INTRODUCTION

> One cannot take away integrity in the search for evidence and honesty in declaring one's results and still have science; one cannot take away a willingness to listen to anyone's scientific theories and findings irrespective of race, creed, or social eccentricity and still have science; one cannot take away the readiness to expose one's findings to criticism and debate and still have science; one cannot take away the idea that the best theories will be able to specify the means by which they could be shown to be wrong and still have science; one cannot take away the idea that a lone voice might be right while all the rest are wrong and still have science; *one cannot take away the idea that good experimentation or theorization usually demand high levels of craft skills and still have science*; and one cannot take away the idea that, in virtue of their experience, some are more capable than others at both producing scientific knowledge and at criticizing it and still have science. These features of science are "essential," not derivative.
>
> —Harry Collins (2004, 156; emphasis added)

IN 1981 I published an article titled, "What Makes a Good Experiment?" (Franklin 1981). At the time I, along with many other philosophers of science, believed that the only significant role of experiment in science was to test theories. Since that time we have recognized that experiment plays many other significant roles in science. These other roles include: exploratory experiments, designed to investigate a subject for which a theory does not exist so that a theory may be formulated; experiments that help to articulate an existing theory; experiments that call for a new theory either by demonstrating the existence of a new phenomenon in need of explanation or by demonstrating that an existing theory is wrong; experiments that provide evidence for the entities involved in our theories or new enti-

ties; experiments that measure quantities that are of physical interest such as Planck's constant or the charge of the electron; and experiments that have a life of their own, independent of high-level theory. An experiment may also correct previous incorrect or misinterpreted results. Experiments may also play a role in providing reasons for pursuit, the further investigation of a theory or a phenomenon. Yet another role for experiment is that of an enabling experiment, an experiment that may give an incorrect result but demonstrates that the quantity of interest can be measured.[1] A related role is that an experiment may demonstrate a successful new experimental technique. Following the suggestions by Pontecorvo (1960) and by Schwartz (1960), a neutrino beam was constructed that led to the discovery of two different kinds of neutrino (see chapter 10). The same beamline technique is used in many neutrino experiments to this day. From this list, it is clear that the roles of experiment include far more than only the testing of theory. I do not, however, believe that this list of the varying roles that experiments play is exclusive or exhaustive. As we shall see, a single experiment can play several roles.

In that 1981 article, I also discussed various ways in which an experiment can be good. I distinguished between "conceptually important" experiments and "technically good" experiments. The former were classified primarily by their relationship to theory. Among the roles played by conceptually important experiments are testing theory, deciding between competing theories (crucial experiments), calling for a new theory, or demonstrating that an existing theory is incorrect.

I might distinguish here between the ways in which an experiment can be good and the role that it played from the attributes that it had. For example, the Michelson-Morley experiment (1887; see chapter 16) was not conceptually important in the genesis of Einstein's special theory of relativity. The results did, however, at least in principle, call for either a new theory or for a significant modification of existing theory. As Oliver Lodge, a supporter of the ether theory remarked, "This experiment might have to be explained away" (1893, 753). Similarly, Millikan's (1916a) measurement of Planck's constant (see chapter 6) is, in retrospect and in textbooks, regarded as providing strong support for Einstein's photon theory of light. At the time, however, it was regarded as confirming Einstein's photoelectric equation.

Technically good experiments are those that measure a quantity of physical interest with greater accuracy and precision than had been done previously. My use of "technically good" is meant to apply either to an experiment in which there have been previous measurements of the quantity of interest and/or to those experiments in which there has been a signifi-

cant advance in experimental technique. An illustration is Millikan's measurement of Planck's constant. In that episode there were several previous measurements and, as we shall see, Millikan made improvements to the experiment so that the measurement was more accurate and precise than any previous measurement. Because of that improvement he was also able to decide between different interpretations of the previous experimental results. Similarly, in Millikan's (1911) experiment to measure the charge of the electron (see chapter 7), he was able to make his measurements on a single oil drop, an innovation that allowed him to obtain a far better value for the electron charge. Previous experiments had only been able to obtain an average value from a cloud of drops.

There are, of course, instances of experiments that are both conceptually important and technically good. Gregor Mendel's experiments on hybridization in pea plants (see chapter 1) and Robert Millikan's measurement of the charge of the electron (see chapter 7) are two such examples.[2] In my earlier paper I also discussed "methodologically good" experiments, those that provided good reasons for belief in their results.

To these I would add "pedagogically important" experiments. These are experiments that play a didactic or explanatory role in textbooks, which they may or may not have played in the actual history. Examples are the Michelson–Morley experiment (see chapter 16), which textbooks often say played a significant role in the genesis of the special theory of relativity but, as we shall see, did not actually play such a role. On the other hand, Mendel's experiments (see chapter 1) established the basic laws of genetics and are extensively discussed in every introductory textbook on genetics that I have seen. I might suggest that other experiments, such as the Ellis–Wooster experiment (1927; see chapter 13), should be pedagogically important, but they are only infrequently mentioned.

I do not believe that either of these lists exhausts the roles that experiment plays in science or the ways in which an experiment can be good. What I will present is a number of examples of experiments that are good in various ways and that play different roles. It is in the details of experiments that we can observe and judge their quality.

If experiments are to play the important roles mentioned above then we must have good reasons to believe their results. I have previously argued that there exists an epistemology of experiment, a set of strategies that can be and are used to argue for the correctness of an experimental result (Franklin 2007, 220–25; Franklin 2002a, chap. 6). It is the use of these strategies that make an experiment methodologically good. 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 (the Sherlock Holmes strategy);[3]
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; and
10. Using "blind" analysis, a strategy for avoiding possible experimenter bias, by setting the selection criteria for "good" data independent of the final result.

This set of strategies is also neither exclusive nor exhaustive. No single strategy, or group of strategies, is necessary to argue for the correctness of an experimental result. Nevertheless, the use of such strategies is, I believe, necessary to establish the credibility of a result. I call experiments that do so "methodologically good" experiments. Most reports of experimental results do, in fact, include such arguments.[4]

"Conceptually important" experiments and "technically good" experiments must, of course, be methodologically good. An experiment cannot be important if we don't have good reasons to believe the result. Not all conceptually important experiments are, however, technically good. Sometimes even a rough measurement may be sufficient. Similarly a measurement may be technically good because it demonstrates that a quantity of interest can be measured. These are enabling experiments.[5] It is also true that experiments that may appear to be conceptually important at the time may not, however, be so in the long run.[6] Experiments sometimes disagree,

an indication that at least one of them must be incorrect. Because virtually all experiments do, I believe, apply the epistemological strategies discussed earlier, some of those applications must be incorrect (for illustrative cases see Franklin 2002a, chaps. 7–10).

I admit that the above lists are rather too dry and abstract. What are needed are the details of actual good experiments. As Stephen Jay Gould remarked, "I concentrate upon details . . . because I don't believe that important concepts should be discussed tendentiously in the abstract. . . . People, as curious primates, dote on concrete objects that can be seen and fondled. God dwells among the details, not in the realm of pure generality. We must tackle and grasp the larger, encompassing themes of our universe, but we make our best approach through small curiosities that rivet our attention—all those pretty pebbles on the shoreline of knowledge" (Gould 1989, 51–52). I will provide such details in the rest of this book.

Although contemporary high-energy physicists require a five-standard-deviation effect (five sigmas [σ]) before they will claim a discovery,[7] I do not believe that satisfying a fixed statistical criterion should be a requirement for a good experiment.[8] As noted earlier, a result needs only to be good enough for the intended purpose. Sometimes even a rough measurement may be sufficient. For example, the first measurement of the K_{e2}^{+} branching ratio, the fraction of all K^{+} mesons that decay into a positron and a neutrino, gave a result of $2.1^{+1.8}_{-1.3} \times 10^{-5}$ (Bowen et al. 1967). This showed that the quantity could be measured, albeit with a large experimental uncertainty. It also provided information about the mathematical form of the weak interaction responsible for this decay. The theoretical predictions for the K_{e2}^{+} branching ratio were explicit. If the interaction was pure axial vector (A) the predicted ratio of K_{e2}^{+} to $K_{\mu 2}^{+}$ decays was 2.6×10^{-5}, corresponding to a branching ratio of 1.6×10^{-5}. Pure pseudoscalar (P) coupling, on the other hand, predicted a K_{e2}^{+} to $K_{\mu 2}^{+}$ ratio of 1.02. If even only a small amount of pseudoscalar interaction were present, along with the dominant axial vector interaction, the K_{e2}^{+} branching ratio would be much larger. For example, adding only one part in a thousand of pseudoscalar interaction to the axial vector interaction would increase the expected branching ratio by a factor of four. Thus, even a rough measurement of the K_{e2}^{+} branching ratio would be a stringent test for the presence of any pseudoscalar interaction in the decay and of the V-A theory in general. The best previous measurement of the $K_{e2}^{+}/K_{\mu 2}^{+}$ ratio had set an upper limit of 2.6×10^{-3}, a factor of one hundred larger than that predicted by V-A theory. This experiment was good enough to help articulate the theory of weak interactions.[9]

"Good enough" is a criterion that may vary with both subject and time. In his mythical experiment on bodies falling from the Leaning

Tower of Pisa,[10] Galileo supposedly found only a hands-breadth difference in the fall of two bodies of very unequal weight. This demonstrated that the Aristotelian law of fall, which stated that the velocity of a falling body was proportional to its weight, was incorrect. Many experimental papers do not even cite statistics. William Wilson (1909; see chapter 12) demonstrated that the absorption of β rays was linear and not exponential by presenting two graphs. Similarly, Pevsner and collaborators (1961; see chapter 9) presented only a graph and the number of total and estimated background events as evidence for their discovery, although their results would have satisfied the current five-sigma criterion.[11]

Two other important experimental results in high-energy physics in the 1960s and 1970s did not use an explicit statistical criterion for a discovery claim. The 1964 discovery of the Ω^- hyperon (Barnes et al. 1964), an important confirmation of the quark model and the eight-fold way, demonstrated that the single observed event fit the expected mass, charge, and strangeness of the particle, the expected production mechanism and a complex decay mode, as evidence for the new particle. Implicitly, the experimenters were stating that such an event was unlikely to be produced by any background process. "In view of the properties of charge ($Q = -1$), strangeness ($S = -3$), and mass ($M = 1686 \pm 12 \rightarrow MeV/c^2$) established for particle 3, we feel justified in identifying it with the sought-for Ω^-" (Barnes et al. 1964, 206).[12]

The 1973 observation of a single neutrino-electron scattering event confirmed the existence of the weak-neutral currents predicted by the Weinberg–Salam unified theory of electroweak interactions. The experimenters estimated, using both measurements and calculations, that the expected background for such an event was 0.03 ± 0.02. This meant that the observation of even a single background event was very unlikely: "We conclude that the probability that the single event observed in the $\bar{\nu}$ film is due to non-neutral current background is less than 3%" (Hasert et al. 1973, 124). This is a far cry from the 2.9×10^{-5} percent probability of a five-sigma effect, or even the 0.27 percent probability of a three-sigma effect. It was, however, sufficient, at the time, to argue for the existence of weak-neutral currents. (To be fair, there was also evidence from another experiment performed at the same time, with the same bubble chamber, that also showed the presence of weak-neutral currents [for details see Galison 1987, chap. 4].)

It is interesting to note that the five-sigma criterion is now being applied in other fields of physics. The recent BICEP2 (Background Imaging of Cosmic Extragalactic Polarization) result, "BICEP2 I: Detection of B-mode Polarization at Degree Angular Scales" (Ade et al. 2014) noted that they found "an excess of B-mode power over the base lensed-ΛCDM

[the standard cosmological model] in the range $30 < l < 150$, inconsistent with the null hypothesis at a significance of $> 5\sigma$" (1). If correct, this is an important result that confirms both inflationary cosmology and the existence of gravitational waves "Although highly successful, the inflationary paradigm represents a vast extrapolation from well-tested regimes in physics. It invokes quantum effects in highly curved spacetime at energies near 10^{16} GeV and timescales less than 10^{-32} s. A definitive test of this paradigm would be of fundamental importance. Gravitational waves generated by inflation have the potential to provide such a definitive test" (Ade et al. 2014, 2). The experimenters remarked that inflation theory predicts the existence of gravitational waves that would produce a polarization pattern: "The detection of B-mode polarization of the CMB [Cosmic Microwave Background] at large angular scales would provide a unique confirmation of inflation" (2). Furthermore, the "observed B-mode power spectrum is well-fit by a lensed-ΛCDM + tensor theoretical model with tensor-scalar ratio $r = 0.20^{+0.07}_{-0.05}$, with $r = 0$ disfavored at 7σ" (1).[13]

A discussion of the five-sigma criterion also formed a significant part of the discussion as to whether gravitational waves had been observed by the Laser Interferometer Gravitational Wave Observatory (LIGO). The discussion was complicated by the fact that the observed signal might have been either a real signal or a blind injection, a simulated signal injected into the data stream to test whether the analysis procedures would detect such a signal (see Collins 2013 for the fascinating details).

I think it's worth stating again that, despite the increasing presence of the five-sigma criterion in physics, I do not believe that any fixed statistical criterion should be a necessary requirement for a good experiment.

In this book I will revisit the question "What makes a good experiment?" and will provide more extensive discussions of various exemplars of the types of good experiment discussed above. These discussions will include the stated purpose of the experiment, how the results were used, and the arguments given for the correctness of the results. My aim is to provide a better and more extensive answer to that question than I did in 1981.

The set of experiments included in this book is not intended as a compilation of the eighteen best experiments. One of them, Peter Thieberger's (1987a, 1987b) experiment on the Fifth Force (see chapter 17), produced a result that is generally regarded as incorrect. Nevertheless, I will argue that it is still a good experiment. These experiments were selected because they illustrated both the various ways in which an experiment can be good and because they also illustrate the different roles that experiment can play.

In presenting the experiments chosen, I have organized them into five parts, according to what I believe is their primary role in science. These

are: Conceptually important experiments, those that lead to significant changes in theory; Experiments that measure a quantity of importance; Experiments that provide evidence for entities; Experiments that provide a solution to a vexing problem; and Experiments that measure nothing, null experiments. In the discussions of the experiments, I will try, as much as possible, to allow the scientists to speak for themselves so that the reader will have the original thoughts and not a latter-day interpretation.