the WINNOWER

SCIENCE AND SOCIETY

Precise, predictive theories without wiggle room: Lessons from Physics

JONA SASSENHAGEN [1]

1. University of Frankfurt, Germany

READ REVIEWS

WRITE A REVIEW

CORRESPONDENCE:
jona.sassenhagen@gmail.com

DATE RECEIVED:
June 22, 2016

KEYWORDS:
#LJAFreproducibility, LIGO, physics

There is a fundamental problem of imprecise theories in many scientific disciplines. Trying to follow the gravitational wave detection recently reported by the LIGO consortium, it appears that in physics, *none* of the problems well-known from e.g. psychology and medicine seem to come up *as problems*. Consider this section from one of the papers reporting on LIGO:

> *Here we perform several studies of GW150914, aimed at detecting deviations from the predictions of GR. Within the limits set by LIGO's sensitivity and b y the nature of GW150914, w e find no statistically significant evidence against the hypothesis that GW150914 was emit- ted by two black holes spiraling towards each other and merging to form a single, rotating black hole [Schwarzschild 1916, Kerr 1963], and that the dynamics of the process as a whole was in accordance with the vacuum Einstein field equations. (LIGO Scientific Collab- oration and Virgo Collaboration (2016): Tests of general relativity with GW150914.)*

Many causes for low reproducibility – and some promising solutions, such as the Registered Report format at *Cortex*, or increased post-publication peer review, or large-scale joint replication projects – are well known. Here, I want to suggest a possibly more fundamental problem and solution: that many scientists have their statistical mainstays the wrong way around. This becomes clear when contrasting it with scientific practice in physics.

From the viewpoint of reproducibility worries, there are a number of things black hole physicists seemingly do horribly wrong. For example, they conduct several analyses on *one single measurement*, and one that cannot be replicated either - the two black holes have collided now, and that's it. Checking the paper, some of the reported statistical tests correspond to an alpha level of 10%. And they celebrate *failing to reject the null hypothesis* ("… we find no statistically significant evidence against the hypothesis …"). Contrast with the perspective of influential scientists from a low-reproducibility theory: "Tests that are not statistically significant should be regarded as indicative of poorly justified, designed, or executed hypothesis-testing studies" (Kraemer 2015, JAMA Psychiatry). LIGO cost on the order of half a billion dollars. Somehow, LIGO scientists

reporting that they have failed to reject the null hypothesis in a sample of *one* does not strike us as indicative of a "poorly … designed, or executed" experiment, a waste of half a billion dollars, but as a great achievement of humanity.

What is the difference? The null hypothesis – **GR** – is *General Relativity*.

The LIGO paragraph cited above references two papers by Schwarzschild and Kerr from 1916 and 1963. Schwarzschild published these solutions just as Freud was giving his first lectures in Vienna. Imagine submitting a paper about a half a billion dollar machine with $n = 1$ and it failed to reject the null hypothesis that there are unconscious drives. Or ego depletion. Or stereotype threat. Or that myelination predicts intelligence. Or neoclassical economics. People would get their pitchforks. Reviewer #2 would declare a state of emergency and begin lobbying for the introduction of the death penalty for sufficiently bad science.

But when the hypothesis is as specific and important as General Relativity, everything changes. First of all, there is much less *wiggle room* regarding what constitutes a confirmation or a falsification. In the Reproducibility Project (Open Science Collaboration 2015, Science), it became clear that it is not even agreed upon what a replication would be. In physics, it can be rather obvious; from General Relativity as formulated by Einstein at the beginning of the last century, a few values for a few fundamental variables of the universe are al- lowed. Any other value would constitute a falsification. Sure, they also rely on consensus, trust and interpretation - to begin with, the LIGO test of General Relativity depends on (rigorously!) establishing first that the observed signal was truly a black hole merger, and only then it can be tested if the way in which these black holes merged is compatible with General Relativity. But building on that, it is rather obvious how relativity could be proved wrong. If, for example, we performed multiple measurements of increased relativistic matter of bodies we had applied energy to, and energy did not correspond to matter times the speed of light squared, but to matter times the speed of light raised to the power of 4, or even just to matter times the speed of light to the power of 2.25, then something in General Relativity would be completely off, and nearly the whole thing might come crumbling down.

In contrast, in low-reproducibility disciplines, hypotheses are much less precise. Typically, the statistical hypotheses tested are of the form “the effect of X on Y is not *precisely* zero”. A lot of theories are compatible with the effect of X on Y being not quite zero. A typical experiment thus carries remarkably little evidential value about any of these theories; these statistical tests are not *severe test* of the theory the researcher worries about. First, because they are not severe; the chance of the true effect being precisely zero is rather low to begin with. Secondly, because they are not *tests of the theory the researcher cares about*, but of the null hypothesis of no difference from (usually) zero. Physicists don’t test if the speed of light is statistically different from zero, they test what the speed of light is – if it agrees with model predictions. In this way, *most* observations hold some evidential value – not only the speed of light, but also how fast an apple thrown off a crooked tower accelerates, or how warm a fire is. In contrast, in a typical psychological experiment, the researcher doesn’t care so much if subjects take 15, 150, or 500 milliseconds longer to respond to apples than to pears; they only care about if this difference is statistically significantly different from zero. Intuitively, it is clear that if we were to observe that subjects reliably take 12 seconds longer to associate words with positive valence with Black than with White faces, something about our understanding of how humans behave would be incredibly off. But the way the researcher has laid out their research plan, as long as the difference is not zero, but (e.g.) positive, at $p < 0.05$, they have “confirmed” their theory.

This likely entails that researchers of low-reproducibility disciplines “turn around” their statistical inference, to unveil their true Popperian power. The null hypothesis should be the theory you actually care about. Rejecting it means your theory was wrong; failing to reject it (with a high-powered design, of course) means your theory has passed a test. Then, the value to be questioned should not be null, but just what your theory predicts the value should be (e.g., c is ~300.000 m/s, g is ~9.8 m/s). If your theory does not in fact predict a precise value (or a narrow range of values), then your theory is not a theory, but just a vague story.

This vagueness allows a near infinite amount of wiggle room when testing and re- testing a theory. Once a researcher of the human mind has first found something suggestive of a neat theory (somehow

this was supported exclusively by discovering that some parameter is not precisely zero), the next step is to conduct a conceptual replication (which unsurprisingly shows that some other parameter is also not precisely zero). Then, three boxes are connected with two arrows, a caveat is added to appease reviewer #2, and the general feeling is that some- thing has been discovered about the mind. The result is a theory where each researcher must decide for themselves, privately, in their own minds, what would constitute a confirmation or falsification of this theory. What predictions does the theory make? At best, that some other parameter is also not zero. This is no basis on which to establish rules of refutation. There is no objective standard of disconfirmation. There is, however, much room for exaggerated interpretations, overhyping, and, eventually, much need for p-hacking, file drawers and defensive blog posts when a grad student with twice the original sample size observes a failure to reject that the parameter in question is zero.

Theories that only this weakly circumscribe the way the world ought to be are nearly impossible to falsify, and the tests used by scientists consequently never are tests of the hypotheses they use, but of the (uninteresting and almost always a priori unbelievable) null hypothesis. I suggest if theories of brain and behavior had the specificity with regards to its parameters as the theories of physics, experimenter wiggle room would no longer make theories untestable, replicability doubtful, and scientific progress glacial.