Introduction

The history of the changing standards for scientific discovery in particle physics discussed in the prologue suggests that it might be interesting to investigate whether other aspects of experiment and the reporting of experimental results in that field have changed with time. If we look at the history of the use of statistics, we find considerable additional change. As Jed Buchwald (2006) remarked in his discussion of the treatment of discrepant experimental results, "well into the eighteenth century experimenters chose to publicize that single golden number which they deemed to be the very best one of all the values that their labor had produced" (566). This statement of results changed over the next several centuries into more standardized treatments of data. Initially, other techniques were sometimes used to deal with discrepant results. For example, some scientists, believing that, as they proceeded with a measurement, the later results were better and more reliable than the earlier ones, used a procedure in which they first took the mean of the first two measurements. They then proceeded to take the mean of this first mean with the third measurement and so on. This procedure had the effect of more heavily weighting the later, and presumably better, measurements. We can see an illustration of the increasing reliability of measurements if we examine Robert Millikan's measurements of the charge of the electron as a function of time (figure I.1). Although Millikan did not use this procedure to calculate his value of the charge of the electron, we see that the spread of the values becomes smaller and that the values converge as he made more measurements. The percentage of events included in his published paper also increased with time.

Isaac Newton was, if not unique, certainly rare in his use of the actual mean of a set of results as the best value, although one could not provide a mathematical justification for this until the development of probability

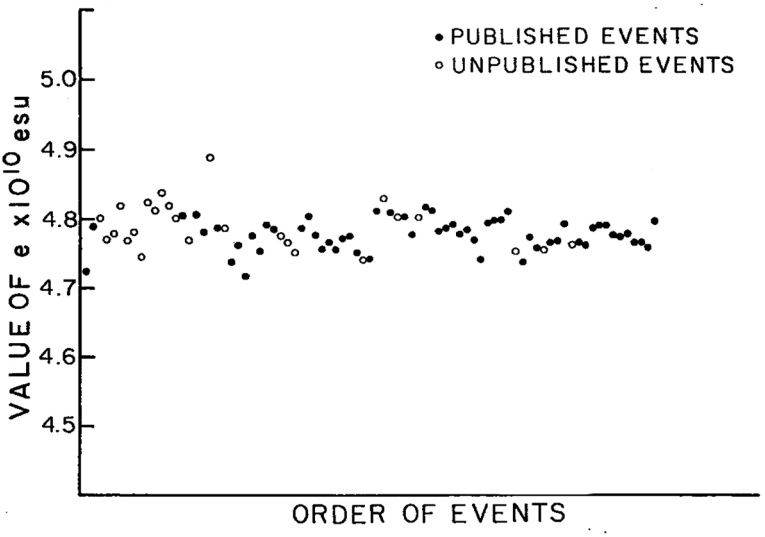

Figure I.1. Millikan's values for *e*, the charge of the electron, as a function of time.

theory and least-squares fitting in the early nineteenth century. This standardization in analyzing results was strengthened in the late nineteenth century with the use of the *standard* deviation and the invention of the χ^2 test by Karl Pearson. The word "standard" is important here. By analyzing data with the same technique, scientists could compare experimental results or at least see whether there was intersubjective agreement. Experimenters throughout the period studied in this book are consistent in providing estimates or calculations of experimental uncertainty,[1] although the mathematical techniques used vary considerably. Thus, Kennelly and Fessenden (see chapter 1), although not providing an uncertainty, presented the maximum and minimum values obtained, which gives an estimate of the experimental uncertainty. Hall and others provide an estimate based on the calculation of the probable error or the standard deviation. The prologue shows the use of standard deviations as both a measure of uncertainty and of significance. More recent experiments calculate the maximum likelihood of the distribution of events given various hypotheses and use Bayesian decision trees, neural nets, and multivariant analyses in presenting an experimental result, estimating the uncertainty, and providing an estimate of the significance of a result.

The use of statistics is only one of the important issues involved in the presentation of experimental results. It seems worthwhile to examine these other issues to see whether they have also changed over time. In this book I examine several of these issues by looking primarily at papers published

in *Physical Review,* which began as the journal of record for the American Physical Society and is now one of the major archival journals.[2] I begin with a paper published in 1894 (volume 1 of *Physical Review* was published in 1893) and continue up to the present, looking at papers at approximately ten-year intervals. Such a history can provide only snapshots of experimental practice at a given time, but, like snapshots of a vacation trip, they can be of value. These snapshots will also provide a feel for the practice of experimental science and for the style of scientific papers at various times. I will discuss, almost exclusively, experiments concerned with elementary particles and their properties. This is because some of the issues, particularly those of scale, are most apparent in that field, although the other issues discussed also apply to other areas of experimental physics.[3] The issues to be discussed include:

1. Exclusion of data and selection of data. These are not the same procedure. Exclusion typically applies to "bad" data, data taken when the apparatus is not working properly. Selection involves "good" data, in which the apparatus is working properly and in which selection criteria have to be applied in order to eliminate background that might mask or mimic the phenomenon under investigation. The exclusion of data is mentioned explicitly in some of the early papers we will study. The process no doubt occurs in later experiments, but it is usually omitted from the papers. The later papers do, however, include discussion of the selection procedures or cuts. We saw this clearly in the discussion of single-top-quark production in the prologue, and we will see it in later discussions. As Peter Galison (1987) has discussed, the elimination or minimization of background is central to modern high-energy physics experiments.

2. Possible experimenter bias. The use of selection criteria raises the possibility that an experimenter may tune the cuts to produce a desired result when the effect of those cuts on the final result is known.[4] That result might be in agreement with existing theory, the experimenter's presuppositions about the phenomenon, or with previous results. We will see examples of all of these in our history.[5]

3. Details of the experimental apparatus. Experimental papers also provide descriptions of both the experimental apparatus and of the procedures used to analyze the data. In early papers the descriptions of the apparatus are quite detailed. In contemporary papers, in which the scale of the experiments has increased dramatically, the description provided in both letters and in short papers is quite limited, usually restricted to a brief discussion of those parts of the apparatus crucial for making the measurement. A full description of the experimental apparatus is provided elsewhere, as are details of the analysis procedures. In some papers there

is, in fact, no description of the apparatus, only a reference to the more detailed account.

The BaBar experiment discussed in chapter 17 cites a 116-page paper describing the apparatus along with a 54-page paper describing the Monte Carlo simulation. For the Collider Detector at Fermilab (CDF), just the overview of the experimental apparatus covered 17 pages in *Physical Review D,* with references to 24 other papers that gave the details of various parts of the apparatus.[6] The experimenters also remarked that for even more detail one should consult the appropriate Fermilab technical reports. This is not to say that there is no interest in the details of the apparatus, but rather that, given the extensive documentation already available, it would be wasteful and unnecessary to include it in every paper published by a group. These large experimental groups may publish fifty or more papers per year.

Early papers are also quite detailed about the procedures used in performing the experiment.

4. The size of the experimental apparatus, the size of the data set, and the number of authors. The explosion of detail, mentioned above, is related to the changes in both the size and complexity of current experiments. This is apparent when we look at the physical size of the experimental apparatuses. Millikan's oil-drop apparatus fit on a table top (figure I.2) and had a volume of approximately 1 m^3. In contrast, the Compact Muon Solenoid (CMS) apparatus (figure I.3) has a volume of approximately 4,000 m^3.[7] Millikan was the sole author of his paper. CMS papers have 2,000 or more authors. Thus, the volume per experimenter has remained approximately constant at about 1 m^3/experimenter.

The increasing complexity of experiments has also increased the number of authors of a paper. Until the early 1950s we typically find one or two authors per paper. The number of authors in high-energy physics papers has gradually increased so that the CMS collaboration at the Large Hadron Collider has almost 3,000 members.[8] This has changed the meaning of what it means to be an author of an experimental paper. A personal anecdote may help here. In 1958, when I was an undergraduate, I worked as an assistant to Eugene Commins on an experiment to measure the nuclear spin of He^6. I assisted in setting up the apparatus, taking the data, and I performed most of the numerical calculations required to obtain the final result. Gene generously offered to allow me to be a coauthor of the paper. A few days later, Polykarp Kusch, the senior author of the paper, called me into his office. He told me that although I had done substantial and valuable work on the experiment I could not be an author of the paper. His reason was that I did not have sufficient knowledge to give a talk about the

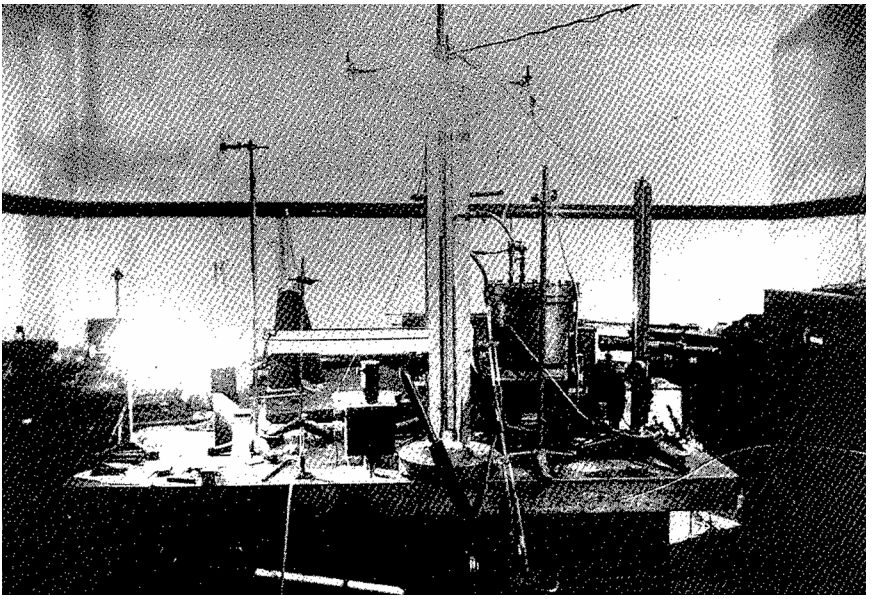

Figure I.2. Millikan's oil-drop apparatus. Courtesy California Institute of Technology Archives.

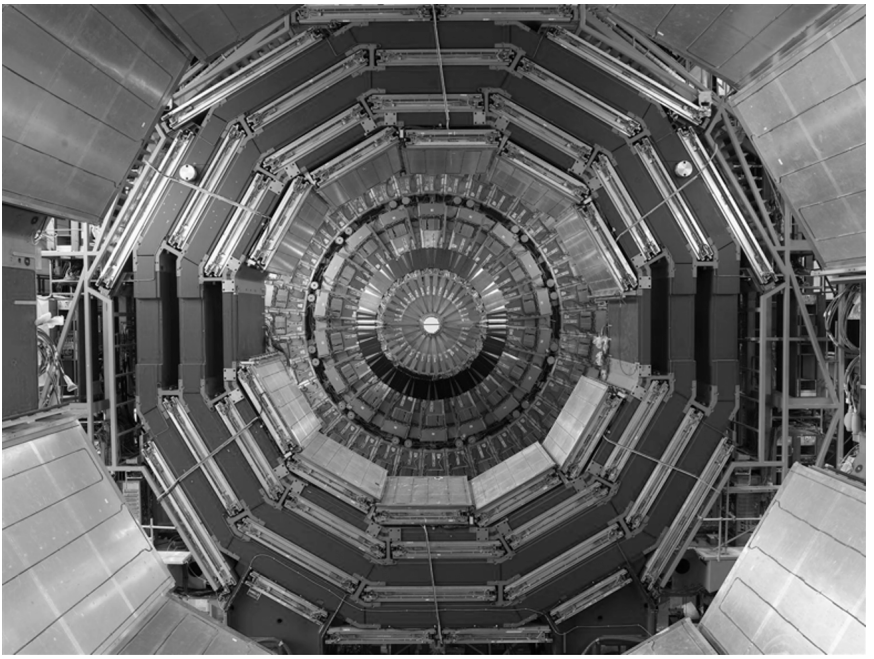

Figure I.3. The Compact Muon Solenoid.

experiment. Although I was disappointed, I could not disagree. It seemed to me, both then and now, to be a reasonable criterion for authorship. In a large contemporary experimental collaboration such a criterion would be far too stringent. It would eliminate a large number of people who had made substantial contributions to the experiment. For example, the analysis of a particular data set obtained in the experiment is usually performed by a relatively small segment of a collaboration, typically 5 to 20 physicists. Thus a vast majority of the collaboration would not know the details of the analysis sufficiently well to be able to give a talk on the result. Yet they may have made considerable contribution to the construction of the apparatus, to the analysis procedures, or to the running of the experiment. There are now rules for authorship. The CMS constitution states that "the authors of CMS physics papers are the physicists, engineers and graduate students who are affiliated with a member institute of CMS and who have spent a significant fraction of their working time for CMS for at least one year since the date of registration with the CMS Secretaria." My colleagues who are members of the CMS have told me that that work must be service to the collaboration and may include helping to construct the apparatus and working on general analysis computer programs. It does not include working on the analysis of data in order to produce a particular experimental result. There are also requirements about how many experimental shifts a group within the collaboration must cover each year. It now takes a village, and one of reasonable size, to do a high-energy physics experiment.

Experiments have not only increased in physical size but also in the size of the data sample collected. This is clear if we look at the data sets for some of the experiments discussed in this book. Millikan (chapter 3) took data on 175 oil drops, of which he published those for 58 drops, only 23 of which were used to determine the value of *e,* the charge of the electron. Alford and Leighton (chapter 8) obtained 134 V^0 events out of a sample 23,000 photographs, of which 74 were used in their measurement of the V^0 lifetime. The K_{e2}^+ branching ratio experiment published in 1967 (chapter 10), had 16,965 events, with a final signal of 6 events. The E791 experiment at Fermilab had 20 billion triggers,[9] and the BaBar experiment produced 467 million B meson pairs.

In order to process such huge amounts of data there must also have been very large improvements in both data taking and in the analysis of data. Contrast this with the experiment on which I did my doctoral research, the photoproduction of ρ^0 mesons (Franklin et al. 1964). The experiment used optical spark chambers, and the data were recorded on film. The data taking rate was limited by the fact that the film in the camera had to move before another event could be recorded, a process lasting on the order of a second. We obtained approximately 12,000 events. The spark chamber

images were then projected onto graph paper. The positions of the sparks were measured, and it took approximately a minute to measure the position of the sparks and to record the data by hand. The data were later put on punch cards and analyzed by computer.[10] At that rate it would have taken approximately 40,000 years to analyze all the triggers obtained by E791, even if the events were as simple as those in the ρ^0 experiment, which was certainly not the case. Clearly the improvements in both data taking and analysis have occurred. Data for contemporary high-energy physics experiments are recorded digitally, and the rate of recorded events can be as large as a few hundred events per second (for an interesting history of the development of some of the new techniques in high-energy physics, see Galison [1997]). As discussed in the conclusion, the rate of events produced in the CMS experiment is 800 MHz.

5. Such data taking rates, along with the analysis of such data, are made possible by advances is electronics and computers. Millikan, for example, used tables of logarithms for his calculations, which were done by hand. By the 1950s, computers, although primitive by contemporary standards, were often used. Their use has increased considerably since then, and their computational power has increased by orders of magnitude.[11] I know of no convenient way to estimate the increase in computing power, but one commentator has noted that if the cost of automobiles had decreased at the same rate as the cost per computer byte, then the price of a new Mercedes Benz would be $0.25. I suspect this is actually an underestimate for the increase in computing power.

6. Distinction between ideal and actual experiments. This takes material form in the descriptions of experiments and in the figures of the apparatus presented.[12] As we shall see, figures of the experimental apparatus become more and more abstract rather than realistic. Millikan's diagram of his apparatus is considerably more realistic than the figure of the apparatus used to search for neutral particles described in chapter 16.

7. History of previous experiments. In papers at the turn of the twentieth century, the authors present a history, sometimes quite extensive,[13] of previous measurements of the same quantity. In more contemporary papers the historical accounts are far shorter, with the exception of review papers or those on controversial subjects. In addition, at least in high-energy physics, the previous history is usually quite short. Thus, in the episode of single-top-quark production, the history spans only from 2000 to 2009 and comprises only nine papers. This number includes five papers by the CDF and D0 collaborations that set limits on the production. If one includes only those papers that report evidence for the production, then the time period is from 2007 to 2009 and includes only four papers. These

experiments could, at the time, only be done by the two experimental groups. There wasn't a lot of history to report.[14]

8. Personal comments and style. In the early papers the authors make judgments about the quality of previous work and, at least on one occasion, about the character of an experimenter. Thus, Edwin Hall remarked that, "moreover, Hooke, a brilliant genius but a somewhat uncertain character, had committed himself in the most open way to the opinion that experiment would reveal a southerly deviation. In a man of his reputation such a bias is not to be overlooked; and yet it is hard to believe that he deliberately lied to his associates in the Royal Society" (Hall 1903a, 182).[15] Such a comment would, I believe, be unthinkable in a contemporary paper. Even in cases of severe controversy, authors may argue that other results are incorrect, but they do so, at least in the published work, rather politely.[16] In private discussions, or in more informal venues, the discussions can sometimes be quite sharp (see the discussion section in the conclusion).

In reading the early papers such as those of Kennelly and Fessenden and of Hall, one gets the feeling that the papers were written by individuals. Later papers are much more generic, tend to use a passive voice, and lack personal style and comments.[17] As John Ziman (1968) noted in *Public Knowledge*, modern papers are written as though they are already part of an archive, as though they are a permanent contribution to science. The reports of experiments also seem to have become more idealized. As we shall see, the later papers have almost no discussion of the experimental apparatus and few, if any, details of the analysis procedures. What is presented are the results.

There are two other issues for which I expect the discussions to remain relatively constant throughout this period. The first involves the many roles of experiment in science. One of its important roles is to test theories and to provide the basis for scientific knowledge. It can also call for a new theory, either by showing that an accepted theory is incorrect or by exhibiting a new phenomenon that needs explanation. Experiment can provide hints toward the structure or mathematical form of a theory and provide evidence for the existence of the entities involved in our theories. There are also exploratory experiments in which a subject of interest is investigated to try to formulate a theory.[18] Experiment can also measure quantities that theory tells us are important or those that have practical importance. Finally, it may also have a life of its own, independent of theory. Scientists may investigate a phenomenon just because it looks interesting, and this will also provide evidence for a future theory to explain. We shall see that experiments discussed play one, or sometimes more than one, of these roles.

If experiment is to play these important roles in science, then we must

have good reasons to believe experimental results. I outline below an epistemology of experiment, a set of strategies that provides reasonable belief in experimental results.[19] In discussing the papers in the chapters that follow, I will examine the arguments offered for the correctness of the experimental results and see whether they match the strategies offered in the epistemology of experiment. Scientific knowledge can then be reasonably based on these experimental results.

It has been more than three decades since Ian Hacking (1981) asked, "Do we see through a microscope?" Hacking's question really asked how do we come to believe in an experimental result obtained with a complex experimental apparatus. How do we distinguish between a valid result and an artifact created by that apparatus? Hacking (1983) provided an extended answer to these questions in the second half of *Representing and Intervening*. He argued that in making observations with a microscope the experimenters intervened. They manipulated the object under observation and predicted what they would observe if the apparatus was working properly. Observing the predicted effect strengthens belief in both the proper operation of the microscope and in the observation. Hacking also discussed the strengthening of one's belief in an observation by independent confirmation.

Hacking's answer is correct as far as it goes. It is, however, incomplete. What happens when one can perform the experiment with only one type of apparatus, such as an electron microscope or a radio telescope, or when intervention is either impossible or extremely difficult? Other strategies are needed to validate the observation. These include

1. Experimental checks and calibration, in which the experimental apparatus reproduces known phenomena. If the check is successful, it provides good reason to believe that the apparatus is working properly and for belief in the result produced. If the check fails, then we have good reason to question the results obtained with that apparatus.

2. Reproducing artifacts that are known in advance to be present. In a sense this is a form of an experimental check.

3. Elimination of plausible sources of error and alternative explanations of the result (the Sherlock Holmes strategy).[20]

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. The support for the theory feeds through to support for the result.

6. Using an apparatus based on a well-corroborated theory. In this case the support for the theory passes on to the apparatus based on that theory.

7. Using statistical arguments. Here one argues that the effect is very unlikely if it is a statistical fluctuation of the background. We have seen this illustrated in the prologue.

8. Using "blind" analysis, a strategy for avoiding possible experimenter bias, by setting the selection criteria independent of the final result (see chapter 6).

These strategies along with Hacking's intervention and independent confirmation provide an epistemology of experiment. I should emphasize that these strategies are neither exclusive nor exhaustive. No one strategy, or even a combination of them, is necessary or sufficient for establishing the correctness of an experimental result. Nor do all of the strategies have equal weight. It depends on the particular experiment. Scientists use those strategies most appropriate for the particular experiment to establish the correctness of their result.

The papers discussed in this book are not a randomly selected sample. Several of them (Millikan's measurement of the charge of the electron, Compton's work on the effect that bears his name, and the discovery of the neutrino by Reines and Cowan) are included because they are historically important experiments. Hall's papers are included because I found them fascinating. All of the papers do, I believe, tell us about the practice of experimental particle physics of their time and allow us to look at any possible changes in that experimental practice.