Воспроизведение эксперимента Часть I[*]

© 2021 г. Р. Лаймон[1**], А. 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: 10.21146/0042-8744-2021-8-116-129

Цитирование: *Лаймон Р., Франклин А.* Воспроизведение эксперимента. Часть I // Вопросы философии. 2021. № 8. С. 116–129 (на английском языке).

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

Replication
Part I[*]

© 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: 10.21146/0042-8744-2021-8-116-129

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

A. Introduction

It is virtually axiomatic that "replication – the confirmation of results and conclusions from one study obtained independently in another is considered the scientific gold standard" [Jasny, Chin et al. 2011]. The underlying argument for this is that if an experiment has succeeded in revealing a real phenomenon or accurately measuring a quantity then that success should reappear when the experiment is repeated under the same circumstances or *more expansively* that the purported phenomenon or quantity measurement be obtained under different circumstances or using an entirely different experimental procedure. The reason for the more expansive requirement should be obvious since simply repeating an experiment may do no more than reproduce the experimental deficiencies, if such there be, that existed in the original experiment. At best, a more or less exact repetition might under the right circumstances reveal the existence of a confounding cause that varied during the course of the experiment. But as discussed more fully below, the better approach is to devise an improved experiment that more specifically targets suspected confounding causes.

[*] 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 and Laymon 2019; Franklin and Laymon 2020].

Putting aside for the moment the more expansive notion of replication, we note that considerable doubt has been expressed whether even when narrowly construed the replication requirement is all that well satisfied in the social sciences. The question of the extent of such replication failure is of considerable importance because if replication (in the narrow sense) is not satisfied, the original experiment was prey to *unrecognized* cofounding causes. With the aim of answering this question of extent, the Open Science Collaboration attempted to replicate 100 experimental results "published in three psychology journals using high-powered design and original materials where available" [Aarts, Anderson et al. 2015, 943]. One immediate problem that became evident was that "there is no single standard for evaluating replication success" [ibid., 943]. Thus, depending on the criteria used, the Collaboration estimated that only either 47% or 39% of the original studies had been successfully replicated[1]. This was in contrast to an expected failure rate of less than 10%. Hence there was a problem.

One need not delve into the statistical weeds (such as insuring adequate sample size and statistical power) in order to appreciate the obvious problems of ensuring relevantly similar, or relevantly superior, initial or test conditions. And since, as we have already noted, there is at best only marginal value in exactly reproducing the original experiment, there is the additional and at bottom fundamental problem of knowing whether differences in initial and test conditions that are claimed to constitute improvements really serve to do so. That this latter concern is paramount is made evident by the fact that, as tellingly realized by [Anderson et al. 2016], it is virtually impossible in the social sciences to conduct an exact repetition of an experiment because of the complex and unruly experimental conditions involved.

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

The reference to *theoretical assessments* is especially noteworthy because *the availability and depth of such assessments* identifies a potentially telling point of difference between the social and the physical sciences. But since the spirit of Anderson's appraisal applies as well to experimentation in physics we have, as indicated earlier, adopted a broad view of replication. It will not be solely performing the experiment again with either the same or a very similar experimental apparatus but also includes experiments that employ different apparatus with corresponding differences in procedure and underlying theoretical assumptions. By way of further expansion of the notion of replication we will also consider experiments that examine different phenomena that bear on the underlying theory or hypothesis involved since such experiments serve the purpose of validating the design and execution of the original experiment[2]. Our broad view of replication was initially developed and applied in [Franklin 2018] which included cases of both successful and failed replications, along with episodes in which there were difficulties in determining in what sense a replication had been achieved.

But before proceeding along these lines, we'll make a brief detour and consider the suggestion made by the Open Science Collaboration that there exists in the social sciences a *research and publication bias* as an additional contributing cause for the problems of replication. The suggestion is that both journals and the scientists themselves value positive results more than negative results and thus may not publish or even submit negative results. This has been called the "file drawer" problem in which negative results are filed away and not submitted for publication. In particular [Anderson et al. 2016] make a persuasive case that the *research and publication bias* operates in three ways: (1) to encourage positive results (*i.e.*, confirmation of the test hypothesis); (2) to discourage publication of failed attempts to confirm the test hypothesis; and (3) to discourage replications of both positive and negative results where a negative result is a failure to confirm the test hypothesis. The last is the idea that the original, or initial, work is more highly regarded than its replication. There is the additional desideratum that positive results are preferred that have large size of effect[3].

The reader may at this stage sense a noticeable difference between the social sciences and the physical sciences, most notably, physics. So, for example, with the physical sciences in mind, Ian Hacking has noted that "no one ever repeats an experiment. Typically serious repetitions of an experiment are attempts *to do the same thing* better – to produce a more stable, less noisy version of the phenomenon" [Hacking 1983, 231]. Jack Steinberger, one of the leaders of a group that performed one of the second set of measurements of η_{+-}, discussed below, concurs. "When we first proposed this experiment, we took it for granted that a more precise measurement of φ_{+-} [the phase of the CP-violating amplitude] might have given a clue on the origin of CP violation, still one of the outstanding problems. This was the physics motivation for constructing the detector. There was another purely experimental: *we saw a way of doing a much better measurement than had been done*" (private communication to Franklin, emphasis added).

In order to measure φ_{+-}, however, one must use an interference technique, which involves both the magnitude and phase of the amplitude, so that, in a sense, the measurement of η_{+-} is free. It is not, however, a requirement that a replication be better. It must simply be good enough to serve as a successful replication.

This notion of doing the same thing only better raises the obvious question of what are the standards for having conducted the better experiment. Stated differently, what is it that gives *credibility* to claims of having made a significant and relevant improvement? Franklin has suggested that this credibility is provided by the use of an epistemology of experiment, a set of strategies used to argue for the correctness of an experimental result [Franklin 2002, 2–6]. These strategies include: 1) experimental checks and calibration, in which the experimental apparatus reproduces known phenomena; 2) reproducing artifacts that are known in advance to be present; 3) elimination of plausible sources of error and alternative explanations of the result; 4) using the results themselves to argue for their validity. In this case one argues that there is no plausible malfunction of the apparatus, or background effect, that would explain the observations; 5) using an independently well-corroborated theory of the phenomena to explain the results; 6) using an apparatus based on a well-corroborated theory; 7) using statistical arguments; 8) manipulation, in which the experimenter manipulates the object under observation and predicts what they would observe if the apparatus was working properly; observing the predicted effect strengthens belief in both the proper operation of the experimental apparatus and in the correctness of the observation; 9) the strengthening of one's belief in an observation by independent confirmation; 10) using "blind" analysis, a strategy for avoiding possible experimenter bias, by setting the selection criteria for "good" data independent of the of the final result. Note that strategy 9, independent confirmation is essentially replication. It is only one of the possible ways of validating an experimental result.

At bottom, much of the concern about the need for replication and more generally scientific objectivity comes from the worry that an experimental result may reflect the influence of confounding factors rather than the underlying fundamental processes that the experiment aims to uncover. Making use of the current terms of art, the problem is how to deal with *systematic uncertainty*. As has been made clear by our case studies what's needed are effective ways of reducing such uncertainty by means of improved experimental apparatus and procedures. And it is here that the strategies discussed above, what we have referred to as the epistemology of experiment, come forcefully into play. Thus, while there may be no harm in simply duplicating an experiment it's much better to think more deeply and aim for the better experiment-assuming of course that the claim of being better has been made *credible*.

For a specific instance where credibility is at issue consider what can be described as the "bandwagon effect". Here experimenter bias comes into play insofar as the experimenter continues his or her manipulation of the experimental particulars until results are obtained that agree with earlier results that have already received the imprimatur of the scientific community[4]. As stated by the Particle Data Group, which assembles the "Review of Particle Physics", the standard reference on particle properties: "The old joke about the experimenter who fights the systematics until he or she gets the 'right' answer (read 'agrees with previous experiments') and then publishes contains a germ of truth" [Kelly, Horne et al. 1980, S286]. One technique to overcome this sort of experimenter temptation and resultant bias is "blind analysis", which made an appearance above in our epistemology of experiment. In short, it's the practice of setting the selection criteria on the data before the final result is calculated and known[5].

There is as should be expected both ambiguity and uncertainty in the application of any requirement of replication. At the top of the list is the very notion of the *result* of an experiment. So, for example, does the demand for replication just deal with the data produced, such as density and time determinations or cloud chamber photographs, or should it be understood more expansively to include higher level theoretical processes and entities such as the replication of DNA or the existence of the positron? And even assuming that the relevant type of result has been agreed upon there is the question of how similar or different the results must be to count as successful or unsuccessful replications. Scientists have offered different answers to this question[6]. A variant of this question has been raised concerning experimental results in high-energy physics, namely, how statistically significant must a result be to be considered a discovery?[7] And as a related question consider whether the same standards should be applied to a discovery and to a confirmation of that discovery?

The case studies discussed here and in our earlier books were undertaken with the aim of discovering *methodologically relevant specificity* regarding the replication of experimental results. In other words, our aim has been to add specific historical instances of the role replication, otherwise only abstractly considered, has played in the development of science.

So in what follows we'll consider (1) a stunning and current case of a what is undoubtedly a successful replication, namely, the near simultaneous discovery of the Higgs boson by two independent research groups; and (2) the more melancholy series of unsuccessful replications of the Universal Gravitational constant. In Part II (to be published in the next issue of this journal) we'll focus on the repetition of null experiments in physics which at first glance stands in stark contrast to what happens in the social sciences where because of the file drawer problem such apparently obsessional behavior carries severe consequences for one's career. In short, measuring nothing repeatedly is not likely to get you very far in the social sciences. We'll also focus attention on cases where replication was not required either as a matter of historical fact or of sound methodology considered from a more philosophical perspective. In short, these are cases where once was enough.

B. The Discovery of the Higgs Boson[8]: A Successful Replication

The recent discovery of the Higgs boson[9] reported by both the ATLAS and CMS collaborations at the Large Hadron Collider (LHC) is a clear example of a successful replication. These experiments used different detectors and were performed by different experimental groups [Aad et al. 2012; Chatrchyan et al. 2012]. They did, however, investigate the same phenomena. They not only provided strong evidence for the existence of a new elementary particle but also provided evidence for the Standard Model, the currently accepted theory of elementary particles. The Higgs boson was the last remaining unobserved piece of that model. The observed effect in each experiment was more than five standard deviations (5 σ) above background[10], providing very strong statistical support for the credibility of the results. The observation of the particle in two different decay modes, $H \rightarrow \gamma\gamma$ and $H \rightarrow ZZ^*$, (See [Franklin 2018], Figures 2.1–2.4)[11] by both experiments supported the conclusion[12]. Both collaborations also found the same mass for the particle, 125.3 ± 0.4 GeV (CMS) and 126.0 ± 0.4 (stat) ± 0.4 (syst) (ATLAS), providing additional support[13]. In addition the CDF and D0 collaborations at Fermilab found a 3 σ effect at the Higgs mass, which although not qualifying as a discovery by contemporary high-energy physics standards, did provide additional support for the existence of the particle[14]. The results reported by both CMS and ATLAS constitute what one might describe as simultaneous and reciprocal replications involving different experiments where the experiments differed in ways that would maximize the identification of confounding causes.

Both the CMS and ATLAS collaborations offered arguments for the credibility of their results. One of these was an extended calibration procedure of both the CMS apparatus and its analysis procedures[15]. Over a period of several years the CMS experiment had replicated much the history of 20th century particle physics. For example, the group observed strong signals for the neutral K mesons as well as other particles. The collaboration constructed a timeline for discoveries in 20th century particle physics along with the timeline for the replications performed by CMS. This procedure demonstrated the ability of the apparatus and the analysis procedures to produce correct results and provides support for the credibility of other results obtained by the collaboration.

A crucial part of the analysis of the Higgs discovery experiments was the determination of the photon energies for the H → γγ decay mode. "A multivariate regression is used to extract the photon energy and a photon-by-photon estimate of the uncertainty in that measurement. The calibration of the photon energy scale uses the Z boson mass as a reference; ECAL showers coming from electrons in Z → ee events are clustered and reconstructed in exactly the same way as photon showers" [Chatrchyan et al. 2012, 34]. Electrons and photons are expected to behave in very similar ways in such a detector. The Z boson has a known mass and was used as a calibration. The mass reconstructed from the two electrons from the decays were required to fit that mass, thus establishing the energy scale, see [Franklin 2018], Figure 2.3.

An important part of the analysis was the need to separate the signal from the background (see [Franklin 2018], Figures 2.1–2.4, [Chatrchyan et al. 2012] and [Aad et al. 2012]). For the decay H → γγ the CMS group stated, "The background is estimated from data, without the use of MC [Monte Carlo] simulation, by fitting the diphoton mass distribution in each of the categories[16] in a range ($100 < m_{\gamma\gamma} < 180$ GeV) extending slightly above and below that in which the search is performed" [Chatrchyan et al. 2012, 34]. (See [Franklin 2018], Figure 2.1.) The collaboration also used a second, independent analysis "using a different approach to the background modelling" [Chatrchyan et al. 2012, 34]. The collaboration remarked that "The observed limit indicates the presence of a significant excess at $m_H = 125$ GeV in both the 7 and 8 TeV data" [ibid., 34–35].

For the decay mode H → ZZ^* the ZZ background was calculated from a Monte Carlo simulation. "Two different approaches are employed to estimate the reducible and instrumental backgrounds from the data. They both used events from a background region, well-separated from the signal region. *Within uncertainties, comparable background counts in the signal region are estimated by both methods*" [ibid., 36]. For both decay modes the two different background estimates agreed. This robustness added credibility to the results, the m_{4l} distribution is shown in Figure 2.3 [Franklin 2018; Chatrchyan et al. 2012]. There is a clear peak at the Z mass where the decay Z → 4l is reconstructed. This feature is well reproduced by the background estimation. The figure also shows an excess above the expected background around 125 GeV (p. 36)". The fact that the figure shows the known Z boson at its known mass added to the credibility of the result. A clear signal is seen at 125 GeV. The mass distribution found by the ATLAS collaboration also shows a significant signal (see [Franklin 2018], Figure 2.4, and [Aad et al. 2012]).

To guard against possible experimenter bias the collaborations used blind analysis. "The new analyses presented herein, …featuring modified event selection criteria, were performed in a 'blind' way: the algorithms and selection procedures were formally approved and fixed before the results from data in the signal region were examined" [Chatrchyan, Khachatryan et al. 2012, 31][17].

There is also what one might call internal replication, within a single experiment. In the cases of the Higgs Boson, as noted above, each of the collaborations reported observing the same particle in each of two decay modes, albeit with a lower statistical significance. We note here that the data sets for each of the decay modes was different[18]. In addition, in each collaboration the analysis of the data was independently performed by several groups, providing robustness for the analysis[19].

The fact that the Higgs boson was seen in both decay modes in each experiment added to the credibility of the result. Further support was added by the fact that for both experiments, the results for each of the decay modes, as well as for the combined result agreed with the predictions of the Standard Model, a theory that already had considerable evidential support.

In sum, the discovery of the Higgs boson is a paradigmatic example of a successful replication where the replication consisted of two simultaneous and reciprocal replications. As clear a demonstration of the value of replication – broadly construed – as could be hoped for.

C. Is There a Universal Gravitational Constant?[20] Failed Replications

Despite its status as perhaps the most venerable of physical constants[21], G, the Universal Gravitational constant is the least well-measured of the important physical constants. Recent measurements of G show a wide variation, with many of the measurements reporting values which differ from the accepted value by far more than their stated uncertainties (see [Franklin 2018], Figure 6.1, [CODATA 2014]). As Mohr and collaborators remarked in their 2016 review of the recommended values of physical constants, the addition of three then new measurements of G (which will be discussed below), obtained with different methods, "have not resolved the considerable disagreements that have existed among the measurements of G for the past 20 years" [Mohr, Newell et al. 2016, 035009–4]. The differences are so large that the Task Group assigned to recommend a value for G used an expansion factor[22] of 6.3 for the initial uncertainties in the reported values "that reduces the normalized residuals of each datum to less than 2…" [ibid.]. This is a clear example of failed replication.

The most recent CODATA value of G is (6.67408 ± 0.00031) G_0, with a relative uncertainty of 4.7×10^{-5}. For simplicity we will write G as a numerical factor multiplying G_0, where $G_0 = 10^{-11}$ m^3 kg^{-1} s^{-2} One might contrast this with the uncertainty in the electron g factor, the ratio of its magnetic moment to its spin[23], of (2.00231930436182 ± 0.00000000000052), with a relative uncertainty of 2.6×10^{-13}.

One possible problem with measuring G is that the gravitational force is quite weak when compared to other forces. For example, the electromagnetic force between the electron and proton in the hydrogen atom is 2×10^{39} times as large as the gravitational force between them. Another is the fact that the gravitational constant is independent of other physical constants. As Mohr and collaborators noted, "because there is no known quantitative theoretical relationship between G and other adjusted constants, they [then recent measurements of G] contribute only to the determination of the 2014 recommended value of G" [ibid.]. For example, the fine structure constant $\alpha = 2\pi e^2/hc$. Thus, measurements of either α, e, h, or c place constraints on the values of these other constants. G has no such relationships.

An interesting attempt to resolve the problem was the 2001 experiment by Quinn and collaborators [Quinn 2001]. They noted that there had been recent measurements of G that gave values closer to the 1998 CODATA value[24], (6.673 ± 0.010) G_0, "particularly the paper by [Gundlach, Merkowitz 2000] that gives a result with the very low uncertainty of 14 ppm [parts per million]. We report here a new determination of G, which has a standard uncertainty of 41 ppm. Our value is unique in that it is based on two results obtained using the same apparatus but with different methods of measurement. Our result does, however, differ from that of Gundlach and Merkowitz by some 200 ppm [ibid., 111101–1]".

Two different methods were used 1) electrostatic servo control and 2) free deflection (Cavendish method). When the source masses were radially aligned with the test masses the gravitational torque was zero. Rotating the source masses by 18.7° in either direction produced a maximum torque. "In the servo-controlled method, the gravitational torque of the source masses is balanced by an electrostatic torque acting directly on the test masses" [ibid., 111101–2]. In the Cavendish method at equilibrium the applied gravitational torque is balanced by the suspension stiffness, $\tau = c\theta$, where c is the stiffness of the suspension wire and θ is the angle of equilibrium. Just as Cavendish had done, the stiffness constant was obtained from the period of free oscillation of the test masses and the moment of inertia.

The final result for the servo method was G = 6.67553 G_0 with a standard uncertainty of 6.0 parts in 10^5. For the Cavendish method G = 6.67565 G_0 with a standard uncertainty of 6.7 parts in 10^5. The final result was G = 6.67559(27) G_0 with a standard uncertainty, which included the effects of correlations between the two methods, of 4.1 parts in 10^5. "In conclusion, the close agreement of the results of our two substantially independent methods is evidence for the absence of many of the systematic errors to which a G measurement is subject. *Nevertheless, the two most accurate measurements of G, this one and that of Gundlach and Merkowitz [6.674215 ± 0.000092 G_0], differ by more than 4 times their combined standard uncertainty*" ([Gundlach, Merkowitz 2000, 111101–3], emphasis added). Both results differ considerably from the 1998 CODATA value. The problem remained[25].

In 2013 Quinn and collaborators reported on their continued efforts to resolve the discordant results. This was the first of the three new values discussed by Mohr and collaborators. They reported a new value for G using the same methods used in their 2001 paper. "The apparatus has been completely rebuilt and extensive tests carried out on the key parameters needed to produce a new value for G" [Quinn, Parks et al. 2013, 101102–1]. They further noted that, "The 2010 CODATA evaluation of the fundamental constants shows a spread in the recent values of the Newtonian constant of Gravitation of some 400 ppm, *more than ten times the estimated uncertainties* of most of the contributing values" ([ibid.], emphasis added). Their reported values of G were 6.67520(41) G_0 and 6.67566(37) G_0 for the servo and Cavendish methods respectively. The weighted mean value was 6.67545(18) G_0. They remarked that their "new value is 21 ppm below our 2001 result which had an uncertainty of 41 ppm but 241 ppm above the CODATA 2010 value" [ibid., 101102–4]. They further stated, "Noting that in each, the result is based on the average of two largely independent methods, taken together, our two results represent a unique contribution to G determinations" [ibid., 101102–5]. The experimenters had successfully replicated their 2001 result but, the discord among all of the measurements of G remained.

The second of the new values reported mentioned by Mohr was that of the European Laboratory for Nonlinear Spectroscopy at the University of Florence [Prevedelli, Cucciapuoti et al. 2014]. "The experiment combines two vertically separated atomic clouds forming a double atom-interferometer-gravity gradiometer that measures the change in the gravity gradient when a well-characterized source mass is displaced" [Mohr, Newell et al. 2016, 035009–39]. Their final value for G was 6.67191(99) G_0, a relative uncertainty of 1.5 × 10^{-4}. As Mohr and collaborators remarked, "Although not competitive, the conceptually different approach could help identify errors that have proved elusive in other experiments" [ibid., 035009–38].

The third new value was the result of a series of measurements, over a period of seven years, taken by a group at the University of California at Irvine headed by Riley Newman [Newman, Bantel et al. 2014]. "A measurement of Newton's gravitational constant *G* has been made with a cryogenic torsion pendulum operating below 4 K in a dynamic mode in which *G* is determined from the change in torsional period when a field source mass is moved between two orientations" [ibid., 2014002–1].

The group used three different torsion fibers and obtained values of 6.674350(97) G_0, 6.67408(15) G_0, and 6.67455(13) G_0, respectively, with relative uncertainties of 1.5 × 10^{-5}, 2.2 × 10^{-5}, and 2.0 x 10^{-5}. The unweighted average of the three values was 6.67433(13) G_0. The experimenters regarded these results as inconsistent. "We have no explanation for the inconsistency of the results" [ibid., 20140025–24].

Although the value of G obtained by Newman and collaborators is reasonably close to the CODATA value of G, the discord between measurements of G remains unresolved to this day. It is fair to say that the measurements of G are a failed sequence of replications.

D. More General Problems in the Determination of the Fundamental Constants

In 1927 Raymond Birge published the first of a series of papers on the value of important physical constants [Birge 1929; Birge 1941^a; Birge 1941^b; Birge 1957]. Those values were generally accepted by the physics community as the most accurate determinations.

In other papers he also discussed the mathematical techniques for calculating the values and their probable errors. In his 1932 paper, "The Calculation of Errors by the Method of Least Squares" [Birge 1932], Birge discussed the distinction between internal and external consistency. (Internal consistency involves a comparison of measurements of a quantity obtained within a single experiment, whereas external consistency involves a comparison among values obtained in different experiments). He noted that both methods yield the same result if only accidental, or statistical errors, are present. "When the ratio of R_e/R_i [external to internal error] exceeds unity by an amount much greater than is to be expected on the basis of statistical fluctuation, one has *almost certain evidence of the presence of systematic errors*" ([ibid., 207], emphasis added). Birge did not, however, specify how large that ratio had to be to justify that conclusion[26].

As an example of bad practice, Birge discussed, "F.W. Clarkes monumental work on the calculation of atomic weights" [ibid., 221]. Clarke used only internal consistency in his calculation of probable errors and in weighting the results obtained by different experimenters. Birge found that the ratio of external to internal consistency in Clarke's calculation averaged approximately ten. "I find, from sample calculations, that the ratio R_e/R_i averages about ten, so that Clarke's stated probable errors average about one-tenth of the most probable values. In certain cases, such a system of analysis leads to a clearly false result for the atomic weight itself, as I have pointed out in a previous paper. Thus, there is occasionally an atomic weight determination by some particular investigator that is quite at variance with all other known results, but that happens to have high internal consistency. Clarke accordingly gives it a high weight, and this weight carries through to the final result, so that the investigation in question, which should have been discarded entirely, produces an appreciable change in the published final result. All of the recent reports of atomic weight committees seem to recognize the fact that the older determinations are nearly all vitiated by constant errors, and as a result the committee makes an arbitrary assignment of weights. To speak bluntly, it gives zero weight to these older determinations, regardless of their apparent probable errors… I have adopted the same policy" [ibid.].

Although it seems possible, even probable, that later measurements of a quantity are more accurate and reliable than earlier measurements because the later experiments have found and corrected earlier systematic errors, this is not necessarily the case. Similarly, it is not necessarily true that a result that is at variance with other results is wrong.

The question of the accuracy of both the values and the uncertainties of the fundamental physical constants has remained an issue until the present. In a 1943 letter to the editor of the *Physical Review*, Frank Benford [Benford 1943] complained that the values of many of the physical constants contained in Raymond Birge's latest compilations [Birge 1929; Birge 1941[a]][27] had changed by far more than the stated probable errors. (Table 1 lists a few of these changes.) For example, [Franklin 2018], Figure 7.1[28], shows a graph of the recommended values of *c*, the velocity of light, as a function of time. The changes shown differ by far more than the stated experimental uncertainty. Benford remarked that, "The main contributing factor to the changes between the 1929 and 1941 list is the new value assigned to the electronic charge. In 1929 it was 4.7700 ± 50 and in 1941 it was 4.8025 ± 10, a difference of 325 as compared to the probable error of 50. The change is 6.5 times as great as the probable error, and the chances against such a change are 100,000 to 1, on the basis of the 1929 figures. It is here again evident that the assigned limits refer to the consistency of the data from which 4.7700 ± 50 was derived, and the present value would, in 1929 have seemed impossible from the internal evidence" [Benford 1943, 212][29].

[Birge 1943] replied that this was a real problem. "In the preceding paper by Dr. Frank Benford, attention is called to the important distinction between consistency and accuracy. The recent large and wholly unanticipated changes in the measured values of many of the general physical constants have been noted and discussed in many papers… For many years now I have attempted to emphasize just this distinction. Unfortunately, the true value of any physical magnitude can never be known. Hence the absolute accuracy of a measured result cannot be determined and we can only note the difference, if any, between the *internal* and *external* consistency of the data" [Birge 1943, 213]. He further noted that "As a result of the diversity of methods now used in determining most of the physics constants and the greater attention being paid to the question of systematic errors, I feel that the values and probable errors of the 1941 list can be accepted with considerably greater confidence than in the case of any earlier list" [ibid.].

Birge was not, however, excessively concerned about the large changes in the value of the fundamental constants. He remarked, "In spite of the delusive word 'Constants' appearing in the title, it is the continual variation in the values of these quantities that furnishes most of the interest in the subject. It would indeed be disheartening to any real scientist to feel that an accepted value of any physical constant would never be changed. The most characteristic feature of science – is its never-ending change" [Birge 1941^{b}, 90]. Cohen and DuMond, who took over the task of compiling the fundamental constants from Birge, agreed. "…having done one's best with the available data, we must all learn not to be too surprised or disappointed if more highly developed methods subsequently reveal the presence of systematic errors unsuspected at the earlier data and of considerably larger magnitude than the earlier estimate of random error" [Cohen, DuMond 1965, 551].

One possible way of avoiding the problem would be for scientists to increase their estimate of the probable error. Cohen and DuMond disagreed. In discussing the compilation of constants "It would be an equally grave mistake to recommend that the reviewer enlarge his error estimates 'to take care of possible but unknown systematic errors'. Systematic errors in physical measurements do not obey any known statistics…[30] We simply have to learn the hard fact that, having arrived at a determination of a physical quantity and its estimated uncertainty in the light of all the best information available at a given epoch, this may prove at a later epoch, when we have more and better information to have been wrong" [ibid., 551–552].

Dunnington had worried about a related, but complementary, problem. This was the existence of a possible "bandwagon effect", in which experimental results tend to agree with either previous measurements or with theoretical calculations. "It is easier than is generally realized to unconsciously work toward a certain value. One cannot, of course, alter or change natural phenomena (for example, the location of the current minimum in the present experiment), but one can, for instance, seek for those corrections and refinements which shift the results in the desired direction" [Dunnington 1933, 416].

Birge agreed that this was a problem and reported an explanation, which he attributed to Ernest Lawrence.

In any highly precise experimental arrangement, there are initially many instrumental difficulties that lead to numerical results far from the accepted value of the quantity being measured. It is, in fact, just such wide divergences that are the best indication of instrumental errors of one kind or another. Accordingly, the experimenter searches for the source or sources of such errors, and continues to search until he gets a result close to the accepted value. *Then he stops.* But it is quite possible that he has still overlooked some source of error that was present also in previous work. In this way one can account for the close agreement of several different results and also for the possibility that all of them are in error by an unexpectedly large amount [Birge 1957, 51].

The existence of, and possible solution to, this problem in shown in [Franklin 2018], Figure 7.2, which shows the values of η_{+-}, the CP violating parameter in $K^{\circ}_{L} \rightarrow \pi^{+}\pi^{-}$ decay, as a function of time, up until 1985. The early measurements were all consistent with one another and gave an average value of $(1.95 \pm 0.03) \text{ x } 10^{-3}$. Similarly, the post-1973 measurements were all mutually consistent, but with an average of $(2.27 \pm 0.022) \times 10^{-3}$. The difference in the averages is more than eight standard deviations, which has a probability of 1.2×10^{-15}, a very unlikely event indeed. The internal agreement of both pre- and post-1973 sets of measurements might indicate a bandwagon effect. The fact that the later experimenters were sufficiently convinced of the correctness of their results that they were willing, indeed eager[31], to publish a result that differed by so much from the world average suggests that the bandwagon can be stopped.

At the time the situation was sufficiently unusual that the Particle Data Group, which compiles results on the properties of elementary particles, felt obliged to comment on it.

There is a very large discrepancy between the old and new results for $|\eta_{+-}|$... The origin of the discrepancy... is not known... We are troubled by this large, unexplained discrepancy. We feel that our normal procedure of averaging and increasing the error by a scale factor S to account for the discrepancy[32] is not adequate for this case. The new results, when combined with the average of earlier results by that procedure give $(2.15 \pm 0.11) \times 10^{-3}$ (S = 6.0)[33]. While this value and error makes some sense in that it nearly spans both incompatible sets of data, we choose not to quote it. Instead, since the newer experiments are in principle superior (higher statistics, better acceptance, easier trigger conditions), we have chosen to average them separately from the earlier results [Trippe, Barbaro-Galtieri et al. 1976, S81].

Although the origin of the discrepancy between the two sets of measurements has never been found, later results have been consistent with the higher value for η_{+-}. The earlier results have now been omitted from the current world average[34] of $\eta_{+-} = (2.232 \pm 0.011) \times 10^{-3}$, with a scale factor of 1.8. There is still some inconsistency in the results. This last episode is an illustration of two apparent sets of successful replications (within the sets) which, unfortunately did not agree (comparing the different sets with each other).

E. Conclusions

(1) Replication in the narrow sense (i.e., exact or nearly exact repetition) is not a necessary condition for scientific acceptability. This because replication in the expanded sense discussed above is a more telling and profitable requirement to apply.

(2) In considering replication as a normative requirement it is essential to keep in mind its overriding purpose which is to ferret out and identify confounding causes. In other words, to reduce systematic uncertainty.

(3) Keeping the second of the above conclusions in mind, it becomes evident that while successful replication may be a “gold standard”, it surely is not *the* “gold standard” for establishing the correctness of a result. There are a host of alternative strategies, what we have described as an epistemology of experiment, that provide similar benefits when it comes to dealing with systemic uncertainty. Unless, of course, one construes replication so widely as to incorporate the entirety of the strategies included in the epistemology of experiment. In that case, it becomes the gold standard by definitional fiat.

(4) As shown by our discussion of actual cases (here and elsewhere), ambiguities and vagueness in application of the replication requirement are, more often than not, resolved when considering *the actual candidates in play* for being confounding causes. Insofar as there are known or newly developed methods for controlling such confounding causes the contours of a telling replication will become apparent.

(5) Finally, as will be shown in Part II, and with due respect to Jacqueline Susann, sometimes once may be enough. And this in the sense that replication (in either a narrow or expanded sense) was not a normative requirement and was not required as a matter of historical practice.

Notes

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

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

[3] In support Anderson et al. argue that: “Low power research designs combined with publication bias favoring positive results together produce a literature with upwardly biased effect sizes. This anticipates that replication effect sizes would be smaller than original studies on a routine basis – not because of differences in implementation but because the original study effect sizes are affected by publication and reporting bias, and the replications are not. Consistent with this expectation, most replication effects were smaller than original results, and reproducibility success was correlated with indicators of the strength of initial evidence, such as lower original P values and larger effect sizes. This suggests publication, selection, and

reporting biases as plausible explanations for the difference between original and replication effects. The replication studies significantly reduced these biases because replication preregistration and pre-analysis plans ensured confirmatory tests and reporting of all results" [Anderson 2015, 3].

[4] For a discussion of this issue see [Franklin 1986], Chapter 8.

[5] See [Franklin 2002a], Chapter 6.

[6] This has been the subject of recent discussions of experiments in psychology. See [Simons 2013; Srivastava 2015]. It is interesting to note that the statistical criterion for a significant effect used in psychology is two standard deviations, whereas particle physics demands a five-sigma effect for a discovery claim.

[7] In high energy physics and in gravity wave physics the statistical criterion for a discovery is that the observed effect be five standard deviations above background. For a discussion and history of the criterion see [Franklin 2013], Prologue.

[8] For more details see [Franklin 2017].

[9] The collaborations did not initially claim to have found the Higgs boson, but rather only a boson. Identifying the particle as the Standard Model Higgs boson would require later work on the branching ratios and coupling constants. Nevertheless, the physics community generally regarded the new particle as the Higgs boson.

[10] We emphasize here that the 5 σ criterion is shorthand for a complex analysis of the data.

[11] [Chatrchyan et al., 2012], Figures 3 and 4 and [Aad et al., 2012], Figures 2 and 4.

[12] In fact, both decay modes were needed to meet the required 5 sigma criterion. In the CMS experiment, for example, the $H \rightarrow \gamma\gamma$ result had a statistical significance of 4.1 σ and the $H \rightarrow ZZ^*$ a significance of 3.2 σ.

[13] The difference in mass is 0.7 ± 0.7 GeV, a 1 standard deviation difference. This is a clear agreement.

[14] These collaborations found evidence for the Higgs boson in yet another decay mode.

[15] This extended calibration procedure was not mentioned in the Higgs discovery paper, but it was known to the community through published papers and talks at conferences and elsewhere.

[16] These categories involved different selection criteria for the Higgs signal and the background.

[17] For a more detailed discussion see [Franklin 2002a], Chapter 6.

[18] Because the LHC produces events at a rate far greater than can be recorded a trigger system is used to select events of interest. A trigger system includes counters and other detectors along with computer programs for making a fast decision on whether to record the detector data for each event.

[19] We are grateful to Keith Ulmer, a member of CMS, for pointing this out.

[20] This will not be a complete discussion of the numerous measurements of G. It will include sufficient discussion to demonstrate a failed replication.

[21] Although Newton did not use such a constant in his statement of his Law of Universal Gravitation and though Cavendish measured the density of the Earth and not G, it seems reasonable to consider G as dating from the time of Newton. Newton did state that the gravitational force between two masses was proportional to the product of the two masses and inversely proportional to the square of the distance between them.

[22] See discussion below concerning the scale factor used by the Particle Data Group.

[23] To be fair the g factor of the electron is among the best measured physical constants.

[24] CODATA values for the physical constants are the accepted values for the physics community.

[25] The experimenters attempted to explain the discrepancy by invoking a failure of Newton's inverse square law of gravity, because the effective distances between the source and test masses were different in the two experiments. They found that the violation required was a factor of three larger than the limit that had been set by [Spero et al. 1980].

[26] It is useful to distinguish between systematic error and systematic uncertainty. The former might be an effect that changes all of the results by the same amount, whereas the latter may introduce an unknown uncertainty. An anecdote may help to clarify the distinction. The British humor magazine, Private Eye, reported that although Princess Margaret and her husband, Lord Snowden, were the same height, in all photographs he appeared taller. The magazine attributed this to his standing on a Lord Snowden Box, which came in 3, 6, and 9-inch sizes. If one measured the height of a group of people all of whom were standing on a Lord Snowden box and one knew which version of the box they were standing on their height would by shifted by 3, 6, or 9 inches. This would be a systematic error. If one didn't know which version of the box they were standing on then we would have a systematic uncertainty of (6 ± 3) inches.

[27] A more detailed account appeared in [Birge 1941b].

[28] See also [Henrion, Bischoff 1986], Figure 2.

[29] The change in the value of e was primarily due to a change in the measured value of the viscosity of air.

[30] Another problem with unknown systematic errors is that they are unknown. The experimenter doesn't know how large a factor by which to increase the reported uncertainty.

[31] Although I was not an author of one of these papers because I was not involved in the analysis of this data, I was a member of the experimental group that took the data. The group had no difficulty in deciding to publish their discordant result. Recall also the comments by Jack Steinberger, the leader of one of the experimental groups. The group thought that their measurement was better and published its result.

[32] In a case in which the reported values differ considerably, the Particle Data Group increases the quoted error by a scale factor, $S = [\Sigma\ \chi 2/ (N-)]1/2$. (See [Amsler, Doser et al. 2008, 16–17] for details).

[33] This was an extraordinarily large scale factor.

[34] This is not an unusual procedure for the Particle Data Group.

Table 1. Some changes in Fundamental Constants [Benford 1943]

	1929		1941		Change 1929 P.E.*	Chance 1 to
Velocity of light	*c* 2.99796	± 4	299776	± 4	5.0	1.3×10^3
Electronic charge	*e* 4.7700	± 50	4.8025	± 10	6.5	1.0×10^5
Planck Constant	*h* 6.5470	± 80	6.6242	± 24	9.6	1.0×10^{10}
Avogadro’s Number	N_o 6.0644	± 60	6.0228	± 11	7.0	4.0×10^5
Boltzmann constant	*K* 1.37089	± 140	1.38047	± 26	6.8	1.0×10^5

* Probable error

References

Aad, Georges, et al. (2012) “Observation of a New Particle in the Search for the Standard Model Higgs Boson with the ATLAS Detector at the LHC”, *Physics Letters*, Vol. B716, pp. 1–29.

Aarts, Alexander A., Anderson, Joanna E., et al. (2015) “Estimating the Reproducibility of Psychological Science”, *Science*, Vol. 349 (6251), p. 943.

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

Benford, Frank (1943) “The Probable Accuracy of the General Physical Constants”, *Physical Review*, Vol. 63, p. 212.

Birge, Raymond T. (1929) “Probable Values of the General Physical Constants”, *Reviews of Modern Physics*, Vol. 1, pp. 1–71.

Birge, Raymond T. (1932) “The Calculation of Errors by the Method of Least Squares”, *Physical Review*, Vol. 40, pp. 207–227.

Birge, Raymond T. (1941^{a}) “A New Table of Values of the General Physical Constants”, *Reviews of Modern Physics*, Vol. 13, pp. 233–239.

Birge, Raymond T. (1941^{b}) “The General Physical Constants: as of August 1941 with details on the velocity of light only”, *Reports on Progress in Physics*, Vol. 8, pp. 90–134.

Birge, Raymond T. (1943) ‘Comments on “The Probable Accuracy of the General Physical Constants”’, *Physical Review*, Vol. 63, p. 213.

Birge, Raymond T. (1957) “A Survey of the Systematic Evaluation of the Universal Physics Contents”, *Nuovo Cimento* (Suppl.), Vol. 6, pp. 39–67.

Chatrchyan, Serguei, et al. (2012) “Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC”, *Physics Letters*, Vol. B716, pp. 30–61.

Cohen, E. Richard, DuMond, Jesse W.M. (1965) “Our Knowledge of the Fundamental Constants of Physics and Chemistry”, *Reviews of Modern Physics*, Vol. 37, pp. 537–594.

Dunnington, Frank G. (1933) “A Determination of e/m for an Electron by a New Deflection Method”, *Physical Review*, Vol. 43, pp. 404–416.

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

Franklin, Allan (2002) *Selectivity and Discord*, University of Pittsburgh Press, Pittsburgh.

Franklin, Allan (2013) *Shifting Standards: Experiments in Particle Physics in the Twentieth Century*, University of Pittsburgh Press, Pittsburgh.

Franklin, Allan (2017) “The missing piece of the puzzle: the discovery of the Higgs boson”, *Synthese*, Vol. 194, pp. 259–274.

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

Franklin, Allan, Howson, Colin (1984) “Why Do Scientists Prefer to Vary Their Experiments?”, *Studies in History and Philosophy of Science*, Vol. 15, pp. 51–62.

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.

Gundlach, Jens, Merkowitz, Stephen M. (2000) “Measurement of Newton's Constant Using a Torsion Balance with Angular Acceleration Feedback”, *Physical Review Letters*, Vol. 85, pp. 2869–2872.

Hacking, Ian (1983) *Representing and Intervening*, Cambridge University Press, Cambridge.

Henrion, Max, Fischoff, Baruch (1986) “Assessing Uncertainty in Physical Constants”, *American Journal of Physics*, Vol. 54, pp. 791–798.

Jasny, Barbara R., Chin, Gilbert, et al. (2011) “Again, and Again, and Again…”, *Science*, Vol. 334, p. 1225.

Kelly, Robert L., Horne, Charles P., et al. (1980) “Review of Particle Properties”, *Reviews of Modern Physics*, Vol. 52, p. S1–S286.

Mohr, Peter J., Newell, David B., et al. (2016) “CODATA recommended values of the fundamental physical constants 2014”, *Reviews of Modern Physics*, Vol. 88, p. 035009, https://doi.org/10.1063/1.4724320.

Newman, Riley, Bantel, Michael, et al. (2014) “A measurement of G with a cryogenic torsion pendulum”, *Philosophical Transactions of the Royal Society*, Vol. 372, pp. 20140021–20140025, DOI:10.1098/rsta.2014.0025.

Prevedelli, Marco, Cucciapuoti, Luigi, et al. (2014) “Measuring of the Newtonian constant G with an atomic interferometer”, *Philosophical Transactions of the Royal Society*, Vol. 372, p. 20140030, DOI:10.1098/rsta.2014.0030.

Quinn, Terry, Parks, Harold, et al. (2013) “Improved Determination of G Using Two Methods”, *Physical Review Letters*, Vol. 111, p. 101102, DOI: 10.1103/PhysRevLett.111.101102.

Quinn, Terence J., Speake, Clive C., et al. (2001) “A New Determination of G Using Two Methods”, *Physical Review Letters*, Vol. 87, p. 111101.

Simons, Daniel (2013) “What Counts as a Successful Replication?”, URL: http://blog.dansimons.com/2013/02/what-counts-as-successful-replication.html

Spero, Robert, Hoskins, J.K., et al. (1980) “Tests of the Gravitational Inverse-Square Law at Laboratory Distances”, *Physical Review Letters*, Vol. 44, pp. 1645–1648.

Srivastava, Sanjai (2015) “What Counts as a Successful or Failed Replication?” URL: https://hardsci.wordpress.com/2012/10/05/what-counts-as-a-successful-or-failed-replication/

Trippe, Thomas G., Barbaro-Galtieri, Angela, et al. (1976) “Review of Particle Properties”, *Reviews of Modern Physics*, Vol. 48, pp. S1–S286.

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.