Jacques Loeb Centre
for the History and Philosophy
of the Life Sciences
Prof. Ute Deichmann, Director

Ben-Gurion University of the Negev
P. O. Box 653, Beer-Sheva, 8410501, Israel
Building 39, Room-112 | Phone: +972-8-6472258
Email: jloebcentre@post.bgu.ac.il | Web: in.bgu.ac.il/en/loeb

The Problem of Irreproducibility in Scientific Experimentation: Is there a "Replication Crisis"?

The Jacques Loeb Centre for the History and Philosophy of the Life Sciences, Ben-Gurion University of the Negev, March 14, 2022

Ilse Katz Institute for Nanoscale Science and Technology (Bldg. 51), Ben Gurion University of the Negev, Marcus Family Campus, Beer Sheva

Abstracts

Anat Ben Zvi, Ben-Gurion University of the Negev, Israel:

Data Variability Can Sometimes Be New Data Hiding in Plain Sight

Applying simple behavioral assays as a phenotypic readout for protein function is a powerful tool to examine the relationship between genes and their biological function in a living multicellular organism. Such a relationship can be measured by following a detectable phenotype associated with a loss-of-function in one gene (or more) in the genetically tractable metazoan, *Caenorhabditis elegans*. One limitation of behavioral assays is data variability that can affect the interpretation of the results. While data variability can be due to stochastic events, there are examples where environmental or genetic factors contribute to this variability. Isolating such factors can promote our understanding of the underlying mechanism driving the observed phenotype. Here, I will discuss several examples exploring data variability unmasked new findings.

Klodian Coko, Ben-Gurion University of the Negev, Israel:

From Replication to Triangulation: towards a more nuanced evaluation of scientific findings

Currently, there is a widespread perception that scientific activity is in the middle of a (so-called) 'reproducibility' or 'replication crisis'. Many important findings published in leading scientific journals have turned out to be difficult or impossible to replicate. The ongoing controversy surrounding the reproducibility of scientific activity threatens to undermine the authority of science. The extent and severity of the 'replication crisis' are being continuously evaluated. It seems, however, that these discussions, rather than revealing the existence of a fatal flaw at the heart of modern scientific activity, show that our general understanding of the complexities surrounding the replication and reproducibility of experimental findings and experimental procedures is rather limited. Even at this moment, there is no agreement on what constitutes a successful replication: there is no agreement on the criteria that sufficiently describe replication success, both in general but also as it pertains to specific scientific disciplines. Scientists often joke that the definition of replication is itself not replicable.

In this presentation, I draw on philosophical analyses of scientific practices to show that replication is a rich complex notion. Replication practices differ with respect to their epistemic import and are conducted for a wide range of purposes. Contrary to the received view, I argue that “replication” is not the only, nor the fundamental strategy for securing the validity of scientific findings. More specifically, I compare replication practices with other methodological strategies that scientists use to confirm and validate their experimental procedures and results, such as robustness analysis (understood as invariance of an experimental result to variations within the same experimental procedure) and multiple determination or triangulation (understood as the determination of ‘the same’ result by multiple, theoretically independent experimental procedures). I argue that multiple determination can provide a way out of the current failure to replicate results in a statistical setting.

Harry Collins, Cardiff University, UK:

The Inevitability of the Replication Crisis and Why the Solution Is Different for Different Kinds of Science

For certain historical reasons an indicator of scientific correctness setting a threshold for publication in many disciplines is a statistical significance test that achieves a certain threshold. In most sciences the threshold is that the odds of the result being due to random sampling error is around one in 20 or less – the 2-sigma criterion. In the 1960s, physics adopted a more demanding criterion: 3-sigma. Over subsequent decades physics raised the bar to 4-sigma and now to 5-sigma. If, in 2016, you wanted a top physics journal to publish your claim to have detected a gravitational wave based on a statistical analysis of the data, you had to establish that the odds against finding such a thing due to chance was less than on in 3.5 million. Physicists had raised the bar several times because they kept finding that lesser criteria led to false results being published, not because every millionth result or so was false but because there were all kinds of subtle systematic errors in such analyses and they were wanted to swamp them with a statistical sledge-hammer. Physics is a long running experiment which shows that the results of ‘the 2-sigma sciences’ are going to be irreproducible so long as they rely on statistics alone. Every now and again, so long as the science is of the right type (what we call hypernormal science), a meta-analysis might improve the situation, but mostly it will continue to be ‘rubbish in rubbish out’, there’ll just be more rubbish, especially in non-hypernormal sciences. What is needed is to take statistics out of the focus and concentrate on scientific understanding.

Ute Deichmann, Ben-Gurion University of the Negev, Israel:

Irreproducibility and Scientific Truth

I show that irreproducibility is not only a problem of various sciences today, but, while dating back at least to the 17th century, was certainly widespread in the first half of the 20th century. I argue that irreproducibility alone, despite having been a serious problem of scientific practice, does not undermine the authority of science for the following reasons:

- Reproducibility does not necessarily lead to good science. I have shown that invalid results and conclusions can be reproduced many times, because the same questionable methods were employed and because of wishful thinking.

- There is more to good science than reproducible experiments and data and novel ideas and theories often arise on the basis of already confirmed knowledge.

- History shows that despite incorrect theories, experimental flaws, and short-lived fashions, much knowledge has remained robust for a long time. With all the spectacular failures (and successes) that are now making journals' headlines in applied sciences, it should not be forgotten that the basis for success is basic research, which is often unspectacular and fallible.

- Following Yemima Ben-Menachem according to whom the occasional failures of science provide a much better argument for realism than its success, I claim that failures, such as irreproducibility, their disclosure and correction, best confirm that the search for truth in terms of reliable knowledge has remained the fundamental goal of science.

Yonatan Dubi, Ben-Gurion University of the Negev, Israel:

War and Pieces of Metal – A Tale from the Scientific Front

This talk will start with an amazing achievement; leading scientists from the nano-chemistry and nanophotonics community demonstrated a revolutionary method for producing green fuels, by using illuminated nano-particles, which were demonstrated to be amazing catalysts to a variety of useful chemical reactions. Undoubtably, this was a magnificent demonstration of science and nanotechnology at their best, leading to a large number of publications in top and glossy journals, and funding galore.

But then – a twist in the plot! In a series of papers, we (myself along with my partners Yonatan Sivan and Josh Baraban, here at BGU) showed that the interpretation of the most famous experiments in the field was actually one big mistake. Together we identified hair-raising flaws in the experiments and showed to the experiments can be explained – and reproduced remarkably well actually much better than the original attempts – by using basic 19th century physics and chemistry.

This is a talk about science, but also about politics and hypes in science. I hope that it will leave you smiling, and maybe even reminded that science is not about glory or reputation, but about seeking simple truths, even if that means walking against the wind.

Gabriel A. Frank, Ben-Gurion University of the Negev, Israel:

The Einstein from Noise Problem: Model Bias and Resolution Criteria in Single-Particle Cryo-EM

Following the groundbreaking "resolution revolution", culminating with the 2017 Nobel Prize in Chemistry, single-particle cryo-EM became the standard method for determining macromolecular structures. The number of cryo-EM structures determined and deposited quickly grows, as cryo-EM is becoming a mainstream field.

Structure determination—reconstruction—by single-particle cryo-EM involves solving an inverse problem, i.e., determining the underlying 3D electron density map, which best explains the set of 2D projection images collected during the experiment. The solution of this inverse problem involves an unusually large-scale optimization (fitting) procedure that currently has no statistical grantees for success or correctness. Moreover, current optimization strategies are strongly dependent on starting assumptions. In my talk, I will provide a visual demonstration of the optimization procedure and describe its theoretical pitfalls and possible practical errors. I will then describe the standard practices established in the field to minimize and prevent such processing errors and the controversy regarding them.

Gerd Gigerenzer, Max Planck Institute for Human Development, Germany:

The Replication Crisis and How We Got There

The replication crisis in psychology and other social sciences has been attributed to misguided external incentives gamed by researchers (the strategic-game hypothesis). I want to draw attention to a complementary internal factor, namely, researchers' widespread faith in the "null ritual" and associated delusions (the statistical-ritual hypothesis). The crucial delusion is that the p-value specifies the probability of a successful replication (i.e., 1 – p), which makes replicability appear to be almost certain and replication studies appear to be superfluous. A review of studies with 839 academic psychologists and 991 students shows that the replication delusion existed among 20% of the faculty teaching statistics in psychology, 39% of the professors and lecturers, and 66% of the students. Two further beliefs, the illusion of certainty (e.g., that statistical significance proves that an effect exists) and Bayesian wishful thinking (e.g., that the probability of the alternative hypothesis being true is 1 – p), also make successful replication appear to be certain or almost certain, respectively. In every study reviewed, the majority of researchers (56%–97%) exhibited one or more of these delusions. Psychology departments need to begin teaching statistical thinking, not rituals, and journal editors should no longer accept manuscripts that report results as "significant" or "not significant."

John P.A. Ioannidis, Stanford University, U.S.A.:

Are Reproducibility and Transparency Improving and Biases Decreasing across Diverse Scientific Fields?

The increasing interest in reproducibility and transparency over the last decade is a welcome sign of a healthy scientific response to the so-called reproducibility crisis. However, is there also empirical evidence that science is becoming more reproducible and more transparent? Also are biases decreasing over time? There are probably major differences across different scientific fields in this regard. There are many initiatives that have tried to made research work in a more open, sharing environment and to place more emphasis on replicable and unbiased results, but efforts are unevenly scattered across science. Several empirical efforts have tried to assess the reproducibility of research findings. Moreover, multiple surveys have assessed the prevalence and evolution over time of research practices that enhance transparency and, hopefully, also reproducibility. These include, but are not limited to, data sharing, code/algorithm sharing, protocol registration, declarations of conflict and funding, and use of better powered studies and better methods. The lecture will present this empirical evidence and discuss the lessons learned from diverse scientific fields.

Sebastian Kozuch, Ben-Gurion University of the Negev, Israel:

Is Quantum Computational Chemistry Reproducible?

Quantum mechanics is taken as the epitome of physical accuracy. And yet, while quantum chemistry has been active for almost a century (arguably since the first calculation of the H2 chemical bond in 1927), it got an aura of a reputable science only since the mid-90s, when DFT took the stage as the state-of-the-art method for molecules. At present, if someone would dare to submit an article on a high-level multidisciplinary chemical journal, including a computational chemistry section might add the prestige points necessary to get it accepted.

However, computational chemistry in general, and quantum chemistry in particular, can only be applicable through several approximations (accuracy of the ab initio or DFT method, basis set size, solvation models, harmonic approximation, gas-phase entropies, etc.). Juggling with all these approaches is a craft, and knowing the confidence interval of a particular value can almost be a mystery (especially for the non-specialist). Due to this, the highly reputable quantum chemistry might actually be a gross approximation, or even fall into a Texas sharpshooter fallacy where everything can be “proven” with the correct method selection (a kind of “everything goes”, paraphrasing Feyerabend).

Therefore, while formal reproducibility is not an issue in quantum chemistry, a value can still be completely irreproducible. In such condition, what validates or invalidates this field? Are its prediction useful? Is it even falsifiable?