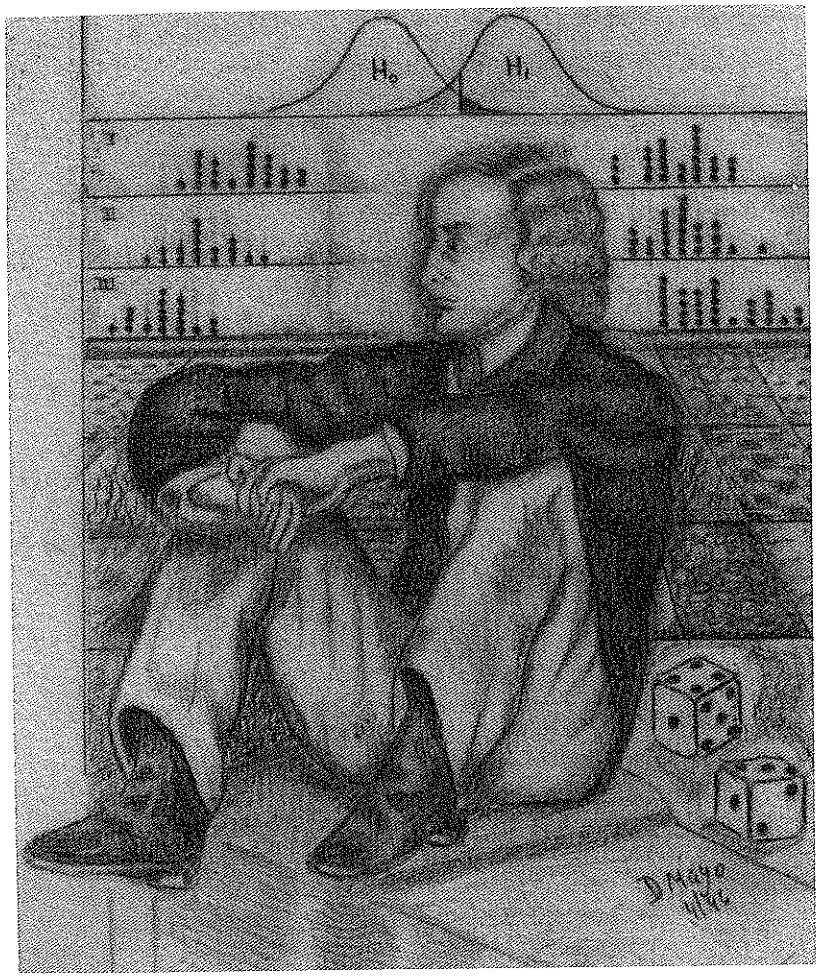
SCIENCE AND ITS CONCEPTUAL FOUNDATIONS

A series edited by David L. Hull



"I might recall how certain early ideas came into my head
as I sat on a gate overlooking an experimental blackcurrant plot . . . ."
—E. S. Pearson, "Statistical Concepts in Their Relation to Reality"

Deborah G. Mayo

Error and the Growth of Experimental Knowledge

THE UNIVERSITY OF CHICAGO PRESS

Chicago and London

for helping to accommodate my leaves and for endorsing this project.

Portions of chapters 3, 8, 11, and 12 have appeared in previously published articles; I thank the publishers for permission to use some of this material: "The New Experimentalism, Topical Hypotheses, and Learning From Error," in *PSA 1994*, vol. 1, edited by D. Hull, M. Forbes, and R. Burian (East Lansing, Mich.: Philosophy of Science Association, 1994), 270–79; "Novel Evidence and Severe Tests," *Philosophy of Science* 58 (1991): 523–52; "Did Pearson Reject the Neyman-Pearson Philosophy of Statistics?" *Synthese* 90 (1992): 233–62; and "The Test of Experiment: C. S. Peirce and E. S. Pearson," in *Charles S. Peirce and the Philosophy of Science: Papers from the 1989 Harvard Conference*, edited by E. Moore (Tuscaloosa, Ala.: University of Alabama Press, 1983), 161–74.

I owe special thanks to Susan Abrams and the University of Chicago Press for supporting this project even when it existed only as a ten-page summary, and for considerable help throughout. I thank David Hull for his careful and instructive review of this manuscript, Madeleine Avirov for superb copyediting, and Stacia Kozlowski for a good deal of assistance with the manuscript preparation.

I obtained valuable feedback on this manuscript from the graduate students of my seminar Foundations of Statistical Inference in 1995, especially Mary Cato, Val Larson, Jean Miller, and Randy Ward. For extremely valuable editorial and library assistance over several years, I thank Mary Cato. For organizational help and superb childcare, I am indebted to Cristin Brew and Wendy Turner.

I am grateful to my father, Louis J. Mayo, for first sparking my interest in philosophy as a child; I regret that he passed away before completion of this book. I am thankful to my mother, Elizabeth Mayo, for understanding my devotion to this project, and to my son, Isaac, for not minding using the backs of discarded drafts as drawing paper for five years. My deepest debt is to my husband, George W. Chatfield, to whom this book is dedicated.

CHAPTER ONE

Learning from Error

> The essays and lectures of which this book is composed are variations upon one very simple theme—the thesis that *we can learn from our mistakes.*
>
> Karl Popper, *Conjectures and Refutations*, p. vii

WE LEARN FROM OUR MISTAKES. Few would take issue with this dictum. If it is more than merely a cliché, then it would seem of interest to epistemologists to inquire how knowledge is obtained from mistakes or from error. But epistemologists have not explored, in any serious way, the basis behind this truism—the different kinds of mistakes that seem to matter, or the role of error in learning about the world. Karl Popper's epistemology of science takes learning from error as its linchpin, as the opening to his *Conjectures and Refutations* announces. In his deductive model the main types of error from which scientists learn are clashes between a hypothesis and some experimental outcome in testing. Nevertheless, Popper says little about what positive information is acquired through error other than just that we learn an error has somewhere been made. Since a great many current approaches take Popper's problems as their starting place, and since I too make learning from error fundamental, I begin by pursuing this criticism of Popper.

1.1 POPPERIAN LEARNING THROUGH FALSIFICATION

For the logical empiricists, learning from experiment was a matter of using observations to arrive at inductive support for hypotheses. Experimental observations were viewed as a relatively unproblematic empirical basis; the task for philosophers was to build inductive logics for assigning degrees of evidential support to hypotheses on the basis of given statements of the evidence. Popper questioned the supposition that experimental data were unproblematic and denied that learning is a matter of building up inductive support through confirming in-

stances of hypotheses. For Popper, learning is a matter of deductive falsification. In a nutshell, hypothesis *H* is deductively falsified if *H* entails experimental outcome *O*, while in fact the outcome is ~*O*. What is learned is that *H* is false.

Several familiar problems stand in the way of such learning. Outcome *O*, itself open to error, is "theory-laden" and derived only with the help of auxiliary hypotheses. The anomaly cannot be taken as teaching us that *H* is false because it might actually be due to some error in the observations or one of the auxiliary hypotheses needed to derive *O*. By means of the *modus tollens*, Popper remarks, "we falsify *the whole system* (the theory as well as the initial conditions) which was required for the deduction of [the prediction], i.e., of the falsified statement" (Popper 1959, 76). We cannot know, however, which of several auxiliary hypotheses is to blame, which needs altering. Often *H* entails, not a specific observation, but a claim about the probability of an outcome. With such a statistical hypothesis *H*, the nonoccurrence of an outcome does not contradict *H*, even if there are no problems with the auxiliaries or the observation.

As such, for a Popperian falsification to get off the ground, additional information is needed to determine (1) what counts as observational (and to decide which observations to accept in a particular experiment), (2) whether auxiliary hypotheses are acceptable and alternatives are ruled out, and (3) when to reject statistical hypotheses. Only with (1) and (2) does an anomalous observation *O* falsify hypothesis *H*, and only with (3) can statistical hypotheses be falsifiable. Because each determination is fallible, Popper and, later, Imre Lakatos regard their acceptance as decisions, driven more by conventions than by experimental evidence.

Lakatos sets out to improve Popper by making these (and other) decisions explicit, yielding "sophisticated methodological falsificationism." Nevertheless, Lakatos finds the decisions required by a Popperian falsificationist too risky, claiming that "the risks are daring to the point of recklessness" (Lakatos 1978, 28), particularly decision 2, to rule out any factors that would threaten auxiliary hypotheses. Says Lakatos: "When he tests a theory together with a *ceteris paribus* clause and finds that this conjunction has been refuted, he must *decide* whether to take the refutation also as a refutation of the specific theory. . . . Yet the decision to 'accept' a *ceteris paribus* clause is a very risky one because of the grave consequences it implies" (ibid., 110). Once the decision is made to reject alternative auxiliary factors, a mere anomaly becomes a genuine falsifier of the theory itself.

Lakatos regards such a decision as too arbitrary. Accepting what is often referred to as the Duhem-Quine thesis, that "no experimental result can ever kill a theory: any theory can be saved from counter-instances either by some auxiliary hypothesis or by a suitable reinterpretation of its terms" (Lakatos 1978, 32), Lakatos believes that an appeal procedure by which to avoid directing the *modus tollens* against a theory is always available.

Moreover, Lakatos, like Thomas Kuhn, finds Popper's picture of conjecture and refutation too removed from actual science, which often lives with anomalies and contains not just falsifications but also confirmations. Attempting to save something of Popper while accommodating Kuhn, Lakatos erects his "methodology of scientific research programmes." Lakatos suggests that there is a *hard core* against which *modus tollens* is not to be directed. In the face of inconsistency or anomaly, we may try to replace any auxiliaries outside this core (the "protective belt"), so long as the result is progressive, that is, predicts some novel phenomena. Determining the progressiveness of the theory change requires us to look not at an isolated theory, but at a series of theories—a *research program*. However, this holistic move does not really solve Popper's problem: as Lakatos recognizes, it enables any theory or research program to be saved—with sufficient genius it may be defended progressively, even if it is false (Lakatos 1978, 111). The cornerstone of the Popperian doctrine against saving theories from falsification is overturned. While Lakatos, like Popper, had hoped to avoid conventionalism, his solution results in making the growth of knowledge even more a matter of convention than did Popper's decisions. It is the unquestioned authority of the conventionally designated hard core, and not "the universe of facts," that decides where to direct the arrow of *modus tollens*. In Lakatos's view:

> The direction of science is determined primarily by human creative imagination and not by the universe of facts which surrounds us. Creative imagination is likely to find corroborating novel evidence even for the most "absurd" programme, if the search has sufficient drive. . . . A brilliant school of scholars (backed by a rich society to finance a few well-planned tests) might succeed in pushing any fantastic programme ahead, or, alternatively, if so inclined, in overthrowing any arbitrarily chosen pillar of "established knowledge." (Lakatos 1978, 99–100)

But let us pull back and recall the problems that set Lakatos off in the first place. Affirming experimental data? Ruling out alternative auxiliaries? Falsifying statistical claims? Why not see if there may not be perfectly good grounds for warranting the information that these tasks require without resorting to conventional decisions in the first place. This is where my project begins.

The key is to erect a genuine account of learning from error—one that is far more aggressive than the Popperian detection of logical inconsistencies. Although Popper's work is full of exhortations to put hypotheses through the wringer, to make them "suffer in our stead in the struggle for the survival of the fittest" (Popper 1962, 52), the tests Popper sets out are white-glove affairs of logical analysis. If anomalies are approached with white gloves, it is little wonder that they seem to tell us only that there is an error somewhere and that they are silent about its source. We have to become shrewd inquisitors of errors, interact with them, simulate them (with models and computers), amplify them: we have to learn to make them talk. A genuine account of learning from error shows where and how to justify Popper's "risky decisions." The result, let me be clear, is not a filling-in of the Popperian (or the Lakatosian) framework, but a wholly different picture of learning from error, and with it a different program for explaining the growth of scientific knowledge.

1.2 DAY-TO-DAY LEARNING FROM MISTAKES

The problem of learning from error in the sense of Popperian falsification, say Lakatos and others, is that learning from error itself is fraught with too much risk of error. But what grounds are there for thinking that such possible errors are actually problematic? How do scientists actually cope with them? It is not enough that mistakes are logically possible, since we are not limited to logic. Unless one is radically skeptical of anything short of certainty, specific grounds are needed for holding that errors actually occur in inquiries, that they go unnoticed, and that they create genuine obstacles to finding things out. No such grounds have been given. If we just ask ourselves about the specific types of mistakes we can and do make, and how we discover and avoid them—in short, how we learn from error—we would find that we have already taken several steps beyond the models of both Popper and Lakatos. Let me try to paint with broad brush strokes the kinds of answers that I think arise in asking this question, with a promise to fill in the details as we proceed.

1. After-trial checking (correcting myself). By "after-trial" I mean after the data or evidence to be used in some inference is at hand. A tentative conclusion may be considered, and we want to check if it is correct. Having made mistakes in reaching a type of inference in the past, we often learn techniques that can be applied the next time to check if we are committing the same error. For example, I have often discovered I was mistaken to think that A caused B when I found that B occurs

even without A. In a subsequent inference about the effect of some factor F, I may deliberately consider what occurs without F in order to check this mistake. Other familiar after-trial checks are the techniques we develop for checking complex arithmetic operations or for balancing checkbooks.

In addition to techniques for catching ourselves in error are techniques for correcting errors. Especially important error-correcting techniques are those designed to go from less accurate to more accurate results, such as taking several measurements, say, of the length of wood or fabric, and averaging them.

2. Before-trial planning. Knowledge of past mistakes gives rise to efforts to avoid the errors ahead of time, before running an experiment or obtaining data. For example, teachers who suspect that knowing the author of a paper may influence their grading may go out of their way to ensure anonymity before starting to grade. This is an informal analogue to techniques of astute experimental design, such as the use of control groups, double blinding, and large sample size.

3. An error repertoire. The history of mistakes made in a type of inquiry gives rise to a list of mistakes that we would either work to avoid (before-trial planning) or check if committed (after-trial checking), for example, a list of the familiar mistakes when inferring a cause of a correlation: Is the correlation spurious? Is it due to an extraneous factor? Am I confusing cause and effect? More homely examples are familiar from past efforts at fixing a car or a computer, at cooking, and the like.

4. The effects of mistakes. Through the study of mistakes we learn about the kind and extent of the effect attributable to different errors or factors. This information is then utilized in subsequent inquiries or criticisms. One such use is to rule out certain errors as responsible for an effect. Perhaps putting in too much water causes the rice to be softer but not saltier.

Knowledge of the effects of mistakes is also exploited to "subtract out" their influences after the trial. If the effects of different factors can be sufficiently distinguished or subtracted out later, then the inferences are not threatened by a failure to control for them. Thus knowing the effects of mistakes is often the key to justifying inferences. In chapter 7 we will see how Jean Perrin debunked an allegation that his results on Brownian motion were due to temperature variations in his experiment. Such variations, he showed, only caused a kind of current easily distinguishable from Brownian motion.

5. Simulating errors. An important way to glean information about the effects of mistakes is by utilizing techniques (real or artificial) to display what it would be like if a given error were committed or a given factor were operative. Observing an antibiotic capsule in a glass of water over several days revealed, by the condition of the coating, how an ulceration likely occurred when its coating stuck in my throat. In the same vein, we find scientists appealing to familiar chance mechanisms (e.g., coin tossing) to simulate what would be expected if a result were due to experimental artifacts.

6. Amplifying and listening to error patterns. One way of learning from error is through techniques for magnifying their effects. I can detect a tiny systematic error in my odometer by driving far enough to a place of known distance. I can learn of a slight movement across my threshold with a sensitive motion detector. Likewise, a pattern may be gleaned from "noisy" data by introducing a known standard and studying the deviations from that standard. By studying the pattern of discrepancy and by magnifying the effects of distortions, the nature of residuals, and so forth, such deviations can be made to speak volumes.

7. Robustness. From the information discussed above, we also learn when violating certain recommendations or background assumptions does not pose any problem, does not vitiate specific inferences. Such outcomes or inferences are said to be robust against such mistakes. These are the kinds of considerations that may be appealed to in answering challenges to an inference. In some cases we can argue that the possibility of such violations actually strengthens the inference. (For example, if my assumptions err in specific ways, then this result is even *more* impressive.) An example might be inferring that a new teaching technique is more effective than a standard one on the basis of higher test scores among a group of students taught with the new technique (the treated group) compared with a group of students taught with the old (the control group). Whereas an assumption of the study might have been that the two groups had about equal ability, discovering that the treated group was actually less able than the control group before being taught with the new technique only strengthens the inference.

An important strategy that may be placed under this rubric is that of deliberately varying the assumptions and seeing whether the result or argument still holds. This often allows for the argument that the inference is sound, despite violations, that inaccuracies in underlying factors cannot be responsible for a result. For, were they responsible we would not have been able to consistently obtain the same results despite variations.

8. Severely probing error. Points 1 through 7 form the basis of learning to detect errors. We can put together so potent an arsenal for unearthing a given error that when we fail to find it we have excellent grounds for concluding that the error is absent. Having failed to detect a given infection with several extremely reliable blood tests, my physician infers that it is absent. The "error" inferred to be absent here is declaring that there is no infection when there is one.

The same kind of reasoning is at the heart of experimental testing. I shall call it *arguing from error.* After learning enough about certain types of mistakes, we may construct (often from other tests) a testing procedure with an overwhelmingly good chance of revealing the presence of a specific error, if it exists—but not otherwise. Such a testing procedure may be called a *severe (or reliable) test,* or a *severe error probe.* If a hypothesized error is not detected by a test that has an overwhelmingly high chance of detecting it, if instead the test yields a result that accords well with no error, then there are grounds for the claim that the error is absent. We can infer something positive, that the particular error is absent (or is no greater than a certain amount). Equivalently, we have grounds for rejecting the hypothesis, H', that the error is present, and affirming H, that it is absent. When we have such information, we say that H has passed a severe test. Alternatively, we can say that the test result is a *good indication* that H is correct.

Is it possible for such humdrum observations to provide a fresh perspective from which to address problems that still stand in the way of a satisfactory epistemology of science? I propose that they can, and that is the underlying thesis of this book.

To turn the humdrum observations into tools for experimental learning, they need to be amplified, generalized, and systematized. As I see it, this is the chief task of an adequate epistemology of experiment. I understand "experiment," I should be clear at the outset, far more broadly than those who take it to require literal control or manipulation. Any planned inquiry in which there is a deliberate and reliable argument from error may be said to be experimental.

How can these day-to-day techniques for learning from error take us beyond Popper's deductive falsification model?

1.3 ACCENTUATE THE POSITIVE, ELIMINATE THE NEGATIVE

I endorse many of Popper's slogans. Like Popper's, the present approach views the growth of knowledge as resulting from severe criticism—from deliberately trying to find errors and mistakes in hypotheses. It likewise endorses his idea that learning about a hypothesis is based on finding out whether it can withstand severe tests. Each of

these slogans, however, is turned into a position where something positive is extracted from the severe criticism; for us, the focus is on *constructive criticism*, on learning from criticizing. It seems incumbent upon anyone mounting such an approach to dispel the ghosts of Popper's negativism right away, or at least sketch how they will be dispelled.

The most devastating criticism of Popper's approach is this: having rejected the notion that learning is a matter of building up probability in a hypothesis, Popper seems to lack any meaningful way of saying why passing severe tests counts in favor of a hypothesis.[1] Popper's account seems utterly incapable of saying anything positive. There are two variants to this criticism, which I shall take up in turn.

a. Why Should Passing Severe Tests Count in Favor of Hypotheses?

If the refuted hypothesis is rejected for one that passes the test it failed, that new hypothesis, Popper says, is preferable. But why should it be preferred? What is it that makes it better? The most Popper can say on its behalf is that it did better in passing the test the previous hypothesis failed and that "it will also have to be regarded as possibly true, since at the time t it has not been shown to be false" (Popper 1979, 14). Popper concedes that there are infinitely many other hypotheses that would also pass the tests that our current favorite has:

> By this method of elimination, we may hit upon a true theory. But in no case can the method *establish* its truth, even if it is true; for the number of *possibly* true theories remains infinite. (Popper 1979, 15)

Popper sees this as a way of stating Hume's problem of induction. Again,

> in my view, all that can possibly be "*positive*" in our scientific knowledge is positive *only* in so far as certain theories are, at a certain moment of time, preferred to others in the light of . . . attempted refutations. (P. 20)

For Popper, we should not *rely* on any hypothesis, at most we should prefer one. But why should we prefer best-tested hypotheses? It is altogether unsatisfactory for Popper to reply as he does that he simply does "not know of anything more 'rational' than a well-conducted critical discussion" (p. 22).

I, too, argue that hypotheses that pass genuinely severe tests gain merit thereby. How do I avoid Popper's problem? Popper's problem

1. It has been raised, for example, by Wesley Salmon (1966), Adolf Grünbaum (1978), and Alan Musgrave (1978).

here is that the grounds for the badge of "best-tested hypothesis of the moment" would also be grounds for giving the badge to an infinite number of (not yet even thought of) hypotheses, had they been the ones considered for testing. If a nonfalsified hypothesis H passes the tests failed by all the existing rivals, then H is best-tested, H gets the badge. Any other hypothesis that would also pass the existing tests would have to be said to do as well as H—by Popper's criteria for judging tests. But this is not the case for the test criteria I shall be setting out. These test criteria will be based on the idea of severity sketched above. A hypothesis H that passes the test failed by rival hypothesis H' (and by other alternative hypotheses considered) has passed a severe test for Popper—but not for me. Why not? Because for H to pass a severe test in my sense, it must have passed a test that is highly capable of probing the ways in which H can err. And the test that alternative hypothesis H' failed need not be probative in the least so far as the errors of H go. So long as two different hypotheses can err in different ways, different tests are needed to probe them severely. This point is the key to worries about underdetermination (to be discussed in chapter 6).

b. Corroboration Does Not Yield Reliability

There is a second variant of the objection that passing a severe test in Popper's sense fails to count in favor of a hypothesis. It is that saying a hypothesis is well tested for Popper says nothing about how successful it can be expected to be in the future.

According to Popper, the more severe the test a hypothesis has passed, the higher its corroboration. Popper regards the degree of corroboration of a hypothesis as "its degree of testability; the severity of tests it has undergone; and the way it has stood up to these tests" (Popper 1979, 18). Not only does Popper deny that we are entitled to consider well-corroborated claims as true, but we are not even to consider them as reliable. Reliability deals with future performance, and corroboration, according to Popper, is only a "*report of past performance.* Like preference, it is essentially comparative. . . . *But it says nothing whatever about future performance, or about the 'reliability' of a theory*" (p. 18). (Nor would this point be affected, Popper adds, by the finding of a quantitative measure of corroboration.)

At least part of the reason for this criticism, as well as for Popper's admission, is the prevalence of the view that induction or ampliative inference requires some assignment of probability, credibility, or other evidential measure to hypotheses. This view is shared by the majority of Popper's critics, and it is one Popper plainly rejects. The present view

of experimental learning, like Popper's, will not be in terms of assigning a degree of probability, credibility, or the like to any hypothesis or theory. But quite unlike Popper's view, this does not preclude our inferring a hypothesis reliably or obtaining reliable knowledge. The needed reliability assignment, I shall argue, is not only obtainable but is more in line with what is wanted in science.

Except in very special cases, the probability of hypotheses can only be construed as subjective degrees of belief, and I will argue that these yield an unsatisfactory account of scientific inference. As C. S. Peirce urged in anticipation of modern frequentists, what we really want to know is not the probability of hypotheses, but the probability with which certain outcomes would occur given that a specified experiment is performed. It was the genius of classical statisticians, R. A. Fisher, Jerzy Neyman, Egon Pearson, and others, to have developed approaches to experimental learning that did not depend on prior probabilities and where probability refers only to relative frequencies of types of outcomes or events. These relative frequency distributions, which may be called *experimental distributions*, model actual experimental processes.

Learning that hypothesis H is reliable, I propose, means learning that what H says about certain experimental results will often be close to the results actually produced—that H will or would often succeed in specified experimental applications. (The notions of "closeness" and "success" must and can be made rigorous.) This knowledge, I argue, results from procedures (e.g., severe tests) whose reliability is of precisely the same variety. My aim will be to show how passing a severe test teaches about experimental distributions or processes, and how this, in turn, grounds experimental knowledge.

The Emerging View of Experimental Knowledge

In summary, let me say a bit more about the view of experimental knowledge that emerges in my approach. I agree with Popper's critics that Popper fails to explain why corroboration counts in favor of a hypothesis—but not because such credit counts only in favor of a hypothesis if it adds to its credibility, support, probability, or the like. The problem stems from two related flaws in Popper's account: First, wearing the badge "best-tested so far" does not distinguish a hypothesis from infinitely many others. Second, there are no grounds for relying on hypotheses that are well corroborated in Popper's sense. I have also sketched how I will be getting around each flaw.

Popper says that passing a severe test (i.e., corroboration) counts in favor of a hypothesis simply because it may be true, while those that

failed the tests are false. In the present view, passing a severe test counts because of the experimental knowledge revealed by passing. Indeed, my reason for promoting the concept of severity in the first place is that it is a test characteristic that is relevant as regards something that has passed the test. To figure out what an experiment reveals, one has to figure out what, if anything, has passed a severe test. The experimental inference that is licensed, in other words, is what has passed a severe test. What is learned thereby can be made out in terms of the presence or absence of an error. Even if the test cannot be regarded as having severely tested any claim, that fact alone is likely to be relevant.

Since the severe test that a hypothesis H passes is at the same time a test that fails some alternative hypothesis (i.e., H's denial), the knowledge gained from passing can also be expressed as learning from failing a hypothesis. (For example, passing H: the disease is present, is to fail H': the disease is absent.) So in failing as well as passing, the present account accentuates the positive.

The centerpiece of my account is the notion of severity involved. Unlike accounts that begin with evidence e and hypothesis H and then seek to define an evidential relationship between them, severity refers to a method or procedure of testing, and cannot be assessed without considering how the data were generated, modeled, and analyzed to obtain relevant evidence in the first place. I propose to capture this by saying that assessing severity always refers to a framework of *experimental inquiry*.

In my account of experimental testing, experimental inquiry is viewed in terms of a series of models, each with different questions, stretching from low-level theories of data and experiment to higher level hypotheses and theories of interest. (I will elaborate in detail in chapter 5.) Whether it is passing or failing, however, what is learned will always be in terms of a specific question in a given model of experimental inquiry. Later we will see how such bits of learning are pieced together.

By Popper's own admission (e.g., Popper 1979, 19), corroboration fails to be an indicator of how a hypothesis would perform in experiments other than the ones already observed. Yet I want to claim for my own account that through severely testing hypotheses we can learn about the (actual or hypothetical) future performance of experimental processes—that is, about outcomes that would occur with specified probability if certain experiments were carried out. This is *experimental knowledge*. In using this special phrase, I mean to identify knowledge of experimental effects (that which would be reliably produced by car-

rying out an appropriate experiment)—whether or not they are part of any scientific theory. The intent may be seen as providing a home for what may be very low-level knowledge of how to reliably bring about certain experimental results. To paraphrase Ian Hacking, it may be seen as a home in which experiment "lives a life of its own" apart from high-level theorizing. But it is to be a real home, not a life in the street; it has its own models, parameters, and theories, albeit experimental ones. And this is so whether the experimental effects are "in nature," whether they are deliberately constructed, or even whether they exist only "on paper" or on computers.

Popper's problems are insurmountable when hypothesis appraisal is considered as a matter of some formal or logical relationship between evidence or evidence statements and hypotheses; but the situation is not improved by appealing to larger units such as Lakatosian research programs. Appealing to an experimental framework and corresponding experimental strategies, I will argue, offers a fresh perspective and fresh tools for solving these problems.

The idea that focusing on experiment might offer new and largely untapped tools for grappling with problems regarding scientific inference is not new; it underlies a good deal of work in the philosophy of science of the last decade. As promising as this new experimentalist movement has been, it is not clear that the new attention to experiment has paid off in advancing solutions to these problems. Nor is it clear that those working in this movement have demarcated a program for developing a philosophy or epistemology of experiment. For sure, they have given us an important start: their experimental narratives are rich in illustrations of the role of experimentation and instrumentation in scientific inference. But something more general and more systematic seems to be needed to show how this grounds experimental knowledge and how this knowledge gets us around the problems of evidence and inference. Where we should look, I will argue, is to the already well worked out methods and models for designing and analyzing experiments that are offered in standard statistical practice.

Experimental knowledge, as I understand it, may be construed in a formal or informal mode. In its formal mode, experimental knowledge is knowledge of the probabilities of specified outcomes in some actual or hypothetical series of experiments. Its formal statement may be given by an *experimental distribution* (a list of outcomes and their associated probabilities), or by a standard "random" process such as a coin-tossing mechanism. Typically, the interest is only in some key characteristic of this distribution—a parameter—such as its arithmetic mean. In its informal mode, the one the practitioner is generally engaged in, experimental knowledge is knowledge of the presence or absence of errors. (For example, a coin-tossing model or its corresponding Binomial distribution might serve as a formal model of an informal claim about a spurious correlation.) I will stress the informal mode.

As we proceed, we will come to see the considerable scope of what can be learned from answers to questions about experimental processes and effects. How far experimental knowledge can take us in understanding theoretical entities and processes is not something that should be decided before exploring this approach much further, further even than I can go in this book. So, for example, I will not argue for or against different realist views. What I will argue is that experimental knowledge is sufficient and, indeed, that it is the key to answering the main philosophical challenges to the objectivity and rationality of science.

1.4 REVISITING THE THREE DECISIONS

The Popperian problems of the last section emanate from the concern with which we began: the three (risky) decisions required to get a Popperian test off the ground. By the various techniques of learning from error, I said, we can substantiate the information needed. This necessitates an approach to experimental learning radically different from Popper's. Nevertheless, it may be of interest to see how the three concerns translate into arguments for checking one or more experimental mistakes. The connections are these:

- The acceptance of observation or basic statements (decision 1) is addressed by arguments justifying the assumptions of the experimental data.
- The elimination of auxiliary factors (decision 2) is addressed by arguments that the experiment is sufficiently controlled.
- The falsification of statistical claims (decision 3) is accomplished by standard statistical tests.

The first two involve justifying assumptions about a specific testing context. In both cases the justifications will take the form either of showing that the assumptions are sufficiently well met for the experimental learning of interest or showing that violations of the assumptions do not prevent specific types of information from being obtained from the experiment. As important as it is to avoid error, the centerpiece of the approach I recommend is its emphasis on procedures that permit a justification of the second type—learning despite errors, or robustness. I will champion a third sort of justificatory argument: even

if a mistake goes undetected, we will, with high probability, be able to find this out.

"The Empirical Basis" Becomes the Assumptions of the Experimental Data

To arrive at the basic (or test) statements for Popper, two decisions are required. The first is to decide which theories to deem "observational," which for Popper means not that they are literally observational, but rather that they may be deemed unproblematic background knowledge for the sake of the test. Such information is often based on well-understood theories of instruments, for example, on theories of microscopes. The second is to decide which particular statements to accept—for example, that the instrument reads such and such.[2] Popper claims that although we can never accept a basic statement with certainty, we "must stop at some basic statement or other which we *decide to accept.*" Otherwise the test leads nowhere. "[W]e arrive in this way at a procedure according to which we stop only at a kind of statement that is especially easy to test. . . . at statements about whose acceptance or rejection the various investigators are likely to reach agreement" (Popper 1959, 104). Nevertheless, for Popper, we can no more rely on these than on other corroborated hypotheses. They, too, are merely conjectures, if at a lower level and easier to test. They are not literally basic statements, but more like "piles driven into a swamp."

But even these singular observation statements are not enough to get a Popperian falsification off the ground. We need, not singular observations, but observational knowledge; the data must warrant a hypothesis about a real or reproducible effect:

> We say that a theory is falsified only if we have accepted basic statements which contradict it. . . . This condition is necessary, but not sufficient; for we have seen that non-reproducible single occurrences are of no significance to science. Thus a few stray basic statements contradicting a theory will hardly induce us to reject it as falsified. We shall take it as falsified only if we discover a *reproducible effect* which refutes the theory. In other words, we only accept the falsification if a low-level empirical hypothesis which describes such an effect is proposed and corroborated. (Popper 1959, 86)

He calls this low-level hypothesis a *falsifying hypothesis* (p. 87).

Here Popper is recognizing what is often overlooked: the empirical data enter hypothesis appraisal in science as a hypothesis about the

2. Basic statements are "statements asserting that an observable event is occurring in a certain individual region of space and time" (Popper 1959, 103). As an example he gives "This clock reads 30 minutes past 3" (Popper 1962, 388).

data. Accounts of hypothesis appraisal that start with evidence e as given vastly oversimplify experimental learning. This recognition, however, only means trouble for Popper. In order for the acceptance of a falsifying hypothesis to be more than a conventional decision, there need to be grounds for inferring a reliable effect—the very thing Popper says we cannot have. We cannot rely on hypotheses about real or reproducible effects for Popper, because they are based on lower-level (singular) observation statements that may themselves be mistaken. "[A]nd should we try to *establish* anything with our tests, we should be involved in an infinite regress" (Popper 1962, 388).

Herein lies a presupposition commonly harbored by philosophers: namely, that empirical claims are only as reliable as the data from which they are inferred. The fact is that we can often arrive at rather accurate claims from far less accurate ones. Scattered measurements, for example, are not of much use, but with a little data massaging (e.g., averaging) we can obtain a value of a quantity of interest that is far more accurate than individual measurements. Our day-to-day learners from error know this fact but, to my knowledge, the only philosopher to attach deep significance to this self-correcting ability is C. S. Peirce.

The present approach rejects both the justificationist image of building on a firm foundation (e.g., protocol statements) and the Popperian image of building on piles driven into a swamp. Instead, the image is one of shrewd experimental strategies that permit detecting errors and squeezing reliable effects out of noisy data. What we rely on, I will urge, are not so much scientific theories but *methods* for producing experimental effects.

Ruling Out Auxiliaries as Arguing for Experimental Control

The need to rule out alternative auxiliary factors (decision 2) gives Popper the most trouble. In order for an effect (e.g., an anomaly or failed prediction) to be attributed to some flaw in a hypothesis H, it is required to affirm a ceteris paribus claim, that it is not due to some other possible factor. As Lakatos notes, one can test a ceteris paribus clause severely by assuming that there are other influencing factors, specifying them, and testing these assumptions. "If many of them are refuted, the *ceteris paribus* clause will be regarded as well-corroborated" (Lakatos 1978, 26). Lakatos found this too risky.

Ruling out auxiliaries is thought to be so problematic because it is assumed that there are always infinitely many causes for which we have not controlled. What is overlooked is the way in which experiments may be designed to deliberately isolate the effect of interest so that only a manageable number of causal factors (or types of factors) may produce the particular experimental outcome. Most important,

literal control is not needed: one need only find ways of arguing so as to avoid the erroneous assignment of the cause of a given effect or anomaly. Several ways were discussed in section 1.2. Another example (under pretrial planning) would be to mimic the strategy of randomized treatment-control studies. The myriad of possible other causes—even without knowing what they are—are allowed to influence the treated and the control groups equally. In other cases, substantive alternative causes cannot be subtracted out in this manner. Then severe tests against hypotheses that these causes are responsible for the given experimental effect must be carried out separately.

In the present approach, ruling out alternative auxiliaries is tantamount to justifying the assumption either that the experiment is sufficiently well controlled or that the experiment allows arguing *as if* it were sufficiently controlled for the purpose of the question of primary interest.

With effective before- and after-trial planning and checking, learning that an anomaly cannot be due to specific background factors may finally show some primary hypothesis to be at fault. Even this rejection has an affirmative side. Precisely because the background checks have been required to be severe, such a rejection pinpoints a genuine effect that needs explaining or that calls for a specific revision. One literally learns from the error or anomaly. Note that no alternative hypothesis to the one rejected is needed in this testing model.

Falsifying Statistical Claims by Statistical Hypothesis Testing

My approach takes as the norm the need to deal with the rejection of statistical hypotheses—decision 3. Even where the primary theory or hypothesis of interest is nonstatistical, a variety of approximations, inaccuracies, and uncertainties results in the entry of statistical considerations in linking experimental data to hypotheses. A hierarchy of models of experimental inquiry will be outlined in chapter 5.

In an interesting footnote, Lakatos remarks that statistical rejection rules constitute "the philosophical basis of some of the most interesting developments in modern statistics. The Neyman-Pearson approach rests completely on methodological falsificationism" (Lakatos 1978, 25, n. 6). Still, neither he nor Popper attempts to use the Neyman-Pearson methods in his approach. By contrast, I shall make fundamental use of this approach, albeit reinterpreted, as well as of cognate methods (e.g., Fisherian tests). My use of these methods, I believe, reflects their actual uses in science and frees them from the confines of the particular philosophies of statistics often associated with them. Thus freed, these methods make up what I call (standard) *error statistics.*

Although Popper makes no explicit attempt to appeal to error statistical methods, his discussion of decision 3 gets to the heart of a fundamental type of error statistical test. While extremely rare events may occur, Popper notes, "such occurrences would not be physical effects, because, on account of their immense improbability, *they are not reproducible at will.* . . . If, however, we find *reproducible* deviations from a macro effect . . . deduced from a probability estimate . . . then we must assume that the probability estimate is *falsified"* (Popper 1959, 203).

The basic idea is this: A hypothesis may entail only that deviations of a certain magnitude are rare, so that an observed deviation from what is predicted does not strictly speaking contradict the prediction. A statistical test allows learning that a deviation is not rare but is *reproducible at will*—that is, can be brought about very frequently. If we learn this, then we have found a real physical effect that is denied by and so, in this sense, contradicts the statistical hypothesis. Rather than viewing this as a conventional decision, it will be seen to rest on solid epistemological grounds. These grounds may be cashed out in two ways: (1) To construe such reproducible effects as unsystematic will very often be mistaken. So unreliable a method would be an obstacle to using data to distinguish real from spurious effects—it would be an obstacle to learning from error. (2) It is extremely improbable that one would be able to regularly reproduce an effect if in fact it was accidental. The hypothesis asserting that it is a "real effect" passes a severe test.

1.5 PIECEMEAL LEARNING FROM ERRORS

My account of the growth of experimental knowledge is a result of having explored the consequences of the thesis that we learn from our mistakes. It is not an attempt to reconstruct after-the-fact scientific inferences or theory changes, but to give an account of forward-looking methods for learning. These methods revolve around tests of low-level or local hypotheses. These hypotheses have their home in experimental models and theories. While experimental hypotheses may be identical to substantive scientific claims, they may also simply be claims about experimental patterns, real or constructed, actual or hypothetical. Local experimental inquiries enable complex scientific problems to be broken down into more manageable pieces—pieces that admit of severe tests. Even when large-scale theories are being investigated or tested, these piecemeal tests are central to the growth of experimental knowledge.[3]

3. Popper puts the burden on the hypothesis to have high information content and so be the most testable. The present approach puts the burden on the experimental test—it is the test that should be severe. The basis for tests with appropriately high severity is the desire to learn the most.

I propose that the piecemeal questions into which experimental inquiries are broken down may be seen to refer to standard types of errors. Strategies for investigating these errors often run to type. Roughly, four such standard or canonical types are

a. mistaking experimental artifacts for real effects; mistaking chance effects for genuine correlations or regularities;
b. mistakes about a quantity or value of a parameter;
c. mistakes about a causal factor;
d. mistakes about the assumptions of experimental data.

These types of mistakes are not exclusive (for example, checking *d* may involve checking the others), nor do they even seem to be on a par with each other. Nevertheless, they often seem to correspond to distinct and canonical types of experimental arguments and strategies.

I suggest that methodological rules should be seen as strategies for conducting reliable inquiries into these standard or "canonical" types of errors. Examples of methodological rules are the use of controlled experiments in testing causal hypotheses, the use of randomization, the preference for novel facts and the avoidance of "ad hoc" hypotheses, the strategy of varying the evidence as much as possible, and the use of double-blind techniques in experimenting on human subjects.

The overarching picture, then, is of a substantive inquiry being broken down into inquiries about one or more canonical errors in such a way that methodological strategies and tools can be applied in investigating those errors. One example (to be fleshed out later) is an inquiry into whether a treatment causes an increased risk of some sort. It might be broken down into two canonical inquiries: first, to establish a real as opposed to a spurious correlation between the treatment and the effect; second, to test quantitatively the extent of the effect, if there is one. One possible methodological strategy would be a treatment-control experiment and an analysis by way of statistical significance tests.

Normative Naturalism in Experimental Methodology

The present model for an epistemology of experiment is both normative and naturalistic. I have in mind this picture of experimental methodology: methodological rules for experimental learning are strategies that enable learning from common types of experimental mistakes. The rules systematize the day-to-day learning from mistakes delineated above. From the history of mistakes made in reaching a type of inference, a repertoire of errors arises; methodological rules are techniques for circumventing and uncovering them. Some refer to before-trial experimental planning, others to after-trial analysis of the

data—or, more generally, learning from the data. The former includes rules about how specific errors are likely to be avoided or circumvented, the latter, rules about checking the extent to which given errors are committed or avoided in specific contexts.

Methodological rules do not rest on a priori intuitions, nor are they matters to be decided by conventions (e.g., about what counts as science or knowledge or rational). They are empirical claims or hypotheses about how to find out certain things by arguing from experiments. Accordingly, these hypotheses are open to an empirical appraisal: their truth depends upon what is actually the case in experimental inquiries. Hence, the account I propose is naturalistic. At the same time it is normative, in that the strategies are claims about how to actually proceed in given contexts to learn from experiments.

Since the rules are claims about strategies for avoiding mistakes and learning from errors, their appraisal turns on understanding how methods enable avoidance of specific errors. One has to examine the methods themselves, their roles, and their functions in experimental inquiry. A methodological rule is not empirically validated by determining whether its past use correlates with "successful" theories.[4] Rather, the value of a methodological rule is determined by understanding how its applications allow us to avoid particular experimental mistakes, to amplify differences between expected and actual experimental results, and to build up our tool kit of techniques of learning from error to determine how successful they have been in past applications.

No assignments of degrees of probability to hypotheses are required or desired in the present account of ampliative inference. Instead, passing a severe test yields positive experimental knowledge by corresponding to a strong argument from error. Accordingly, progress is not in terms of increasing or revising probability assignments but in terms of the growth of experimental knowledge, including advances in techniques for sustaining experimental arguments. Such features of my account stand in marked contrast to the popular Bayesian Way in the philosophy of science.

Next Step

The response to Popper's problems, which of course are not just Popper's, has generally been to "go bigger," to view theory testing in terms of larger units—whole paradigms, research programs, and a variety of other holisms. What I have just proposed instead is that the

4. This is essentially Larry Laudan's (1987, 1990b, 1996) normative naturalism.

lesson from Popper's problems is to go not bigger but smaller. Moreover, I propose that this lesson is, in a sense, also Thomas Kuhn's, despite his being a major leader of the holistic movement. Let us therefore begin our project by turning to some Kuhnian reflections on Popper.

CHAPTER TWO

Ducks, Rabbits, and Normal Science: Recasting the Kuhn's-Eye View of Popper

SHORTLY AFTER the publication of his enormously influential book *The Structure of Scientific Revolutions*, Thomas Kuhn offered "a disciplined comparison" of his and Popper's views of science in the paper "Logic of Discovery or Psychology of Research?" It begins with these lines:

> My object in these pages is to juxtapose the view of scientific development outlined in my book [*Structure*], with the better known views of our chairman, Sir Karl Popper. Ordinarily I should decline such an undertaking, for I am not so sanguine as Sir Karl about the utility of confrontations. . . . Even before my book was published two and a half years ago, I had begun to discover special and often puzzling characteristics of the relation between my views and his. That relation and the divergent reactions I have encountered to it suggest that a disciplined comparison of the two may produce peculiar enlightenment. (Kuhn 1970, 1)

"Peculiar enlightenment" is an apt description of what may be found in going back to Kuhn's early comparison with Popper and the responses it engendered. What makes my recasting of Kuhn peculiar is that while it justifies the very theses by which Kuhn effects the contrast with Popper, the picture that results is decidedly *un*-Kuhnian. As such I do not doubt that my recasting differs from the "peculiar enlightenment" Kuhn intended, but my task is not a faithful explication of what Kuhn saw himself as doing. Rather it is an attempt, at times deliberately un-Kuhnian, to see what philosophical mileage can be gotten from exploring the Kuhnian contrast with Popper. This exercise will serve as a springboard for the picture of experimental knowledge that I want to develop in this book.

Kuhn begins by listing the similarities between himself and Popper that place them "in the same minority" among philosophers of science of the day (Kuhn 1970, 2). Both accept theory-ladenness of observation, hold some version of realism (at least as a proper aim of science), and reject the view of progress "by accretion," emphasizing instead