\*\*masterP\*\*

*Possible title 1:* Why Non-Significant Results Matter. A Critique of Statistical Heaven

*Possible title 2:* Significant meta-level results with non-significant individual results

*Possible title 3:*  I Find Your Lack of Power Disturbing: A Critique of Statistical Heaven

*Possible title 4:* Too Good to be False: A Critique of Statistical Heaven

Chris H.J. Hartgerink1

Marcel A.L.M. Van Assen1

Jelte M. Wicherts1

1 Tilburg University, the Netherlands

WORD COUNT: XXXX words

Author note

Abstract

\*placeholder\*

*Keywords:* p-value, significance, underpowered, effect size, fisher

**Too Good to be False: A Critique of Statistical Heaven**

Statistical hypothesis testing is the crux of contemporary psychological science, and the validity of a study is regularly assessed by the results of individual hypothesis tests. There is much value in hypothesis testing, but the application of individual hypothesis testing is subject to unconscious human fallacies (e.g., law of small numbers, confirmation bias, hindsight bias) and overlooks the probabilistic nature of statistical results (Murdoch, Tsai, & Adcock, 2008). The unconscious human fallacies and the probabilistic nature of statistical results are at interplay, resulting in a fallacy of its own accord. In the current paper, we explain the origins of this fallacy, which we call the fallacy of *statistical heaven*. Subsequently, we investigate how this fallacy of statistical heaven has affected research in psychological science, and propose a method to decrease itsThese unconscious fallacies make the interpretation of results arduous and can lead to seemingly paradoxical conclusions. Additionally, hypothesis testing has intricacies that are often misconceived (see Schmidt & Hunter, 1997), which adds . In the current paper, we propose a method that

One of the major problems is the dichotomization of results: either there is an effect, or there is not. In light of Popper’s () hypothetico-deductivism, such a dichotomy seemingly makes sense, as Popper’s demarcation criterion for science is falsification. Hypothesis testing embodies this demarcation criterion, and is oft seen as a way of attaining objectivity of results (). This could be true, were it not for the probabilistic nature of statistics.

This probabilistic nature makes the reasoning underpinning falsification inherently problematic, as Cohen () aptly illustrated. Take for example the following syllogism as taken from Cohen ():

“If a person is an American, then he is not a member of Congress. (WRONG!)

This person is a member of Congress.

Therefore, he is not an American.” (p. 23, Cohen, 1997)

This example shows correct logic, but an incorrect premise — the conclusion, given the premise, is correct however. If we subsequently correct the premise, and make it probabilistic, we attain the following:

“If a person is an American, then he is probably not a member of Congress (TRUE, RIGHT?)

This person is a member of Congress.

Therefore, he is probably not an American.” (p. 24, Cohen, 1997)

The premise is now correct, but the resulting conclusion does not make logical sense. In fact, this conclusion is formally the same as saying, if the data are unlikely under the null-hypothesis, the null is probably not true.

These explorations of the logic of hypothesis testing indicate that falsification can serve as a demarcation criterion for science, but that the probabilistic implementation of it as done by hypothesis testing does not hold. This provides a serious problem for the idea of *hypothesis testing as objective criterion.* But why does hypothesis testing prevail despite of these strong, logical counterarguments?

We propose the notion of *statistical heaven* to be the main driver behind this prevailing of hypothesis testing, and more importantly, the firm persistent belief in the value of hypothesis testing. Statistical heaven is the place where research makes no errors, and has a 100% chance of finding an effect if there actually is one (i.e., α = 0; 1-β = 1). This notion easily arises, as research has repeatedly indicated that humans lack randomness of mind, i.e., they have bad intuition for probabilities. Take for example the persistent belief in the law of small numbers (), the Monte Hall problem (), the conjunction fallacy () or simply the generation of random data (). Power and Type I error rates (or other results for that matter) remain probabilities, and therefore are subject to that same bad intuition. Since researchers fail to take into account these random elements to hypothesis testing, they resort to thinking their results indicate certainties: there is an effect, or there is none.

Seeing how this ignorance of the probabilistic elements of hypothesis testing reduces findings from probabilistic to deterministic, it is easy to see why hypothesis testing prevails. The probabilistic syllogism, as illustrated by Cohen (), was clearly false. However, if we remove the probabilistic element, the logic reduces to that which makes full sense. In other words: probabilistic hypothesis testing makes no sense, as it contains flawed logic, but deterministic hypothesis testing makes very much sense. Too bad hypothesis testing is never deterministic and will remain probabilistic, no matter how much we will it otherwise.

* Short history on the debate in psychological science
  + Differentiate between theoretical context (content that is being translated to hypotheses) and statistical context (the results of the hypotheses being tested)
  + Significance testing and philosophy of hypothetico-deductivism and objectivism
    - NHST as embodiment of the falsification criterion
    - Effect/no effect as pseudo-objectivism
* Why a more optimistic view can be taken of non-significant studies
  + Statistical heaven
    - How it arises
      * Confirmation bias (Rosenthal, 1964)
      * Belief in the law of small numbers (Tversky & Kahnemann)
    - Researchers are human and lack randomness of mind — statistics uses probability and if humans cannot see such probabilities properly (Wason task; Monte Hall problem) than they should not be allowed to filter the probabilistic resul
  + Why it is incorrect
* Talk about α and β error control (Fiedler 2012)
  + We focus on β error control and try to show that α error control is non-vital in practical research (very important in theoretical research on methods, i.e., specificity of tests)
    - More specific hypotheses
* Lack of power in psychological science throughout the last decades

Method

Results

Discussion

* Recommendations