AARHUS UNIVERSITY

Coversheet

This is the accepted manuscript (post-print version) of the article.
Contentwise, the post-print version is identical to the final published version, but there may be differences in typography and layout.

How to cite this publication
Please cite the final published version:

Schneider, J. W. (2018). NHST is still logically flawed. *Scientometrics*, 115(1), 627-635.
https://doi.org/10.1007/s11192-018-2655-4

Publication metadata

Title:	NHST is still logically flawed
Author(s):	Jesper Wiborg Schneider
Journal:	*Scientometrics*, 115(1), 627-635
DOI/Link:	https://doi.org/10.1007/s11192-018-2655-4
Document version:	Accepted manuscript (post-print)

General Rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

- *Users may download and print one copy of any publication from the public portal for the purpose of private study or research.*
- *You may not further distribute the material or use it for any profit-making activity or commercial gain*
- *You may freely distribute the URL identifying the publication in the public portal*

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

This coversheet template is made available by AU Library
Version 1.0, October 2016

NHST is still logically flawed

Jesper W. Schneider

Danish Centre for Studies in Research and Research Policy,

Department of Political Science,

Aarhus University, Denmark

Jesper W. Schneider

Department of Political Science, Bartholins Alle 7,

Aarhus University, 8000-DK, Denmark

jws@ps.au.dk

Abstract

In this elaborate response to Wu (xxxx), I maintain that null hypothesis significance testing (NHST) is logically flawed. Wu (xxxx) disagrees with this claim presented in Schneider (2015). In this response, I examine the claim in more depth and demonstrate that since NHST is based on one conditional probability alone and framed in a probabilistic *modus tollens* framework of reasoning, it is by definition logically invalid. I also argue that disregarding this logically fallacy, as most researchers do, and treating the p-value as a heuristic value for dichotomous decisions against the null hypothesis, is a risky business that often leads to false-positive claims.

NHST is still logically flawed

I thank the Editor of *Scientometrics* for giving me the opportunity to respond to Wu's (xxxx) commentary to my article "*Null hypothesis significance tests. A mix-up of two different theories: the basis for widespread confusion and numerous misinterpretations*" (Schneider, 2015), and my discussion of it in Wuhan, China in October 2017 during the 16th ISSI conference in connection with the article winning the inaugural ISSI Paper of the Year Award[1].

In the article, I outline the historical roots of what is known today as "null hypothesis significance testing" (NHST). While NHST is the dominating statistical inference practice in the life, medical, behavioral and social sciences, it is a strange patchwork of two fundamentally different and discordant inference models rooted in frequentist probability theories. The article outlines these conflicting issues and shows how they influence the endless number of misunderstandings and misuses of NHST found in the published scientific

[1] http://issi-society.org/blog/posts/2017/october/issi-paper-of-the-year-award-2017/

literature. However, the article also discusses some intrinsic logical problems of NHST as an inference model, and as far as I understand Wu's commentary, he disagrees with the claim that NHST is logically flawed. He argues that if there is a logical flaw in NHST it should be banned, but his commentary sets out to demonstrate that there indeed is no such logical flaw, despite evidence to the contrary put forward for many decades (e.g., Berkson, 1942; Lindley, 1957; Hacking, 1965; Meehl, 1967; Edwards, 1972; Pollard & Richardson, 1987; Cohen, 1994; Falk & Greenbaum, 1995; Royall, 1997; Nickerson, 2000; Hofmann, 2002; Sober, 2008; Schneider, 2015; Szucs & Ioannidis, 2017).

Wu claims that in Schneider (2015), I present a "Bayesian view of NHST" and therefore he will stick to the same approach to demonstrate that NHST is not logically flawed and thus sets out to do so using Bayes' theorem. What follows in Wu's commentary is to me somewhat confusing and presumably based upon some misunderstandings? To me at least, it seems that Wu is "kicking in an open door" under a false premise. In the following, I will elaborate on my impressions, but first I will outline what I presume is Wu's main points.

It seems that the main line of reasoning is presented in section two of his commentary (Wu, xxxx, p. 4), titled "the full conditions of NHST", where Wu claims that:

> "… it is the relative relation between *P* (Data |H_0) and *P* (Data | H_1) that really makes a difference and one should consider both but not one of the two values".

And Wu continues:

> "In this sense, we fully agree with part of the argument in Schneider (2015) that from *P* (Data | H0) ≈ 0.0 one cannot conclude *P* (H1 | Data) ≈ 1.0. However, we do not agree the conclusion drown from this observation that NHST has intrinsic logical problem, it only requires to examine the full conditions in inequality (5) or sometimes the sufficient but not necessary condition in Eq. (6), instead of inequality (3). Furthermore, we also understand why inequality (3) often works although neither sufficient nor necessary in principle".

Clearly, Wu is acknowledging the well-known inverse probability problem inherent in NHST, and like others (e.g., Pollard & Richardson, 1987; Trafimow, 2003), Wu essentially promotes a Bayesian solution. I fully support such line of reasoning, but I seriously doubt that frequentists will do the same. The paradox here is that NHST is firmly based in frequentist orthodoxy, where there is no room for prior probabilities such a p(H) which is required in Bayes' theorem in order to calculate posteriors such as p(H|D). No matter how useful Bayes' theorem might be, and Fisher certainly acknowledged this, the "full condition" in frequentist NHST only contains one conditional probability based on relative frequencies and that is p(D|H). The inference logic is therefore dependent on this conditional probability alone.

It is unclear to me whether Wu acknowledges this premise for NHST, yet he subsequently argues, again using Bayes' theorem, that under certain conditions we can expect very low probabilities of p(D|H) to indicate complementary high probabilities for p(H|D), the latter representing the (research) hypothesis which we are really interested in. So, essentially Wu is arguing that p(D|H) is associated with p(H|D). If my understanding of Wu is correct, then there is also some truth in this claim, but not in the sense that it can fix the logical invalidity of the NHST inference model. It seems to me therefore that Wu is arguing from a false premise in as much as p(D|H) is all we got in NHST, an alternative based on Bayes' theorem is no longer frequentist or NHST.

Both of the issues that Wu addresses and their potential solutions are well-known, in fact I address both of them briefly in my article. I discuss the inverse probability fallacy and suggest the use of Bayesian inference as an alternative to NHST, and I briefly discuss the link between p-values and the posterior probability of H_0. However, contrary to what Wu seems to argue, none of this removes the inherent logical fallacy from the NHST inference model, as I will outline below.

The NHST inference model: Fisher’s disjunction, probabilistic modus tollens and the inverse probability problem

First let me make one thing clear, I do not present a “Bayesian view of NHST” in my article as Wu claims. I do use some simple notation to demonstrate the differences between two kinds of conditional probabilities, i.e. $p(D|H) \neq p(H|D)$, which resembles traditional notation used in Bayes’ theorem. D stands for Data and H for Hypothesis, and the p-value is the probability of collecting data *at least as* extreme as the ones observed, given that the null hypothesis and all other assumptions are satisfied, simply written as $p(D+|H_0)$, where + indicates more extreme unobserved events in the probability space included in the calculation of p-values[2]. As mentioned above, what really should interest researchers are the evidential value for their hypothesis given the data they have observed, or simply $p(H|D)$. This also seems to be Wu’s argument. In conditional probabilities order matters, hence $p(D|H) \neq p(H|D)$. This fact is not very well understood by users of NHST and many misinterpretations and false claims follow from conflating the former with the latter. This is the inverse probability fallacy also known as the “fallacy of the transposed conditional” (Gigerenzer, 1993) and this is the focal point in relation to the logical fallacy in the NHST inference model. We are interested in $p(H|D)$ but NHST only gives us $p(D|H)$, and to paraphrase Cohen (1994), we so much want to know $p(H|D)$ that, out of desperation, we come to believe that $p(D|H)$ gives us this knowledge. Unfortunately, this is not so and Wu obviously recognizes this problem and tries to solve it in his commentary. But we cannot “solve” the inverse probability problem from within NHST when the epistemological basis denies us the use of Bayes’ theorem.

[2] The latter in itself can be seen as a logical flaw, as a valid measure of strength of evidence should not include the probabilities of unobserved outcomes (Jeffreys, 1939; Berger & Delampady, 1987; Berger & Berry, 1988; Royall, 1997; Goodman, 1999), but this is not the main logical flaw of interest here.

We need to recall the NHST inference model and its *modus operandi* for making inferences based on deductive reasoning, in order to demonstrate why it is logically flawed and can easily lead to wrong conclusions, as the current so-called reproducibility crisis testifies to (Ioannidis, 2005; Ioannidis, Stanley, & Doucouliagos, 2017).

Notice, it is important to realize that NHST is a ritualistic inference practice based on an inaccurate mix-up of ideas from Fisher and Neyman-Pearson basically following these steps (Gigerenzer, 2004):

1. A researcher has a theory or hypothesis to be (hopefully) supported known as the “alternative” hypothesis” (H_A).
2. A statistical null hypothesis of some parameter $\theta = 0$ is proposed (H_0), (hopefully) to be rejected. Notice, contrary to Neyman-Pearson’s idea about testing two complementary rival hypotheses (H_1 and H_1) using the most powerful test, in NHST, only the proposed “straw man” point null hypothesis, $H_{0,}$ is defined and tested in line with Fisher’s “nullifying” ideas.
3. Data are collected and the probability of obtaining more extreme data given that H_0 is true is computed using a test statistic, this gives us the p-value, or $p(D+|H_0)$.
4. $p(D+|H_0)$ is compared to an arbitrarily chosen significance level α, which by convention is often set to 5% in the soft science.
5. If $p(D+|H_0) \leq .05$, then H_0 is rejected as the null hypothesis is assumed to be unlikely, and most researchers implicitly infer that its complement, the “alternative” hypothesis, H_A, is therefore supported (this is not in line with Fisher). Notice, H_A is never tested even though it represents the research hypothesis or theory for which one seeks supports, it is merely the complement of a point null hypothesis.

Two things are evident from this ritual: 1) only one conditional probability is applied, $p(D+|H_0)$, and 2) the significance levels (α) needed for rejection decisions are defined by convention and have no scientific basis whatsoever.

The *modus operandi* for making inferences in the NHST ritual outlined above is the deductive syllogistic reasoning known as *modus tollens*. As I demonstrate in Schneider (2015, p. 422), the syllogistic argument if *P* then *Q* (premise 1), not-*Q* (premise 2), then not-*P* (conclusion), leads to formally valid conclusions when statements are absolute. *P* is proved by contradicting *Q*, i.e. the falsehood of *P* follows from the fact that *Q* is false.

Seemingly, the inference model in NHST follows the same reasoning. Its rationale comes from *Fisher's disjunction*: "[e]ither an exceptionally rare chance has occurred, or the theory [= the null hypothesis] is not true" (Fisher, 1956, pp., 39). In other words, Fisher argues that a *very unlikely* result under H_0, $p(D+|H_0) < \alpha$, undermines the objective tenability of the null hypothesis. It is of course a reasonable question to ask what would constitute a "very unlikely result"? As it turns out, no objective probability cutoff for rejection can be established (see Sober, 2008, p. 53). P-values are continuous measures and are perceived to be evidentiary measures against the $H_{0.}$ This leads to fallacies such as $p(D|H) = p(H|D)$. But most importantly, despite being used for dichotomous discrete decisions based on arbitrary thresholds, p-values are probabilistic and not absolute statements; hence NHST becomes a probabilistic version of *modus tollens* that can be outlined as follows:

Premise 1: If *P* (H_0 is true), then *Q* is highly likely (i.e. the test statistic will most likely fall in the nonrejection region, $p(D+|H_0) > \alpha$)

Premise 2: Not-*Q* (i.e. the test statistic is in the rejection region, $p(D+|H_0) \leq \alpha$)

Conclusion: Hence *P* is highly unlikely (H_0 is highly unlikely)

An argument can be formally valid only if its conclusion is true whenever its premises are true. By making an argument probabilistic, it becomes possible that its conclusion is false although all of its premises are true. Therefore, a syllogism that is probabilistic by definition does not meet criteria for formal validity because the conclusion is not necessary true whenever its premises are true. It is therefore a fallacy to assert that obtaining data that are *unlikely* under H_0 implies that H_0 is *likely* false. Almost a contradiction of H_0 does not imply that H_0 is almost false. No matter what Wu claims, this fundamental fact means that NHST rests on a flawed logical basis simply because once *modus tollens* is made probabilistic, the formal deductive inference is no longer valid and can easily lead to wrong conclusions. Clearly, sensible conclusions can be reached as well, albeit not as a result of formal deductive reasoning. Formal validity defines deductive reasoning and if it is false a logical fallacy is inherent in the inference model.

The whole point about NHST is that it is practiced as a ritual where presumed logical dichotomous decisions follow from the *modus operandi*. Hardly ever in the published literature universe do we see reasoning that strictly separates $p(D+|H_0)$ from $p(H_0|D)$, on the contrary, acceptance of (H_A) almost always follow implicitly from $p(D+|H_0) \leq \alpha$, and this is not logically valid. As Wu shows in his commentary, Bayes' theorem can demonstrate this. Judgments about evidential meaning are essentially contrastive. To decide whether an observation is evidence against H, we need to know what the alternative hypotheses are; to test a hypothesis requires testing it against alternatives. Probabilistic *modus tollens* needs to be replaced by the law of likelihood which is part of Bayes' theorem.

Inverse probability and Bayes' theorem

Bayes' theorem formalizes the matter of inverse probability and demonstrates why the syllogistic reasoning in NHST does not hold. Consider the likelihood version of Bayes' theorem:

$$p(H|D) = \frac{p(H) \times p(D|H)}{p(H) \times p(D|H) + p(\sim H) \times p(D|\sim H)}. \qquad (1)$$

Or specifically in relation to the null hypothesis case:

$$p(H_0|D) = \frac{p(H_0) \times p(D|H_0)}{[p(H_0) \times p(D|H_0) + p(\sim H_0) \times p(D|\sim H_0)]}. \qquad (2)$$

By definition $p(\sim H_0)$ is equal to the probability of the alternative hypothesis $p(H_A)$, which is equal to $1 - p(H_0)$, so the following equation is also true:

$$p(H_0|D) = \frac{p(H_0) \times p(D|H_0)}{\left[p(H_0) \times p(D|H_0) + \left(1 - p(H_0)\right) \times p(D|\sim H_0)\right]}. \qquad (3)$$

According to the last equation, we need to know three probabilities to calculate the desired posterior probability $p(H_0|D)$. These are $p(D|H_0)$, $p(H_0)$ and $p(D|\sim H_0)$. The first we recognize as the outcome of NHST. The latter two probabilities, on the other hand, are not part of the frequentist inventory. They are the prior probability of the null hypothesis and the probability of the data if the null hypothesis is not true. Following equation 3 above, when $p(D|H_0)$ is zero then so will $p(H_0|D)$ essentially reproducing the result from a syllogistic reasoning with absolute statements. However, when D is merely *unlikely*, the posterior probability of H_0 crucially depends on the prior probability of the null hypothesis and the probability of the data if the null hypothesis is not true. So, knowing that the data are *unlikely* under H_0 is of little use unless one determines whether or not they are also *unlikely* under H_A (Sellke, Bayarri, & Berger, 2001).

In Schneider (2015) I used the famous Congress-example (see Pollard & Richardson, 1987; Cohen, 1994), to demonstrate the invalid logic in the probabilistic version of *modus tollens* (below is a slightly revised version inspired by Szucs and Ioannidis (2017)):

> **Premise 1:** If Ronald is American, then he is very unlikely to be a member of Congress
>
> **Premise 2:** Ronald is a member of Congress
>
> **Conclusion:** Then Ronald is most likely not American

First, consider this from the NHST perspective. The null hypothesis is that Ronald is American. The complement "alternative" hypothesis is that Ronald is not American. The null model in premise 1 is therefore: If Ronald is American (H_0), then he is very unlikely to be a member of Congress. The observed data (D) are: Ronald is a member of Congress, and since $p(D|H_0)$ is very low, i.e. we have observed a rare event under the null hypothesis, our inference logic leads to a rejection of H_0, i.e. Ronald is American, and implicitly support H_A, that Ronald is most likely not American. The conclusion is obviously mistaken and we are being misled by the flawed logic of NHST.

Now consider the Bayesian perspective. Here we need prior probabilities of H_0 and H_A (i.e. $(1 - p(H_0))$). Now suppose we do not know Ronald's nationality, we therefore assign an uninformative prior of 0.5 to H_0 and by definition also 0.5 to H_A. Now for this illustrative example we can also assign arbitrary but plausible probabilities to the conditional probabilities of the data under the two hypotheses. Following Szucs and Ioannidis (2017), the probability that Ronald is member of Congress given he is American is very low, $p(D|H_0) = 10^{-7}$, very few Americans are member of Congress; and the probability that Ronald is member of Congress given he is not American is zero, $p(D|\sim H_0) = 0$, you need an American citizenship to be a member of Congress. Its clear from these reasonable probabilities, that while data is certainly rare under the null hypothesis (corresponding to a very low p-value), they are actual impossible under the alternative hypothesis. So, the proper conclusion between these two hypotheses is to accept H_0 instead of rejecting it as inferred from the NHST syllogistic logic. Using Bayes' theorem in equation 3, we can explicitly examine $p(H_0|D)$:

$$p(H_0|D) = \frac{0.5 \times 10^{-7}}{[0.5 \times 10^{-7} + 0.5 \times 0]} = \frac{0.00000005}{0.00000005} = 1$$

Hence, with the Bayesian framework we can show the logically fallacy inherent in NHST and calculate probabilities of the hypotheses given the data and compare them relative to each other, which obviously leads to the intuitive conclusion in this example that the null hypothesis, i.e. Ronald is American, must be "true".

The example is clearly extreme, however, more generally, if $p(D+|H_0) = 5\%$, it is not difficult to choose values for $p(H_0)$ or $p(D|\sim H_0)$ that will result in $p(H_0|D)$ being larger than the 5% cutoff level dominating rejection decisions in the soft sciences (Trafimow, 2003).

Bayes' theorem clarifies at least two things in relation to the inherent logically fallacy in NHST. First, the well-known inverse probability fallacy of equating $p(D+|H_0)$ with $p(H_0|D)$. Second, $p(D|H_0)$ is one of three factors needed to derive $p(H_0|D)$, we cannot therefore totally ignore this conditional probability, as Wu also states. Accordingly, to some extent we can expect that $p(D+|H_0)$, i.e. the p-value, is associated with the posterior probability of the null hypothesis. So, we seem to have a paradox, on the one hand, we cannot draw logically valid conclusions from $p(D+|H_0)$ about $p(H_0|D)$, but, on the other hand, we know that the two conditional probabilities are associated. Knowing that data needed to obtain $p(H_0|D)$ is either practically or epistemically unobtainable for many researchers, could we then simply settle for $p(D+|H_0)$ for practical reasons (Krueger, 2001)? Clearly, most practitioners of NHST do that, and as far as I can see Wu is also suggesting that this is a solution, at least under some limiting conditions, but as I will outline below, such associations are too vague to be trusted as a heuristic value in inductive inference.

NHST as a suitable heuristic despite being logically flawed?

While, some supporters of NHST recognize that the ritual is logically invalid, they often downplay the fallacy by arguing that the pragmatic value of NHST far outweighs these rather dubious philosophical concerns (Krueger, 2001; Hofmann, 2002; Krueger & Heck, 2017). Their main common-sense argument is that $p(D+|H_0)$ tends to correlate with $p(H_0|D)$. They do indeed correlate, but as several authors have shown theoretically and through simulations, even in optimal settings (i.e. no error or noise), such associations are vague and too often leads to "over-confident" statements about the evidence against H_0 (e.g., Berger & Sellke, 1987; Berger & Berry, 1988; Goodman, 1999; Ioannidis, 2005; Hubbard & Lindsay, 2008; Trafimow & Rice, 2009; Colquhoun, 2014; Krueger & Heck, 2017). Obviously, factors such as prior probabilities and statistical power influences the associations, but what these studies generally find is that p-values overstate the evidence against H_0, especially in the interval between significance levels of 1% and 5%. Much lower p-values such as 0.1% are needed for p-values to converge with proper evidential measures against H_0. Even that is not a guarantee. The traditional cut off level at 5% have been shown to give posterior probabilities of at least 30% in support of $H_{0.}$ Analogous, Trafimow and Rice (2009) demonstrate unimpressive correlations for p-values at 5% to predict p(H_0|D) with r^2 of only 16%, and furthermore, as the significance rule becomes more stringent (e.g., 1%, 0.1%), the correlation actually decreases. The univocal conclusion from these studies is that the p-value cannot function alone as a heuristic value for inductive inference, which is exactly what it does when applied in NHST for dichotomous decisions concerning the null hypothesis. Under favourable circumstances (e.g. sufficient power and well-controlled studies), NHST may be able to minimize Type I errors, but many false conclusions will still follow when we try to address the "truthfulness" of a hypothesis in a single study as the current replicability crisis testify to (Ioannidis, 2005; Colquhoun, 2017).

Summary

I am not so sure to what extent Wu (xxxx) and I disagree. To me it seems that Wu's commentary is based upon a misunderstanding or perhaps a false premise. NHST is a ritual that is based on frequentist principles. As a consequence, we only have one conditional probability to rely upon, p(D+|H_0). Placed in a probabilistic *modus tollens* framework, NHST by definition becomes logically invalid. The question is to what extent we can ignore this logical fallacy and still rely on p(D+|H_0) as a heuristic value for dichotomous decisions against H_0. As I have argued above, this is certainly a risky business. The fact that a rare finding, given the null hypothesis, has been obtained does not justify the conclusion that the null hypothesis is likely to be false, and certainly not through some kind of valid logic.

The only known solution to the inverse probability problem is to use Bayes' theorem as Wu, and others before him, demonstrates. But the conundrum is that this involves commitments that many frequentist inclined researchers are not willing or able to make. Hence, while a Bayesian approach is certainly *one* solution, it does not solve the logical fallacy inherent in NHST. This will remain as NHST is dependent on one conditional probability alone and framed as probabilistic *modus tollens* reasoning ritual. In that sense, it seems to me that Wu is “kicking in an open door” under a false premise.

References

Berger, J. O., & Berry, D. A. (1988). The relevance of stopping rules in statistical inference. *Statistical decision theory and related topics IV, 1*, 29-47.

Berger, J. O., & Berry, D. A. (1988). Statistical Analysis and the Illusion of Objectivity. *American Scientist, 76*(2), 159-165.

Berger, J. O., & Delampady, M. (1987). Testing Precise Hypotheses. *Statistcial Science, 2*(3), 317-352.

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis - the irreconcilability of p-values and evidence. *Journal of the American Statistical Association, 82*(397), 112-122.

Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association, 37*(219), 325-335.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49*(12), 997-1003.

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science, 1*(3).

Colquhoun, D. (2017). The Reproducibility Of Research And The Misinterpretation Of P Values. *bioRxiv*. doi:10.1101/144337

Edwards, A. W. F. (1972). *Likelihood*. Cambridge: Cambridge University Press.

Falk, R., & Greenbaum, C. (1995). Significance tests die hard: the amazing persistence of a probabilistic misconception. *Theory Psychology, 5*, 75 - 98.

Fisher, R. A. (1956). *Statistical Methods and Scientific Inference*. New York, NY: Hafner.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: methodological issues (pp. 311–339).*

Hillsdale: Erlbaum. Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*(5), 587-606.

Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: The P value fallacy. *Annals of Internal Medicine, 130*(12), 995-1004.

Hacking, I. (1965). *Logic of Statistical Inference*. Cambridge: Cambridge University Press.

Hofmann, S. G. (2002). Fisher’s fallacy and NHST's flawed logic. *American Psychologist, 57*(1), 69-70.

Hubbard, R., & Lindsay, R. M. (2008). Why P values are not a useful measure of evidence in statistical significance testing. *Theory and Psychology, 18*(1), 69-88.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), 696-701.

Ioannidis, J. P. A., Stanley, T. D., & Doucouliagos, H. (2017). The Power of Bias in Economics Research. *The Economic Journal, 127*(605), F236-F265.

Jeffreys, H. (1939). *Theory of Probability*. Oxford, UK: Clarendon Press.

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist, 56*(1), 16-26.

Krueger, J., & Heck, P. R. (2017). The Heuristic Value of p in Inductive Statistical Inference. *Frontiers in Psychology, 8*(908).

Lindley, D. V. (1957). A Statistical Paradox. *Biometrika, 44*(1-2), 187-192.

Meehl, P. E. (1967). Theory-Testing in Psychology and Physics: A Methodological Paradox. *Philosophy of Science, 34*(2), 103-115.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*(2), 241-301.

Pollard, P., & Richardson, J. T. (1987). On the probability of making Type I errors. *Psychological Bulletin, 102*(1), 159-163.

Royall, R. (1997). *Statistical Evidence: A Likelihood Paradigm.* London: Chapman & Hall.

Schneider, J. W. (2015). Null hypothesis significance tests. A mix-up of two different theories: the basis for widespread confusion and numerous misinterpretations. *Scientometrics, 102*(1), 411-432.

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of rho values for testing precise null hypotheses. *The American Statistician, 55*, 62 - 71.

Sober, E. (2008). *Evidence and Evoluation. The Logic Behind Science*. Cambridge, UK: Cambridge University Press.

Szucs, D., & Ioannidis, J. P. A. (2017). When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment. *Frontiers in Human Neuroscience, 11*(390).

Trafimow, D. (2003). Hypothesis testing and theory evaluation at the boundaries: surprising insights from Bayes's theorem. *Psychological Review, 110*(3), 526.

Trafimow, D., & Rice, S. (2009). A Test of the Null Hypothesis Significance Testing Procedure Correlation Argument. *The Journal of General Psychology, 136*(3), 261-270.

Wu, J. (xxxx). Is there an intrinsic logical error in null hypothesis significance tests? Commentary on: “Null hypothesis significance tests. A mix-up of two different theories: the basis for widespread confusion and numerous misinterpretations”. *Scientometrics.*