Compatibility ranges as a practical alternative to the significant and non-significant statistical dichotomy

Alessandro Rovetta[1]

AFFILIATION
1 R&C Research, Research and Disclosure, Bovezzo, Italy

CORRESPONDENCE TO
Alessandro Rovetta. R&C Research, Research and Disclosure, Via Brede Traversa II, 25073, Bovezzo, Italy. E-mail: rovetta.mresearch@gmail.com
ORCID iD: https://orcid.org/0000-0002-4634-279X
KEYWORDS
compatibility, confidence interval, hypothesis testing, magnitude fallacy, nullism, surprisal

Received: 4 October 2023, **Revised:** 27 May 2024, **Accepted:** 31 May 2024
Public Health Toxicol. 2024;4(2):6
https://doi.org/10.18332/pht/189530

ABSTRACT

Science can be defined as a social system built on the concept of critical agreement on evidentiary states. The latter must be achieved through rational thinking, communication, and the so-called 'scientific method', which involves a series of procedures aimed at ensuring the replicability of investigative experiments. In this regard, the dichotomization of statistical results into 'significant' and 'non-significant' has led to a long series of replication failures and fostered the misleading expectation that a mere numerical criterion can replace analytical reasoning. Especially in fields like toxicology and public health, such misuse can have serious consequences. Indeed, no study can prove that a result is (not) significant since uncertainty is always part of scientific research. At most, based on the above considerations and a comprehensive analysis of costs, risks, and benefits, it can be decided whether a certain phenomenon meets the threshold of scientific evidence required to undertake concrete actions. In light of this, the present manuscript proposes and discusses alternative concepts to statistical dichotomies, such as ranges of compatibility and effect size. Furthermore, it emphasizes the necessity to investigate the compatibility of the experimental data with all the relevant target hypotheses (not just the null one) and all the background assumptions. Finally, it proposes a compact framework for a complete presentation of results, including effect size. In this regard, the adoption of multiple confidence/compatibility intervals or surprisal intervals is recommended.

INTRODUCTION

Context

The concept of statistical significance is widely employed in medical research, especially in clinical and pharmacological studies and, at the same time, it is one of the most controversial, debated, and misunderstood topics since its original formulation[1-3]. In particular, it is often mistakenly believed that statistical testing can provide objective evidence about the real significance of phenomena (e.g. their existence or relevance). On the contrary, such a procedure is based on various hypotheses assumed to be true *a priori* and choices conditioned by an ineliminable margin of subjectivity[1-8]. Although the ambiguous concept of 'significance' was discussed by previous authors (e.g. William Sealy Gosset, otherwise known as Student), it was Sir Ronald Fisher who made it particularly famous in the 1920–1930 decade[9]. After selecting an appropriate investigative methodology and assuming a priori that mere chance is the only phenomenon at play, a researcher can calculate the probability of obtaining an experimental result (the test statistic, e.g. the *t* of Student's t-test) as or more extreme than that obtained in the experiment. This probability, referred to as the p-value, corresponds to the expected frequency of the statistical event within an infinite (very large) population of valid applications (i.e. where all background assumptions hold). In this regard, it is important to clarify some fundamental aspects. Firstly, Fisher's approach involves establishing a series of hypotheses (the so-called statistical model) that are assumed to be perfectly met (true). These include underlying hypotheses (e.g. random sampling,

normal distribution, linearity, etc.) and the null hypothesis of no effect or association. Moreover, as stressed by Greenland[5], such background assumptions also involve human aspects (e.g. transparency, honesty, collaboration, competence, etc.). Once this is done, supposing *a priori* that no causal mechanism exists or we are exclusively in presence of a set of cofactors behaving randomly, it is a matter of assessing how 'surprising' (i.e. 'statistically significant') the obtained result compared to the null hypothesis prediction (previously set as true). Fisher (1920s) initially suggested a heuristic threshold of 5% to consider the outcome as unexpected (significant, $p<0.05$) or expected (non-significant, $p\geq0.05$)[10]. This threshold was intended to work decently in most real applications. Subsequently, he regretted his own proposal, emphasizing that the p-value should be used as a graded measure of the strength of evidence against the null hypothesis[2,11].

Egon Pearson and Jerzy Neyman (1933), critics of the idea of statistically evaluating the true significance of a hypothesis (a valid point), proposed instead a novel decision-theoretical approach (rule of behavior)[12]. The conventional goal, strictly conditional on the same underlying assumptions described above, is to establish two contrasting simple hypotheses: the null hypothesis of an exactly zero effect and the alternative hypothesis of non-exactly zero effect. If the experimental test statistic (e.g. Student's t) is more extreme than a predetermined critical value (e.g. $t_c=1.96$ in a very large sample), then the null hypothesis is arbitrarily rejected in favor of the alternative hypothesis; otherwise, the null hypothesis cannot be rejected (but neither accepted, although Neyman and Pearson originally used such a word). As specified by Neyman and Pearson themselves, the choice of this threshold is an open problem that, according to the latter Neyman, must be grounded in the evaluation of costs, risks, and benefits (as well as the selection of the hypothesis to be examined)[13].

Applicability in scientific investigations

As explained by Fisher in 1955, the Neyman-Pearson approach (NP) can be useful in well-defined, limited contexts (e.g. inference regarding the proper functioning of a population of light bulbs produced by a factory), but it is generally not-recommendable in the scientific scenario[14]. In modern terms, Neyman-Pearson inference can be summarized as follows[2,3]: The critical region (e.g. $z > z^* = 1.96$) can be defined in terms of decision p-values (e.g. $p< \alpha=0.05$). Assuming the process is iterated in numerous equivalent applications (i.e. all background hypotheses are met in each of these), it amounts to committing a total of $\alpha\cdot100\%$ type I errors or 'false positives' (sometimes written as α%, i.e. the percentage version of α) and, if power $(1-\beta)\cdot100\%$ is also fixed, of $\beta\cdot100\%$ type II errors or 'false negatives'. In other words, the p-value is a mere decision-making index devoid of direct scientific meaning (i.e. if $\alpha=0.05$, $p=0.049$ and $p=0.001$ are decisionally equivalent as they lead to the same decision). The so-called statistical confidence is based on the concept of coverage probability: only in numerous equivalent applications, $(1-\alpha)\cdot100\%$ (e.g. 95%) of the confidence intervals of the form $(1-\alpha)\cdot100\%$ (e.g. 95%) will contain the population parameter. Thus, the first essential aspect is that such a framework never informs decisions on individual studies (e.g. it is incorrect to think that a 95% confidence interval has a 95% probability of containing the true value) since it is mathematically structured to operate merely on high numerosity under ideal conditions[2,3,7,11,12]. In addition, as evidenced by the (re)current 'replication crisis', equivalent conditions cannot be guaranteed in practice due to sources of scientific uncertainty that are not only difficult to model (e.g. researchers' attention, confounding factors, proper sampling, etc.) but are also often unknown[1-7,15-17]. This leads to decisions that are inconsistent with the predetermined goal (due to what are sometimes called 'Type III errors')[16]. According to the recommendations of some of the leading global authorities in the field – including the American Statistical Association – and recent initiatives like the International Committee Against the Misuse of Statistical Significance (ICAMSS), the p-value should therefore be employed in a neo-Fisherian manner[1,2,18-24]. Specifically, the p-value is a continuous measure of the compatibility of the statistical result with the target hypothesis (e.g. the point null hypothesis of an exactly zero effect), whose interpretability in this sense is conditional on the background assumptions. The notion of 'compatibility' – which has been traced back to Karl Pearson[25] (father of Egon) in 1900 – to indicate the degree of agreement of the data with the target hypothesis as evaluated by the chosen test, is a much more moderate expression than 'support' and it is not conceived to make terminal decisions. Indeed, supporting a hypothesis means assigning greater plausibility to the latter compared to others; on the contrary, showing a certain degree of compatibility with a hypothesis does not exclude the presence of other hypotheses that are equally or even more consistent with the data (as evaluated by the chosen statistical model) or the scientific phenomenon. Concerning mere statistics, p-values close to 1 indicate high compatibility, while p-values close to 0 indicate low compatibility. Hence, confidence intervals become compatibility intervals: for instance, a 95% compatibility interval of the form (x, y) contains all hypotheses whose p-value is greater than 0.05, meaning they are more compatible with the data than hypotheses predicting effects 'x' and 'y' (as conditionally assessed by the statistical test)[7,22,26,27].

Common errors in public health

As extensively documented in the literature, there is a growing need to raise awareness within the medical community about the correct use of the aforementioned frequentist-inferential methods[1]. In light of the costs and

risks linked to investigations in public health, it is essential to provide an overview of the most common errors and seek both short-term and long-term solutions. The first common flawed approach is the so-called null hypothesis significance testing (NHST), where only the point hypothesis of zero effect is considered and evaluated in dichotomous terms of 'significance' and 'non-significance'[24,28,29]. Even in the utopian scenario where all background assumptions are perfectly met, a large p-value for the null hypothesis only indicates a high degree of compatibility of the latter with the data (as conditionally evaluated by the test) but does not in any way support such a hypothesis over others. An easy counterexample is as follows: Let (1–9) be an 80% compatibility interval associated with the best point estimate of a hazard ratio HR=3. The p-value for the mathematically null hypothesis HR*=1 is thus equal to p=0.20 (as HR=1 is the first limit of the 80% compatibility interval). Many would wrongly classify this outcome as '(statistically) non-significant' only because the p-value for the null hypothesis is greater than 0.05; however, under the conditions described above, the data have exactly the same statistical compatibility with the decidedly non-null hypothesis HR*=9 (p=0.20, as HR=9 is the other limit of the 80% compatibility interval). But that is not all: the hypothesis most compatible with the data is not HR*=1 but HR*=3 (p=1, since HR=3 is the best point estimate). Thus, conditionally on the background assumptions, we can only conclude large statistical uncertainty and not the absence of any significance: indeed, this outcome is highly compatible with hypotheses of both low and broad effect[7,8]. The second issue, closely related to nullism (mere interest in the point null hypothesis), is the lack of distinction between large and small effect sizes. For instance, we could encounter situations where two (or more) hypotheses consistent with a low-magnitude phenomenon (e.g. HR*=1 and HR*=1.2) lead to quite different degrees of compatibility according to the adopted test (e.g. p=0.01 and p=0.05, respectively). If the best point estimate is consonant with a non-negligible effect (e.g. HR=2.4), such a scenario also signals high uncertainty[7,8,20]. The third issue is that, often, many authors mix Neyman and Pearson's rule of behavior with Fisher's significance testing, even though these approaches are based on mathematically and epistemologically incompatible formulations[3]. Therefore, this article discusses a possible approach to mitigate such misunderstandings.

METHODOLOGICAL APPROACH

Foundations

Human psychology – and thus all biases and inevitable sources of uncertainty that it carries with it – is an integral component of scientific investigations[3,7,30,31]. Since its earliest Bayesian formulations, modern statistics has been modeled on human perception, taking into account cognitive and even cultural aspects (e.g. Good 1952)[32]. In this regard, the reasons behind the vast success of NHST should be searched in university education and cognitive distortions aimed at oversimplifying complex concepts[7,18,24,33]. As a remedy, Rafi and Greenland[26] propose to explain the ambiguous and unclear concept of 'statistical significance' through familiar statistical phenomena such as flipping an unbiased two-headed coin. The so-called 'surprisal' (or 'S-value') thus represents, conditionally on the background assumptions, the number of consecutive heads one would need to obtain – by flipping an unbiased two-headed coin – to match the statistical surprise of the result calculated in the experiment (the test statistic). This approach, subsequently extended to statistical compatibility via surprisal intervals[7,8], resolves some thorny issues not only regarding the interpretation of p-values but also their mathematical-statistical utilization. Indeed, the p-value has been widely adopted by the neo-Fisherian statisticians as a graded/continuous measure of the refutational evidence against one or more hypotheses[11,14,34]. As recently demonstrated by Greenland[3,5], such an interpretation is legitimate within this framework. However, this 'divergence' p-value (even if intended as a mere descriptive indicator of the discrepancy between observed data and the predictions of the statistical model) possesses counterintuitive properties. For instance, the difference in information content between p_1=0.05 and p_2=0.10 is larger than that between p_3=0.95 and p_4=1, despite 1 - 0.95 = 0.10 - 0.05 = 0.05[7]. This occurs because the ratio p_2/p_1 is 2, while the ratio p_4/p_3 is less than 1.1 (i.e. the probabilities are quite different in the first pair and very similar in the second). If we compare p to the probability of obtaining S consecutive heads, by flipping an unbiased coin (a phenomenon of which we have immediate perception), we get that $S= -\log_2 p$ (from $p=0.5^S$). Thus, the S-values in the four preceding cases are, respectively: S_1=4.3, S_2=3.3, S_3=0.07, and S_4=0. In other words, the first statistical result is as surprising as about 4 consecutive heads, the second is as surprising as about 3 consecutive heads, and the third and fourth are markedly less surprising than getting head when flipping an unbiased coin (compared to the model prediction). By doing so, the difference in information becomes evident ($S_2 - S_1 = 1$ while $S_4 - S_3 = 0.07$). Nevertheless, the current scenario is consistent with a widespread rejection of methodologies that are too innovative or complex. Accordingly, this study proposes and discusses a graded scale of statistical compatibility whose ranges are based on the information (surprisal) contained within.

Graded compatibility

Consistently with Fisherian indications, Muff et al.[35] recently proposed a graded scale to read p-values as measures of evidence against a hypothesis. Although such an attempt has been subject to criticism, Amrhein and Greenland[21] argue that the proposal of Muff et al.[35] does '*more good than harm*' since it contrasts the dichotomous, overconfident interpretation of statistical significance. Nevertheless, the term 'evidence' – that could be defined, according to the Oxford English Dictionary, as '*facts or observations adduced*

in support of a conclusion or statement' – transcends the actual epistemological capabilities of the p-value when drawing practical conclusions. Indeed, since a statistical hypothesis (SH) is just our attempt to represent an empirical hypothesis (EH) on a mathematical level, p-values could measure evidence against 'SH' but not 'EH': it all depends on our ability to select a proper SH based on EH. In this regard, we should also acknowledge that a real phenomenon might be too complex to be well-represented by simple statistical hypotheses. For these reasons, a framework to elaborate a graduated scale of mere compatibility is proposed here (Table 1). This should be constructed based on two main aspects: 1) the information contained within the range (surprisal), and 2) the predetermined scientific objective. Specifically, the first point aims to realize the gradation of the scale according to the degree of surprise (incompatibility) of the results compared with the fixed hypothesis (which does not necessarily have to be the null hypothesis of no effect) as conditionally assessed by the chosen test. The second point emphasizes that there is no absolute or unique way to evaluate a statistical result and that, as stated by Neyman[13] in 1977, even the choice of the statistical hypothesis to investigate must be calibrated to the scientific objective.

The goal is to establish various thresholds – thus counteracting the dichotomous view – and, at the same time, to prevent the adoption of incorrect and misleading expressions such as '(non) significant'. Some heuristics for setting the various thresholds could be as follows: E1) Each subsequent threshold is half of the previous one (so that the S-value increases by 1 bit of information for each jump). For example, α_1=0.250, α_2=0.125, α_3=0.063, α_4=0.032, and α_5=0.016; and E2) The thresholds are consistent with common ones (e.g. α_1=0.20, α_2=0.10, α_3=0.05, α_4=0.01, and α_5=0.001).

However, it must be clear that, being a purely descriptive approach, the specification of these ranges aims only to simplify communication and limit overstatements. Moreover, the publication of a specific pre-study protocol lends much more weight – even in the eyes of editors and reviewers – to the issue of threshold selection and the evaluation of statistical compatibility in relation to the research scope.

Table 1. Compatibility ranges protocol. All the multiple thresholds should be established and published (with a digital object identifier, or DOI) before conducting the experiment

The ranges of p-values	Compatibility range	Surprisal range
$\alpha_1 \leq p \leq 1$	Marked	Minimal
$\alpha_2 \leq p < \alpha_1$	High	Weak
$\alpha_3 \leq p < \alpha_2$	Moderate	Marginal
$\alpha_4 \leq p < \alpha_3$	Marginal	Moderate
$\alpha_5 \leq p < \alpha_4$	Weak	High
$0 < p < \alpha_5$	Minimal	Marked

Compatibility distributions and intervals

The selection of a specific target hypothesis concerning a single-point effect is generally insufficient to properly inform a scientific conclusion, since it does not allow us to evaluate the consistency of the experimental scenario with all relevant hypotheses. Recent literature proposes various ways to address this issue, like representing the so-called 'p-distributions' or 'S-distributions' (i.e. 'compatibility distributions' and 'surprisal distributions', respectively) to observe the (in)compatibility of the data with the set of all possible target hypotheses[20,26]. In this regard, some authors propose adding pre-study protocols to divide such hypotheses into different groups based on the effect size[36]. A practical example of application is provided in the Supplementary file. Nonetheless, these modalities of presentation are confined within manuscripts – as they are hardly communicable in introductory or summary sections such as abstracts – and are difficult to implement when dealing with several outcomes within the same study. To solve this problem, a novel convention for reporting multiple compatibility and surprisal intervals can be adopted[7,8]. According to the E2 protocol (Table 1), we could choose three compatibility intervals associated with the thresholds α_1=0.20 (80% CI), α_3=0.05 (95% CI), and α=0.01 (99% CI) as follows: 80|95|99% CI = (a–b|c–d|e–f). For instance, considering a calculated best point estimate of 10, if 80% CI = (6–14), 95% CI = (3–17), 99% CI = (0–20), we can write 80|95|99% CI = (6–14|3–17|0–20). This tells us that all hypotheses that predict an effect between 6 and 14 are, at least, highly compatible with the data (p>0.20, i.e. S<2.3). At the same time, all hypotheses between 3 and 17 are, at least, marginally compatible with the data (p>0.05, i.e. S<4.3). Finally, all hypotheses between 0 and 20 are, at least, weakly compatible with the data (p>0.01, i.e. S<6.6) or, equivalently, all hypotheses outside (i.e. those <0 or >20) are minimally compatible with the data (p<0.01, i.e. S>6.6).

DISCUSSION

We need descriptive approaches to reach causal inference

There are various non-descriptive methods for attempting to solve the problem of testing single-point hypotheses. Among these, the so-called 'equivalence testing' involves setting a target hypothesis in the form of a range and then adopting a dichotomous decisional rule of behavior. For example, when dealing with adverse events related to LDL cholesterol levels, a certain research group could define an effect as practically null when the average change falls between -5 and 5 mg/dL (range null hypothesis). However, as shown by Greenland[3], this procedure does not escape the criticalities that permeate the standard Neyman-Pearson approach; rather, it introduces additional ones. Firstly, the

dichotomous decision 'rejection versus non-rejection', to be made in individual studies, becomes extremely conditional on the choice of the initial null range (to be maintained in all studies). On this point, it is particularly complex to establish the width of this interval based on the scientific context and the associated costs, risks, and benefits – which could become clearer only over the course of various experiments – while taking into account biases and financial interests[3,5,28,30]. Secondly, the dichotomization of hypotheses creates regions of statistical equivalence that do not correspond to regions of scientific equivalence: for instance, an increase of 6 mg/dL is not scientifically equivalent to an increase of 30 mg/dL despite both point hypotheses belonging to the statistical alternative hypothesis (which assumes the form $h < -5$ or $h > 5$) within the same side ($h > 5$). Third, in many epidemiological situations such as the COVID-19 crisis, the whole scientific landscape is constantly changing (e.g. the occurrence of viral mutations can substantially alter the duration and symptoms of the disease), thus invalidating the equivalence request (e.g. the risk-benefit ratio can change drastically). This further underscores the necessity of engaging in critical thinking rather than relying on mere numerical criteria. Additionally, the generally overlooked aspect is that scientific inference requires consistency not only among statistical studies (which should always include randomized experiments to provide causal evidence) but also among extra-statistical evidence (analysis of biochemical mechanisms, clinical and medical observations, etc.). When this multidisciplinary set of reasoned, epistemic evidence converges in the same direction, causal inference can be claimed[1,2,7,8,15,17,23,30]. However, since we are forced to make dichotomous final decisions, such as approving or rejecting drugs, it is important to establish guidelines that serve as a good compromise while realizing inference. In valid repetitions, one should expect a therapeutic effect within an optimal pre-defined range in most cases, although the size of this effect must be assessed continuously or, at least, through a graduated scale; multiple ranges of effect (e.g. small, medium, large, etc.) could be defined in order to avoid dichotomization (Supplementary file). The expression 'valid repetitions' emphasizes the need to approach equivalence conditions as much as possible (indeed, minimizing sources of uncertainty remains fundamental) without the implausible expectation to perfectly achieve them. Concerning pharmacological development, there may be circumstances where the dose to administer must be lower, the treatment must be implemented for a shorter duration, or the initial clinical conditions are simply different. In such situations, it is appropriate to recalibrate the descriptive protocol or, if possible, establish a pre-study protocol of protocols that encompass various optimal, graded ranges based on different scenarios. The ultimate goal is to properly inform the so-called 'non-terminal decisions' (e.g. these findings are consistent with the treatment effectiveness, which justifies further research), which still require a broader clinical assessment (e.g. the absence of severe side effects, sustainable invasiveness, etc.).

CONCLUSIONS

This article discusses the epistemological, scientific, and statistical reasons supporting the descriptive approach over theoretical-decisional frameworks in public health. In particular, the strong dependence of the latter on assumptions that are too often violated, such as the absence of sources of uncertainty – including variability and bias – makes the latter not recommended in the medical field (e.g. replication crisis). In this regard, the proposed protocol for graded assessment of statistical compatibility aims to mitigate overstatement and bias as well as to avoid the dichotomization of scientific results into 'significant' and 'non-significant' based on a mere numerical criterion. The adoption of multiple compatibility or surprisal intervals can serve as a compromise between completeness and conciseness. It is recommended to adopt this or similar descriptive methods for scientific investigations in the soft sciences.

REFERENCES

1. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature. 2019;567(7748):305-307. doi:10.1038/d41586-019-00857-9
2. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016;31(4):337-350. doi:10.1007/s10654-016-0149-3
3. Greenland S. Divergence versus decision P-values: a distinction worth making in theory and keeping in practice: or, how divergence P-values measure evidence even when decision P-values do not. Scand J Statist. 2022;50(1):54–88. doi:10.1111/sjos.12625
4. Gelman A, Hennig C. Beyond subjective and objective in statistics. Journal of the Royal Statistical Society: Series A (Statistics in Society). 201;180(4):967–1033. doi:10.1111/rssa.12276
5. Greenland S.. Connecting simple and precise P-values to complex and ambiguous realities (includes rejoinder to comments on "Divergence vs. decision P-values"). Scand J Statist. 2023;50(3):899–914. doi:10.1111/sjos.12645
6. Mansournia MA, Collins GS, Nielsen RO, et al. A CHecklist for statistical Assessment of Medical Papers (the CHAMP statement): explanation and elaboration. Br J Sports Med. 2021;55(18):1009-1017. doi:10.1136/bjsports-2020-103652
7. Rovetta A. S-values and surprisal intervals to replace p-values and confidence intervals. REVSTAT-Statistical Journal. Accessed May 27, 2024. https://revstat.ine.pt/index.php/REVSTAT/article/view/669
8. Rovetta A. Multiple Confidence intervals and surprisal intervals to avoid significance fallacy. Cureus. 2024;16(1):e51964. doi:10.7759/cureus.51964
9. Lehmann EL. Fisher, Neyman, and the creation of classical

statistics. Springer New York; 2011.
10. Fisher RA. Statistical methods for research workers (11th ed. rev.). Oliver and Boyd: 1925. Accessed May 27, 2024. https://psycnet.apa.org/record/1925-15003-000
11. Amrhein V, Korner-Nievergelt F, Roth T. The earth is flat (p > 0.05): significance thresholds and the crisis of unreplicable research. PeerJ. 2017;5:e3544. doi:10.7717/peerj.3544
12. Neyman J, Pearson ES. On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character. 1933:231:289–337. doi:10.1098/rsta.1933.0009
13. Neyman J. Frequentist probability and frequentist statistics. Synthese 1977:36:97–131. doi:10.1007/BF00485695
14. Fisher R. Statistical methods and scientific induction. Journal of the Royal Statistical Society: Series B (Methodological). 1955;17:69-78. doi:10.1111/j.2517-6161.1955.tb00180.x
15. Amrhein V, Trafimow D, Greenland S. Inferential statistics as descriptive statistics: there is no replication crisis if we don't expect replication. The American Statistician. 2019;73(sup1):262–270. doi:10.1080/00031305.2018.1543137
16. Rubin M. "Repeated sampling from the same population?" A critique of Neyman and Pearson's responses to Fisher. European Journal for Philosophy of Science. 2020;10(3):42. doi:10.1007/s13194-020-00309-6
17. Ting C, Greenland S. Forcing a Deterministic Frame on Probabilistic Phenomena: A Communication Blind Spot in Media Coverage of the "Replication Crisis". Science Communication. doi:10.1177/10755470241239947
18. Wasserstein RL, Lazar NA. The ASA statement on p-Values: context, process, and purpose. The American Statistician. 2016;70(2):129–133. doi:10.1080/00031305.2016.1154108
19. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "p < 0.05." The American Statistician. 2019;73(sup1):1–19. doi:10.1080/00031305.2019.1583913.
20. Amrhein V, Greenland S. Discuss practical importance of results based on interval estimates and p-value functions, not only on point estimates and null p-values. Journal of Information Technology. 2022;37(3):316-320. doi:10.1177/02683962221105904
21. Amrhein V, Greenland S. Rewriting results in the language of compatibility. Trends Ecol Evol. 2022;37(7):567-568. doi:10.1016/j.tree.2022.02.001
22. Mansournia MA, Nazemipour M, Etminan M. P-value, compatibility, and S-value. Glob Epidemiol. 2022;4:100085. doi:10.1016/j.gloepi.2022.100085
23. Rovetta, A, Mansournia MA, Vitale, A. ICAMSS Statement on Statistical Testing - v.01. Figshare. doi:10.6084/m9.figshare.25850044.v1
24. McShane BB, Bradlow ET, Lynch JG, Meyer R.J. "Statistical Significance" and Statistical Reporting: Moving Beyond Binary. Journal of Marketing. 2024;88(3):1-19. doi:10.1177/00222429231216910
25. Pearson K. 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 Philosophical Magazine and Journal of Science. 1900;50(302):157–175. doi:10.1080/14786440009463897
26. Rafi Z, Greenland S. Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise. BMC Med Res Methodol. 2020;20(1):244. doi:10.1186/s12874-020-01105-9
27. Greenland S, Mansournia MA, Joffe M. To curb research misreporting, replace significance and confidence by compatibility: a Preventive Medicine Golden Jubilee article. Prev Med. 2022;164:107127. doi:10.1016/j.ypmed.2022.107127
28. Gelman A. The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. Personality & social psychology bulletin. 2018;44(1):16–23. doi:10.1177/0146167217729162
29. Gelman A, Greenland S. Are confidence intervals better termed 'uncertainty intervals'?. BMJ. 2019:l5381. doi:10.1136/bmj.l5381
30. Harvard Chan School Department of Epidemiology. There's Not Much Science in Science Addressing the Psychosocial Gap in Methodology. Accessed May 27, 2024. https://www.youtube.com/watch?v=N7-yn5dd7Hg
31. McShane BB, Gal D. Blinding us to the obvious? The effect of statistical training on the evaluation of Evidence. Management Science, 2016;62(6):1707–1718. doi:10.1287/mnsc.2015.2212
32. Good IJ. Rational decisions. Journal of the Royal Statistical Society. Series B (Methodological). 1952;14(1):107-114.
33. Kühberger A, Fritz A, Lermer E, Scherndl T. The significance fallacy in inferential statistics. BMC research notes, 2015;8:84. doi:10.1186/s13104-015-1020-4
34. Cox DR, Spjøtvoll E, Johansen S, et al. The role of significance tests [with discussion and reply]. Scandinavian Journal of Statistics. 1977;4(2):49–70.
35. Muff S, Nilsen EB, O'Hara RB, Nater CR. Rewriting results sections in the language of evidence. Trends in Ecology & Evolution. 2022;37(3):203–210. doi:10.1016/j.tree.2021.10.009
36. Rovetta A, Mansournia MA. P > 0.05 is good: the NORD-h protocol for several hypotheses analysis based on known risks, costs, and benefits. Center for Open Science; 2024.

CONFLICTS OF INTEREST
The author has completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none was reported.

FUNDING
There was no source of funding for this research.

ETHICAL APPROVAL AND INFORMED CONSENT
Ethical approval and informed consent were not required for this study.

DATA AVAILABILITY
Data sharing is not applicable to this article as no new data were created.

PROVENANCE AND PEER REVIEW
Not commissioned; externally peer reviewed