DRAFT

Chapter from a book, *The Material Theory of Induction*, now in preparation.

Reproducibility of experiment.

John D. Norton
Department of History and Philosophy of Science
Center for Philosophy of Science
University of Pittsburgh
http://www.pitt.edu/~jdnorton

The general idea is simple and instantly compelling. If an experimental result has succeeded in revealing a real process or effect, then that success should be reproducible when the experiment is done again, whether it is done by the same experimenter in the same lab or by others, elsewhere, using equivalent procedures. It is, at base, the same idea that leads us to the near universal reaction when a conjurer makes a coin vanish. "Do it again!" we demand. And this time, we will watch more closely.

What kind of an inductive notion is reproducibility? Is it possible to state it as a general principle? A good start is something like this:

> Reproducibility of an experiment is a good indicator of a veridical experimental outcome; failure of reproducibility is a good indicator of a spurious experimental outcome.

This is far from self-contained. Each term needs further explication. The more straightforward are the notions of real and spurious experimental outcomes:

> A veridical experimental outcome is one that properly demonstrates the process or effect sought by the experimental design.
>
> A spurious or artefactual experimental outcome fails to do so; it arises from an unintended disruption to the experimental design.

This is a rich enough characterization for us to proceed, even though many details are left open. For example, the terms "repeatability," "reproducibility" and "replicability" are often used

loosely and interchangeably. In some contexts, repeatability indicates as exact a replication of all conditions as possible; whereas reproducibility is looser allowing some change of conditions.[1]

1. Failure of Formal Analysis

Can we give a formal analysis of the requirement of reproducibility? Is it a universal inductive principle, perhaps an inductive analog of the universal, formal principles of deductive logic. In asking, we should bear in mind what the latter are like. One such universal deductive principle is the law of the excluded middle. It asserts:

For any proposition P, either P is true or P is false.

This deductive principle is a schema: we can insert any proposition we like for "P" and recover a truth, the application of the principle to that proposition. It is self-contained. There are no tacit conditions limiting just which propositions can be substituted for "P"; and there is no ambiguity in what is meant by the truth or falsity attributed to the proposition (Or at least there are none beyond the usual evasions made by philosophers when they have to use these terms.)

It is quite different with characterization of reproducibility of experiment above. That characterization includes many notions that require elaboration if the characterization is to rise to the level of precision of the law of the excluded middle. Just what is "a process or effect sought by the experimental design"? Just when is a second experiment reproducing an earlier

---

[1] In the narrower context of standardized measurement, the *International Organization for Standardization* has decreed (ISO 21748:2010(E), p. 3): "Repeatability conditions include: the same measurement procedure or test procedure; the same operator; the same measuring or test equipment used under the same condition; the same location; repetition over a short period of time. Reproducibility requires only that the measurement must reappear under changed conditions. That is, (ISO 21748:2010(E), p. 3): "reproducibility conditions[:] observation conditions where independent test/measurement results are obtained with the same method on identical test/measurement items in different test or measurement facilities with different operators using different equipment[.]"
Source: "Guidance for the use of repeatability, reproducibility and trueness estimates in measurement uncertainty estimates," Publication ISO 21748: 2010(E).

experiment as opposed to being a different experiment that looks similar to it? Elaborating on these and similar questions is likely to be tedious and unlikely to yield a formulation of the precision of the law of the excluded middle.

Rather than pursuing all these aspects of the characterization, it is sufficient to scrutinize one aspect to show that a formal analysis faces insuperable barriers. The characterization speaks of "good indicators." This is an inherently vague notion. Presumably there is some idea that multiple successful repetitions are better than just one. How much better are they? Is there a point of diminishing returns? When there are some successes of replication and some failures, how do we trade them off to come to our final assessment? Even in the case of a single successful or failed replication, the strength of the indication can have the widest range, as will displayed in the examples that follow shortly. Somehow the formal analysis will need to formulate a principle that can express the precise conditions under which all these differing levels of strength obtain.

2. A Material Analysis

How strong is the inductive import of a successful or failed replication? A general principle is unable to answer this question and others like it, raised above. Rather, these questions are answered differently in different contexts. The more we narrow the context, the more precise the answers become. This is what we would expect from a material approach to induction. What appeared as a universal principle is really only a resemblance among many distinct inductive inferences that vary in details according to their domains. No universal principle of inductive logic provides a warrant for these individual inferences. They are warranted by the particular facts prevailing in each domain.

In principle, each case will be unique with its own analysis. In practice, there will be similarities across many cases. Successful replication is generally favorable; and failures to replicate are generally unfavorable. This broad similarity across many cases supports the illusion that there is some general inductive principle concerning reproducibility at work.

The situation is not so different from the case of enumerative induction. In many domains, we find the background facts warranting an inference from some individuals bearing a property to all individuals in that class bearing the property. I argued above that these cases must

be treated individually. The different background facts obtaining in each case will specify which individuals in the domain are subject to the generalization, for which properties and to which strength. Nonetheless, as a looser gloss, the inferences will look something like a progression from "Some As are B." to "All As are B." It can be glossed loosely as enumerative induction, but all efforts to find a single inductive schema implemented in all the cases fails. Their unity is superficial.

At this similar level of superficial unity, we can identify two classes of background facts that serve to license the inferences associated with reproducibility. They are:

A. Experimental conditions: these background facts specify conditions under which the effect or process of interest will manifest in an experimental outcome.[2]

B. Confounders: these background facts specify the conditions conducive to confounding effects. These effects simulate what might be mistaken as a successful outcome, when the effect or process is not present; or may interfere sufficiently to produce an unsuccessful outcome, when the effect or process is present.

Inductive inferences associated with the replication of experiments will be warranted by facts in both classes A and B. The first, A. Experimental conditions, are required to enable an inference from the experimental outcome to the effect or process of interest, whether the experiment is the original attempt or a replication. The facts under B. Confounders are of special relevance to replication. For successful replication generally establishes that confounding effects were not responsible for the outcome; and unsuccessful replication suggests the confounding effects were responsible. An inductive inference of this type will be licensed by facts about the confounders, such as in B.

A common way in which spurious results arise is from errors in the original experiments, in the details of the design or its execution. A common goal of replication is to replace all the specific conditions of the original experiment with a novel set, so that confounders associated with that part of the original experiment are no longer present.

The four examples below are intended to display how a material theory deals with successes and failures of replication. The four examples have been chosen specifically to display the huge range in strength of inductive import associated with success and failure of replication.

---

[2] This is sometimes called "construct validity."

We will see a case in which replication is a strong indicator of a veridical outcome; and one in which it is regarded as having no evidential import. We shall see a case in which failure to replicate is a strong indicator of a spurious outcome; and one in which its inductive import is discounted. These combinations are illustrated in the table:

	Requirement of reproducibility upheld	Requirement of reproducibility discarded
Successful replication	H. Pylori Stomach Ulcers (result accepted as veridical)	Intercessionary prayer (result rejected as spurious)
Failed replication	Cold fusion (result rejected as spurious)	Miller experiment contradicts relativity (relativity upheld as veridical)

In the first column, the requirement of reproducibility is applied as expected; in the second, it is discarded and results directly contradicting the requirement are upheld.

It remains unclear how this range of strengths could be incorporated into a formal principle. However, in the material approach, the differences are recovered from differing background facts.

3. H. Pylori Stomach Ulcers: Successful Replication

In 2005, Barry Marshall and Robin Warren won the Nobel Prize in Physiology or Medicine with the citation reading "for their discovery of the bacterium Helicobacter pylori and its role in gastritis and peptic ulcer disease." Prior to their work, it had been assumed that stomach ulcers were caused by stress and spicy food. The idea that a bacterium may be involved was discounted. The stomach is highly acidic and bacteria do not tolerate such environments well.

By taking biopsies from 100 participant patients, as reported in their initial letter (Marshall and Warren, 1983), they were able to demonstrate an association between the presence of the bacterium Helicobacter pylori and gastritis and ulcers, with 100% association for duodenal ulcers. The importance of replication even at this early stage became clear when they sought to publish a more complete account. Warren (2005, pp. 301-302) recounts the decisive moment.

> We sent our definitive paper to the *Lancet* in 1984 ([Marshall and Warren, 1984]). Although the editors wanted to publish, they were unable to find any reviewers who believed our findings. Our contact with Skirrow became crucial here. We told him of our trouble, and he had our work repeated in his laboratory, with similar results. He informed the *Lancet* and shortly afterwards they published our paper, unaltered.

Contrary to a persistent myth, the new work was assimilated and rapidly repeated. As part of an account debunking this myth, Atwood (2004) reported:

> Within a couple of years of the original report, numerous groups searched for, and most found, the same organism. Bacteriologists were giddy over the discovery of a new species. By 1987—virtually overnight, on the timescale of medical science—reports from all over the world, including Africa, the Soviet Union, China, Peru, and elsewhere, had confirmed the finding of this bacterium in association with gastritis and, to a lesser extent, ulcers.

One replication was more of a media stunt than controlled science. To prove the association, Marshall drank a beaker of Helicobacter pylori and subsequently succumbed to gastritis.

This is a "text book" case of the proper functioning of replication and there is little in it to distinguish formal and material approaches. The earlier reluctance to accept Marshall and Warren's work is readily explained materially. As long as it was taken as a background fact that bacteria do not live well in the highly acid environment of the stomach, there are insufficient facts in the background to support for the facts in class A. Detection of bacteria can only be through some coincidental contamination.

4. Cold Fusion: Failed Replication

The episode of controlled fusion is traditionally presented as one in which an avenue of research was closed because of failure of replication. At the most superficial level, that may be a correct description. However a closer look at the episode reveals something more complicated than the application of some principle of reproducibility. There certainly were many failed attempts at replication reported. However there were also many successful replications also reported. This has lead to a bifurcation in the community into those who discard the idea of cold

fusion (the establishment view) and those who continue to pursue it (a dissident minority). No simple inductive principle concerning reproducibility of experiment can capture the inductive reasoning associated with this bifurcation. It derives essentially from differences in the background assumptions of the groups and talk of replication is really a gloss on more complicated inferences, as the material theory of induction indicates.

Traditional nuclear power generation derives from the fission—the splitting apart—of radioactive Uranium or Plutonium atoms. This fission is distinct from the nuclear reactions that power stars like our sun. They are driven by fusion—the joining together—of atoms of hydrogen and other light elements to form heavier elements. In the process, prodigious quantities of energy are released. It has long been a goal of the nuclear power industry to adapt fusion reactions to power generation. Their present terrestrial use has been limited to the uncontrolled fusion in hydrogen bombs. The difficulty is that enormously high temperatures are needed to smash the hydrogen atoms together sufficiently energetically to ignite a fusion reaction. Materials at these high temperatures are difficult to control in a power station and practical, fusion-based nuclear power generation remains a distant dream.

In March 1989, chemists Martin Fleischmann and B. Stanley Pons announced in a press release from the University of Utah that they had found a way of carrying out fusion reactions on a laboratory bench at ordinary temperatures. Their experiments did not use hydrogen but a heavier isotope of hydrogen, deuterium, in the form of deuterium oxide, also known as "heavy water." They electrolyzed the heavy water using palladium electrodes. Over a lengthy electrolysis, one of the palladium electrodes, the cathode, would become saturated with deuterium and, as a result, the individual deuterium atoms would be driven closely enough together to ignite a nuclear fusion reaction. At least, that is what they claimed had happened, on the basis of the large quantities of heat produced. These quantities were greater than could be recovered from chemical changes, they asserted. In one burst, the released heat had melted and vaporized part of the electrode, destroying some of the equipment. Then, Steven Jones, working at nearby Brigham Young University, revealed that he had been working largely independently on a similar cold fusion project and had experimental results involving not the generation of heat, but neutrons, a familiar signature of nuclear reactions.

Whether the researchers succeeded in igniting fusion reactions remains debated. However there is no doubt that they ignited a scientific and popular frenzy. The principal trigger was the

possibility of a new process that would revolutionize the power generation industry. There was a scramble to replicate the cold fusion experiments in the US and internationally. The resulting episode was complex and fascinating on many levels. Cold fusion, if affirmed, would be a scientific discovery of the highest order. That lofty pinnacle was overshadowed by the possibility of new technology for a major industry and its lucrative patent rights. These financial motivations lent an uncommon urgency in what was otherwise the realm of arcane specialists. There were other tensions, such as the professional rivalry of physicists and chemists. Here were physicists failing to tame nuclear fusion with enormous, expensive devices; and now some chemists succeed with a project plotted in one of their kitchens and funded personally. Then there was a soap-opera quality to the rivalry between the Fleischmann/Pons and Jones projects. They had planned to coordinate their communications, but the arrangements had misfired and Fleischmann and Pons took the unusual course of announcing their discovery through a press release without Jones' knowledge.

Let us set all these complications aside and focus on the inductive relations. While there was initially considerable confusion over the inductive import of the experiments, that confusion resolved within a year into two views and it has largely remained so bifurcated. The establishment response was that the experiments failed to demonstrate fusion on the lab bench and that only modest resources should be assigned to further research. The minority, dissident view was that a great discovery had been made and all efforts should be put into developing it.

We find a clear statement of establishment view in the November 1989 report of the Energy Research Advisory Board to the US Department of Energy (ERAB, 1989). It concluded in its Executive Summary that

> The Panel concludes that the experimental results on excess heat from calorimetric cells reported to date do not present convincing evidence that useful sources of energy will result from the phenomena attributed to cold fusion. In addition, the Panel concludes that experiments reported to date do not present convincing evidence to associate the reported anomalous heat with a nuclear process.

The Board was reserved in its recommendation for action:

> The Panel recommends against the establishment of special programs or research centers to develop cold fusion. However, there remain unresolved

> issues which may have interesting implications. The Panel is, therefore, sympathetic toward modest support for carefully focused and cooperative experiments within the present funding system.

The dissident community continued its research and, in 2004, was successful in pressing the US Department of Energy to reopen its evaluation. The community supplied a document, "New Physical Effects in Metal Deuterides," that was subjected to peer review and discussion. It was found (DOE, 2004) that "…the conclusions reached by the reviewers today are similar to those found in the 1989 review." The bifurcation remained unbreached.

Both sides deferred to reproducibility as a guiding standard. The 1989 Advisory Board report (ERAB, 1989) commences its preamble by noting the failure of reliable replication:

> Ordinarily, new scientific discoveries are claimed to be consistent and reproducible; as a result, if the experiments are not complicated, the discovery can usually be confirmed or disproved in a few months. The claims of cold fusion, however, are unusual in that even the strongest proponents of cold fusion assert that the experiments, for unknown reasons, are not consistent and reproducible at the present time.

However mere problems of reproducibility cannot be the principal basis for the solidly negative conclusions reached by the Advisory Board. For their report documents both successful and failed replications of the various types of experiments aimed at testing cold fusion. For example, in relation to experiments yielding excess heat, the report's Table 2.1 lists five experiments that found excess heat and thirteen that did not. While the ratio of five to thirteen certainly favors the no-heat result, it is hardly sufficient to dismiss the effect, especially when its reality, if demonstrated, would be of great utility.

The deeper grounding for the negative report is laid out early in the report (pp. 6-8) when answers are offered to the rhetorical question "Then why the skepticism?" The first reason is developed only in a few sentences: many researchers have been unable to replicate the excess heat effect; and these calorimetric measurements are technically rather difficult. The two remaining reasons are developed in some detail and amount to conflicts between the particulars of the positive experiments and the accepted science of nuclear reactions.

The second reason was summarized as:

> the discrepancy between the claims of heat production and the failure to observe commensurate levels of fusion products, which should be by far the most sensitive signatures of fusion.

The nuclear reactions proposed for cold fusion involve fusion of two deuterium atoms to produce other atoms. Various reactions were possible and they would yield tritium, isotopes of helium or other products. The quantities of these fusion products detected did not match the quantities of heat reported. It was as if one burns wood in a fire. From the heat generated, one can determine how much wood ash must fall through the grate. The positive experiments were not finding the right amounts of ash.

The most important discrepancy was in neutron production. The most likely fusion reactions would produce neutrons and in large quantities. The report noted:

> The initial announcement by Pons and Fleischmann in March 1989 exhibited the discrepancy between heat and fusion products in sharp terms. Namely, the level of neutrons they claimed to observe was 10^9 times less than that required if their stated heat output were due to fusion.

This discrepancy was noted very early by critics and, by itself, was deemed sufficient for instant dismissal of the claims of cold fusion. Here is how one popular narrative from 1989 reported the problem (Peat, 1989, p. 82)

> According to Robert L. McCrory of the University of Rochester's Laboratory of Laser Energetics, for example, if nuclear fusion was really taking place, then the only way to make sense of all that heat was to have a trillion neutrons being emitted each second—enough to kill everyone in the room.
>
> By now the following joke had begun to circulate around the world's laboratories:
>
> FIRST SCIENTIST: Have you heard about the dead-graduate-student problem?
>
> SECOND SCIENTIST: No, what's that.
>
> FIRST SCIENTIST: There are no dead graduate students.

The third reason was summarized as "cold fusion should not be possible based on established theory." Deuterium does not undergo fusion reactions under normal conditions because the electrostatic repulsion of the nuclei prevent its atoms approaching closer than about 0.1 nanometers, which is too great a separation for a nuclear reaction to start. The hope of the

cold fusion researchers was that a palladium electrode could be so densely laden with deuterium that sufficiently close approaches would occur. The report, however, disputed these hopes. The closest approach of deuterium atoms in palladium is just 0.17 nanometers. That is over twice the distance (0.074 nanometers) separating two deuterium atoms in molecular deuterium, D_2. The cold fusion researchers would be bringing the deuterium atoms closer if they merely left them in the form of free molecular deuterium.

Supporters of cold fusion also defer to the idea of reproducibility. Sturms (2007, p. 49) initiated the discussion of the challenges to cold fusion with the resounding affirmation:

> Replication is the gold standard of reality. If enough people are able to make an effect work, the consensus of science and the general public accept the effect as being real and not error or figment of imagination.

He affirmed that replication has been successful:

> A Myth has formed about cold fusion not being duplicated, being based on error, and being an example of "pathological science", […] i.e. wishful thinking. None of this description is correct. The basic claims have been duplicated hundreds of times and continue to be duplicated by laboratories all over the world, although success is difficult to achieve.

However he also allowed that the replication has not been uniformly successful (p. 117):

> Replication occurs when other people observe the same effects using essentially the same conditions. Unfortunately, in the case of cold fusion, the required conditions are not known. Occasionally, when a lucky combination of conditions has been created, the effects are observed. These effects have been seen many times, as the results listed throughout the book demonstrate, but not always on command. This failure of the effects to occur every time they are sought has become a major issue for the field and needs to be examined in detail because some confusion exists about what replication actually means.

The record of successful replication was reinforced with massive tables listing many successes. The table listing experiments that report successful "anomalous power" production spans nearly ten pages (pp. 52-61).

Sturms came to very different conclusions than the Advisory Board concerning cold fusion. He regarded cold fusion as established fact to be announced with text-book like certainty (p. 190):

> The phenomenon of cold fusion or low energy nuclear reaction occurs in an unusual solid or even within complex organic molecules. A variety of nuclear reactions are initiated, depending on the atoms present. Some of these reactions occur at a rate sufficient to make measurable heat. The most active reaction produces ^{4}He when deuterium is present. Other reactions occur at lesser rates, but rapidly enough to accumulate detectable nuclear products.

Where the Advisory Board report found the existing theory of nuclear fusion secure and unfavorable to cold fusion, Sturms inverted the relation and impugned the theory for its failure to accommodate experiment.

His treatment of neutron emissions illustrates this inversion. Standard nuclear physics allows for deuterium to fuse in several ways. The most probable reactions yield high neutron and proton emissions. The reaction favored by cold fusion supporters was the fusing of two deuterium atoms to yield a ^{4}He atom, for that reaction involved only gamma ray emission, but no neutrons. The difficulty is that the neutron free reaction is weaker by a ratio of 10^7 in cross-section than the other reactions. Somehow the novel environment of the cold fusion experiment would need to bring about a great enhancement of this reaction. This, the Advisory Board, found to be a fatal problem (ERAB, 1989, Sect. B.2):

> We know of no way whereby the atomic or chemical environment can effect such an enhancement, as this ratio is set by nuclear phenomena and is on a length scale some 10^4 times smaller than the atomic scale.

The point is mildly stated, but the idea is powerful. Fusion reactions involving deuterium had been well researched and well understood. Proponents of cold fusion had to argue that this established theory fails for some as yet unknown reason when the fusion reaction occurs within a palladium electrode. Effects of this type were otherwise unknown and implausible because fusion requires a closeness of approach of the deuterium atoms at which scales the palladium atoms are distant spectators.

Sturms (2007, p. 13) took a different view:[3]

> If theory and observation are in conflict, theory wins [in the skeptics view]. In this case, the absence of neutrons proved that the effect does not occur even when tritium and extra heat are measured, because theory requires neutrons be produced. In their minds, the extra heat must be a measurement error and the tritium must be contamination. Evidence to the contrary was simply ignored. This is how faith-based science operates, but not the kind of science we are taught to respect. On the other hand, reality-based science acknowledges what nature reveals and then attempts to find an explanation. Rejection occurs only if a satisfactory explanation cannot be demonstrated. This demonstration is still in progress for cold fusion.

In sum, the real basis of the varying appraisals of cold fusion lay in inductive inferences grounded by background facts of class A. These facts specified the conditions under which cold fusion would manifest experimentally. In the establishment view, these facts called for rates of neutron and other fusion production not reported in the experiments; and, in addition, these facts denied that deuterium saturated electrodes could bring the deuterium atoms close enough to ignite fusion in the first place. Hence these facts warranted the inference that the experiments had failed. The dissidents, however, were willing to conjecture looser background theories, including some undeveloped or even unknown theories that would warrant the inference from the experimental results to cold fusion. Both deferred to the idea of reproducibility. Yet, with the same record of experiment, they came to different conclusions.

My proposal is that they are not calling upon a universal principle of reproducibility that resides within some abstracted, logic of induction. Rather, the idea of reproducibility is merely a gloss on inferences that are quite specific to the case at hand and dependent essentially on

---

[3] I have not found an establishment response to this argument, but it is not too hard to imagine its content: the establishment view is not rejecting evidence, but considering a larger class that includes the experiments and observations in other arenas that support the standard theory of fusion reactions.

background assumptions. It is exactly because the two groups differ in their background assumptions that they can come to judge different inferences warranted.[4]

5. The Miller Experiment: Failed Replication with no Inductive Import

How are we to deal with a case in which there are multiple successful replications of an experiment, but a prominent, well-executed failure? Understood as a formal principle, reproducibility gives us no real guidance. It cannot authorize us simply to dismiss the one failure of replication as inductively inert. Or at least it cannot do so without extensive elaboration on just what conditions distinguish those cases in which the failure carries import and those in which it does not. Such elaborations are not at hand and not likely to be forthcoming.

A material analysis of cases like this, however, faces no such general problems. For approached materially, there is no universal principle implemented. There are only particular cases, each of which is ultimately to be analyzed individually.

Here is a celebrated example. Nineteenth century electrodynamics had given center stage to the ether, the medium that carries light and electric and magnetic fields. It surrounds the earth and the earth's motion through it creates currents that blow past us, much as a car's motion creates a headwind. Famously, the Michelson-Morley experiment of 1887 had failed to detect this ether wind. The experiment employed an extremely sensitive interferometer that split a light beam into two folded pathways and then recombined the beams. The results were read from changes in the interference patterns formed by the recombined beams as the interferometer was slowly rotated. While its importance in Einstein's pathway to special relativity remains debated (see Norton, forthcoming), the null result of the experiment is foundational for special relativity. Had this experiment detected an ether wind or ether drift, it would have detected the absolute motion of the earth, in contradiction with the principle of relativity.

---

[4] According to the material theory, that does not mean that both inferences are sound. The situation is little different from the corresponding case of deductive logic. If two scientists employ the same premises but different deductive schema to arrive at contradictory conclusions, at least one of the schema is a fallacy. Correspondingly, if two scientists arrive at differing conclusions by inductive inference, at least one has a false warranting fact presumed.

On December 29, 1925, Dayton C. Miller (1926), addressed the American Physical Society in Kansas City. He recounted his own efforts to replicate the Michelson-Morley experiment, and reported the results of his latest efforts of 1925, when his apparatus was set up on Mount Wilson near the Observatory in California. He had found a positive result of 10 km/sec for the ether drift. It was less than the 30 km/sec or so that might otherwise be expected from the motion of the earth. Yet it was not a null result. This replication of the Michelson-Morley experiment had failed.

This was not a failure to be taken lightly. Now, over a hundred years after the discovery of special relativity, we classify experiments challenging special relativity with circle squaring and perpetual motion machines. That dismissal was not so easy in 1926, especially in light of who Dayton C. Miller was. He was then the President of the American Physical Society; and he was employed by the Case School of Science, in Cleveland, the site of the famous Michelson-Morley experiment of 1887. His experiments had a venerable lineage. In 1902 to 1904, he had collaborated on ether drift experiments with Michelson's original collaborator, Edward Morley. They had reused parts of the apparatus of the original 1887 experiment. These parts included the iron trough that held the mercury in which the interferometer floated and the original circular wooden float. These parts, Miller (1933, p. 209), noted with some pride of ownership in his later review, "have been continued in use by the writer to the present time."

While there were other ether drift experiments from the time, Miller's used one of the longest folded pathways for light, which would give his one of the greatest sensitivities.[5] The experiments of 1926 built on the experience with Miller's earlier collaboration with Morley and successive refinements of the apparatus and experimental design through multiple experiments in a new series starting in 1921. It was feared, for example, that a basement in Cleveland, a mere 300 feet above the level of Lake Erie, may be too shielded from the ether current. For this reason, the entire apparatus was relocated to a mountainside next to the Mount Wilson Observatory, at an elevation of about six thousand feet. Miller's (1926, 1933) recounts the elaborate cautions undertaken to avoid and control all imaginable sources of error.

---

[5] For a compendium of other ether drift experiments from that time, see Miller, 1933, pp. 239-40 and Shankland et al., 1955, p. 168.

The report of Miller's positive result produced great interest in both scientific and popular circles. Miller was even awarded a $1000 prize by the American Association for the Advancement for Science for a related article. Einstein soon succumbed to popular pressure to respond. He wrote a short note for the popular press, published January 26, 1926, in the *Vossische Zeitung*, a well-known liberal newspaper in Berlin.[6] His remarks included:

> There is, however, in my opinion *practically no likelihood* that Mr. Miller is right. [Einstein's emphasis]. His results are irregular and point rather to an undiscovered source of error than to a systematic effect. Furthermore, Miller's results are in and of themselves hardly credible, because they assume a strong dependence of the velocity of light upon the height above sea level. Finally a German physicist (Tomaschek) recently performed an electrical experiment also at a considerable height above the sea (the Trouton-Noble experiment), the result of which speaks against Miller's results insofar as it supports the absence of an "ether wind" at great altitudes.

From our perspective, what is notable about Einstein's response is that it invokes no matters of general inductive principle. Had Miller's claims somehow contravened an identifiable, universal inductive principle, it would have been easy for Einstein merely to point that out, much as one might identify a deductive fallacy. Rather, Einstein proceeds precisely as one would expect from the material theory. He gets the sharpest image of the inductive import of Miller's work by looking most narrowly at it.

Einstein's critique draws on facts in classes A and B above. For example, he complained that Miller's results are "irregular." Einstein did not elaborate, but, presumably, his concerns are similar to those expressed by Hans Thirring later in a June 1926 communication to *Nature*. In explaining his complete disagreement with Miller's interpretation of the experimental results, Thirring (1926) noted several irregularities within Miller's data. Since the ether wind will come from one direction in space, the direction detected by the interferometer should rotate through all

---

[6] This article was found by Klaus Hetschel (1992) from whose paper the translation of the text is drawn. See Hetschel (1992) for more details of the scientific and popular reaction to Miller's experiments.

points of the compass in the course of a day, as the daily rotation of the earth rotates the apparatus once per day in space. Yet, Thirring (p. 82) found:

> …an effect pointing towards the north-west quadrant of the compass in about ninety-five per cent. of all observations. This fact seems to be fatal to the assumption of an ether drift of constant direction towards a certain point of the heavens…

The facts at issue here are those in class A, which specify the conditions under which the process of interest manifests an experimental outcome. Under the supposition of an ether theory, the process of interest, the earth's motion through the ether, would manifest as an ether wind of a definite direction in space. That was not found, so that these background facts could not license the inference from the experimental outcome to the ether current.

Einstein then conjectured an "an undiscovered source of error." He did not specify what this source may be. However Einstein was quite direct in his private notes to correspondents. He wrote to his friend and confidant, Michele Besso, on December 25, 1926:[7] "I think that the Miller experiments rest on an error in temperature. I have not taken them seriously for a minute." He pressed this concern in a subsequent correspondence with Miller later in 1926, with Miller dismissing it by describing the elaborate corrections put it place to control temperature effects.[8] Einstein's doubts may have had a firmer foundation than the brevity of his *Vossische Zeitung* remarks suggest, for he had long taken a keen interest in Miller's experiment. During Einstein's 1921 visit to the US, he had taken the trouble visit Miller and, on Miller's report, had spent over an hour and a half discussing the ether drift experiments.[9] Einstein's suspicions were affirmed when Shankland et al. (1955) later performed a painstaking re-analysis of Miller's results, finding that positive results were associated with temperature variations in apparatus.

This second set of inferences drew on facts in class B. Einstein and Shankland and his colleagues had a sense of the processes that could produce a confounding result and, as

---

[7] As quoted in Holton (1969, pp. 185-86).

[8] For details, see Hentschel (1992, p. 608). Einstein noted that temperature changes of as little as 1/10th degree in the air of the light path would be sufficient to generate results of the magnitude of Miller's.

[9] As affirmed by a letter of Miller's quoted in Holton (1969, p. 186).

Shankland and his colleagues affirmed, the pattern of results, in conjunction with these facts supported the conclusion of the thermal original of Miller's results.

6. Intercessionary Prayer: Successful Replication with No Inductive Import

The converse case is also possible: the successful replication of experiments, yet those successes are nonetheless regarded as inductively inert. Once again no formal account of reproducibility of experiment can accommodate this unless it specifies the conditions under which successful replication does and does not have inductive import. Approached materially, each case is treated individually and we face no insurmountable problems of general principle.

In intercessionary prayer, one entreats a deity or supernatural power to intervene in mundane affairs. The entreaty is most commonly for well-being and health and especially the speedy recovery of the sick. In the nineteenth century, two leading scientists, John Tyndall and Francis Galton, proposed that the efficacy of prayer could be assessed by objective tests of the type routinely employed in science.[10] If the sick do indeed fare better when they are prayed for, that good effect out to be discernible through simple statistical analysis. They were skeptical. Galton had been collecting data for what amounted to a rather fragile retrospective study. He displayed a table of the mean lifetimes of males who survived past 30 years. Recalling that sovereigns in every state are the subjects of public prayer, such as "Grant her in health long to live," he observed of his table (Galton, 1872):

> The sovereigns are literally the shortest lived of all who have the advantage of affluence. The prayer has therefore no efficacy, unless the very questionable hypothesis be raised, that the conditions of royal life may naturally be yet more fatal, and that their influence is partly, though incompletely, neutralized by the effects of public prayers.:

The proposal, as one might expect, evoked derision from theological circles. James M'Cosh (1872, pp. 777-778) retorted

[10] For a brief history, see Brush (1974).

> We laugh at Rousseau's method of settling the question of the existence of God: he was to pray and then throw a stone at a tree, and decide in the affirmative or negative, according as it did or did not strike the object. The experiment projected by Professor Tyndall's friend is scarcely less irrational.

The mood had changed by the later twentieth century. Controlled studies of intecessionary prayer were conducted and continue to be conducted. Randolph Byrd (1988), for example, reported a prospective randomized double-blind trial of the effects of intercessionary prayer on the recovery of patients in a coronary care unit. He reported statistically significant improvements in recovery among those in the test group receiving prayer. Harris et al. (1999) performed a similar study on cardiac patients, again finding prayer to be associated with improvements in recovery. While not all experimental tests of intercessionary prayer have produced positive results, there are sufficient for meta-level surveys to be written. Astin et al. (2000) report the two studies above as the only ones producing positive results among the five surveyed. However, in the broader category of "distant healing," 57% of the studies reported positive results, which supported the final conclusion that the field "merits further study." A later review (Roberts et al., 2009)[11] was less optimistic. They found the results among the ten trials surveyed to be equivocal and recommended against further investigation.

Most of these reports are of little use in our efforts to understand what grounds inductive inference in relation to the reproducibility of experiment. Both surveys grapple awkwardly with the problem of some successful and some failed replication and, from them, arrive abruptly at a synoptic judgment. We are given little insight into how the analysts balanced the competing inductive import of the successes and failure to arise.

There is a subgroup, however, who make clear that they regard successful replication of the intercessionary prayer experiments inductively inert, for they do not believe that these studies have any inductive powers at all. Their analysis conforms with the material approach to reproducibility. For successful replication requires the facts in classes A and B above to be

---

[11] Curiously, this report included positive results from the spoof Leibovici (2001) study. It also noted a later critic who pointed out their error, but nonetheless did not disavow the study, concluding: "The Leibovici 2001 was not in jest. It is a rather serious paper, intended as a challenge." (pp. 56-57).

hospitable. This skeptical group does not find facts in class A supporting an inference from the experimental outcome to the supernatural intervention proposed. Hence replication adds nothing to an outcome that was already inductively inert.

Needless to say, this group includes atheist polemicists like Richard Dawkins. He remarks in his *God Delusion* (p.86) that "the very idea of doing such experiments is open to a generous measure of ridicule…" Theists also have traditionally been skeptical of such experiments. Their analyses can be more measured and thus prove more illuminating. The three authors of Chibnall et al. (2001), a Catholic, a Protestant and a Jew, describe how they set out to perform an experimental test of distant prayer. They "became convinced that the very idea of testing distant prayer scientifically was fundamentally unsound." In a telling, detailed analysis, they argue powerfully that, in effects, the requisite facts of the class A do not obtain: we have no good reason to expect the effect or process of interest (supernatural intervention) to be manifested in the experimental outcome (statistics of recovery rates among patients). They ask:

> If prayer is a metaphysical concept linked to a supernatural being or force, why would its efficacy vary according to parameters such as frequency, duration, type, or form? The very concept of prayer exists only in the context of human intercourse with the transcendent, not in nature. The epistemology that governs prayer (and all matters of faith) is separate from that which governs nature. Why, then, attempt to explicate it as if it were a controllable, natural phenomenon?
>
> … there is no reasonable theoretical construct to which to link prayer because of, we would argue, its very nature. No model guides our understanding of intercessory prayer as a treatment in the way we know that drug pharmacokinetics, type, dose, schedule, interactions, and treatment length are critical to an antibiotic as a treatment. In fact, we believe no scientific model can guide it.

Perhaps one of the most revealing of all the intercessionary prayer studies was reported in the December 2001 issue of the *British Medical Journal*. Leibovici (2001) collected all reports of patients who were detected with blood infections in a university hospital in Israel (Rabin Medical Center, Beilinson Campus) in 1990-1996. In 2000, he randomized the cases and arranged for prayer for a test group. The results show no improvement in mortality among the test group but a

statistically significant shortening of both hospital stay and fever duration. The results were "retrospective" in the sense that these outcomes had already happened at the time the prayers were administered. It was suggested that we should not assume that "God is limited by a linear time, as we are."[12]

This peculiar report produced the uproar one might expect. Letters to the editor in the April 27, 2002, issue of the *British Medical Journal* covered a wide range of complaints; and it was at times hard to tell if they were written in the same spirit at the original article. They included a defense of the laws of physics against breakage and protests over the ethics of experimenting on subjects whose consent could no longer be secured at the time of the experiment. The letters were concluded with an "Author's Reply," in which Leibovici admitted that the paper was really a spoof, but with a deeper purpose:[13]

> The purpose of the article was to ask the following question: Would you believe in a study that looks methodologically correct but tests something that is completely out of people's frame (or model) of the physical world—for example, retroactive intervention or badly distilled water for asthma?

Of three possible answers, Leibovici endorsed the third:

> To deny from the beginning that empirical methods can be applied to questions that are completely outside the scientific model of the world. Or in a more formal way, if the pre-trial probability is infinitesimally low, the results of the trial will not really change it, and the trial should not be performed. This, to my mind, turns the article into a non-study, although the details provided in the

---

[12] I learned of this bizarre paper from a talk by John Worrall.

[13] Fact can be stranger that fiction. Over a year after the scam was admitted, Olshanky and Dossey (2003) published a note in the same journal that dismissed Leibovici's disavowal. In a narrative laden with pleas for open-minds, Einstein, Stephen Hawking, quantum mechanics, string theory and consciousness, they urged that we should subject these non-local, anomalous effects to serious study. This paper gives me great confidence in humanities' ability to turn every stone, for clearly no idea, no matter how absurd, lacks proponents.

publication (randomization done only once, statement of a wish, analysis, etc) are correct.

Leibovici's assessment expresses in miniature why a formal account of controlled trials fails, where a material account succeeds. He notes that one can have a trial that meets all the requisite formal conditions. That was how he set up the study his article. Nonetheless the study has no inductive import. This situation is inexplicable if one adheres to a general, formal account of the reproducibility of experiment. The material approach faces no such problems. In it, the trial can have inductive import only if requisite background facts are hospitable. This, Leibovici asserts, is not the case here.

References

Astin, John A. et al. (2000), "The Efficacy of 'Distant Healing': A Systematic Review of Randomized Trials," *Annals of Internal Medicine*, **132**, pp. 903-910.

Atwood, Kimball C. (2004), "Bacteria, Ulcers, and Ostracism? H. Pylori and the Making of a Myth," *Skeptical Inquirer*, **28.6** (November/December 20004).

Brush, Stephen G. (1974), "The Prayer Test," American Scientist, **5**, pp. 561-563; **63** (1975), pp. 6-7.

Byrd, Randolph C. (1988), "Positive Therapeutic Effects of Intercessory Prayer in a Coronary Care Unit Population," *Southern Medical Journal*, **81**, pp. 826-29.

Chibnall, John T. et al. (2001) "Experiments on Distant Intercessory Prayer: God, Science, and the Lesson of Massah," *Archives of Internal Medicine*, **161**, pp. 2529-36.

Dawkins, Richard (2008), *The God Delusion*. Boston: Mariner.

DOE (2004), *Report of the Review of Low Energy Nuclear Reactions*. Downloaded as http://newenergytimes.com/v2/government/DOE2004/DOE-CF-Final-120104.pdf

ERAB (1989), Cold Fusion Research, November 1989: A Report of the Energy Research Advisory Board to the United States Department of Energy. Washington, DC. DOE/S-0073 DE90 005611

Galton, Francis (1872), "Statistical Inquiries into the Efficacy of Prayer" *Fortnightly Review*, **12**, pp. 125-35.

Harris, William S. et al. (1999), "A Randomized, Controlled Trial of the Effects of Remote, Intercessory Prayer on Outcomes in Patients Admitted to the Coronary Care Unit," *Archives of Internal Medicine*, **159**, pp. 2273-78.

Hentschel, Klaus (1992), "Einstein's Attitude Towards Experiments: Testing Relativity Theory 1907-1927," *Studies in History and Philosophy of Science,* **23**, pp. 593-624.

Holton, Gerald (1969) "Einstein, Michelson, and the "Crucial" Experiment," *Isis*, **60**, pp. 132-197.

Leibovici, Leonard (2001), "Effects Of Remote, Retroactive Intercessory Prayer On Outcomes In Patients With Bloodstream Infection: Randomised Controlled Trial," *British Medical Journal*, **323**, No. 7327 (Dec. 22 - 29, 2001), pp. 1450-1451.

M'Cosh, James (1872), "On Prayer. III" *Contemporary Review*, **20**, pp. 777-782

Marshall, Barry and Warren, J. Robin (1983), "Unidentified curved bacilli on gastric epithelium in active chronic gastritis" *Lancet* 1(8336) (June 4), pp. 1273—1275.

Marshall, Barry and Warren, J. Robin (1984), "Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration," *Lancet* 1 (8390)pp. 1311–15

Miller, Dayton C. (1926) "Significance of the Ether-Drift Experiments of 1925 at Mount Wilson," *Science*, **63**, pp. 433-443.

Miller, Dayton C. (1933), "The Ether-Drift Experiment and the Determination of the Absolute Motion of the Earth," *Reviews of Modern Physics*, **5**, pp. 203-42.

Norton, John D. (forthcoming) "Einstein's Special Theory of Relativity and the Problems in the Electrodynamics of Moving Bodies that Led him to it." in *Cambridge Companion to Einstein*, M. Janssen and C. Lehner, eds., Cambridge University Press.

Olshansky, Brian and Dossey, Larry (2003) "Retroactive Prayer: A Preposterous Hypothesis," *British Medical Journal*, **327**, pp. 1465-1468.

Peat, F. David (1989), *Cold Fusion: The Making of a Scientific Controversy*. Chicago: Contemporary Books.

Roberts, Leanne, et al. (2009), "Intercessory prayer for the alleviation of ill health (Review)," *The Cochrane Collaboration* in *The Cochrane Library*, 2009, Issue 3, John Wiley & SonsSoddy, Frederick (1907) "Radioactivity," pp. 311-343 in *Annual Reports of the Progress in Chemistry. 1906."* Vol. III. London: Guerney and Jackson.

Shankland, R. S. et al. (1955), "New Analysis of the Interferometer Observations of Dayton C. Miller," *Reviews of Modern Physics*, **27**, pp. 167-78.

Sturms, Edmund (2007), *The Science of Low Energy Nuclear Reaction: A Comprehensive Compilation of Evidence and Explanations about Cold Fusion*. Singapore: World Scientific Publishing.

Thirring, Hans (1926), "Prof. Miller's Ether Drift Experiments," *Nature*, **118**(No. 2595), pp. 81-82.

Warren, J. Robin (2005) "Helicobacter—The Ease and Difficulty of a New Discovery," Nobel Lecture, December 8, 2005. http://nobelprize.org/nobel_prizes/medicine/laureates/2005/warren-lecture.pdf