Norbert Schwarz publicly criticized the "False Positives" symposium at the most recent SPSP. I write to rectify the factually false statements he made about my research, and to refocus our attention to addressing the serious credibility problems our field, along many other fields, faces.

You can read a .pdf version of this posting here: http://opim.wharton.upenn.edu/~uws/SPSP/post.pdf

WHY AM I RESPONDING ONLY NOW?

On February 1st, after Norbert sent his original posting to this list serve, but before it was actually distributed (he had the wrong list serve address), I emailed him, pointed out his reservations about my research were entirely unfounded, and invited him to chat on the phone. I insisted on having a phone-call on February 2nd, 3rd and 8th. On the 10th we finally spoke (for 40 minutes). We agreed on drafting a joint statement highlighting areas of agreement and disagreement, trying to move things forward in a constructive manner. Three days later I sent Norbert a draft, inviting him to edit it to his satisfaction. I asked again on the 18th and 26th of February. On the 28th I indicated that if I did not hear from him by Sunday March 4th, I would rectify the erroneous impressions he created about my work through a sole-author response. He did not reply to that email. These two pages are that sole-author response.

You can read this entire exchange here: http://opim.wharton.upenn.edu/~uws/SPSP/email.pdf

BACKGROUND

For those absent from my SPSP talk, two useful definitions,
p-curve: the distribution of significant p-values across a set of findings
p-hacking: exploiting –perhaps unconsciously- researcher degrees-of-freedom until $p<.05$.

In my SPSP talk I presented joint work with Leif Nelson and Joe Simmons showing that *p*-curve can be employed to assess if a set of statistically significant findings are likely to be false-positive.
[We have not yet completed the first draft of our paper].

NORBERT'S BASELESS AND ERRONEOUS CRITIQUE

Norbert proposed in his posting, matter-of-factly, that the number of p-values needed for making credible inferences with *p*-curve is large, that *p*-curve may hence be useful for meta-analysis but not for the applications I presented in my talk (e.g., assessing if a finding documented in four studies would replicate, or if a given researcher *p*-hacks intensely). As a reminder of his tone: "So a belief in small numbers is fine when it generates the right outcome? Ignoring the prerequisites of an analysis is fine if it serves a good goal? And a sufficient data base is only a requirement for others?"

In no uncertain terms let me state the following: NORBERT'S CLAIMS ARE WRONG.

P-curve ***can*** provide informative answers with a small number of p-values, and when such number is in fact too small, the confidence intervals around *p*-curve's assessments will reflect it. Notably, I explained this to Norbert in my first email to him, 6 hours ***before*** he resent his message to the list serve.

In our paper we will propose two methods for assessing the statistical significance of p-curve (e.g., to test if it is significantly upward sloping, which occurs when findings are *p*-hacked).
The ***less*** powerful of the two tests is a simple 50:50 binomial. This means that, when *p*-hacking is sufficiently intense, the null of a flat p-curve could be rejected with just five studies (.5 to the fifth leads to $p<.05$). The other test could do so with fewer p-values. For instance, if a paper reports (only) three p-values, all $p=.049$, the null of a flat p-curve is rejected with $p<.001$.

Norbert's stat-gut based power analysis of *p*-curve, which he so fervently broadcasted, and publicly questioned my research integrity with, is both baseless and incorrect.

WE ARE ALL P-HACKERS, THOSE OF US WHO REALIZE IT WANT CHANGE.

As we stated with Leif and Joe in our paper, talks, interviews and private communications, we think most people p-hack, including the three of us. I don't know of anybody who runs a study, conducts one test, and publishes it no matter what the p-value is.

Norbert, in contrast, stood up during the symposium, loudly and rudely interrupted the first speaker, and proclaimed that he, for one, only drops measures for statistically sound reasons. He re-insinuated such claim in writing, to the list-serve.

This self-exculpatory proclamation, however, is not accurate.

I know for a fact that in a study Norbert published, whose original materials I obtained to attempt a replication a long time ago, he dropped several measures with no defensible justification. Exclusions sometimes fall in a grey area. A variable with too little variation, or too distantly related to the current research, for example, may be left off the manuscript for the sake of clarity.

This is not a grey-area exclusion. It is a pants-on-fire one: (i) the included and excluded variables are perfectly interchangeable, (ii) moreover, one of the excluded variables is a more reasonable variable to include than the one that made it into the study that "worked," and (iii) for the little that is worth, some of the measures reported in the paper came AFTER, in the original study, those that were shrouded . I do not provide further details to avoid singling out co-authors who have not claimed to be holier than thou. Norbert can post the original materials if he disagrees with my assessment of the justifiability of the exclusions.

Now, I don't bring this up to wag a moral finger at Norbert. He did not act immorally when he dropped measures. Doing so is regrettably accepted in our field, and often, actually encouraged or even required by associate editors and referees.

I bring it up because if the person putting his reputation on the line to publicly claim he definitely does not *p*-hack, publishes research with such blatant ex-post variable selection, this is not an isolated problem we should assume away. It is one we should keep in mind when we debate about the things that researchers do and do not do to obtain publishable results.

I bring this example up because it so vividly shows how difficult the status-quo makes it for readers to make informed inferences regarding the validity of published research. We must not only trust authors to do the right thing. We must trust that they share our view of what the right thing is (and there is no written code for what that is). Moreover, we must trust that people they trust to run their studies share our research standards and are themselves trustworthy. That's an unacceptably unreliable safeguard against excesses that may challenge the credibility of our science.

In "False-Positive Psychology" we proposed greater disclosure as a solution to this problem. Our solution would have ensured that all readers of Norbert's paper, not just me, knew of the dropped measures. It would have deservedly increased the scrutiny its claims were subjected to.

Without disclosure about data collection and analysis, our p-values are mere lip-service; we pretend to be interested in ruling out chance, but we are not really doing that. Readers cannot tell if findings are inconsistent with sampling error (i.e., significant) if they don't know the kind of sampling behind them.

This is not more a matter of opinion than the law of large numbers is.

We can work together to make psychology a beacon of credibility for other sciences to emulate. Let's start requiring authors to properly disclose the details of the data collection and analysis behind their findings so that readers can make evidence-based decisions regarding the relative credibility of different claims.

As scientists, that's our job description.

Uri Simonsohn