproducibility crisis.

2. Relevant Theory and Technical Foundations

2.1 The Origins, Applications, and Limitations of Statistical Significance

The modern framework for statistical significance testing integrates the concepts of significance testing and hypothesis testing. Its core tool, the p-value, is widely operationalized as the probability of observing the current sample data or more extreme data under the null hypothesis [1]. Despite controversies surrounding its underlying frequentist philosophical foundation and the logical incompatibility between Fisher's p-value and the Neyman-Pearson hypothesis testing framework, the threshold of $p < 0.05$ for "statistical significance" has become an almost unquestioned "golden rule" across disciplines and a de facto passport for publication.

Academia widely acknowledges that statistical significance serves as a probabilistic decision-making tool addressing the question of "existence." It provides quantitative evidence for researchers to reject the null hypothesis, fundamentally answering: "Could the observed effect or difference stem solely from random sampling error?" However, it inherently contains no information about the effect's direction, magnitude, or real-world significance [2]. This simplified decision rule has led researchers to overlook its underlying probabilistic logic, resulting in inherent limitations that have been deeply criticized by the academic community. On one hand, in studies with massive samples, even effects of negligible magnitude—those with no theoretical or practical significance—can achieve statistical significance ($p < 0.05$) due to the extreme inflation of statistical power. Conversely, an effect with potentially significant practical implications may fail to cross this artificial threshold due to limited sample size and insufficient statistical power, leading to its erroneous omission in decision-making and creating a risk of "false negatives." Furthermore, a common misconception is the direct misinterpretation of p-values as "the probability that the research hypothesis is true" or "the probability of false discovery." P-values do not represent the magnitude or importance of an effect, nor can they serve as a measure of the probability that a hypothesis is true [3]. Cognitive biases have led to a widespread tendency in scientific practice to equate "statistical significance" simplistically with "scientific truth."

2.2 The Essence, Measurement, and Value of Practical Significance

Practical significance shifts the focus from purely random-

ness-based judgments to the importance, value, and impact of research findings within specific disciplinary contexts and the real world. It addresses not merely "whether something is non-random," but the substantive scientific question of "whether it matters"—inevitably involving domain expertise and value judgments. Emphasizing practical significance reflects science's profound return from "discovering any association" to "discovering meaningful associations." [4]

The measurement and evaluation of practical significance employ a richer, more interpretive set of metrics. First is the effect size. Measures such as Cohen's d, correlation coefficient r, and effect size ratios directly quantify the strength of an effect or association, enabling comparability across different research contexts. Second is the confidence interval, which provides an intuitive range for the estimated population parameter, reflecting the magnitude of the effect while expressing the precision and uncertainty of the estimate [5]. A narrow confidence interval entirely within the range of practical significance provides stronger evidence of practical significance than a simple $p < 0.05$. Third are domain-specific metrics. In practical applications, significance must be contextualized. For instance, in clinical medicine, this may manifest as the minimum clinically important difference or the number needed to treat; in economics and social policy, it could involve cost-benefit analysis or potential impacts on social welfare. These metrics anchor abstract statistical findings within specific decision-making contexts, completing the crucial transformation from "statistical results" to "scientific evidence."

2.3 Academic Consensus and Dialectical Relationship Between Statistical Significance and Practical Significance

Statistical significance concerns the strength of evidence against the null hypothesis, essentially serving as a mathematical tool; practical significance, however, pertains to judgments about scientific or practical value, fundamentally constituting a process of scientific reasoning. The fundamental distinction between statistical and practical significance directly leads to common disconnects and paradoxes in research practice. One such disconnect is statistical significance without practical significance. In large-sample observational or omics studies, an effect of negligible magnitude—theoretically or practically insignificant—can easily achieve $p < 0.05$ due to its minuscule standard error. Blind pursuit and overinterpretation of such "significant" findings generate vast amounts of reproducible yet substantively meaningless literature, creating academic noise. For instance, a gene variant might influence height by merely 0.1 centimeters [6]. Conversely, a finding may be practically significant yet statistically insignificant. In highly exploratory fields with heterogeneous samples or extreme acquisition difficulties, a potentially groundbreaking discovery with a large effect size may appear "statistically insignificant" due to insufficient statistical power [7]. The current academic publishing system's preference for positive results often buries such highly exploratory yet "non-significant" findings in "file drawers," leading to systemic loss of scientific knowledge and delayed innovation.

An effect is statistically significant when it is proven unlikely to arise from random chance, and simultaneously demonstrated to possess sufficient magnitude or value (practical significance). However, existing literature indicates that current academic incentives and publication systems almost exclusively favor the former, while systematically neglecting and under-rewarding the latter. This deep-seated methodological imbalance constitutes a critical flaw rendering much research fragile and irreproducible. Therefore, a profound understanding and rebalancing of their dialectical relationship is the primary and essential theoretical step to guide scientific practice out of its current reproducibility crisis.

3. Mechanism Analysis: How the Worship of P-Values Breeds Irreproducible Science

3.1 "P-Value Manipulation" in Research Practice

"P-value manipulation" refers to a series of operational strategies emerging in research practice, driven by the "$p < 0.05$" golden rule for scientific discovery. These strategies aim to optimize p-values rather than explore scientific truth. At its core, this practice transforms data analysis from a rigorous testing process into an exploratory data fitting exercise designed to secure "statistically significant" outcomes. First, p-hacking involves researchers conducting extensive, exploratory, and unregistered model comparisons and data selection during analysis until statistically significant results emerge [8]. Without a predefined analysis plan, this involves: - Combining different subsets of continuous variables - Repeatedly attempting to exclude or retain potential outliers - Continuously adding or removing covariates in regression models to observe p-value changes - Selectively reporting significant results from a set of related outcome variables This process "mines" statistically significant results from purely ran-

ISSN 2959-6157

dom data. Second, data dredging occurs when researchers conduct multiple interim analyses or continuous monitoring during data collection instead of performing a single final analysis after completing the prespecified sample size. Data collection ceases immediately upon any interim analysis crossing the $p < 0.05$ threshold. This violates sequential analysis principles by failing to correct for multiple testing, artificially inflating the false positive rate [9]. Finally, selective reporting occurs when researchers, reviewers, and journals favor submitting and publishing only statistically significant findings while shelving or concealing complete studies or partial analyses yielding non-significant results that fail to support the research hypothesis. This creates a highly biased sample of published literature, severely distorting perceptions of the true body of evidence on a scientific issue.

P-hacking, data dredging, and selective reporting collectively form a mechanism that systematically distorts the production of scientific evidence, dramatically inflating the actual probability of Type I errors (false positives) far beyond the nominal 5% threshold. This implies that a substantial proportion of published research findings are essentially embellished coincidences or noise, artificially inflated through repeated attempts and data-dependent modeling. These findings themselves are products of statistical test abuse. Their inherent fragility and instability ensure they will inevitably fail under independent, standardized replication, directly fueling the reproducibility crisis.

3.2 Distorting Effects of Incentive Structures and Publication Bias

P-value manipulation in research practice is not an isolated instance of misconduct but is deeply embedded within a distorted academic incentive and dissemination ecosystem. This system systematically encourages and reinforces the pursuit of statistical significance through its inherent reward structure. Specifically, within contemporary academic frameworks, researchers' career trajectories heavily depend on the quantity and frequency of publications in high-impact journals. Performance evaluation systems create a "production pressure" that makes generating novel, attention-grabbing positive results a critical necessity for professional survival [10]. Simultaneously, high-impact journals, driven by the need to attract citations and attention, exhibit a pervasive preference for novelty, drama, and positive outcomes. A clean, unambiguous statistical significance result ($p < 0.05$) is regarded as the core element of a complete, credible scientific narrative, while complex, ambiguous, or negative findings are often dismissed as "lacking contribution" or "unpersuasive," making them difficult to publish. This academic culture and journal preference for significance form a self-reinforcing vicious cycle. To survive and succeed, researchers must pursue p-values to publish in high-impact journals → Journals consequently become saturated with (potentially manipulated) positive results → This in turn fosters a false perception across academia that "negative results are extremely rare or worthless," raising the comparative benchmark for subsequent research → This environmental pressure further compels the next generation of researchers to adopt p-value manipulation strategies earlier and more proactively, intensifying the system's overall worship of p-values. This incentive structure creates a conflict where "making publishable choices" sometimes clashes with "making scientifically sound choices."

3.3 Cognitive Biases and Lack of Scientific Literacy

The misuse and overreliance on p-values are deeply rooted in widespread cognitive biases and insufficient training in scientific methodology. On one hand, despite clear definitions in statistics textbooks, severe misinterpretations of p-values are common in practice. The most common errors include interpreting p-values as "the probability that the research hypothesis is true" (which is actually $P(D|H_0)$ rather than $P(H_1|D)$), or equating "$p < 0.05$" with "evidence that the probability of the finding being true exceeds 95%" or "evidence of a large effect." This cognitive bias—equating "evidence strength" with "probability of hypothesis being true"—causes both researchers and readers to overestimate the certainty represented by a single p-value [11]. On the other hand, inadequate statistical power analysis during exploratory phases or in resource-constrained fields leads to insufficient sample sizes. Low-power studies not only struggle to detect genuine effects (high Type II error rates) but, more critically, yield relatively low true positive probabilities even when achieving $p < 0.05$ significance. This counterintuitive insight implies that published low-power studies are disproportionately likely to report false positives, directly exacerbating the reproducibility crisis and incentivizing selective reporting. Furthermore, current academic reporting norms overemphasize the p-value as a singular metric while systematically neglecting the reporting of effect sizes and their confidence intervals. Reporting only p-values prevents readers from discerning whether a "significant" effect represents a truly meaningful discovery or a statistical artifact arising from large sample sizes [12]. Simultaneously, omitting confidence intervals obscures the inherent substantial uncertainty in estimates, hindering the scientific community's ability to assess and integrate the true value of research

findings. This results in literature flooded with statistically significant yet practically insignificant discoveries, further diluting scientific reliability and information content.

In summary, this complex structural issue—a confluence of distorted research practices, systemic incentive biases, and widespread cognitive and literacy deficits—is not merely a technical statistical fallacy but a symptom of contemporary research ecosystems. It guides researchers to prioritize "statistically publishable" outcomes over "scientifically robust" ones, ultimately producing vast quantities of seemingly polished yet fundamentally fragile scientific literature. This constitutes the core mechanism generating the reproducibility crisis.

4. Solutions and Paradigm Shift

4.1 Methodological Innovation: Adopting a More Comprehensive Evidence Evaluation System

A single p-value cannot convey the full information required for scientific inference. Methodological innovation drives the adoption of richer, more robust evidence evaluation standards. First, mandate reporting of effect sizes and confidence intervals. Effect sizes quantify the magnitude of phenomena, enabling comparability of research findings across different contexts. Confidence intervals provide the range of effect estimates, intuitively illustrating precision and uncertainty. This shifts scientific discourse from the binary question of "does an effect exist?" (based on p-value judgments) to the more nuanced inquiries of "how large is the effect?" and "how confident are we in our estimate?" For instance, a narrow confidence interval far from zero provides both strong evidence and implications of practical significance; conversely, a wide interval—even if containing a "significant" result—warns of conclusion instability. Second, it advocates Bayesian statistical methods. Bayesian approaches directly quantify the strength of data support for different hypotheses by calculating Bayes factors or posterior probability distributions, offering an alternative evidence evaluation paradigm. Furthermore, Bayesian methods offer continuous evidence metrics, allowing researchers to state that "the data supports hypothesis H1 ten times more strongly than H0" rather than merely "rejecting the null hypothesis." Additionally, Bayesian approaches naturally incorporate prior knowledge and provide a more comprehensive characterization of parameter uncertainty, helping overcome limitations inherent in relying solely on p-values. Third, pre-registration and registration reports curb "p-value manipulation" and "selective reporting" at their source. Pre-registration requires researchers to register their study hypotheses, experimental designs, variable measurement methods, and planned statistical analysis protocols on a public, time-stamped platform before data collection begins. Journals, after peer review, accept papers for publication based on the significance of the research question and the rigor of the methodology, regardless of whether the final results are "significant." This decouples the value of research from the "significance" of results, greatly reducing the motivation for p-hacking and hiding negative results. It encourages honest data reporting and transparent differentiation between exploratory and confirmatory analyses.

4.2 Reforming Research Culture and Incentive Mechanisms

Methodological innovation requires supporting research culture and incentive mechanisms to achieve widespread adoption. The goal is to establish an ecosystem rewarding rigorous, transparent, and cumulative contributions. First, the bias favoring "originality above all" must be dismantled, granting replication studies their rightful academic standing. Academic journals should establish dedicated sections or launch new publications to provide formal channels for publishing reproducible studies. Second, reform academic evaluation systems by incorporating open science practices as key indicators of research quality. Encourage or mandate the sharing of research data and analysis codes to facilitate verification and replication by others. Recognize and publish rigorously reviewed negative results to fill knowledge gaps caused by "file drawer effects." Promote open access to research materials. Furthermore, journals should lead by example in reforming their manuscript acceptance criteria. Editors and reviewers should prioritize the significance of research questions, the rigor of methodological design, the transparency of analyses, and the honest reporting of results. By publishing submission guidelines that encourage reporting effect sizes, confidence intervals, and adherence to open science practices, journals can powerfully guide the entire research community toward better practices.

4.3 Improving Science Education and Communication

The paradigm shift ultimately depends on elevating the cognitive level of researchers and the public, which requires fundamental change through education and communication. Therefore, training for undergraduates, graduate students, and early-career researchers must move beyond the mechanical application of statistical testing procedures. Teaching should emphasize cultivating "sta-

ISSN 2959-6157

tistical thinking"—how to reason and make decisions amid uncertainty—rather than merely instructing software operation and formula calculations. Simultaneously, researchers and science communicators must develop critical skills to interpret findings. For the public, science communication should avoid simplifying "research findings" into definitive 'yes' or "no" conclusions, instead highlighting science's cumulative, incremental nature and the inherent uncertainty of new knowledge.

A paradigm shift constitutes a systemic scientific revolution involving methodology, institutions, and culture. It necessitates adopting more comprehensive evidence evaluation systems, reforming incentive structures to reward rigor and transparency, and fundamentally enhancing the methodological literacy of the scientific community. This will gradually build a more resilient and trustworthy scientific enterprise, thereby genuinely overcoming the reproducibility crisis and propelling science forward with robustness.

5. Conclusion

The reproducibility crisis in modern scientific research is a crisis of scientific quality triggered by methodological imbalance. The epistemological core of this crisis lies in the narrow pursuit of statistical significance coupled with the systematic neglect of practical significance. This imbalance manifests not only in individual researchers' data analysis practices but is also entrenched in the incentive structures governing academic publishing, career advancement, and resource allocation. Consequently, academic literature is flooded with research findings that are statistically significant yet scientifically fragile, unable to withstand replication. Through systematic argumentation, we demonstrate: First, as a finite probabilistic tool, the inherent limitations of p-values and their widespread misinterpretation in practice render them incapable of serving as the sole arbiter of scientific discovery. Second, judgments of statistical significance divorced from effect size and practical context are not only meaningless but potentially harmful, distorting the process of scientific inquiry and squandering precious research resources. Finally, resolving the reproducibility crisis cannot be achieved through piecemeal technical fixes but requires a paradigm shift driven by coordinated advances in methodology, research culture, and scientific education. Methodologically, alternative frameworks like Bayesian statistics should be promoted, and pre-registration with registered reports should become standard practice across many disciplines. Institutionally, academic evaluation systems must genuinely reward rigor, transparency, and cumulative contributions, transforming open science from an optional practice into a core value. Educationally, cultivating critical statistical thinking should take precedence over teaching computational skills. The ultimate goal of this paradigm shift is to restore science to its essence: pursuing truth and serving society, rather than relying on the "$p < 0.05$" benchmark.

References

[1] Wasserstein, R. L. (2016). ASA statement on statistical significance and P-values. American Statistician, 70(2), 131-133.

[2] Cohen, J. (1994). The earth is round (p<. 05). American psychologist, 49(12), 997.

[3] Cumming, G. (2014). The new statistics: Why and how. Psychological science, 25(1), 7-29.

[4] Goodman, S. (2008, July). A dirty dozen: twelve p-value misconceptions. In Seminars in hematology (Vol. 45, No. 3, pp. 135-140). WB Saunders.

[5] Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. European journal of epidemiology, 31(4), 337-350.

[6] Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, G. B. (2015). The fickle P value generates irreproducible results. Nature methods, 12(3), 179-185.

[7] Ioannidis, J. P. (2005). Why most published research findings are false. PLoS medicine, 2(8), e124.

[8] Muff, S., Nilsen, E. B., O'Hara, R. B., & Nater, C. R. (2022). Rewriting results sections in the language of evidence. Trends in ecology & evolution, 37(3), 203-210.

[9] Májovský, M., Černý, M., Kasal, M., Komarc, M., & Netuka, D. (2023). Artificial intelligence can generate fraudulent but authentic-looking scientific medical articles: Pandora's box has been opened. Journal of medical Internet research, 25(1), e46924.

[10] Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., ... & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. Annual review of psychology, 73(1), 719-748.

[11] Althubaiti, A. (2023). Sample size determination: A practical guide for health researchers. Journal of general and family medicine, 24(2), 72-78.

[12] Kim, J. H. (2022). Moving to a world beyond p-value< 0.05: a guide for business researchers. Review of managerial science, 16(8), 2467-2493.