IOP Concise Physics

Measuring Nothing, Repeatedly

Null experiments in physics

Allan Franklin and Ronald Laymon

Chapter 1

Introduction

As indicated by our title, our aim is to highlight and examine an important species of scientific experiment, namely, those that deliver a null or 'zero' result. The importance of such experiments derives from what is often their central role in the development of theory and their associated deep connections at a foundational level. As also indicated by our title, we intend to focus on why, as a feature of scientific practice in physics, frequent replication of such experiments occurs and what the conditions are for meaningful replication.

It is virtually axiomatic that 'replication—the confirmation of results and conclusions from one study obtained independently in another—is considered the scientific gold standard.' (Jasny *et al* 2011). The underlying argument for this is that if an experiment has succeeded in revealing a real phenomenon or accurately measuring a quantity then that success should reappear when the experiment is repeated under the same circumstances or when it is reproduced in a different experiment.

By way of providing a revealing comparison with the null results of physics, we note that considerable doubt has been expressed that this replication requirement is satisfied in the social sciences. Thus, for example, the Open Science Collaboration attempted to replicate 100 experimental results, which appeared in three leading psychology journals. 'We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered design and original materials where available' (Aarts *et al* 2015, p 943). They noted that 'there is no single standard for evaluating replication success' (p 943). Depending on the criteria used, they estimated that either 47% or 39% of the original studies had been successfully replicated[1]. This was in contrast to an expected failure rate of less than 10%. Hence there was a problem.

[1] The collaboration used significance, *P* values, effect sizes, subjective assessments of replication teams, and meta-analysis of effect sizes.

doi:10.1088/2053-2571/ab3918ch1

This work, however, was criticized by Gilbert and his collaborators:

> A paper from the Open Science Collaboration attempting to replicate 100 published studies suggests that the reproducibility of psychological science is surprising low. We show that this article contains three statistical errors and provides no support for such a conclusion. Indeed, the data are consistent with the opposite conclusion, namely that the reproducibility of psychological science is quite high. (Gilbert *et al* 2016)

Questions were raised as to whether the attempted replications were sufficiently similar to the original experiments to count as failed replications. Using results from the 'Many Labs' project, they concluded that 'a full 85% of the original studies were successfully replicated' (Gilbert *et al* 2016).

The Open Science Collaboration responded:

> Reproducibility Project: Psychology indicates high reproducibility, given the study methodology. Their very optimistic assessment is limited by statistical misconceptions and by causal inferences from selectively interpreted data. Using the Reproducibility Project: Psychology data, both optimistic and pessimistic conclusions about reproducibility are possible, and neither are yet warranted. (Anderson *et al* 2016)

The discussion continues. One need not delve into the statistical weeds to acknowledge the obvious problems of ensuring relevantly similar, or relevantly superior, initial or test conditions. Add to this the deeper methodological problems, such as insuring adequate sample size and statistical power, and problems with replication, even if not of crisis proportion, are inevitable. And since there seems little point in exactly reproducing the original experiment, there is the additional problem of knowing whether proposed improvements really serve to do so. Thus, as aptly summarized by Anderson *et al*:

> More generally, there is no such thing as exact replication. All replications differ in innumerable ways from original studies. They are conducted in different facilities, in different weather, with different experimenters, with different computers and displays, in different languages, at different points in history, and so on. What counts as a replication involves *theoretical assessments* of the many differences expected to moderate a phenomenon. (Anderson *et al* 2016, emphasis added)

The reference to *theoretical assessments* is especially noteworthy because the availability and depth of such assessments identifies a potentially telling point of comparison between the social and the physical sciences. But since the spirit of Anderson's appraisal applies as well to experimentation in physics, we will adopt a broad view of replication. It will not be solely performing the experiment again with either the same or a very similar experimental apparatus but also includes

experiments that employ different apparatus. We will also consider in this regard experiments that examine different phenomena that bear on the same theory or hypothesis, since such experiments serve the purpose of validating the design and execution of the original experiment[2]. This sort of broad view was developed and applied in Franklin (2018) which included cases of both successful and failed replications, along with episodes in which there were difficulties in determining in what sense a replication had been achieved.

The Open Science Collaboration suggested the existence of a *research and publication bias* as an additional contributing cause for the problems of replication. The suggestion is that both journals and the scientists themselves value positive results more than negative or null results and thus may not publish or even submit negative results. This has been called the 'file drawer' problem in which negative results are filed away and not submitted for publication. In particular Anderson *et al* make a persuasive case that the *research and publication bias* operates in three ways: (1) to encourage positive results (i.e. confirmation of the test hypothesis); (2) to discourage publication of failed attempts to confirm the test hypothesis; and (3) to discourage replications of both positive and negative results where a negative result is a failure to confirm the test hypothesis. There is the additional desideratum that positive results are preferred that have a large size of effect[3].

A recent more general review of replication in the social and other sciences by Randall and Welser reinforces the contention that replicating the results of others is not as highly regarded or rewarded as original work:

> Modern science's professional culture prizes positive results, and offers relatively few rewards to those who fail to find statistically significant relationships in their data. It also esteems apparently groundbreaking results far more than attempts to replicate earlier research. PhDs, grant funding, publications, promotions, lateral moves to more prestigious universities, professional esteem, public attention—they all depend upon positive results that seem to reveal something new. A scientist who tries to build his career on checking old findings or publishing negative results isn't likely to get very far. Scientists therefore steer away from replication studies, and they often can't help looking

[2] This broad view of replication is further motivated by the fact, as argued by Franklin and Howson (1984), that 'different' experiments provide more support for a hypothesis or an experimental result than narrowly conceived replications of the 'same' experiment.

[3] In support, Anderson *et al* argue that 'low power research designs combined with publication bias favoring positive results together produce a literature with upwardly biased effect sizes. This anticipates that replication effect sizes would be smaller than original studies on a routine basis—not because of differences in implementation but because the original study effect sizes are affected by publication and reporting bias, and the replications are not. Consistent with this expectation, most replication effects were smaller than original results, and reproducibility success was correlated with indicators of the strength of initial evidence, such as lower original *P* values and larger effect sizes. This suggests publication, selection, and reporting biases as plausible explanations for the difference between original and replication effects. The replication studies significantly reduced these biases because replication preregistration and pre-analysis plans ensured confirmatory tests and reporting of all results.' (Anderson 2016, at 3)

> for ways to turn negative results into positive ones. If those ways can't be found, the negative results go into the file drawer.
>
> Common sense says as much to any casual observer of modern science, but a growing body of research has documented the extent of the problem. As far back as 1987, a study of the medical literature on clinical trials showed a publication bias toward positive results. Later studies provided further evidence that the phenomenon affects an extraordinarily wide range of fields, including the social sciences generally, climate science, psychology, sociology research on drug education, research on informational technology in education, research on 'mindfulness-based mental health interventions,' and even dentistry. (Randall and Welser 2018, p 35)

While we doubt that practitioners in the physical sciences are entirely immune from this sort of research and publication bias, we suspect that its force and pervasiveness is limited[4]. It is, in part, to test this appraisal that we have embarked on our study of the replication of null results in physics. Our title, *Measuring Nothing, Repeatedly*, is meant to invoke the question: why bother to repeat the measurement of 'nothing'? In the social sciences, the answer is in large part: don't bother, it does no good for your career. That, however, is in the main not true in the physical sciences. Or so we aim to substantiate in our study of null results in physics.

Being not true in the main, however, does not mean never being true. Hence there are express exceptions, such as contained in the work of Edwin Hall, who obtained a null result on the question of whether falling bodies move south as well as east, where he remarked that:

> … granting that every experimenter probably wanted to find some deviation, *a positive result in such research being far more interesting than a negative one*, granting that in a case like this, which presents great difficulties and uncertainties, *a prejudice in favor of this or that result* may lead the experimenter to look farther, so long as his expectation is not fulfilled—granting all this, the writer finds himself unable to remain quite content with the theory that these conditions have been created out of nothing the general evidence in favor of a southerly deviation. (Hall 1903, pp 189–90, emphasis added)

In addition to the possible and general effect of research and publication bias on experimentation in physics, there are several more particular reasons for our focus on null results. First, there is the fact of the numerous episodes involving null results and *replications* of those results[5]. Furthermore, because of their typical centrality to

[4] The late Stuart Freedman, a member of the National Academy of Sciences, for example, told Franklin that during his career he had performed 27 null experiments (private communication to the author). He measured nothing, but measured it very well.

[5] For example, almost all tests of conservation laws, such as energy, momentum, angular momentum, and charge, can and have been formulated as null experiments such as: is the difference in energy before and after an interaction equal to zero? Such experiments can also be formulated to give a non-zero result. Is the ratio of the energies before and after an interaction equal to one?

theoretical advancement, their design and validation makes heavy use of applicable and available *theoretical assessments* which greatly facilitate the process—often including, given their 'zero' result, theoretically significant underlying symmetries. In short, they are *deeply embedded in a well-developed theoretical context*. This feature, at least as a relative matter, is not the case in the social sciences.

We'll need in our examination of the historical cases to be more precise when it comes to claiming a 'nothing' or 'null' result as opposed to a result that is in some way or other 'negative' but not necessarily null or nothing. Taking, for the moment, Michelson's interferometer results as a paradigmatic example of a null result in physics, a result may be said to be null when it not detected by the measuring devices employed. Roughly speaking, the value returned by the measuring instrumentation is 'zero.' Of course, it is very rarely the case that an unadulterated zero result will occur since there will almost always be measurable, small interfering causes and resultant noise at play—as there was in the Michelson–Morley experiment. So a better description of a null result is that it is 'zero' plus small though annoying residual variations, i.e. low level pollution of the purity of a true zero. Thus, to describe the result as *effectively zero* is to indicate that the residual variations from zero are of no consequence and have been or likely to be explained away. In addition, experimenters may include an estimate of the uncertainty in reporting their result.

We will have much to say in our case studies about the *credibility* of claims to have explained away or otherwise rendered harmless such residual variations. In this regard see (Franklin 2004) where he has specifically discussed the strategies used by experimenters to establish the correctness of their null result. These include the use of surrogate signals as well as the use of blind injections, the insertion of simulated events into the data stream to judge whether those injections would be detected. More generally, the requisite credibility is provided by the use of an epistemology of experiment, i.e. a set of strategies used to argue for the correctness of an experimental result (Franklin 2002, pp 2–6, chapter 6)[6].

Returning to the social sciences, it must be acknowledged that there is not a straightforward correspondence between what we have described as the

[6] These strategies include (1) experimental checks and calibration, in which the experimental apparatus reproduces known phenomena; (2) reproducing artifacts that are known in advance to be present; (3) elimination of plausible sources of error and alternative explanations of the result; (4) using the results themselves to argue for their validity. In this case, one argues that there is no plausible malfunction of the apparatus, or background effect, that would explain the observations; (5) using an independently well-corroborated theory of the phenomena to explain the results; (6) using an apparatus based on a well-corroborated theory; (7) using statistical arguments; (8) manipulation, in which the experimenter manipulates the object under observation and predicts what they would observe if the apparatus was working properly. Observing the predicted effect strengthens belief in both the proper operation of the experimental apparatus and in the correctness of the observation; (9) the strengthening of one's belief in an observation by independent confirmation; (10) using 'blind' analysis, a strategy for avoiding possible experimenter bias, by setting the selection criteria for 'good' data independent of the final result. As will be shown below, the use of these strategies is often an important part of determining whether a replication has been successful or not. One can argue for the rationality of these strategies by embedding them within a Bayesian approach (Franklin and Howson 1988).

paradigmatic null results of physics and what, somewhat misleadingly, may be described as the 'null' results of the social sciences. Thus, for example, Anderson *et al* in their review of the replication 'crisis' in the social sciences, do *not* employ the expression 'null result' but rather speak of the 'null hypothesis of no effect' (see, for example, Anderson 2016). Showing, however, that the statistical 'null hypothesis' survives statistical examination, and that the test hypothesis does not, is not the same thing as our paradigmatic instance of a null result in physics, i.e. reading a 'zero' off one's instrumentation. Accordingly, Anderson *et al* instead speak of *positive* and *negative* results, where (as noted above) a positive result is the confirmation of the test hypothesis and a negative result is a failure to confirm the test hypothesis. Thus, negative results (i.e. what *might* be referred to as a 'null' result) in the social sciences are, as it were, *locked into the particular hypothesis being tested.*

By contrast, null results in physics are *chameleon* in character in the sense that they may serve to confirm one theory and disconfirm a competing theory, albeit not necessarily at the same time[7]. Not surprisingly then, null results and replications have played such important roles in physics as deciding between discordant results, deciding between hypotheses or theories, demonstrating that a previous result is incorrect, and confirming a theory. Still, the essential character of the research and publication bias remains even in the case of physics where the bias can be expected to be against the 'mere' replication of an already 'established' zero result or where, by extension, the null result is a consequence of well-established theoretical considerations.

This difference between the null results of physics and the survival after testing of null hypotheses of the social sciences, while initially striking, must be tempered in at least two respects. First, the analysis of the complex instrumentation used in modern physics incorporates highly sophisticated forms of statistical analysis. The days of simply reading the measurement result off the instrument dial are long gone. In short, the path from an instrument reading to measurement value will be mediated by extensive statistical analysis. Thus, considering the heavy use of statistics in modern physics, it would be better to say (as will be shown in our historical case studies) that there is more to a null result than just a registration of zero with a wiggle of noise that can be directly seen on the instrumentation.

Second, nothing said by Anderson *et al* prohibits the use of a confirmed test hypothesis to either confirm or disconfirm a higher level theory in the social sciences. In such cases, a confirmed test hypothesis will be *chameleon* in ways analogous to those of confirmed null results in physics and more generally to non-null results as well. The possibilities, however, for such embedded application are, we believe, less extensive in the social as compared with the physical sciences.

[7] For example, Galileo's experiment on falling bodies at the Leaning Tower refuted Aristotle's theory that objects fall at speeds proportional to their weight. It later confirmed Newton's theory that all bodies fall at the same rate. In 1957, three experiments demonstrated that the class of theories that conserved parity, or space reflection symmetry, was refuted. At the same time, they confirmed the class of theories that violated parity conservation. No specific theory was involved. For details, see Franklin (1986, chapter 1).

Moreover, and continuing in the same vein, while in contemporary physics as in the social sciences, the use of statistical analysis is *de rigeur*; there is this related difference because of the deeper and more developed theoretical background in the physical sciences, namely, the central fact that in the physical sciences *very small differences* can have *very large consequences* for whether proposed theories fail or pass their experimental examination. In other words, results in the physical sciences can be *extremely sensitive* to small variations in experimental results. This means that in physics and especially with respect to a purportedly null result, explaining away small residual variations from 'zero' will rarely be a simple matter. Consequently, there exists in the physical sciences *an urgency and strong motivation for replication*, since experimental results are in a sense 'up for grabs' given the difficulties and uncertainties in dealing with what are characterized as statistical error and systematic uncertainty. Bluntly stated, this sort of urgency and motivation is largely absent in the non-physical sciences because of the relative absence of such sensitivity and associated measurement techniques employed in those sciences.

With these preliminaries in hand, we now embark—with a promise to further elaborate and clarify those preliminaries—on our examination of the historical cases of null results in physics and their replication.

References

Aarts A A and Anderson J E *et al* 2015 Estimating the reproducibility of psychological science *Science* **349** 943

Anderson C J and Bahnik S *et al* 2016 Response to comment on 'Estimating the reproducibilty of psychological science' *Science* **351** 1037-c

Franklin A 1986 *The Neglect of Experiment* (Cambridge: Cambridge University Press)

Franklin A 2002 *Selectivity and Discord* (Pittsburgh, PA: University of Pittsburgh Press)

Franklin A 2004 Doing much about nothing *Arch. His. of Exact Sci.* **58** 323–79

Franklin A 2018 *Is It the Same Result? Replication in Physics* (Bristol: IOP Publishing)

Franklin A and Howson C 1984 Why do scientists prefer to vary their experiments? *Stud. Hist. Phil. Sci.* **15** 51–62

Franklin A and Howson C 1988 It probably is a valid experimental result: a Bayesian approach to the epistemology of experiment *Stud. Hist. Phil. Sci.* **19** 419–27

Gilbert D T and King G *et al* 2016 Comment on 'Estimating the Reproducibility of Psychological Science' *Science* **351** 1037-b

Hall E H 1903 Do falling bodies move south? *Phys. Rev.* **17** 179–90

Jasny B R and Chin G *et al* 2011 Again and again and again *Science* **334** 1225

Randall D and Welser C 2018 *The Irreproducibility Crisis of Modern Science* (New York: National Association of Scholars), pp 1–70