Воспроизведение эксперимента
Часть II*

© 2021 г. Рональд Лаймон[1**], Аллан Франклин[2***]

[1] Государственный университет Огайо, философский факультет,
281 W Lane Ave, Columbus, OH 43210, USA.

[2] Университет Колорадо, физический факультет,
390 UCB University of Colorado, Boulder, CO 80309-0390, USA.

[**] E-mail: laymon.1@osu.edu

[***] E-mail: allan.franklin@colorado.edu

Поступила 20.05.2021

Существует немалое число дискуссий и исследований о природе и масштабах того явления, которое многие считают кризисом воспроизводимости в психологии и других социальных науках, а также (возможно, в меньшей степени) в медицинских науках. Наш подход к природе и значению воспроизводимости основан на той идее, что главная цель воспроизведения – выявить и идентифицировать совместно действующие причины, иными словами, уменьшить систематическую неопределенность. Это ведет нас к пониманию воспроизведения в более широком, чем обычно, смысле. Мы подробно разработали этот подход в трех недавних книгах, которые включают как абстрактный анализ, так и различные конкретные исследования, в основном из области физики, но не только из нее. Мы, например, рассмотрели сложность принятия решения о том, было ли воспроизведение опыта успешным или неудачным, о роли нулевых экспериментов и эпизодов, в которых одного эксперимента было достаточно для решения или дальнейшего исследования проблемы. В этой статье мы рассмотрим и обобщим наш подход и его результаты.

Ключевые слова: воспроизведение, эксперимент, физика, психология, социальные науки.

DOI: 10.21146/0042-8744-2021-9-118-131

Цитирование: *Лаймон Р., Франклин А.* Воспроизведение эксперимента. Часть II // Вопросы философии. 2021. № 9. С. 118–131 (на английском языке).

[*] Настоящая стать в двух частях представляет собой изложение трех книг на темы воспроизведения; см. подробнее: [Franklin 2018; Franklin, Laymon 2019; Franklin, Laymon 2020].

Replication
Part II*

© 2021 Ronald Laymon[1**], Allan Franklin[2***]

[1] *The Ohio State University Department of Philosophy,*
281 W Lane Ave, Columbus, OH 43210, USA.

[2] *University of Colorado Department of Physics,*
390 UCB University of Colorado, Boulder, CO 80309-0390, USA.

[**] *E-mail: laymon.1@osu.edu*

[***] *E-mail: allan.franklin@colorado.edu*

Received 20.05.2021

There has been considerable debate and analysis about the nature and extent of what many believe to be a replication crisis in psychology and other social sciences. And perhaps to a lesser degree in the medical sciences. Our approach to the nature and value of replication has been based on the idea that the overriding purpose of replication is to ferret out and identify confounding causes. In other words, to reduce systematic uncertainty. This has led us to understand replication in a broader sense than ordinarily understood. We have developed this approach in considerable detail in three recent books which include both abstract analysis and many case studies drawn primarily, but not exclusively, from physics. We have, for example, considered the difficulty of deciding whether a replication has been successful or has failed, the roles of null experiments, and episodes in which a single experiment has been sufficient to decide, or to further investigate, an issue. In this two-part essay we will review and summarize our approach and results.

Keywords: replication, experiment, physics, psychology, social sciences.

DOI: 10.21146/0042-8744-2021-9-118-131

Citation: Laymon, Ronald, Franklin, Allan (2021) "Replication. Part II", *Voprosy Filosofii*, Vol. 9 (2021), pp. 118–131.

A. Introduction

In Part I of this essay [Laymon, Franklin 2021], we gave a brief account of our views about replication, and considered what constitutes replication, and what is its value for scientific practice. We questioned whether it really is the "gold standard" that many have claimed, and how expansively the concept should be understood. In short, our answers were: It's not *the* gold standard; and in any case it should be construed broadly. Most importantly our claim is that in considering replication as a normative requirement it is essential to keep in mind that its overriding purpose is to ferret out and identify confounding causes. In other words, to reduce systematic uncertainty.

While the initial motivation for our interest in replication comes from what many believe to be a replication crisis in psychology and other social sciences, we have decided here to restrict our attention to the role replication plays in the physical sciences. Thus, in Part I we considered in some detail the role played by replication in the near simultaneous discovery of the Higgs boson by two independent research groups; and (2) the series of unsuccessful

* This two-part essay is a brief summary of three books we have written on the subject of replication. For more details see [Franklin 2018; Franklin, Laymon 2019; Franklin, Laymon 2020].

replications of claimed values for the Universal Gravitational constant. Here in Part II our focus will be on replication with respect to (1) the role played by *null experiments* in physics, and (2) whether there have been cases where *once can be enough* in the sense that replication was not required either historically or as a matter of good scientific practice. As in Part I, we will have much to say in our Part II case studies about the *credibility* of claims to have explained away or otherwise rendered harmless such residual variations. And again, as in Part I, we'll make application of what we referred to there as an *epistemology of experiment*.

B. Null Experiments

The importance of null experiments in physics derives from what is often their central role in the development of theory and their associated deep connections at a foundational level. We'll focus on why frequent replication of such experiments occurs and what the conditions are for meaningful replication.

As an initial matter we'll need to be clear 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 the sought-after phenomenon or effect is 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's "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 in which case a null result is zero within the experimental uncertainty.

Returning to the social sciences, it must be acknowledged that there is not a straightforward correspondence between what we have described as the 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, e.g., [Anderson, Bahnik et al. 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. speak instead speak of *positive* and *negative* results, where 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[1]. 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. Insofar, however, that a published or otherwise known null result carries with it a measure of its systematic uncertainty, replications that seek to narrow that uncertainty may depending on the circumstances take on considerable importance – as will be shown by the case of the Michelson – Morley experiment to which we now turn.

C. The Michelson – Morley Experiment

Our story begins in the late 19th century when the wave theory of light was generally accepted by the physics community. Such a theory required a medium in which the waves could propagate. That medium was what became known as the luminiferous ether.

The question arose as to how one might measure the relative velocity of the earth and the theoretically required ether. James Clerk Maxwell made the first suggestion [Maxwell 1878][2]. Assuming that light propagated in the ether with the same velocity in all directions. Maxwell realized that all feasible methods of measuring the relative velocity of the earth relative to the ether would require a roundtrip for the light and thus, that first order effects would almost completely cancel. There would, however, be a second order effect, proportional to v^2/c^2, where v is the velocity of the earth relative to the ether and c is the speed of light. For the earth's 30 km/sec orbital velocity: "The change in the time of transmission of the light on account of a relative velocity of the aether equal to that of the earth in its orbit would be only one hundred-millionth part of the whole time of transmission, and would therefore be quite insensible" [ibid.]. That was the problem that Albert Michelson would solve.

The Michelson – Morley experiment was designed to measure the velocity of the earth relative to the luminiferous ether, the presumed medium through which light and electromagnetic radiation were propagated. Michelson's basic insight was that this measurement could be achieved by determining the *change in the interference pattern* produced by two light beams, one traveling parallel to the velocity of the earth relative to the ether, and the other perpendicular to that velocity, when the experimental apparatus was rotated through 90°. Assuming the velocity of the earth relative to the ether was the earth's orbital velocity of 30 km/sec, the predicted effect for the apparatus employed was only four tenths of a fringe.

The design of the experimental apparatus can be described as follows. Light from the source strikes a half-silvered mirror, with half the light being reflected to mirror perpendicular to the initial light beam, and the other half transmitted to a mirror along the beam line. The light reflected from those mirrors again strike the half-silvered mirror, and some of the light is transmitted or reflected toward an observer, who will observe an interference pattern. If the earth is moving relative to the ether this pattern should change when the apparatus is rotated through 90°.

The derivation of the change in the interference pattern is as follows[3]. The lengths of the two arms were made approximately equal so that white light could be used which would create a clear central fringe that could serve as a zero point and would thus facilitate readjustment. Let D be the length of each of the arms, and c be the velocity of light relative to the ether and v ~~is~~ the velocity of the earth relative to the ether, or the velocity of the ether relative to the earth. The time, T_1, for light to travel back and forth along the beam line is

$$T_1 = D/(c - v) + D/(c + v) = 2Dc/(c^2 - v^2)$$

The distance d_1 traveled by the light in this time is

$$d_1 = 2Dc^2/(c^2 - v^2) \approx 2D(1 + v^2/c^2) \text{ to first order in } v^2/c^2.$$

For the perpendicular path, remembering that the mirror is moving while the light is traveling between the mirrors, the distance d_2 traveled by the light is

$$d_2 = 2d\sqrt{(1 + v^2/c^2)} \approx 2D(1 + v^2/2c^2)$$

The path difference $\Delta = d_1 - d_2 = Dv^2/c^2$

When the apparatus is rotated through 90° the difference $\Delta' = d_2 - d_1 = Dv^2/c^2$. Thus, one expects a total fringe shift of $2Dv^2/c^2$ [4].

The problem Michelson faced was to construct an apparatus sufficiently sensitive to detect the small effect expected, yet robust enough to be immune to background effects caused by vibration or by the rotation of the apparatus, which might mimic or mask the predicted effect. Michelson faced several difficulties, particularly the effects of temperature and of the mechanical stress caused by rotation of the instrument. "The principal difficulty, which was to be feared in making these experiments, was that arising from changes of temperature of the two arms of the instrument. These being of brass whose coefficient of expansion is 0.00019 and having a length of about 1000 mm. or 1 700 000 wave-lengths, if one arm should have a temperature only one-hundredth of a degree higher than the other, the fringes would thereby experience a displacement three times as great as that which would result from the rotation. On the other hand, since the changes of temperature are independent of the direction of the arms, if these changes were not too great their effect could be eliminated" [Michelson 1881, 125]. Michelson covered the arms with long paper boxes to guard against such changes in temperature.

Michelson's first performed the experiment in 1881 where the final results were –0.004 fringe units for the N-S, E-W directions and –0.015 for the NE-SW directions. According to Michelson these were "simply error of experiment", and thus, were no more than a measure of systematic uncertainty, and that moreover were dramatically smaller than the anticipated result. Thus, Michelson concluded: "The interpretation of these results is that there is no displacement of the interference bands. The result of the hypothesis of a stationary ether is thus shown to be incorrect, and the necessary conclusion follows that the hypothesis is erroneous" [ibid., 128][5].

Michelson was encouraged to continue his work by both Lorentz and by Rayleigh. In a letter to Rayleigh, Michelson remarked: "I have never been fully satisfied with my Potsdam experiment [1881], even taking into account the correction which H.A. Lorentz points out… I have repeatedly tried to interest my scientific friends in this experiment without avail, and the reason for my never publishing the correction was (I am ashamed to confess it) that I was discouraged at the slight attention the work received and did not think it worthwhile" (Michelson to Rayleigh, March 6, 1887, full text in [Shankland 1964, 29]).

Michelson and Morley prepared a more robust version of the experimental apparatus and reported their results in 1887 [Michelson, Morley 1887]. In this report Michelson and Morley drew attention to their earlier error that Lorentz had noted, the omission of the motion of the apparatus relative to the ether and pointed out that once this error was corrected the predicted effect was reduced by a factor of two and that Michelson's 1881 conclusion might well be questioned.

They then went on to derive the size of the effect expected. They also noted the problems of vibration and rotation that had plagued their 1881 experiment "were entirely overcome by mounting the apparatus on a massive stone floating on mercury". And that the size of the expected fringe shift could be enlarged "by increasing, by repeated reflection, the path of the light to ten times its former value [ibid., 336–337]".

Michelson and Morley fitted a straight line to the data and took the differences between the data and the fitted line as their residuals. The result of this was that there was no obvious displacement of the fringes. Michelson and Morley conservatively concluded: "Considering the motion of the earth in its orbit only, this displacement should be $2Dv^2/c^2 = 2D \times 10^{-8}$. The distance D was about eleven meters, or 2×10^7 wavelengths of yellow light; hence the displacement was expected to be 0.4 fringe. The actual displacement was certainly less than the twentieth part of this, and probably less than the fortieth part. But since the displacement is proportional to the square of the velocity, the relative velocity of the earth and the ether is probably less than one sixth of the earth's orbital velocity, and certainly less than one fourth" [ibid., 340–341].

In other words, the velocity of the earth relative to the ether, if any, cannot be more than a small fraction of its orbital velocity. Michelson and Morley noted, however, that in their analysis: "…only the orbital motion of the earth is considered. If this is combined with the motion of the solar system, concerning which but little is known with certainty, the result would have to be modified; and it is just possible that the resultant velocity at the time of the observations was small though the chances are much against it. The experiments will therefore be repeated at intervals of three months, and thus all uncertainty will be avoided" [ibid., 341]"[6].

The Michelson – Morley experiment was repeated numerous times in the early twentieth century, including measurements made at higher altitude, even in a balloon. All of the replications gave results consistent with those of Michelson and Morley. The only notable exception was the work of Dayton Miller, discussed in the next section.

D. The Work of Dayton Miller[7]

By 1924 Miller had repeated the Michelson-Morey experiment, using increasingly refined apparatus and procedures, at various locations, including as a last stop the Mount Palomar observatory. For Miller this line of experimentation constituted what could naturally be described as a series of *successful replications* showing that "there has persisted a constant and consistent small effect which has not been explained" [Miller 1933, 222]. And for Miller the explanation of that *persistence* was not that the results were due to some continuing set of confounding causes (this being unlikely Miller believed because of the variations in apparatus and procedure) but rather that the results yielded a reliable measure of a real effect due the motion of the earth through the ether.

There was, however, another line of experimentation that focused attention on using smaller interferometers that could be better insulated from variations in temperature and would be more stable with respect to mechanical distortions. Using increasingly more refined interferometers and procedures, proponents of this approach obtained results that grew increasingly closer to zero in a way that corresponded to the incremental improvements in apparatus and procedure. Thus, [Kennedy 1926; Illingworth 1927; Joos 1934] could justifiably claim that they had obtained a serious of *successful replications* of a null result and that the increasingly smaller residual shift was just a measure of the remaining systematic uncertainty.

How was this conflict of competing replications to be adjudicated? Historically it seems to have been the case that Einstein's theory of special relativity along with the numerous successful replications of the Michelson – Morley result weighed the scales decisively in favor of the null result. Still, this leaves open the question of whether there was an experimental variation that by itself would tip the scales one way or the other – i.e., one that did not beg the question by appealing to a well-regarded and more fundamental theory.

The problem here, as Miller was quick to insist, was that the very thorough insulation of the interferometers of Kennedy and company also served to insulate their interferometers from the otherwise distorting effect of motion through the ether. In other words, that those experiments showed only that the elaborate insulation employed was successful in blocking the flow of the ether through the apparatus and thereby preventing the ether from having any significant differential effect as the interferometer was rotated.

So, the issue, at least from a purely experimental point of view, was at an apparent stalemate. It was at this point that Miller, in response to a suggestion by Gustaf Strömberg, one of the astronomers at Mount Wilson, realized that a determination of the phase shifts as his unshielded interferometer was made to rotate during four epochs (in Miller's case April, August, and September 1925, and February 1926) could be used to coherently locate the apex of the earth's motion on the celestial sphere. And from this, a determination of the ethereal velocity could be made. We note here that Miller's proposal was just a variation of a traditional problem in spherical astronomy that dated back at least to Herschel. Thus, there was in fact a potentially *crucial experiment* that could determine which of the competing series of replications, Miller's or those of Kennedy and company, got it right. Where, importantly, one did not need to beg the question by making an appeal to the Special Theory of Relativity.

Sad to say, at least from Miller's perspective, his results did not lead to a coherent determination of the apex of the earth's motion on the celestial sphere. And so, the issue could have been decided against Miller by relatively neutral experimental means. Moreover, not only were Miller's results incoherent but they were irredeemably so because there were no likely confounding causes that Miller could appeal to in order to save the day.

There remained, however, the question of what explains his earlier apparently positive results. The leading contender, of course, was the existence of variations in ambient temperature.

It was virtually impossible, however, to develop such an explanation in any detail because of the paucity of the necessary temperature data for his experiments. But there also was a more specific explanatory problem that was made evident by the following interaction between Miller and Joos. As noted by Joos: "…if, assuming a length of the light path of 30 m, one calculates what difference in temperature of the two branches of the interferometer produces a displacement of 1/10 of a fringe (this is the order of magnitude observed). One gets the astonishing result that a difference of 1/500° is sufficient. The mere warmth of the body of the observer who, in Mr. Miller's experiments, stands near the interferometer can produce such an effect" [Joos 1934, 114].

Miller made the reasonable response: "It should be borne in mind that the ether-drift observation… depends upon a regularly periodic variation in the position of the entire fringe system, and the period is twenty-five seconds throughout. The temperature would have to increase and decrease, with periodic regularity in each twenty-five seconds! to produce the results. Any irregular fluctuation will be eliminated in the long series of turns. [And because the] observer maintains a constant relation to the apparatus and if the warmth of the observer's body is effective, it would be a continual heating effect which produces a *continuous drift of the fringes, which is of no effect in the calculated results*" [Miller 1934, 114] (emphasis added).

In other words, the most direct explanation in terms of temperature variations yields the wrong periodicity. And, in fact, an acceptable and specific explanation of Miller's results was not obtained until Thomas Roberts in 2006 came up with a penetrating account, based on contemporary signal theory, of the biasing effect of the data analysis used by Miller. Still, speaking on Miller's behalf, Roberts offered this bit of mitigation: "Dayton Miller was a prisoner of his time. In the 1920s and 30s digital signal processing was unknown, and the serious flaws of the data reduction algorithm used by all such experiments went unnoticed. Also, the use of errorbars and quantitative error analyses were in their infancy. These aspects of the state of scientific knowledge combined to permit him to be fooled into thinking his interferometer measurements did indeed determine the "absolute motion of the earth". Even in 1955, Shankland et al. did not have knowledge of these aspects of Miller's analysis" [Roberts 2006, 7].

For a somewhat similar case, but with interesting variation, where there was no adequate explanation as to why a vanquished competitor had failed, we refer the reader to our discussion of what became known as the hypothesis of the existence of a "Fifth Force". Here the first two experiments performed, respectively by Peter Thieberger and the Eöt-Wash group, gave conflicting results as to the existence of such a force. The experiments used rather different apparatus and procedure, but both were, as far as could be determined, exemplars of well-conceived design and execution. When it came to replication, however, there was a consistent set of independent replications of the Eöt-Wash null result but none for Theiberger's positive result. But while this was held sufficient for the rejection of the Theiberger experiment, there was no real and specific explanation for why and how it had gone wrong. Sometimes there are limits on how far systematic uncertainty can be reduced[8].

E. The Discovery of Parity Nonconservation

In our book *Once Can Be Enough* we examined a number of experiments and their surrounding context where replication as a matter of historical fact was not required and even in retrospect is not required as a normative requirement in the sense of acceptable scientific practice. Here we'll briefly review one such experiment conducted by Chien-Shiung Wu and her collaborators [Wu et al. 1957]. What prompted the experiment was the proposal by Tsung-Dao Lee and Chen-Ning Yang that a long-standing problem, known as the "$\theta - \tau$ puzzle", would be solved if parity was not conserved in the weak interactions, a startling proposal [Lee, Yang 1956]. We can do no more here than give a very brief account of what parity conservation requires.

The laws of classical mechanics and electrodynamics are all invariant with respect to what is known as the parity operation which is defined as the reflection of all spatial coordinates through the origin, that is, where x is replaced by $-x$, y by $-y$, and z by $-z$. When it comes to quantum mechanics the operation is made to apply to the wave function whereby if $\psi(\boldsymbol{r}) = -\psi(-\boldsymbol{r})$ the wave function said to have odd or negative parity. If $\psi(\boldsymbol{r}) = \psi(-\boldsymbol{r})$ the parity is even or positive. Parity conservation requires then that in an interaction the parity of the final state be equal to that of the initial state. An elementary particle can have an intrinsic parity. For example, the pion has negative parity. In addition, if a quantum state has angular momentum l its parity is -1^l.

Historically, there was little if any doubt that parity conservation applied to the laws of quantum mechanics and indeed there was convincing experimental evidence that this was the case with respect to the strong interactions. But there wasn't such supporting evidence when it came to the weak interactions. And here matters were brought to a head by the $\theta - \tau$ puzzle which dealt with two of the K mesons, the θ and the τ. On one set of criteria, namely mass and lifetime, they appeared to be the same particle. On another set of criteria, spin and parity, they seemed to be different particles. More specifically, the puzzle arises with respect to a difference in decay modes where the θ^+ decays into two pions ($\theta^+ \rightarrow \pi^+\pi^0$) and the τ^+ into three pions ($\tau^+ \rightarrow \pi^+\pi^0\pi^0$ or $\tau^+ \rightarrow \pi^+\pi^+\pi^-$).

If parity conservation is assumed in weak interactions then these are different particles, but if it is assumed – as suggested by Lee and Yang – not to hold then θ^+ and τ^+ are merely two different decay modes of the same particle. And so, the puzzle dissolves. All well and good but what's the experimental evidence for the postulated nonconservation of parity? As noted by Lee and Yang, the simplest and most direct approach was to determine whether β decay from an oriented, i.e., polarized nucleus, was distributed in a way that did not satisfy the parity transformation requirement.

We spent more than a few pages in our book *Once Can Be Enough* explaining the theoretical basis for such an experiment as well as how those possibilities were actualized in the case of the Wu experiment [Wu et al. 1957][9]. In brief, the method of polarization was to prepare certain structurally advantageous salts containing Co^{60} nuclei that under significant cooling could be polarized by the imposition of a magnetic field which caused electrons to exit from the Co^{60} in an up or down direction roughly perpendicular to the plane of the magnetic field[10]. Thus, the degree of polarization could be detected by measuring the anisotropy of the succeeding electrons, i.e., the difference between the number going up and the number going down. But because of the apparatus employed, a "cryostat", only the number of electrons going up could be measured. To determine the number that would have gone down if the apparatus had so allowed, the magnetic field was reversed so that down now became up. The experiment convincingly revealed the requisite anisotropy for nonconservation and was universally agreed to be a great success.

In part this appraisal reflected the fact that Wu and her colleagues made a very thorough investigation regarding the systematic uncertainty involved. So, for example, the fact that the warm counting rates (i.e., for no polarization) were independent of the polarizing field direction argues against any significant instrumental asymmetry. Similarly, for the concordance between the β asymmetry and the associated gamma anisotropy, which measured the polarization of the nuclei, also persisted despite the change in polarizing field. It was also possible that the demagnetization field used to cool the sample might have left a remnant magnetization that caused the β-ray asymmetry. This confounding possibility was eliminated by noting that the observed asymmetry did not change sign with the reversal of the field direction. A last systematic uncertainty that had to be dealt with was that there might have been a small magnetic field perpendicular to the polarizing field due to the fact that the Co^{60} crystal axis was not parallel to the polarizing field. Eliminating this possibility and a related companion uncertainty involved the introduction of small additions and variations to the apparatus which would have the effect of magnifying the effect of variations from a parallel orientation.

There still remained, however, the problem that: "In order to evaluate [the asymmetry coefficient] α, accurately, many supplementary experiments must be carried out to determine the various correction factors. It is estimated here *only to show the large asymmetry effect*" [Wu et al. 1957, 1414], emphasis added.

In other words, because the estimated value of the lower limit of the asymmetry was so large, that approximation was more than enough to decisively demonstrate parity nonconservation. For unlike the case of a null experiment where increasingly greater accuracy is required, such excruciating precision was not required here. In sum, once was enough for the Wu experiment to have demonstrated the nonconservation of parity. No replication was needed[11].

F. When Once Came Close To Being Enough

While in the case of the Wu experiment once was enough, in the case of a 1928 experiment conducted by the American physicist Richard Cox, once was *almost but not quite enough* to have demonstrated the nonconservation of parity. Consequently, the resulting historical process was highly convoluted and contentious, with missed opportunities, and experiments left for dead only to be later resurrected. Add to this a gestation period of more than twenty-five years and what we have provides an *informative contrast* to the relatively quick historically made determination regarding the success of the Wu experiment[12].

In 1927, C.G. Darwin realized that "just as there are two independent polarized components in a wave of light, so there are two independent components in the wave of an electron" [Darwin 1927, 230]. Which meant that the spinning electron could serve as the *surrogate* for the electric and magnetic fields of ordinary light and thereby provide the basis for a coherent concept of electron polarization. Shortly thereafter the American physicist Richard Cox took up the challenge of developing an experiment that would demonstrate the existence of such polarization. And here he hit upon the idea of using Charles Barkla's earlier demonstration that X-rays could be polarized as the basis for an analogous test of electron polarization.

The apparatus devised and constructed by Cox and his colleagues consisted of a radium source that could be rotated with respect to two lead targets such that the electron beam would be perpendicularly reflected off the first target and then off the second target, after which the intensity of the resulting beam would be determined by a Geiger counter, which was replaced later by a more reliable and sensitive electroscope. The radium source was in turn successively rotated 90° such that electron counts were collected for each of the 0°, 90°, 180° and 270° orientations where the existence of polarization would be revealed in comparisons of the intensity asymmetries between the 0° and 180°orientations, and the 90° and 270° orientations [Cox et al. 1928].

After considerable refinement and secondary testing of the apparatus for the existence of confounding factors, Cox and his student Carl Chase were able to achieve sufficient control over systemic uncertainty so to convincingly demonstrate the existence of a polarization of around 4% when comparing the 90°/270° asymmetry. Their demonstration, however, of a similar 0°/180° asymmetry was admittedly less convincing [Chase 1930].

At this stage one might have thought that with respect to the experimentally determined 90°/270° asymmetry, *once was enough* to have shown that electrons could be polarized. But that was not to be because of Neville Francis Mott's contrary theoretical determination (based on the Dirac equation) that the "greatest asymmetry" was to be found at the 0°/180° positions and that "the scattering *is* symmetrical" at the 90°/270° positions [Mott 1929, 431]. This confrontation between experiment and theoretical prediction did, however, serve to encourage and justify further pursuit of relevant experiment and theory. But given the preponderance of the later reported null results for the 0°/180° asymmetry, such pursuit came to be restricted to the conflict of those null results with Mott's theory.

And here it must be remembered that the path taken by Mott from the Dirac equation to the tangible reality of electron polarization had to be mediated by a great many assumptions, idealizations, approximations, and massively complex computations. This meant that any conflict between Mott's theory and experiment could be explained away by placing the blame on the experiment or on one or more of the many steps that connected the Dirac equation with Mott's final predictions about the asymmetries. In the process Chase's positive result for the 90°/270° asymmetry ceased to be of interest. Once was deemed not enough to justify pursuit of that conflict with Mott's theory. The reason for this neglect is not entirely clear but we think it had to have been significant that Mott's argument for the null value here was considerably simpler and more direct than for the claimed positive 0°/180° asymmetry.

The nearly ten-year conflict between Mott's theory and the null experimental results for the 0°/180° asymmetry was finally resolved given Cox's discovery in 1940 of a hitherto unnoticed difference in the intensities in reflected and transmitted electron beams. And because transmission was significantly more efficient than reflection, this opened the door for the more efficient and convincing 1943 experiment by Shull, Chase and Meyers that confirmed the existence of the 0°/180° asymmetry as predicted by Mott [Shull, Chase et al. 1943].

But still lost in the shuffle were both Cox's and Chase's results for the 90°/270° asymmetry. So whatever justification there might have been in 1943 for further research on electron polarization did *not* extend to solving the mystery of the 90°/270° asymmetry. That case remained closed. After all, on this Mott's theory had predicted a null result and its prediction of a positive result for the 0°/180° asymmetry had been confirmed by Shull, Chase and Meyers.

The discovery of parity nonconservation in 1957 provided the theoretical background that led to the realization that a longitudinally polarized electron would by itself be an instance of parity violation. That was enough to justify the development of experimental methods to test whether β-rays were longitudinally polarized. And so, they were, as shown by A. De Shalit, H.J. Lipkin and their collaborators [De Shalit, Cuperman et al. 1957]. As a final historical irony, it was only realized several years later by L. Grodzins that this confirmation of longitudinal polarization provided the explanation for Cox and Chase's original 90°/270° asymmetry [Grodzins 1959]. This because Cox and Chase had incorrectly assumed that their radium source produced unpolarized electrons.

Had Cox and Chase realized in 1928 that β-rays were polarized and that transmission through the foil was significantly more efficient than reflection, and had they acted on that, then *once would have been enough both historically and justifiably so*. They did, however, come tantalizingly close. When looking for "some explanation" of the "wide divergence" among his data Cox brought up the possibility of "some asymmetry in the electron itself". And in this regard he considered: "The supposition that the beam of β-particles *undergoes a polarization* in passing through the mica windows, similar to the polarization of light in passing through a tourmaline crystal. *This effect was in fact looked for* carefully in an experiment auxiliary to the present investigation but *it was not found*" ([Cox et al. 1928, 548], emphasis added).

Unfortunately, Cox did not provide an account of the nature of this "auxiliary" experiment and so we can only speculate as to what it might have been. In any case, Cox's auxiliary experiment did not reveal that the β-rays were already polarized. More squarely in the category of a squandered opportunity is the fact that Cox and Chase could have run Mott's argument (for a null result at the 90°/270° orientations) backwards to conclude that β-rays were not unpolarized. But then again nobody thought to do this, not even Mott. Thus, while Cox and Chase got close, once wasn't really enough.

G. Conclusions[13]

In this two-part essay we have provided very brief summaries of our discussions of issues relevant to what some have called the "replication crisis" in psychology and other social sciences. Our more detailed discussions are contained in our three books: *Is It the Same Result? Replication in Physics* [Franklin 2018]; *Measuring Nothing Repeatedly: Null Experiments in Physics* [Franklin, Laymon 2019]; and *Once Can Be Enough: Decisive Experiments, No Replication Required* [Franklin, Laymon 2020]. We have restricted our discussion to physics, with an occasional excursion to genetics, because that is the science we know best. We have found important differences between physics and the social sciences and a few similarities. We argued that, contrary to popular belief, replication is not *the gold standard* for establishing the credibility of an experimental result, but rather only one of a set of strategies. As discussed below, one experiment, alone, can perform significant roles in science. Our discussions included the difficulty of deciding whether a replication has been successful or has failed. In contrast to the social sciences in which null or negative results are disfavored and are often not published or even submitted, we have shown that null experiments play an important role in physics. Several of these discussions are discussed in more detail below.

1. Null Experiments as Sequences of Improved Replications. Our paradigm example, of the Michelson-Morey experiment, provides a compelling illustration of the fact that the central feature of null experiments is the development of a *sequence of improved replications*. Michelson and later Miller made improvements with respect to increased length of path, suspension, and materials used. As against this sequence of improvements there was the alternative approach followed by Kennedy, Illingworth and Joos whereby attention was focused on using smaller interferometers that could be better insulated from variations in temperature and would be more stable with respect to mechanical distortions. And because there were these competing sequences of improved replications, there was a corresponding difference in how the parties viewed the difference between the measured fringe shift and a strictly "zero" value.

For Miller the *persistence* of a fringe shift indicated that the variation from a "zero" result yielded a reliable measure of a real effect due the motion of the earth through the ether. For Kennedy, Illingworth and Joos, the *ever-decreasing variation* from a "zero" result yielded a reliable measure of the decreasing systematic uncertainty. And as we have seen the stalemate was resolved by Miller's failure to coherently locate what would have been the apex of the solar system's motion through the ether. But note that while this looks like the resolution of a simple conflict of competing theoretical predictions, at a more nuanced level it comes down to a determination of how to interpret the residual variation from a "zero" measurement.

The resolution of the conflict between the experiments of Peter Theiberger and the Eöt-Wash group concerning the existence of a "Fifth Force", illustrates a different pattern because here it was the case that there existed a sequence of improved replications for the Eöt-Wash experiment but not for the Theiberger experiment. That decided the issue even though an explanation for Theiberger's false positive was never determined.

2. Experiments and Systematic Uncertainty. As shown by our case studies, both here and in our books, experiments, more often than not, include – as typically made clear in their published reports – what are best regarded as *built-in* attempts to deal with *systematic uncertainty*. And by so doing the experimenter wards off, at least in part, demands for replication because of the possibility of unaccounted for confounding causes. In short, the experiment has already taken account of the replications needed to at least minimize systematic uncertainty.

A clear example of what we have in mind is the Wu experiment with its very complete accounting of possible confounding causes and the determination of their extent by careful experimentation conducted on the apparatus itself. Similarly, one problematic aspect of Cox's original data was that although the average of the observed asymmetries was consistent with polarization there was variation in the direction of the asymmetries detected in the individual experimental runs. This suggested that there was a confounding cause at work that varied during the experimental runs. The response was to separately test the consistency of the Geiger counter used and then replace it with a more consistent electroscope.

3. Interaction of Theory and Experiment. Experimental investigations may be undertaken in response to a determinative theoretical prediction, in which case their purpose is to test that prediction, or they may have been undertaken in an effort to search for answers to questions that arise because of an absence or incompleteness of applicable theory. In the latter case they take on the character of being exploratory.

The case of the Cox and Chase experiments illustrates this exploratory aspect of experimentation. While a "theoretical prediction" was involved from the very start it was "theoretical" only in a very attenuated sense. Cox recognized that given de Broglie's wave proposal there was an *analogy* to be made between the spinning electron and the electric and magnetic fields of ordinary light – and moreover that this analogy could serve as the basis for an experimental test for the existence of electron polarization. But the requisite underlying theory was never developed by Cox and the experiments were accordingly *exploratory* in character.

Moreover, as we have seen, the true significance of the Cox and Chase experiments was not fully grasped until Lee and Yang proposed that parity was not conserved and the proposal was experimentally confirmed by the Wu experiment. And here we note that the proposal by Lee and Yang was not a theory derived prediction but rather a proposal that if accepted would resolve the $\theta - \tau$ puzzle.

4. Replication, Independent Testability, and the Epistemology of Experiment. That the results of a scientific experiment must be reproducible appears at first to be of singular axiomatic significance. But what sorts of "results" are to be replicated? It is certainly not a necessary condition that the raw data result of an experiment be reproduced in order to pass scientific muster. As amply shown by our historical examples, it is much better by far to replicate not the raw data but rather the higher-level experimental result by means of experimental apparatus and procedures designed specifically to deal with the uncertainties involved in the original experiment. To replicate a result in this indirect way without having to reproduce the original experiment obviously opens the door to a wide range of experimental possibilities. Once, however, the constraints on replication are relaxed this way, any requirement for replication merges into the desideratum that experimental results should be *independently testable*, i.e., testable in ways that rely on different underlying theoretical and operational suppositions. And by way of adding some specificity here, we suggest that the experimental procedures included in what we have described as an *epistemology of experiment* should be and have in fact been utilized to provide for such independent testability.

5. Once was Enough but for What Purpose? Now while it might be thought that even though replication in the narrow sense of an exact, or nearly so, repeat of an experiment is not a necessary condition for the acceptability of an experimental result, replication in the broad sense that we have endorsed is such a necessary condition. And in fact, replication if construed broadly merges into the truism that at bottom there should be independent testability of the essential assumptions of an experiment and the interpretation of its results.

But here we take a stand for the contrary view that while replication in a wide sense is desirable, it is not necessary because, as we have shown, in some cases *once is enough*. We do not mean to be claiming that once will always, no matter what, be understood as having been enough only that all foreseeable arguments for assuming otherwise have been exhausted. Of course, our contrary view must be tempered by the fact, as noted above, that good experiments typically include arguments that their own results do not need replications.

We've relied on the expression "once was enough" as a shorthand for our claim that there was little if any point in replicating an identified result of an experiment – where the result could be the raw data or that involving a higher level of theoretical involvement. But any judgment that once is enough must take into account the question of for what *purpose* is once enough. This because the reliability and range of an experimental result will constrain what purpose or purposes can be served by the result.

There is a useful distinction to be made – which we have only alluded to – between two sorts of purpose broadly conceived. First, that once is enough for accepting a result as well confirmed and thus unlikely to have that status easily changed. Second, that once is enough for accepting the result as worthy of further investigation and development even though it does not fall into the first category. In short, the distinction is between *acceptance* and *pursuit*.

So, for example, the Wu experiment was clearly a case where once was enough for acceptance. On the other hand, the Cox and Chase experiments were tantalizingly close for the purpose of *acceptance* and surely, one would have thought, for the purpose of *pursuit* (in terms of extensions and refinement) of the experiments. But as sketched out above, the actual history here was rather more complex where the question of what was worthy of pursuit had to take into account Mott's conflicting theoretical predictions and the predominance of null values where Mott had predicted a positive result.

In summary, we have suggested that while replication can be a good method for establishing the credibility of an experimental result, it is not a necessary requirement.

Notes

[1] 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.

[2] There were other theories of the ether. Stokes, for example, proposed a theory in which a layer of the ether was dragged along with the earth [Stokes 1846]. Thus, there would be no velocity of the earth relative to the ether.

[3] We follow the derivation given in [Michelson, Morley 1887]. We use c for the velocity of light, rather than Michelson's V. The derivation in [Michelson 1881] is, as pointed out by Lorentz, incorrect.

[4] Michel Janssen has pointed out that, "The derivation of the prediction for the experiment does make it clear that the standard treatment of the experiment contains some dubious assumptions. Looking at the stretched-out interferometer in the figure, one clearly sees that the stretched-out mirrors, even though they are at rest in the ether, do not reflect light according to the standard law of reflection from geometrical optics if the light waves are to travel along the arms of the interferometer as was assumed in the derivations of both equations" (private communication).

[5] Michelson added a comment noting that Stokes' ether drag theory was compatible with Michelson's null result.

[6] Michelson never performed such experiments, although Miller ultimately would.

[7] For more details of this complex episode see [Franklin, Laymon 2019], Chapter 8.

[8] For details see [Franklin, Laymon 2019], Chapter 5.

[9] See [Franklin, Laymon 2020, 133–156].

[10] The polarization could be measured by the anisotropy in the number of gamma rays emitted in the polar or equatorial directions.

[11] To be fair, there were two other experiments, published at the same time that also demonstreated parity nonconservation [Garwin, Lederman et al. 1957] and [Friedman, Telegdi 1957]. In our view the experiment of Wu and collaborators was the most convincing (see [Franklin, Laymon 2020], Chapter 7).

[12] This case is dealt with inconsiderably more detail in [Franklin, Laymon 2020, 103–131].

[13] For a more complete statement of these conclusions along with many more supporting examples see [Franklin, Laymon 2019, 11-1 – 11-12] and [Franklin, Laymon 2020, 171–181].

References

Anderson, Christopher J., Bahnik, Štěpán, et al. (2016) 'Response to Comment on "Estimating the reproducibilty of psychological science"', *Science*, Vol. 351, p. 1037-c.

Chase, Carl T. (1930) "The Scattering of Fast Electrons by Metals II", *Physical Review*, Vol. 36, pp. 1060–1065.

Cox, Richard T., McIlwraith, et al. (1928) "Apparent Evidence of Polarization in a Beam of b-Rays", *Proceedings of the National Academy of Sciences (USA)*, Vol. 14, pp. 544–549.

Darwin, Charles G. (1927) "The Electron as a Vector Wave", *Proceedings of the Royal Society (London)*, A 116, pp. 227–253.

De Shalit, Amos, Kuperman, Stanley, et al. (1957) "Detection of Electron Polarization by Double Scattering", *Physical Review*, Vol. 107, pp. 1459–1460.

Franklin, Allan (1986) *The Neglect of Experiment*, Cambridge University Press, Cambridge.

Franklin, Allan (2018) *Is It the Same Result? Replication in Physics*, Morgan and Claypool, San Rafael, CA.

Franklin, Allan, Laymon, Ronald (2019). *Measuring Nothing, Repeatedly*, Morgan and Claypool, San Rafael, CA.

Franklin, Allan, Laymon, Ronald (2020) *Once Can Be Enough: Decisive Experiments, No Replication Required*, Springer, Heidelberg.

Friedman, Jerome L., Telegdi, Valentine L. (1957) “Nuclear Emulsion Evidence for Parity Nonconservation in the Decay Chain π+ – μ+ – e+”, *Physical Review*, Vol. 105, pp. 1681–1682.

Garwin, Richard L., Lederman, Leon M., et al. (1957) “Observation of the Failure of Conservation of Parity and Charge Conjugation in Meson Decays: The Magnetic Moment of the Free Muon”, *Physical Review*, Vol. 105, pp. 1415–1417.

Grodzins, Lee (1959) “The History of Double Scattering of Electrons and Evidence for the Polarization of Beta Rays”, *Proceedings of the National Academy of Sciences (USA)*, Vol. 45, pp. 399–405.

Illingworth, K.K. (1927) “A Repetition of the Michelson-Morley Experiment Using Kennedy's Refinement”, *Physical Review*, Vol. 30, pp. 692–696.

Joos, Georg (1934) *Theoretical Physics*, Blackie and Son Limited, London.

Kennedy, Roy J. (1926) “A Refinement of the Michelson – Morley Experiment”, *Proceedings of the National Academy of Arts and Sciences*, Vol. 12, pp. 621–629.

Laymon, Ronald, Franklin, Allan (2021) “Replication. Part I”, *Voprosy Filosofii*, Vol. 8 (2021), pp. 116–129.

Lee, Tsung-Dao, Yang, Chen Ning (1956) “Question of Parity Nonconservation in Weak Interactions”, *Physical Review*, Vol. 104, pp. 254–258.

Maxwell, James C. (1878) “Ether”, *Encyclopedia Britannica*, Edinburgh, Vol. VIII, pp. 568–572.

Michelson, Albert A. (1881) “The Relative Motion of the Earth and the Luminiferous Ether”, *American Journal of Science*, Vol. 22, pp. 120–129.

Michelson, Albert A., Morley, Edward W. (1887) “On the Relative Motion of the Earth and the Luminiferous Ether”, *American Journal of Science*, Vol. 34, pp. 333–345.

Miller, Dayton C. (1933) “The Ether-Drift Experiment and the Determination of the Absolute Motion of the Earth”, *Reviews of Modern Physics*, Vol. 5, pp. 205–242.

Mott, Nevill F. (1929) “Scattering of Fast Electrons by Atomic Nuclei”, *Proceedings of the Royal Society (London)*, A124, pp. 425–442.

Roberts, Thomas J (2006) ‘An Explanation of Dayton Miller's Anomalous “Ether Drift” Result’, URL: https://arxiv.org/vc/physics/papers/0608/0608238v2.pdf, pp. 1–23.

Shankland, Robert S. (1964) “Michelson-Morley Experiment”, *American Journal of Physics*, Vol. 32, pp. 16–35.

Shull, Clifford G., Chase, Carl T., et al. (1943) “Electron Polarization”, *Physical Review*, Vol. 63, pp. 29–37.

Stokes, George G. (1846) “On the Constitution of the Luminiferous Aether Viewed with Refernce to the Phaenomenon of the Aberration of Light”, *Philosophical Magazine*, Vol. 29, pp. 6–10.

Wu, Chien-Shiung, et al. (1957) “Experimental Test of Parity Nonconservation in Beta Decay”, *Physical Review*, Vol. 105, pp. 1413–1415.

Сведения об авторах

ЛАЙМОН Рональд –
профессор философского факультета
Государственного университета Огайо, США.

ФРАНКЛИН Аллан –
профессор физического факультета
Университета Колорадо, США.

Authors’ Information

LAYMON Ronald –
Professor, Department of Philosophy,
The Ohio State University, USA.

FRANKLIN Allan –
Department of Physics,
University of Colorado, USA.